HIGH DIMENSIONAL ANALYSIS OF GENETIC DATA FOR THE CLASSIFICATION OF TYPE 2 DIABETES USING ADVANCED MACHINE LEARNING ALGORITHMS

Basma Taleb Abdulaimma

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

February 2019

ABSTRACT

The prevalence of type 2 diabetes (T2D) has increased steadily over the last thirty years and has now reached epidemic proportions. The secondary complications associated with T2D have significant health and economic impacts worldwide and it is now regarded as the seventh leading cause of mortality. Therefore, understanding the underlying causes of T2D is high on government agendas. The condition is a multifactorial disorder with a complex aetiology. This means that T2D emerges from the convergence between genetics, the environment and diet, and lifestyle choices. The genetic determinants remain largely elusive, with only a handful of identified candidate genes. Genome-wide association studies (GWAS) have enhanced our understanding of genetic-based determinants in common complex human diseases. To date, 120 single nucleotide polymorphisms (SNPs) for T2D have been identified using GWAS. Standard statistical tests for single and multi-locus analysis, such as logistic regression, have demonstrated little effect in understanding the genetic architecture of complex human diseases. Logistic regression can capture linear interactions between SNPs and traits however it neglects the non-linear epistatic interactions that are often present within genetic data. Complex human diseases are caused by the contributions made by many interacting genetic variants. However, detecting epistatic interactions and understanding the underlying pathogenesis architecture of complex human disorders remains a significant challenge.

This thesis presents a novel framework based on deep learning to reduce the high-dimensional space in GWAS and learn non-linear epistatic interactions in T2D genetic data for binary classification tasks. This framework includes traditional GWAS quality control, association analysis, deep learning stacked autoencoders, and a multilayer perceptron for classification.

Quality control procedures are conducted to exclude genetic variants and individuals that do not meet a pre-specified criterion. Logistic association analysis under an additive genetic

model adjusted for genomic control inflation factor is also conducted. SNPs generated with a p-value threshold of 10^{-2} are considered, resulting in 6609 SNPs (features), to remove statistically improbable SNPs and help minimise the computational requirements needed to process all SNPs. The 6609 SNPs are used for epistatic analysis through progressively smaller hidden layer units. Latent representations are extracted using stacked autoencoders to initialise a multilayer feedforward network for binary classification. The classifier is fine-tuned to discriminate between cases and controls using T2D genetic data. The performance of a deep learning stacked autoencoder model is evaluated and benchmarked against a multilayer perceptron and a random forest learning algorithm. The findings show that the best results were obtained using 2500 compressed hidden units (AUC=94.25%). However, the classification accuracy when using 300 compressed neurons remains reasonable with (AUC=80.78%). The results are promising. Using deep learning stacked autoencoders, it is possible to reduce high-dimensional features in T2D GWAS data and learn non-linear epistatic interactions between SNPs while enhancing overall model performance for binary classification purposes.

Acknowledgements

In the Name of Allah, the Most Gracious, the Most Merciful

At the beginning, Praise and Great Thanks be to Allah for providing me with the ability, strength, patience and determination to complete this study.

It is my pleasure to extend my sincere gratitude to all the people who believed in me and helped bring this PhD to life. I am deeply indebted to my supervisor, Dr Paul Fergus, for his patience, invaluable assistance, continuous support and guidance throughout each step of this PhD. Without his unlimited support, I would not be able to complete this research work given that I had no prior experience of the subject. I would like to take this opportunity to express my appreciation of his infinite support that helped me to face the challenges more successfully. Lastly but not the least, I want to say it has been an honoured and privileged to be one of his students.

I would like to give my deepest gratitude to my second supervisor Dr Carl Chalmers who was abundantly helpful. Without his assistance, the implementation part would not have been successful. I would also like to convey a special thanks to Prof Mehmet Dorak for the precious time he dedicated to my research and for providing fundamentally important information that helped in the completion of this research. Special thanks go to Mrs Tricia Waterson and all the staff and technicians in the Department of Computer Science for their assistance and guidance. I would also like to thank Liverpool John Moores University for offering me the PhD scholarship and sponsoring me through this time.

Warm appreciation and thanks to my family for their support, patience and understanding throughout this journey. Specially, my father Dr Talib Majeed, my mother Mrs Jazaer Al-Ward, my uncle Dr Faisal Majeed and my auntie Mrs Janan Al-Beer for their support and advice through some difficult times during this PhD – this has helped me to stay positive and to continue with my journey. My warm thanks go to my wonderful husband Mr Bashar Fadhil for his continuous help and patience and more importantly for looking after our children which gave me more time to work on this PhD. Finally, to my friends' Dr Raghad Al-Shabandar, Dr Casimiro Aday Curbelo Montañez, and Dr Jade Hind for their constant encouragement and support and for keeping me grounded over the last four years.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

ACRONYMS

AUC	Area Under the Curve
ANNs	Artificial Neural Networks
AE	Autoencoder
BMI	Body Mass Index
CG	Candidate Gene
CHR	Chromosome
CMP	Combinatorial Partitioning Method
CD-CV	Common Disease-Common Variant Hypothesis
CI	Confidence Interval
CNV	Copy Number Variations
dbGap	Database of Genotypes and Phenotypes
DL	Deep Learning
DNA	Deoxyribonucleic Acid
GLM	Generalized Linear Model
GWAS	Genome-Wide Association Studies
GC	Genomic Control
GDM	Gestational Diabetes Mellitus
HWE	Hardy-Weinberg Equilibrium
HPFS	Health Professionals Follow-up Study
IBD	Identity by Descent
IBS	Identity by State
IDF	International Diabetes Federation
KL	Kullback-Leibler
LD	Linkage Disequilibrium
MODY	Maturity-Onset Diabetes of the Young
MSE	Mean Squared Error
mRNA	Messenger Ribonucleic Acid
MAF	Minor Allele Frequency
MDR	Multifactor Dimensionality Reduction
MLP	Multilayer Perceptron
NCBI	National Center for Biotechnology Information
NHGRI	National Human Genome Research Institute

NDM	Neonatal Diabetes Mellitus
NGS	Next Generation Sequencing
NHS	Nurses' Health Study
OR	Odds Ratio
PAR	Population Attributable Risk
PCA	Principal Component Analysis
QC	Quality Control
RF	Random Forest
RGS	Random Grid Search
RJ	Random Jungle
ROC	Receiver Operating Characteristic
ReLU	Rectifier Linear Unit
RFE	Recursive Feature Elimination
RPART	Recursive Partitioning and Regression Trees
RPM	Restricted Partition Method
RNA	Ribonucleic Acid
RAF	Risk Allele Frequency
SNP	Single Nucleotide Polymorphism
SAE	Stacked Autoencoder
GBM	Stochastic Gradient Boosting
SVM	Support Vector Machine
T2D	Type 2 Diabetes
WTCCC	Wellcome Trust Case Control Consortium
WHO	World Health Organization
XHE	X-Chromosome Homozygosity Rate

LIST OF PUBLICATIONS

Journal Papers

Paul Fergus, Aday Curbelo Montañez C, **Basma Abdulaimma**, Paulo Lisboa, Carl Chalmers, Beth Pineles. "*Utilising Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women"*, IEEE/ACM Transactions on Computational Biology and Bioinformatics. (2018)

Basma Abdulaimma, Paul Fergus, Aday Curbelo Montañez C, Carl Chalmers. *"Extracting Epistatic Interactions in Type 2 Diabetes Genome-Wide Data Using Stacked Autoencoder"* (2019) (In Preparation)

Conference Papers

Basma Abdulaimma, Paul Fergus, Carl Chalmers, Aday Curbelo Montañez C. "*Deep Learning and Genome Wide Association Studies for the Classification of Type 2 Diabetes"* International Joint Conference on Neural Networks (IJCNN 2019) (Under Review)

Basma Abdulaimma, Abir Hussain, Paul Fergus, Dhiya Al-Jumeily, Paulo Lisboa, Huang D-S, Naeem Radi. "*Improving Type 2 Diabetes Phenotypic Classification by Combining Genetics and Conventional Risk Factors"* 2018 IEEE Congress on Evolutionary Computation (IEEE CEC 2018)

Basma Abdulaimma, Abir Hussain, Paul Fergus, Dhiya Al-Jumeily, Aday Curbelo Montañez C, Jade Hind. "*Association Mapping Approach into Type 2 Diabetes using Biomarkers and Clinical Data"* Lecture Notes in Computer Science, Vol. 10362, pp. 325–336, International Conference Intelligence Computing (ICIC 2017)

Aday Curbelo Montañez C, Paul Fergus, Abir Hussain, Dhiya Al-Jumeily, **Basma Abdulaimma**, Jade Hind. "*Machine learning approaches for the prediction of obesity using publicly available genetic profiles"*, 2017 International Joint Conference on Neural Network (IJCNN 2017)

Paulo Lisboa, Jade Hind, Abir Hussain, Dhiya Al-Jumeily, **Basma Abdulaimma**, Aday Curbelo Montañez C. "*A Robust Method for the Interpretation of Genomic Data*", 2017 International Joint Conference on Neural Networks (IJCNN 2017)

Aday Curbelo Montañez C, Paul Fergus, Abir Hussain, Dhyia Al-Jumeily, **Basma Abdulaimma**, Haya Al-askar. "*A Genetic Analytics Framework for Risk Variant Identification to Support Intervention Strategies for People Susceptible to Polygenic Obesity and Overweight"*, Lecture Notes in Computer Science, 2016 International Conference on Intelligent Computation (ICIC 2016)

Chapter 1 Introduction

1.1 Preamble

The prevalence of Type 2 Diabetes (T2D) throughout the world has reached epidemic proportions. In 2012, the World Health Organization (WHO) (World Health Organization 2016) estimated that 1.5 million deaths were directly attributed to diabetes, and by 2030 diabetes will be the seventh leading cause of mortality worldwide (Mathers & Loncar 2006). T2D is the most predominant form of diabetes (World Health Organization 2016) and is regarded as a multifactorial disorder, caused by the convergence of genetics, the environment, and a sedentary lifestyle (Lyssenko et al. 2008). There is strong evidence that genetic factors play a significant role in T2D susceptibility (Prasad & Groop 2015). T2D is a polygenetic disorder that is caused by a complex interaction among multiple genes. As such, an in-depth investigation into the T2D pathogenetic architecture is needed to help researchers and professionals understand the aetiology of T2D.

1.2 Genome-Wide Association Studies (GWAS)

Genes influence all human diseases and yet the genetic foundation of many complex diseases is still unknown. With the availability of cheaper genotyping technologies (Behjati & Tarpey 2013), genome-wide association studies (GWAS) have seen widespread use within genetic research. In recent years, GWAS have succeeded in identifying genetic variants that demonstrate evidence of increased susceptibility in a wide range of complex diseases, including Schizophrenia, Epilepsy, Obesity, Cardiovascular Disease, Hypertension and T2D (Guo et al. 2014; Bush & Moore 2012). GWAS have also been used to detect the genetic effects associated with phenotypes (disease trait) in population-based studies using single-locus statistical tests. In these studies each single nucleotide polymorphism (SNP) is explored separately for association with particular diseases or traits (Clarke et al. 2011). The genetic variants identified so far have helped to explain a relatively small proportion of

heritability, however, the question remains about how missing heritability can be better explained (Blanco-Gómez et al. 2016; Manolio et al. 2009). More importantly, it is generally believed that the underlying cause of complex human diseases does not rely on single genetic variations but instead on a contribution of many interactions between genetic loci (Morris et al. 2012; Robinson et al. 2014; Lee et al. 2012); referred to as epistasis.

In this thesis the term epistasis refers specifically to the latent interactions between multiple SNPs and their effects (Wei et al. 2014). This is a topic studied in molecular biology, particularly genetic biomolecules. The primary goal is to understand the underlying pathogenesis architecture linked with common complex disorders. Epistasis arises due to non-linear interactions between genetic variants. Detecting epistatic interactions and genetic interactive effects, however, remains a significant challenge in large-scale GWAS data. This is due to various factors that include genetic heterogeneities, low penetrance, small epidemiology sample sizes, polygenic inheritance, and the large number of genetic variants often considered in GWAS studies.

Consequently, complex non-linear relationships between genotypes and the phenotype are not investigated in GWAS. Standard parametric multi-variable statistical approaches, such as logistic regression, which is used in GWAS, are more suited to capturing linear interactions between genotypes and phenotype in much simpler diseases like cystic fibrosis which is known to only have one associated SNP (Cutting 2015).

Existing studies using GWAS data have focused on the use of data mining and machine learning algorithms (Botta et al. 2014; López et al. 2018; Nguyen et al. 2015; Chen et al. 2008). These techniques have been used to model complex relationships and interactions between features (SNPs) and their association with phenotypes. Data reduction approaches like multifactor dimensionality reduction have also been successfully applied to detect putative interactions between loci for a wide variety of human diseases (Barna et al. 2018; Andrew et al. 2008; Oh et al. 2012; R De et al. 2015). Ensemble methods, such as the random

forest algorithm, have been broadly applied for genomic data analysis to detect SNP correlations (Botta et al. 2014), disease risk prediction (López et al. 2018), and feature selection (Nguyen et al. 2015). Support Vector Machines (SVMs) have been used to detect gene-gene interactions (Chen et al. 2008) and disease classification (Vanitha et al. 2015). Artificial Neural Networks (ANNs) have been utilized to detect SNP correlations as demonstrated in (Koo et al. 2013; Motsinger-Reif et al. 2008). Although, these machine learning algorithms are competent in handling complex correlations and interactions among a small number of features, they do not scale to a very larger number of SNPs, which is often the case in GWAS (genotypes of almost one million SNPs and thousands of samples). In particular, using machine learning algorithms for epistatic analysis with a few hundred loci becomes computationally very difficult. Furthermore, traditional machine learning algorithms suffer with multicollinearity (Waaijenborg & Zwinderman 2009) and the curse of dimensionality (Sharma & Saroha 2015).

Therefore, an alternative approach to model high-dimensional GWAS data and handle non-linear epistatic interactions between SNPs is needed. In this thesis we investigate the use of unsupervised deep learning (DL) since it can deal with big data and the detection of complex features and associated non-linear interactions. More specifically we explore the use of deep learning stacked autoencoders as a way of learning the epistatic interactions that exist between SNPs. To evaluate the approach, learned features are used to initialise the weights of a fully connected multilayer perceptron (MLP) before it is fine-tuned to classify observations as either case or control in a T2D GWAS dataset.

1.3 Computational Biology

A fundamental challenge in molecular biology networks, particularly in genetics, is to identify and understand the underlying interactions between genetic variants (SNPs) and how they contribute to human disease and complex phenotypic traits. The main goal is to

pinpoint genetic markers that can be used to predict an individuals' predisposition to developing a particular disorder.

In large-scale GWAS data, despite the use of advanced statistical methods and computational strategies to detect SNPs interactions, they cannot deal with large combinatorial analysis, scalability, and low statistical power. As advances in high-dimensional GWAS data generation continues, it is becoming increasingly more important to develop more powerful methodologies to analyse and examine epistatic interactions in complex, unstructured, and large datasets.

1.4 Scope of Research

The research question is whether complex interactions between SNPs (epistasis) can be learnt using deep learning stacked autoencoders to classify T2D risk in humans. The approach follows a traditional GWAS quality control and association analysis methodology where the most significant SNPs are selected and used in subsequent analysis. This helps to manage computational demands. Stacked autoencoders are implemented as a feature extraction/learning technique to capture the salient relationships that exist between SNPs, thus capturing epistatic interactions. The final set of features is used to initialise the weights of a fully connected multilayer perceptron (MLP) which is then fine-tuned to classify case and control GWAS observations.

1.5 Aims and Objectives of the Thesis

The main aim of this thesis is to investigate the aetiology of T2D through effective use of bioinformatics and state-of-the-art machine learning algorithms. The approach provides a robust framework for processing high-dimensional genetic data to model and classify case-control individuals using a GWAS dataset. More precisely, the framework allows us to capture the genetic architecture of epistasis in T2D genomic data and to investigate its

influence in disease susceptibility. In order to fulfil the research aims, several key objectives have been set:

- Investigate open source databases including the Genotype and Phenotype (dbGap) database which contains genetic and clinical information for case-control individuals.
- Identify and remove low quality genetic markers and samples to produce a reliable subset for subsequent association analysis.
- Apply Genome-wide association analysis to test for associations between genetic markers and T2D in a population-based study.
- Filter genetic markers (SNPs) using a simple statistical approach to select a subset of SNPs for subsequent interaction analysis. SNPs are selected based on the strength of independent effects and are ranked using pre-specified thresholds.
- Perform non-linear dimensionality reduction to retain important SNPs and learn the cumulative non-linear epistatic interactions between them using deep learning stacked autoencoders.
- Classify and evaluate T2D high-dimensional genetic data using advanced machine learning techniques.
- Classify and evaluate genetic and non-genetic (environmental, sociodemographic and clinical) risk factors using linear and non-linear traditional machine learning algorithms and explore the contribution and the effects of these factors in T2D susceptibility.
- Design and implement a framework for the proposed project to produce an effective data analytic system to fulfil the aims of this study.

1.6 Novel Contributions

This thesis presents a novel framework for the binary classification of high-dimensional T2D using case-control GWAS data. Using deep learning stacked autoencoders we can extract

SNPs and latent relationships in large scale biological data structures. Acting as a feature learning step, features are used to initialise the weights of a fully connected multilayer feedforward softmax classifier and fine-tune it to classify T2D observations. To the best of our knowledge, this is the first comprehensive study of its kind that uses stacked autoencoders to capture the epistatic interactions between SNPs in T2D GWAS data.

Existing studies in the genomic field depend heavily on manual feature engineering using labelled data. The greedy layer-wise learning algorithm solution performed with stacked autoencoders in this thesis is based on training the network layer-by-layer using unlabelled data. The results show that this is a very efficient way to convert high-dimensional GWAS data into low-dimensional data to allow us to discover the non-linear structures that exist between SNPs. These reduced, compressed features act as an abstract representation of the original feature space. The ability to automatically extract latent representation of SNPs related to T2D GWAS data enhances the quality of experimental investigations, allowing researchers to discover and investigate the pathogenesis architecture of T2D further.

1.6.1 Literature Review

In this thesis, an up to date biomedical literature review of current works in the field of T2D GWAS study is collected from PubMed, the United States National Library of Medicine and the National Center for Biotechnology Information (NCBI) that provides resources for genomic, genetic and biomedical research.

T2D with its aetiology in addition to the application of GWAS in T2D in different cohorts and ethnic groups are reviewed. The state-of-the-art in machine learning approaches used to predict risk susceptibility to T2D and to detect and explore SNPs correlations is also presented.

Furthermore, parts of the materials and results presented in this thesis have contributed to the literature as shown in (Abdulaimma et al. 2017; Abdulaimma et al. 2018) where an

association mapping approach was investigated with our dataset to identify potential candidate SNP to T2D predisposition. In addition, the investigation is conducted in (Abdulaimma et al. 2018) to evaluate the predictive capacity of several machine learning algorithms in discriminating between cases and controls in T2D GWAS Data. Collectively, this review and our publications contribute to current genetic research in T2D which provides up to date information in the biomedical and bioinformatics research fields.

1.6.2 Stacked Autoencoders

A stacked autoencoder, which is an unsupervised learning process, is adopted in this thesis. Stacked autoencoders offer a method to automatically learn features from unlabelled data. This is an efficient method to reduce and compress high-dimensional GWAS data, producing an abstract representation of the original data space. Stacked autoencoders can discover the non-linear structures in complex, large, unstructured data as is the case in GWAS. This allowed us to extract the non-linear epistatic interactions between SNPs which is an important topic in understanding missing heritability and predisposition in many complex disorders.

Our work has been published in IEEE/ACM Transaction on Computational Biology and Bioinformation, which demonstrates stacked autoencoders can be applied successfully to learn the abstract representation of SNP data and to study epistatic interactions between SNPs (Fergus et al. 2018). The results are encouraging and show that stacked autoencoders are an effective method for dealing with high-dimensional GWAS data and detecting epistatic interactions between SNPs.

1.6.3 Combined Framework

Quality control, logistic regression association analysis, and deep learning stacked autoencoders were combined to constitute the components of our proposed methodology. Various stringent quality control assessment steps followed by logistic regression and association analysis adjusted for genomic control were performed for single-SNP analysis.

Statistically significant SNPs identified via association tests were used as input features for deep learning stacked autoencoders. The output from the stacked autoencoders comprised an abstract representation of the input features which were in turn used to initialise the weights of a fully connected multilayer feedforward softmax classifier, fine-tune it to classify T2D observations.

1.6.4 Decision Support Tool for Early Detection of T2D Susceptibility

Our GWAS classification method could be considered as an early screening tool for the identification of people with a genetic disposition to T2D. This would aid physicians to identify pre-diabetic individuals with high-risk of developing the condition much earlier thus allowing appropriate actions to be administered to mitigate long-term effects.

Early detection could reduce premature death and the risk of developing secondary complications associated with the condition. A study conducted by Herman et.al (Herman et al. 2015) investigated the benefits of early screening, diagnosis, and treatment of T2D and compared the results with those who had no screening and late treatment using the ADDITION-Europe population. The study found that cardiovascular risk, which is one of the common complications associated with T2D, can be reduced with early screening and diagnosis. In another study (Olafsdottir et al. 2016) also revealed that cumulative retinopathy prevalence and severity could be reduced with early detection of T2D.

The current protocol used by physicians in hospitals and clinics is based on a blood sugar and/or oral glucose tolerance test (American Diabetes Association 2018). Physicians make their decision based on plasma glucose criteria, even though the test is normal this may not eliminate the possibility of T2D. Therefore, adopting our GWAS classification system could act as an early screening intervention to provide physicians with an additional source of information alongside existing tests to aid decision making.

1.7 Thesis Structure

The remainder of this thesis is structured as follows. Chapter 2 provides a brief overview of human genetic structures, components, mechanisms and functionalities. A discussion on T2D including its aetiology and risk factors are also presented including a comprehensive discussion relating to genome-wide association studies and associated quality control procedures used in T2D analysis. This chapter is concluded with a comprehensive literature review of existing GWAS in T2D studies.

Chapter 3 introduces bioinformatics and advanced machine learning algorithms. The chapter begins with a discussion on epistasis and its challenges, followed by a comprehensive literature review of existing epistatic applications before artificial neural networks are described. This is followed by a discussion on the state-of-the-art in deep learning. This includes a brief overview on supervised and unsupervised learning across six machine learning algorithms and their use in T2D studies.

Chapter 4 introduces the framework and proposed methodology. This includes a discussion on data acquisition and a description on the data quality control procedure, and association testing with genetic variants. The discussion emphasises the novel contributions made in the proposed methodology along with the theoretical aspects of deep learning and stacked autoencoders. Furthermore, this chapter examines the clinical and genetic factors on the predictive discriminatory power of T2D modelling using machine learning. Finally, the performance metrics for each of the machine learning models used are evaluated.

Chapter 5 presents the results for the various experiments conducted in the investigation. While chapter 6 discusses the results and draws on conclusions and recommendations derived from the study. The thesis is concluded in Chapter 7 before the future directions for this study are presented.

Chapter 2 Background

2.1 Introduction

This chapter begins with an overview on human genetics followed by a discussion on T2D, including the diseases aetiology and associated risk factors. This chapter also discusses genome-wide association studies and includes the quality control steps needed and the statistical methods used. The chapter is finally concluded with a review of genome-wide association studies in T2D.

2.2 Human Biology Background

In biology, ***genome*** is defined as a cell's total genetic information (Alberts et al. 2015). A ***cell*** is a fundamental and basic unit of life (Alberts et al. 2014). Living organisms are divided into two types including unicellular organisms and multicellular organisms. Multicellular organisms, like humans, are made up of a large number of specialized cells that work together to perform different functions. Human bodies are composed of millions of cells and each one contains a complete copy of an individual's genetic information (Alberts et al. 2015).

In each cell, there are 23 pairs of ***chromosomes*** and they are situated in the cell nucleus (Alberts et al. 2015). The chromosome consists of very long strands of ***Deoxyribonucleic Acid (DNA)*** along with the proteins responsible for folding and packaging the DNA string into a compact structure. The DNA is a molecule that carries most of the genetic information and is the hereditary material found in all living organisms (Alberts et al. 2014). DNA is made of four chemical monomers known as ***nucleotides***. Each nucleotide contains deoxyribose (sugar with phosphate) and a base. This base is adenine (A), guanine (G), cytosine (C), and thymine (T) and they are linked together in a long linear sequence to form a DNA strand that is known as the ***polynucleotide***. DNA molecules consist of two antiparallel polynucleotides joined together through the process of complementary base

pairing, where A pairs with T and C pairs with G, to form the DNA double helix which encodes all genetic information (Alberts et al. 2015). Figure 2.1 illustrates the human DNA structure, from the cell through the chromosome to the DNA components.

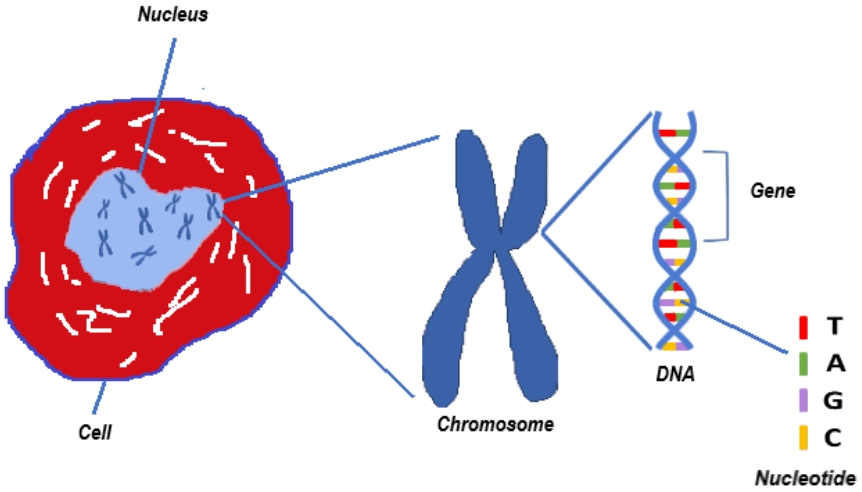


Figure 2.1: Human DNA Structure. Cell, Chromosome, DNA

DNA molecules contain a linear sequence of many genes. Each ***gene*** is a segment of DNA and represents a functional unit for the production of specific proteins (Alberts et al. 2015). The human genome contains over 3 billion base pairs (nucleotides). Only a small percentage of the entire DNA is composed of genes (International Human Genome Sequencing Consortium 2004). There are over 21,000 genes in the entire human genome (International Human Genome Sequencing Consortium 2004), and these contain the information necessary to produce proteins. An alternative form of a gene is known as an ***allele*** (Alberts et al. 2015). Each gene contains two alleles, a ***dominant allele*** and a ***recessive allele***. Each allele pair is located at a similar locus on homologous chromosomes (one chromosome comes from the male parent and the other one comes from the female parent). The dominant trait is expressed if the gene is heterozygous, i.e. possesses both dominant and recessive alleles. The recessive trait is expressed if the gene is homozygous, i.e. both alleles are recessive.

The combination and pairing of alleles for a specific gene is referred to as a ***genotype*** (Alberts et al. 2015). A genotype is either homozygous or heterozygous as explained previously. The genotype is responsible for expressing an organisms' characteristics in the

form of a ***phenotype*** (Alberts et al. 2015). The phenotype focuses on a trait, which is expressed as the appearance, behaviour or medical condition of an individual.

The central dogma in molecular biology defines the flow of genetic information in cells from Deoxyribonucleic Acid (DNA) through Ribonucleic Acid (RNA) to proteins (Alberts et al. 2015). This transformation process occurs thousands of times every second in all living cells. This describes the mechanisms by which cells copy segments of DNA into RNA, through a process called transcription, followed by the synthesis of proteins from RNA through a process called translation as illustrated in Figure 2.2.

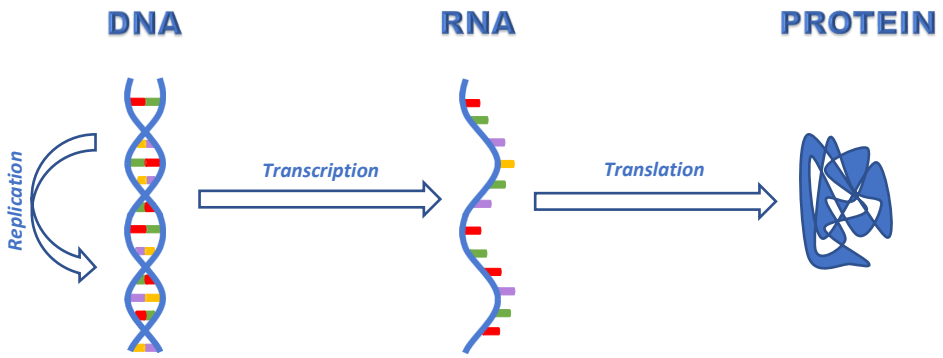


Figure 2.2: The Central Dogma of Molecular Biology

There are 20,000 proteins made in humans and they are responsible for regulating the structure of the cell and executing the majority of the functions cells provide (Alberts et al. 2015). Proteins determine the biological instructions contained in DNA that are necessary for building and maintaining an organism.

2.3 Human Genetic Variations

All humans have small variations in their genetic code (Alberts et al. 2014) and it is not possible for any two people to have the same genomic sequence. Since the completion of the Human Genome Project in 2003, researchers have confirmed that among the 3 billion base pairs that comprise DNA, 99.9% are very similar (International Human Genome Sequencing Consortium 2004). However, the remaining 0.1% makes each individual unique (Alberts et

al. 2014). More importantly, this variation explains the differences among people and their susceptibility to particular diseases.

Genetic variations, also called mutations, can occur due to the substitution of a single base-pair (nucleotide) and this is termed a ***Single Nucleotide Polymorphism (SNP)***. Typically, an SNP is defined as single base-pair change in the genetic code (DNA sequence) and it is the main cause of human genetic variability (Durbin et al. 2010). Figure 2.3 illustrates the genetic variation in the same region of the genome for three different individuals. Another source of genetic variation can result from duplications, deletions and insertions of large segments of the DNA molecule. These types of mutation are known as ***Copy Number Variations (CNVs)*** (Alberts et al. 2015) which have been implicated in several human traits, including hypertension, and colour blindness.

Figure 2.3: Genetic Variation (SNPs) among Three Individuals

Most of these mutations are common and have no functional significance, thus they are relatively harmless (Alberts et al. 2015). However, there are single nucleotide changes that can alter gene production and change regulatory DNA sequences. When this occurs it can have a profound effect on human health, behaviour, and physiology and can be the cause of serious diseases (Alberts et al. 2014). While there are a large number of variants, a relatively small number affect us functionally. The challenge in human genetics, however, is to discover those that are harmful to us.

More importantly, the genetic roots for common complex diseases is more difficult to understand (Mitchell 2012). Instead of a single allele or single gene, many complex disorders, referred to as polygenetic conditions, stem from the interactions and contributions

of multiple SNPs or genes. For these types of conditions, which include Diabetes, Schizophrenia, Epilepsy, Obesity, Cardiovascular Disease, and Hypertension, understanding SNP interactions and environmental risk factors is fundamentally important. Typically, environmental factors have vital effects from the outset and they significantly influence the severity of conditions (Korkiakangas et al. 2009; Cooper et al. 2012). By investigating the effects of these multiple factors, it will help us to improve both medicine and our understanding of human biology.

2.4 Diabetes

The World Health Organisation (WHO) reported that over the past few decades, both Diabetes' cases and prevalence have been progressively growing (World Health Organization 2016). Diabetes is a serious, chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. According to the International Diabetes Federation (IDF) the number of diabetic people worldwide is expected to rise from 366 million in 2011 to 552 million by 2030 (Whiting et al. 2011). One in 11 adults had diabetes in 2015 with this figure expected to be one in 10 adults by 2040.

Additionally, in 2015, Diabetes UK[1] announced that there were 3.9 million people in the UK living with diabetes. This figure shows that there were approximately 125,000 more adults with diabetes compared with the previous year. This indicates that there is a dramatic increase in diabetic cases. Diabetes is one of the leading causes of death (2.7%) worldwide. In 2012, the WHO revealed that diabetes killed 1.5 million people worldwide (World Health Organization 2016). The main types of diabetes are type 1 and type 2, and gestational diabetes (GDM). However, nine other subtypes do exist (World Health Organization 2016).

[1] https://www.diabetes.org.uk/About_us/News/39-million-people-now-living-with-diabetes/

Type 2 diabetes is the most predominant form of diabetes around the world and is the category studied in this thesis.

2.4.1 Key Facts about Type 2 Diabetes

According to the WHO, T2D accounts for the vast majority of people with diabetes worldwide (World Health Organization 2016). It is estimated that people diagnosed with T2D constitute 90% of all reported diabetic cases. Until recently, T2D was recognized only in people over the age of 40 but have now found in young children (Farsani et al. 2013).

T2D remains the leading cause of serious long-term health conditions. It is responsible for most cases of blindness (Diabetic retinopathy), kidney failure and lower limb amputation. Moreover, high glucose levels (raised blood sugar levels) or Hyperglycaemia in the bloodstream can damage blood vessels which increases the likelihood of atherosclerosis (cardiovascular disease) and stroke and can cause nerve damage (Inzucchi et al. 2012). In the UK, the annual direct cost of T2D to the National Health Service (NHS) in 2035 is estimated to be £15 billion - the indirect costs will be close to £20.5 billion (Hex et al. 2012).

2.4.2 Type 2 Diabetes Phenomena

T2D, which is known as insulin resistance, is a chronic disease that occurs when the pancreas does not produce enough insulin or the insulin produced does not interact with the body's cells (World Health Organization 2016). Consequently, glucose remains in the blood and the body cannot effectively use it for energy. Researchers believed that T2D is a multifactorial disorder with a complex aetiology (Lyssenko et al. 2008). The condition is said to result from the convergence of genetics, the environment, diet and lifestyle risk factors (Lyssenko et al. 2008). These risk factors include obesity and overweight (with a body mass index (BMI) of 30 or more), family history, old age (people over the age of 40), ethnicity, and physical inactivity (Lewis et al. 2010)

There is a complementary role for conventional factors modulating the genetic predisposition of such a complex disease, that emerged from the National Health Service Diabetes

Prevention Program (DPP) (Wise 2018). In a large randomized cohort of lifestyle interventions including weight loss, exercise and dietary modification, 58% of the overweight adults with mean BMI 31kg/m^2 achieved a reduction in the incidence of T2D (Tudies et al. 2012). In another study results from lifestyle intervention school-based programs suggested that the reduction in the prevalence of overweight and obesity among adolescents may decrease the risk of childhood-onset of T2D (The HEALTHY Study Group 2010).

Twin studies have shown that the concordance rate of T2D in monozygotic twins is approximately 70% compared with 20% to 30% in dizygotic twins (Medici et al. 1999). Furthermore, the lifetime risk of developing the disease in individuals if one parent is affected is about 40%, while it increases to 70% if both parents are affected (Köbberling & Tillil 1982). In addition, a study of parental transmission of T2D showed that the influence of first-degree relatives in the risk of developing T2D is varied. The risk of developing the disease in offspring who have one diabetic parent is about 3.5-fold higher and is 6-fold higher if both parents are affected compared to the general population (offspring without parental diabetes) (Meigs et al. 2000). However, these risk ratio figures vary in different cohort and population studies. The studies performed in (Al-Sinani et al. 2014; Medici et al. 1999; Köbberling & Tillil 1982; Meigs et al. 2000) indicate that there is consistent evidence to show which genetic determinants are an important factor in modifying an individual predisposition to T2D. Thus, the potential influence of genetics on T2D risk is significant with predicted heritability between 20 and 70 percent.

A relatively small proportion of diabetic cases occur due to a mutation in a single gene. These cases are classified as either monogenic diabetes, neonatal diabetes mellitus (NDM), or maturity-onset diabetes of the young (MODY) (Philippe et al. 2015). T2D on the other hand is known to be a polygenic disorder. This indicates that T2D occurs due to complex interactions between multiple SNPs or genes. Over the past decade, advances in genotyping technology have made it possible to discover the genetic constituents associated with T2D.

Several loci, identified before the widespread use of genome-wide association studies (GWAS), include calpain 10 (*CAPN10*) and transcription factor 7 like 2 (*TCF7L2*) - genes that were discovered using linkage analysis (Prasad & Groop 2015). These were the only two genes associated with T2D. Linkage analysis failed to detect genes involved in complex polygenic disorders. In candidate gene studies several genes have been found to be associated with T2D including Peroxisome proliferator-activated receptor gamma (*PPARG*), Insulin receptor substrate 1 (*IRS1*) and (*IRS-2*), potassium inwardly rectifying channel, subfamily J, member 11 (*KCNJ11*), and Wolfram syndrome 1 (*WFS1*) (Ali 2013). These two approaches have detected a number of T2D risk genes. However, alternative techniques are required to detect variants that candidate gene and linkage analysis cannot identify.

To date there are more than 120 susceptibility loci for T2D that have been identified using GWAS (Prasad & Groop 2015; Wang et al. 2016). A review conducted by Prasad and Groop (Prasad & Groop 2015) provides a complete list of T2D risk SNPs. Genetic markers identified in pre-GWAS studies have also been confirmed by GWAS. *TCF7L2*, which was proved to be associated with T2D via linkage studies, is the most significant and repeatedly replicated gene discovered via GWAS (Ali 2013). Several other genes have been consistently identified among multiple populations as being associated with T2D such as Hematopoietically-expressed homeobox (*HHEX*), Solute carrier family 30 (zinc transporter) member 8 (*SLC30A8*), Cyclin-Dependent Kinase Inhibitor 2A/B (*CDKN2A/B*), and Insulin-like growth factor 2 mRNA-binding protein 2 (*IGF2BP2*) (Tudies et al. 2012).

The discovery of these genes has served as a rigorous foundation to understand the regulation of glucose metabolism and the development of T2D. It is hoped that these investigations could yield a comprehensive understanding of the mechanisms that regulate insulin secretion and action and help to understand the changes that cause an increased risk to T2D. These findings may ultimately lead to improve diagnostic testing, prevention of disease onset, and

future treatments underpinned by advances in personalized medicine. This could help mitigate the progression of the disease and its complications.

2.5 Genetic Association Studies

Genetic association studies are used to detect genetic susceptibility (or susceptibility loci) to specific medical disorders (Lewis & Knight 2012). There are several approaches in genetic association studies: linkage studies (Ott et al. 2015), candidate gene (CG) studies (Foulkes 2009), and genome-wide association study (GWAS) (Bush & Moore 2012). All approaches are based on the co-inheritance of genetic markers associated with disease allele.

2.5.1 Linkage Studies

Linkage studies focus on identifying rare alleles (variants) correlated with the phenotype of interest within a pedigree (Ott et al. 2015). The study design for this approach is family-based association which uses genotypes of candidate individuals with his/her parents. This type of study is more costly than other approaches and parents need to be part of the study (Ott et al. 2015). Despite these limitations, family-based association studies are immune to population stratification (SNP allele frequencies vary among different population ancestry) that occurs in other approaches. Family-based association studies can offer a method to assess mendelian genetic errors (Teare & Koref 2014).

2.5.2 Candidate Gene Studies

Candidate gene studies (CG) focus on identifying risk alleles associated with a particular disease within population studies (Patnala et al. 2013). The study design for this approach is based on case-control subjects. In case-control studies, the investigators compare DNA samples of individuals who have a disease (cases) with individuals who do not have the disease (controls). The candidate gene approach uses genes previously identified and thus this approach is initiated with prior knowledge of gene function (Patnala et al. 2013). While CG has proved to be useful, it fails to discover new genes or combinations and interacting genes (Amos et al. 2011). In addition, the fact that unrelated case-control samples are

recruited makes the study more susceptible to population stratification issues that occur due to variable ancestral backgrounds. This can lead to false positive outcomes.

2.5.3 Genome-Wide Association Studies

With the completion of the Human Genome Project in 2003 (Green et al. 2015) and the International HapMap Project in 2005 (Gibbs et al. 2003; Manolio & Collins 2009), genome-wide association studies are more widely used in genetic studies. GWAS have been used in a broad range of disease type studies to detect statistically significant SNPs and investigate the genetic architecture of human disease in the entire genome (Bush & Moore 2012). GWAS are a population-based approach where the study design utilises unrelated case-control observations. In this thesis GWAS data is utilised, consequently a more in-depth discussion on GWAS is presented below.

2.6 Genome-Wide Association Studies Overview

The primary objective in GWAS is to identify genetic risk factors for common complex diseases (Bush & Moore 2012). Proponents claimed that GWAS would significantly enhance our understanding of genetic-based determinants for common complex diseases, such as, T2D, Schizophrenia, Epilepsy, Obesity, Cardiovascular Disease, and Hypertension (Bush & Moore 2012; Guo et al. 2014). More specifically, to determine if SNPs occur more frequently in individuals affected with a particular disease, than in individuals unaffected by the disease. In other words, GWAS was developed to discover ***direct*** and ***indirect*** associations between SNPs and specific diseases (Bush & Moore 2012; Balding 2006). Direct (causal) association refers to the SNP that directly influences the biological configurations found to be statistically associated with a phenotype (Balding 2006). Indirect (non-causal) association describes influential SNPs that are not directly genotyped (Balding 2006). There are other SNPs known as tag SNPs are genotyped and statistically associated to the trait, located in a region of high ***linkage disequilibrium (LD)*** with the influential SNPs (Bush & Moore 2012). LD is a non-random association between allelic variants at different

loci on the same chromosome in a given population, typically the two alleles are either inherited or correlated (Lewis & Knight 2012). Therefore, significant SNPs from GWAS are not always assumed to be causative variants but instead they may require further investigation to map the actual location of influential SNPs. In other words, significant SNPs in a genetic association study are more likely to be indirect. Figure 2.4 illustrates direct and indirect associations between SNPs and the disease phenotype.

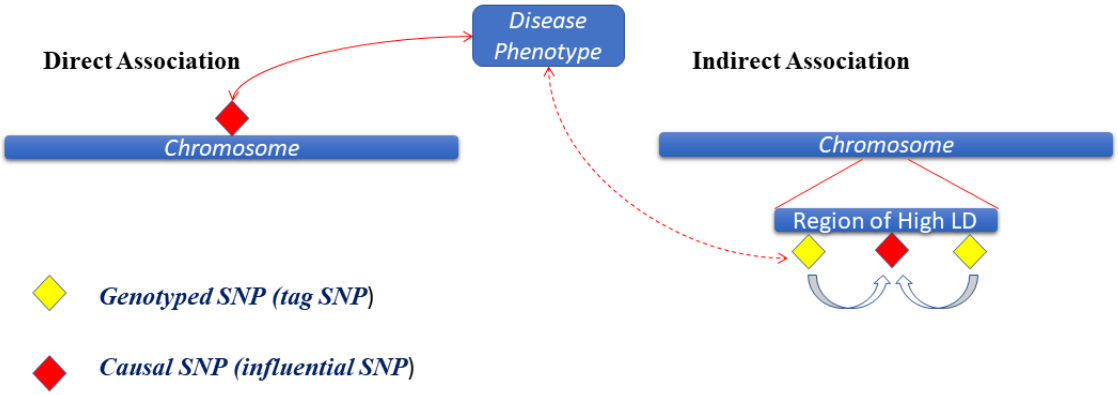


Figure 2.4: Direct and Indirect Association

Since GWAS is a population-based method that consists of a large number of unrelated samples (case-control), most GWAS are well developed to find associations with common variants (>5%) and less for detecting low allele frequency variants (Sebastiani & Solovieff 2011; Fadista et al. 2016). This highlights the ***common disease - common variant (CD-CV)*** hypothesis (Shields 2011) indicating that common diseases are probably influenced by genetic markers that are relatively common in the population. Under this hypothesis, phenotype associated alleles are more likely established using common genetic markers, specifically SNPs that have been detected and compared with affected and unaffected samples. However, there is disagreement among researchers as they suggested that common diseases cannot be caused by common alleles but rather they are influenced by rare variants (Cirulli & Goldstein 2010).

Genotyping technology has facilitated rapid progress in genome-wide association studies. These technologies have been specifically designed to assay more than one million SNPs.

However, it is now possible, to sequence the entire human genome within a single day (Behjati & Tarpey 2013). The most recent DNA sequencing technology is Next Generation Sequencing (NGS) (Behjati & Tarpey 2013), which provides tools to sequence DNA and RNA. NGS is cost effective with rapid performance compared to the previously used Sanger Sequencing Technology (Pareek et al. 2011; Xuan et al. 2013). Currently, there are two platforms utilized in GWAS - the Illumina and Affymetrix platforms (Bush & Moore 2012). Each technique offers a different approach to measure and detect genomic variation (alleles).

2.6.1 Choice of Significance Test

GWAS studies often test millions of independent SNPs for associations with particular diseases. Thus, to find SNPs that are statistically significant and to limit type I errors (false-positives) a very stringent statistical threshold is used $p < 5 \times 10^{-8}$ (Panagiotou & Ioannidis 2012). This threshold is called a Bonferroni-corrected genome-wide significance threshold (Panagiotou & Ioannidis 2012) and it has become a standard in most GWAS. The Bonferroni correction offers a method to control family-wise error rates (FWER) (Zeng et al. 2015). FWER is the probability of rejecting at least one null hypothesis when all the nulls are correct (Zhang et al. 2012). An SNP is considered statistically significant if its p-value is less than the Bonferroni-corrected genome-wide significance level. Bonferroni correction is a conservative threshold (Zeng et al. 2015) and it may be highly likely that none of the SNPs under investigation reach such a small threshold. Therefore, as recommended in Duggal's study (Duggal et al. 2008) a suggestive association threshold $p < 1 \times 10^{-5}$ should be utilized. A suggestive threshold is less conservative, and it is generally used to detect SNPs for consideration in follow-up studies.

2.6.2 Challenges Associated to GWAS Approach

Although, GWAS have significantly impacted the field of human genetics, there are still challenges associated with computational and statistical methods. These challenges include scalability, missing markers and complex traits (Zhang et al. 2012). Usually GWAS datasets

contain millions of SNPs across thousands of individuals. Therefore, to perform GWAS, the algorithms need to be extremely efficient and scalable to avoid issues with computational resources and to minimise the time required to conduct GWAS. In addition, missing markers need to be appropriately handled. One approach to handle missing markers is to use imputation (Howie et al. 2012) to impute unidentified markers using an accessible SNPs database such as the 1000 Genome (Adam 2015) and International HapMap Projects (Gibbs et al. 2003).

Another major limitation with GWAS is that, while being successful at detecting single SNPs relating to phenotype traits, its ability to find SNPs associated with complex traits/diseases. Complex traits are more likely to be affected by multiple SNPs which separately may have a weak association with the disease but cumulatively have a much more important part to play in the development of complex diseases. In this case it is extremely difficult for an SNP with low marginal effects to be identified using single-locus methods. Consequently, an alternative approach such as multi-locus analysis needs to be conducted (Bush & Moore 2012).

2.6.3 Hypothesis Testing for GWAS

In biomedical research the most popular tool for statistical analysis is ***hypothesis testing*** (Penrod & Moore 2014). Hypothesis testing is used to determine if the evidence that is available in the data is adequate to conclude that a particular condition (the question being asked) is true for that population (Taeger & Kuhnt 2014). There are two contrasting hypotheses relating to the population - the ***null hypothesis*** and the ***alternative hypothesis***. The null hypothesis is tested, and based on the outcome, is either acceptance or rejection. The alternative hypothesis on the other hand is the hypothesis that challenges the null hypothesis (Taeger & Kuhnt 2014).

Generally, significance testing, also known as a p-value, is conducted. The p-value is defined as the probability of seeing a value of a test statistic as equal to or larger than the one that

was observed in a dataset, assuming the null hypothesis is true (Bush & Moore 2012). In academic research the significance threshold (α) of 0.05 is widely used. Typically, this indicates that the analysis probably has a type I error rate (false positive) of 5%. This in turn means that there is a 5% chance of making an error in rejecting the null hypothesis when it was in fact true. Meanwhile, type II errors (false negatives) also need to be measured to calculate the probability of accepting the null hypothesis when it is in fact false. The statistical power of the study is calculated using the formula (1-type II error). In general, if the power of the study is 80% or more, this indicates that the study is sufficiently powered. More specifically, saying that the power is 80%, means that 80% of the time the null hypothesis will be rejected when it is false (Penrod & Moore 2014).

For GWAS the null hypothesis represents a situation where there is no association between the genotype and phenotype of interest. The alternative hypothesis on the other hand indicates that there is at least a single SNP (genotype) associated with the disease of interest in the given dataset (Bush & Moore 2012).

2.6.4 False Discovery Rate in GWAS

The traditional multiple hypothesis testing based on FWER provides a strong control on false positives however it is too conservative - using very small p-value threshold. In GWAS setting, the main goal is to identify as many true positive findings as possible, while controlling against any single false positive occurring. The false discovery rate (FDR) method proposed by Soric (Soric 1989) is designed to measure such type of trade-off. FDR is used to evaluate the statistical significance of multiple hypothesis tests based on the proportion of false positives among the claimed rejected hypotheses (positives). Storey and Tibshirani (Storey & Tibshirani 2003) proposed the q-value statistical method to estimate FDR based measure of significance. They defined q-value to be the minimum FDR at which the particular test (p-value) is called significant: $q(p_i) = \min_{t \geq p_i} FDR(t)$ where t is the threshold and all $t \geq p_i$.

The FDR and its related estimation method (q-value) have been widely used in GWAS analysis (Hind et al. 2017; Kotnik et al. 2018; LeBlanc et al. 2016; Heller & Yekutieli 2014) in which a number of individual hypothesis tests are performed simultaneously, resulting in combinations of true and false null hypotheses.

2.6.5 Visual Presentation for GWAS

Data visualization tools in GWAS facilitate the interpretation of genome-wide association study outcomes. Various visual tools for GWAS have been developed which include Manhattan and Q-Q plots (Turner 2018).

2.6.5.1 Manhattan Plot

Manhattan plots are designed to visualize GWA significance levels (p-values) by chromosome position. This plot highlights any regions of significance. Manhattan plots are generated by plotting the p-value in the vertical axis which represents the $-log_{10}$ scale and the physical position of the SNPs in each chromosome in the horizontal axis. This plot uses a Bonferroni corrected genome-wide significance threshold to highlight statistically significant SNPs, which highlight potential disease-associated SNPs.

2.6.5.2 Quantile-Quantile Plot

Quantile-Quantile (Q-Q) plots are used to show the relationship between the expected distribution of p-values (null hypothesis) and the observed distribution of p-values in test statistics. Q-Q plots are typically used to detect if there is any evidence of systematic bias such as population stratification. Doing so is good practice in robust analysis that assures the quality and the validity of the data used in the study. Q-Q plots are produced by plotting the observed p-value obtained in test statistics (Chi-Squared statistic or logistic regression test) against the theoretical expected values under the null hypothesis of no association. The plot should go along the diagonal linearly with a slight deviation towards the top. In a scenario where there is evidence of population stratification the plot may deviate too early from the diagonal.

2.7 Quality Control and Filtering for GWAS Data

Quality control (QC) is used in GWAS to identify and eliminate low quality DNA samples and markers prior to association analysis (Laurie et al. 2010). QC is a critical element in GWAS analysis, and it is essential to avoid spurious GWAS results. There are two fundamental areas of QC: Individual-Based Quality Control measures and Marker-Based Quality Control measures (Perreault et al. 2013) as explained in the following sections.

2.7.1 Individual-Based Quality Control

Individual-based QC is performed to select and discard subjects (individuals) who do not meet specific criteria for GWAS analysis. There are four essential measures required which include Gender Ambiguity (inconsistency) check, Missingness Rate per Individual, Duplicated or Relatedness Individuals, and Population Stratification.

2.7.1.1 Gender Ambiguity Check

Gender ambiguity typically arises from sample handling errors. Homozygosity rate calculation can be used to detect individuals, who have been reported as male/female, but were their existing sex information does not match with genotype gender information. This calculation is applied across all X-chromosome markers for each individual in the study and compared to the expected homozygosity rate (less than 0.2 for female, more than 0.8 for male) (Anderson et al. 2010).

2.7.1.2 Missingness Rate per Individual

Missingness rate per individual also known as an individual call rate or genotyping efficiency per individual is an indicator of individual DNA quality. The call rate per individual presents the percentage of SNPs genotyped in each sample (S. Turner et al. 2011). A low genotyping call rate describes an issue with a poor quality DNA sample or low sample concentration. Samples with poor genotyping efficiency need to be removed. The recommended call rate threshold is between 98 and 99 percent (S. Turner et al. 2011). This threshold is an approximation and the exact threshold depends on various factors (i.e.

genotyping platform and DNA sample quality) and this may vary between different studies. The call rate threshold depends on the objective of the study whereby a balance between increasing genotypic efficiency and sample size is considered.

2.7.1.3 Individuals Duplicated or Relatedness

Duplicated and sample relatedness is measured to examine the identity and pedigree integrity between individuals by comparing genomic data with self-reported relationships among subjects in the study. The family relationship between two samples can be quantified by estimating the degree of identity-by-descent (IBD) - in other words the extent to which alleles among relatives are shared (Anderson et al. 2010). IBD is defined as the segments of the genome that come from the same ancestral source - they are copies of the same ancestral chromosome (Thompson 2013). Typically, the expected IBD sharing degree for a related pair is estimated based on their pedigree relationship. Thus, duplicated samples or monozygotic twins share two alleles, first degree relatives are more likely to share half of their alleles, second degree relatives share 0.25, third degree relatives share 0.125 and unrelated samples share zero alleles (Duggirala et al. 2015; Browning & Browning 2012).

In population based case-control association studies, independence between observations is assumed (Bush & Moore 2012) - in other words the observed genotypes come from unrelated samples. If duplicated, or first or second-degree relatives are found then the distribution of the samples' genotypes will not be appropriately represented within the population. This over representation of genotypes may cause bias in the study and increase type I and type II errors. Therefore, the extent of relatedness in the entire population must be reduced to second degree relatives (0.25) (Anderson et al. 2010).

2.7.1.4 Population Stratification

Population stratification occurs when case-control study samples contain multiple groups of individuals who do not share the same genetic ancestry (S. Turner et al. 2011). When this is the case, studies carry different allele frequencies due to population diversity as each

population has a unique genetic fingerprint. Thus, allele frequency diversity between individuals is not necessarily associated with any specific disease causing spurious associations (Cardon & Palmer 2003). This is the major cause of confounding factors in GWAS analysis (Anderson et al. 2010). Therefore, in order to avoid introducing bias to the study due to population stratification, it is important to conduct the analysis using a dataset from a relatively homogenous population (Bush & Moore 2012).

There are a number of methods to detect and characterise population stratification in GWAS. These include Genomic Control (GC), Structured Association, and Principal Component Analysis. The GC (S. Turner et al. 2011) method is based on calculating and estimating an inflation factor λ and dividing and adjusting all of the test statistics downward by this inflation factor. Inflation factor values greater than 1 indicate inflation, therefore population stratification exists, and correction is applied to bring the value closer to 1.

The structured association (Sebastiani & Solovieff 2011) method is a model-based clustering technique that groups samples into clusters using a subset of SNPs and performing association tests among each inferred group. The method can identify individuals that do not cluster with the majority of samples and eliminate these individuals from the study.

Principal component analysis (PCA) (Hotelling 1933) is a multivariate statistical approach used to summarise and produce principal components of uncorrelated variables obtained from a data matrix consisting of samples with a number of potentially correlated variables. PCA is a widely used method in GWAS due to its computationally convenient manner (Anderson et al. 2010). Typically, a PCA model is constructed using genotype data obtained from populations of known ancestry such as the reference panel of HapMap phase III data which contains four different ancestral populations including Europe, Asia (Chinese and Japanese populations), and Africa. The method is used to cluster samples from GWA data in terms of ancestry alongside the HapMap samples to produce principal component scores for GWA samples.

2.7.2 Marker-Based Quality Control

Marker-based quality control also consists of several key steps: identifying SNPs with excessive missing genotype, SNPs showing a significant deviation from Hardy-Weinberg Equilibrium (HWE), and finally identifying markers with very low Minor Allele Frequency (MAF). Removing SNPs from the study is critical as each SNP may correlate with disease risk (Laurie et al. 2010). Therefore, caution needs to be taken when deciding what thresholds to use to remove SNPs from the study.

2.7.2.1 Missingness Rate Per Marker

Missingness rate per marker also known as marker genotyping efficiency or call rate is an informative indicator of marker quality. The call rate per marker represents the proportion of individuals with a genotype call for each SNP (Weale 2010). Typically, this step is conducted to remove SNPs if they are missing in a large number of samples. This is a good indicator for a poor quality marker that is more likely to induce false associations. The authors in (Donaldson et al. 2016) indicate that the recommended threshold for removing markers with low call rates is 98-99 percentage. This means that if the SNP is missing in more than 1 or 2 percentage of samples, it will be removed from the study. However, this recommended threshold may vary between studies.

2.7.2.2 Minor Allele Frequency (MAF)

Minor allele Frequency refers to the frequency of the less common allele at a given SNP (Bush & Moore 2012). More specifically, if a particular SNP (for example C) appears in 30% of a population that means this SNP is classified as a minor allele, while the more common allele (major allele) can be found in 70% of the same population (Bush & Moore 2012). Filtering SNPs based on MAF is an important step toward increasing statistical power. Generally, the statistical power for rare SNPs is considerably low therefore it has been recommended to exclude any extremely rare SNPs (Winkler et al. 2014). For instance, if an SNP demonstrates variation in only 1 of the 82 samples, this proportion is inadequate statistically and should be discarded from the study.

Furthermore, to remove SNPs with MAF, the threshold limit is chosen by considering the samples size in the study. In some instances, SNPs have been removed for which the MAF is less than 1% while in other studies with a small sample setting a higher threshold such as 5% as a cut-off point is chosen (Tabangin et al. 2009).

2.7.2.3 Hardy-Weinberg Equilibrium (HWE)

Hardy-Weinberg Equilibrium assumes that, allele and genotype frequencies remain constant from one generation to the next, in the absence of other evolutionary influences such as mutation, natural selection, migration, and associative mating (Wigginton et al. 2005). Departure from this equilibrium can indicate the occurrence of potential genotyping errors, and the existence of population stratification (Graffelman & Weir 2016). In study-based case-control approaches, it is necessary to conduct HWE in controls separately as a departure in cases can be indicative of true association to the trait under investigation (Anderson et al. 2010). In the literature, various significance thresholds between 0.001 (McCaughan et al. 2013) and 5.7×10^{-7}(Burton et al. 2007) to identify markers in HWE have been reported. However, values do vary between studies. Checking markers for HWE is the last step in quality control analysis and is a common practice to remove SNPs that show deviation from HWE.

Table 2.1 presents the commands used in PLINK (Purcell et al. 2007) to fulfil QC for samples and markers prior to association analysis.

Table 2.1: QC Command for Samples and Markers

Command	**Description of the Command**
--check-sex	Check for sample identity problems
--genome	Examine pedigree integrity
--missing	- Check genotype efficiency for each sample - Check genotype efficiency for each marker
--mind	Remove samples with low call rate
--geno	Remove markers with low call rate
--freq	Report minor allele frequency for each marker

--maf	Remove extremely rare markers
--hardy	Examine markers for Hardy-Weinberg Equilibrium
--hwe	Remove markers showing departure from Hardy-Weinberg Equilibrium

2.8 Association Analysis

Association analysis in case-control studies compares the frequency of alleles or genotypes at genetic marker loci (SNP) between cases and controls in a given population (Clarke et al. 2011). This analysis is used to detect statistically significant differences in the frequency of alleles between individuals in the study. These alleles (genetic markers) are used to test associations with the phenotype (disease trait) (Clarke et al. 2011). In other words, association analysis is a series of single-locus statistical tests, that explore each SNP separately and their likely association with a particular phenotype.

Genetic association mapping can be performed using several statistical methods including Pearson's Chi-Squared test (x^2), Fisher's exact test, linear model test, logistic test, and transmission/disequilibrium test (TDT) (Cortes et al. 2013). The use of one of these tests depends on the type and the size of a dataset where the dataset is either family-based or population-based (Zhang et al. 2012). For example, Fisher's exact test is more appropriate with small sample sizes (Zhang et al. 2012) compared with Pearson's Chi-Squared test (x^2) which is often used with much bigger sample sizes. TDT (Montana 2006) is used in family-based association testing whereas, for population-based association (unrelated samples), Pearson's Chi-Squared testing (x^2) and linear/logistic regression are used.

2.8.1 Statistical Methods of a Case-Control Study

In a case-control study, the association between a single SNP and disease status can be based on standard contingency table tests for independence (Balding 2006). Contingency tables are widely used to display genetic markers (SNPs) in the format of genotype or allele frequency by disease status (case-control) (Clarke et al. 2011).

For each single SNP:

$$\begin{cases} a\ is\ a\ minor\ allele\ in\ N\ case-control \\ A\ is\ a\ major\ allele\ in\ N\ case-control \\ for\ a\ and\ A, \exists\ F\ where\ F\ is\ a\ contingency\ table\ of\ genetic\ model \end{cases}$$

F can be represented (X. Wang et al. 2016) as:

$$\begin{cases} 2 \times 3\ \ table\ of\ N\ case-control\ by\ genotype\ counts\ (AA, Aa, aa) \\ 2 \times 2\ table\ of\ 2N\ case-control\ by\ allele\ counts\ (A, a) \end{cases}$$

The contingency table for case and control analyses using genotypic and allelic genetic models of penetrance is summarized in Table 2.2, where DF represents the degrees-of-freedom in genetic models and is calculated based on the (number of rows in the contingency table – 1) × (number of columns in the contingency table – 1) (Bland 2015).

Whereas O_{ij} refers to the observed frequency of individuals in cases and controls, i refers to the row number and j represents the column number. For example, in a genotypic model test O_{11} refers to the observed frequency of individuals in cases when genotype *aa* occurs.

Table 2.2: Contingency Table for Genetic Models

Test	**DF**	**Contingency table representation**			
Genotypic test	2		***aa***	***Aa***	***AA***
		Cases	O_{11}	O_{12}	O_{13}
		Controls	O_{21}	O_{22}	O_{23}
Allelic test	1		***a***	***A***	
		Cases	O_{11}	O_{12}	
		Controls	O_{21}	O_{22}	

2.8.2 Association Analysis Method

The principal formulation for association testing is defined in Definition 1.

Definition 1. Let $\{X_1, \dots, X_u\}$ be a set of U SNPs for N individuals. Let phenotype $= \{y_1, \dots, y_n\}$. Assume the genomic data for each SNP has minor allele a and major allele A. To represent the homozygous major allele AA, heterozygous allele Aa and homozygous minor allele aa, numbers such as 0, 1, and 2 are used respectively. Consequently, $X_{un} \in \{0, 1, 2\}, (1 \leq u \leq U, 1 \leq n \leq N)$.

For case-control studies the phenotype can be represented as a binary variable, 0 referring to controls and 1 referring to cases. The association test within genetic data is to test for the null hypothesis (no association between the SNP and phenotype of interest (disease status)) in the contingency table. Pearson's Chi-Squared test (x^2) can be used to test for association. The principle of Chi-Squared test (x^2) is to compare the distributions of observed and expected values with their contingency tables (Zhongxue Chen et al. 2014). Chi-Squared test summarises the differences between the observed frequency values and the expected frequency values at single genetic marker loci (SNP) across cases and controls. The calculation of a Chi-Squared test (x^2) is formulated in Definition 2.

Definition 2. The standard Chi-Squared test for the independence of rows and columns in the contingency table considering a genotypic model for association (X. Wang et al. 2016) is defined as:

$$x^2 = \sum_i \sum_j \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{2.1}$$

where E_{ij} is the expected frequency of the allele or genotype in cases and controls and is defined as:

$$E_{ij} = \frac{O_{i.} O_{.j}}{N}$$

where $$O_{i.} = \sum_j O_{ij} \tag{2.2}$$

and $$O_{.j} = \sum_i O_{ij}$$

where O_{ij} refers to the observed frequency of individuals whose $X_u$equals i and Y equals j.

Following the calculation of a Chi-Squared test, the p-value for Chi-Squared is determined based on the degrees of freedom used in the test. Formally, the p-value is defined as the probability of seeing a value (Chi-Squared statistic test) as equal to or larger than the one that was observed in a given dataset, assuming the null hypothesis (no association) is true

(Bush & Moore 2012). More specifically, the p-value represents the degree of association between the SNP and the phenotype across the entire sample set.

2.8.3 Logistic Regression

Logistic regression (Cox 1958) is defined as a statistical method for predicting binary outcomes. Logistic regression modelling can be used to analyze the contingency table for independence, where disease status is a binary trait (0/1) with 0 indicating a control and 1 indicating a case (Clarke et al. 2011). Let $Y \in \{0,1\}$ be a binary variable for case and control status and let $X \in \{0,1,2\}$ be a genotype at a particular SNP. Assuming that 0, 1, 2 represent homozygous major allele AA, heterozygous allele Aa and homozygous minor allele aa respectively. Logistic regression modelling is therefore given as (X. Wang et al. 2016):

The conditional probability of $Y = 1$ is

$$\theta(X) = P(Y = 1|X) \tag{2.3}$$

The logit function, which is the inverse of the sigmoidal logistic function, is represented as:

$$logit\big(\theta(X)\big) = \ln\left(\frac{\theta(X)}{1-\theta(X)}\right) \tag{2.4}$$

The logit is given as a linear predictor function as follows

$$logit\big(\theta(X)\big) = \beta_0 + \sum_i \beta_i . X_i \tag{2.5}$$

where β_0 represents the intercept and β_i denotes the regression coefficient.

Logistic regression modelling is a predominant method for investigating each SNP separately and to capture the linear associations between SNPs and the phenotype. Logistic regression can be readily expanded to allow for covariates such as other SNPs, sociodemographic and clinical factors. Other genetic models including Allelic, Genotypic, Dominant, and Recessive are available however logistic regression is the preferred approach.

2.8.4 Odds Ratio of Disease for Case-Control Study

In a case-control study, the strength of an association is measured by the odds ratio (OR) (Clarke et al. 2011). OR is the ratio of the odds of disease in the exposed group (risk mark-

positives) compared with those in the non-exposed group (risk mark-negatives) (Clarke et al. 2011). For example, based on the variables provided in Table 2.2, the allelic OR measure for the association between disease and allele, is the odds of disease if allele *A* (major allele) is carried compared with the odds of disease if allele *a* (minor allele) is carried. The following formula is used to estimate the allelic OR for allele *A* (Li 2007).

$$OR_A = \frac{odds\ of\ disease\ with\ A\ allele}{odds\ of\ disease\ with\ a\ allele} \tag{2.6}$$

Based on the variables in the contingency table for an allelic test, the OR is estimated as follows:

$$OR_A = \frac{(O_{12}/O_{22})}{(O_{11}/O_{21})} \tag{2.7}$$

therefore

$$OR_A = \frac{O_{12}O_{21}}{O_{11}O_{22}}$$

The strength of the association for allele *A* is estimated based on the value of OR as explained in Table 2.3. An OR $= 1$ signifies that the condition under study appears equally in both groups (case and control). However, an OR > 1 indicates that the condition occurs in the case group more than in the control group. An OR < 1 indicates that the condition is more likely in the control group.

Table 2.3: The Description of Odds Ratio Numerical Value

Odds Ratio	Description of the OR Value
$OR_A = 1$	Indicates no association between genotype and disease
$OR_A > 1$	Indicates that there is a risk association between allele *A* and disease
$OR_A < 1$	Indicates a protective association for allele *A*

2.9 The Application of GWAS into T2D

Recently, several GWAS and meta-analysis studies have been performed in different cohorts and/or ethnic groups. The studies have described associations between genetic variants and

T2D in different populations. In Qiu's study (Qiu et al. 2014), association analysis was performed in a case-control study to investigate the role of potassium inwardly-rectifying-channel, subfamily-J, member 11 (KCNJ11) variation particularly E23K polymorphism (rs5219) in susceptibility to T2D. In this study, 56,349 T2D cases, 81,800 controls, and 483 family trios were collected from 48 published studies. The statistical methods used within the approach included The Standard Q-statistic test, and subgroup analysis (ethnicity, sample size, BMI, age and sex) to explore whether variation in these studies was due to heterogeneity. Furthermore, the odds ratio with its 95% confidence interval of KCNJ11 E23K polymorphism was calculated to measure the association with T2D. Dominant and Recessive genetic models were applied to examine the association between the KCNJ11 E23K polymorphism and T2D risk. The results suggest that the KCNJ11 E23K allele for rs5219 (OR = 1.12, $p < 10^{-5}$) was significantly associated with T2D risk. For heterozygous and homozygous alleles with (OR = 1.09, $< 10^{-5}$) and (OR = 1.26, $p < 10^{-5}$) respectively, a significant increase of T2D risk was observed. This study suggested that there is a modest but statistically significant effect of the 23K allele of the rs5219 polymorphism in susceptibility to T2D, particularly in East Asians and Caucasians. The contribution of these genetic variations to T2D in other ethnic populations (e.g. Indian, African, American, Jews, and Arabian), appear to be relatively low. For Dominant and Recessive genetic models, similar results were obtained.

Seven novel T2D susceptibility loci were identified in Mahajan's work (Mahajan 2014) using several published meta-analysis GWAS. The studies contain 26,488 cases and 83,964 controls of East Asian, South Asian, European, Mexican, and Mexican American ancestry. By combining GWAS across ancestry groups using Trans-ethnic meta-analysis, it was possible to observe significant improvement in the detection of novel complex trait loci for the disease. Furthermore, with this approach, there was an enhancement in the fine-mapping

resolution of causal variants by leveraging differences in local linkage disequilibrium structure between ethnic groups.

In Phani's study (Phani et al. 2014), the authors performed a case-control study using 400 T2D cases and controls from a South Indian population to analyse and outline the association of Potassium inwardly rectifying channel, subfamily J, member 11 (KCNJ11) genes and the risk of T2D. The study conducted a systematic review and meta-analysis for KCNJ11 (rs5219) polymorphism in 3,831 cases and 3,543 controls that were aggregated from 5 published reports from South Asian and East Asian populations. In this case OR was used as a measure of association of KCNJ11 polymorphisms (rs5219, rs5215, rs41282930, rs1800467) and T2D with its corresponding 95% Confidence Interval (CI). Moreover, Cochran's Q, I^2 statistics were utilized to assess heterogeneity within and between the eligible studies. The resulting evidence therefore showed that KCNJ11 rs5215, C-G-C-C haplotype and two loci (rs5219 vs rs1800467) had a significant association with T2D. However, Copy Number Variations (CNV) analysis did not show significant variation between T2D case and control subjects. Furthermore, meta-analysis of the study suggested that KCNJ11 (rs5219) polymorphism is associated with the risk of T2D in East Asian and Global populations but not in the South Asian population.

In Cheema's work (Cheema et al. 2015) the authors performed a case-control study to investigate the differences in the association of peroxisome proliferator activated receptor, gamma, coac- tivator 1 alpha (PPARGC1A) genes and T2D risk among populations from African origins. The study includes adults aged >30 years old from African Americans (cases = 124, controls = 122) and Haitian Americans (cases = 110, controls = 116). The statistical methods used within this study included standard summary statistics such as the Chi-Squared goodness-fit test that was employed to check genotype counts for each SNP for Hardy-Weinberg Equilibrium. Furthermore, the t-test was used to compare demographics (age, sex, BMI, smoking status) between cases and controls, and clinical information. Logistic

regression was also used to calculate adjusted and unadjusted OR with a 95% CI. The results indicated that SNP rs7656250 (OR = 0.22, p-value = 0.005) and rs4235308 (OR = 0.42, p-value = 0.026) showed protective association with T2D in Haitian Americans using adjusted logistic regression. While in African Americans, SNP rs4235308 (OR = 2.53, p-value = 0.028) showed significant risk association with T2D. Furthermore, the study concluded that the differences in genetic associations of PPARGC1A with T2D among Haitian Americans and African Americans were due to the contribution of differences in ancestry (Black race).

The reproducibility of previously identified single SNP associations in case-control studies of T2D among the Singapore Chinese population was conducted in Chen's study (Zhanghua Chen et al. 2014). The study contained 2338 T2D cases and 2339 controls with 507,509 genotyped SNPs. The statistical methods employed included two sample t-tests to compare the mean differences for variables with normal distributions, the Wilcoxon rank sum test to compare median differences for variables with skewed distributions, and Pearson's Chi-Squared test (x^2) to test the different frequency distributions for categorial variables between T2D cases and controls. Furthermore, the authors interrogated the combined effects of several loci on disease risk using the National Human Genome Research Institute (NHGRI) GWAS Catalog to identify SNPs associated with T2D. Among the 55 indexed SNPs obtained from the NHGRI GWAS Catalog, 15 SNPs were replicated (at p-value < 0.05). Moreover, Conditional fine-mapping analysis was used to search regions near GWAS alleles for additional and new disease associations. SNPs in regions $\pm$ 100 kb around each index SNP were interrogated for associations with T2D. The results highlight two SNPs located in linkage disequilibrium close to rs10923931 and 5 new candidate SNPs located close to rs10965250 and rs1111875. Nonetheless, these SNPs only explain a small proportion (2.3%) of the disease risk in the Singapore population.

In another work conducted by Li (Li et al. 2013), the authors performed a case-control GWAS and replication study in the Chinese Hans population. The study comprised three-

stage GWAS independent sample sets. In the first GWAS, 1999 T2D case and 1976 nondiabetic control subjects of Chinese Hans from Shanghai and Beijing ethnic populations were included in the study with 657,366 genotyped SNPs. In the second study, a replication study was conducted using 96 SNPs selected from the first GWAS analysis. In phase three, an independent Chinese Hans population from Beijing, Guizhou, and Hubei with 6570 T2D cases and 6947 controls subjects were used. In the last stage, 10 candidate SNPs selected based on the findings of a large-scale GWAS that combined the first and second phase analysis were used for a second replication study. The study contained 3410 T2D cases and 3412 controls of Chinese Hans from Shanghai in addition to 6952 T2D cases and 11865 controls from an East Asian population. The initial association analysis was implemented using logistic regression under an additive genetic model adjusted for age, sex, BMI, and the first two principal components in PCA analysis. Genomic control inflation factor to adjust for potential population stratification was performed. The SNP selection for the remaining analysis was based on the smallest p-values and a set of SNPs in linkage disequilibrium at $r^2 \geq 0.1$ with the most associated SNPs. In addition, the Cochran Q statistic was used to assess the heterogeneity across studies. Two novel T2D loci were identified in this study including rs10886471 in the G-protein-coupled receptor kinase 5 (*GRK5*) gene with p-value=7.1×10^{-9} and rs7403531 in the RAS guanyl releasing protein1 (*RASGRP1*) gene with p-value=3.9×10^{-9}. The authors further confirmed seven established T2D loci and they concluded that their study not only contributes to the pathophysiology of T2D but may emphasise and highlight the ethnic differences in T2D susceptibility.

Two novel T2D susceptibility variants were identified in Tsai's study (Tsai et al. 2010). The authors conducted a two-stage genome-wide association analysis in a Han Chinese population, in which 995 T2D cases and 894 controls with 516737 genotyped SNPs were used in GWAS analysis. For the replication stage, 1803 T2D cases and 1473 controls with a set of SNPs identified by the initial GWAS analysis (p-value $< 10^{-5}$) were considered. T2D

association analysis was carried out using various genetic models including genotypic, allelic, trend, dominant, and recessive. The study chose the most significant test statistic attained from the five genetic models and the SNPs with p-value $< 2 \times 10^{-8}$ were considered to be statistically significant. The study also applied Fisher's exact test to combine p-values for joint analysis. The first significant new variant was found for rs17584499 located in and around protein tyrosine phosphatase receptor type D gene (*PTPRD*) with (p-value $= 8.54 \times 10^{-10}$, OR $= 1.57$, 95% confidence interval [CI]$= 1.36 - 1.82$). The second significant variant was rs391300 (p-value $= 3.06 \times 10^{-9}$, OR $= 1.28$, 95% confidence interval [CI]$= 1.18 - 1.39$). The results suggest that identifying two novel T2D susceptibility variants in the Han Chinese population may lead to a better understanding of the ethnic differences in the molecular pathogenesis of T2D.

Despite the success of GWAS in revealing genetic variants that are associated with complex disorders in populations, GWAS is still in its infancy and further studies to explore the genetic components of complex diseases are needed.

2.10 Summary

T2D has reached epidemic proportions. Therefore, understanding the underlying causes of T2D is of significant importance. A strong body of evidence has suggested that genetic factors contribute significantly to the predisposition of T2D. GWAS have succeeded in identifying genetic variants that show evidence of increased susceptibility to T2D, however GWAS is more suitable for capturing linear interactions between genetic variants ignoring the non-linear interactions (epistatic interactions), of multiple genetic variants that exist in polygenetic disorders like T2D.

This chapter presented an overview of T2D aetiology and associated risk factors. This is followed by a discussion on genome-wide association analysis and its contribution to T2D studies. The next chapter will provide a comprehensive review of epistatic interactions and

the approaches used for detecting epistasis in the context of GWAS. Furthermore, artificial neural networks, which is the adopted approach posited in this thesis to detect epistasis in T2D GWAS, will be discussed.

Chapter 3 Computational Biology

3.1 Introduction

In computational biology, the accumulation of biological data and the need for its storage, analysis, annotation, interpretation, visualization, systematization, and integration into database management systems and biological networks is the main reason ***bioinformatics*** emerged. Bioinformatics is a rapidly evolving, multidisciplinary field that provides applications, analysis tools, and methods to explore and understand biological data (i.e. genomic and proteomic) (Abdurakhmonov 2016; Gauthier et al. 2018; Bartlett et al. 2017). Bioinformatics brings expertise from different fields, such as biology, chemistry, physics, mathematics, computer science, statistics and engineering to develop theoretical models for biological data analyses (Can 2014; Searls 2010).

Nowadays there is a movement from traditional biostatistical approaches towards a more integrated approach that provides advanced methods to handle the complexity of biological data analysis as well as the structural interactions between biomolecules. The application of bioinformatics in biomedical research has become fundamentally important to advance research within the genomic domain (Abdurakhmonov 2016).

In biomedical research, understanding the aetiology of complex diseases is complicated. It has been thought that complex diseases involve multiple genetic constructs and the interactions that occur between them (Wang et al. 2014). Genetic factors do not act independently but rather in conjunction with other factors such as the environmental and sociodemographics. Moore et.al (Moore et al. 2010) suggested that traditional parametric statistical approaches, such as linear modelling frameworks (i.e. logistic regression), have limited power for modelling the complexity of non-linear interactions between SNPs or genes. Yet, these non-linear interactions are necessary to discover the aetiology of complex diseases. More specifically, the linear modelling framework examines each SNP

independently to discover associations with the phenotype, while ignoring epistatic interactions (SNP-SNP interactions) and environmental exposure. Consequently, the challenges associated with traditional approaches have led to alternative methods, particularly those that incorporate machine learning techniques.

Advances in machine learning algorithms have enabled further development and improvement in the genomic research domain. Using advanced machine learning techniques allows us to model the non-linear interaction between genetic variants, the environmental and clinical factors. Thus, enhancing our understanding of molecular biology, and complex disease susceptibility.

In this chapter, we introduce the concept of epistatic interactions which is one of our main features considered in this thesis. This is followed by the theoretical discussion on artificial neural networks, more specifically the state-of-the-art in deep learning which is the adopted approach posited in this thesis for detecting epistatic interactions between SNPs. The chapter also presents a discussion on six traditional machine learning algorithms used in our methodology as a comparison to discover the contribution of non-genetic risk factors to help explain disease susceptibility.

3.2 Understanding Epistasis

The genetic architecture in complex diseases is not caused by an individual allele or gene. It is increasingly apparent that in order to understand the genetic contributions in complex disorders, the interactions between SNPs and genes must be considered. This type of latent interaction between genetic markers is called ***epistasis***. Epistatic analysis has been reported in the literature since it was first coined by William Bateson in 1909 (Bateson 1909). Since then, there has been no clear explanation on the meaning of the word "***interaction***" which varies between scientists in the field of biology and statistics. Consequently, the two most

common ways to describe epistasis is ***biological epistasis*** and ***statistical epistasis*** (Evans 2011).

Biological epistasis was defined originally by Bateson to describe the masking effect one gene has on another. This means a variant at one locus masks the phenotypic expression generated by another locus. A simple example is the coat colour of a dog. Assume that there are two primary loci that control the coat colour of a dog - black/brown locus (B/b) and a white locus (W/w) (Cordell 2002). The black allele is dominant to a brown allele therefore if the dog possesses heterozygous genotype (Bb) at this locus the dog will have a black coat colour. However, the phenotype expression at the black/brown locus is also controlled by the genotype at the white locus. If the dog possesses a homozygous recessive genotype for the allele "w" at the white locus, the dog will have a white coat colour regardless of their genotype at another locus. This implies that the homozygous recessive genotype of the white locus masks the effect of the black/brown locus. In other words, the white locus is said to be epistatic to the effect at the black/brown genotype.

Statistical epistasis was defined by Ronald Fisher in 1918 (Fisher 1918) to describe the joint effect of risk alleles at both loci in which the effect is much larger or smaller than implied by their individual single-locus additive effect. This simply means any statistical departure from the additive combined effect of two loci. For example, consider two genes G1 and G2 that cause increased body weight. The contribution of each gene separately is a 1-pound increase in body weight. Suppose that there is an individual carrying both genes who shows a 2-pound increase. This means interaction between the two genes does not exist. In addition, the effect of both genes on the phenotype implies a normal additive model of inheritance. However, if the joint effect of both genes on an individual showed a 5-pound weight gain or even weight loss we could conclude that epistatic interaction must exist.

Phillips (Phillips 2008) reviewed the essential role of gene interactions in the structure and evolution of genetic systems. In his review three different forms were highlighted to describe

the concept of epistatic interactions - functional epistasis, compositional epistasis, and statistical epistasis. Functional epistasis addresses molecular interactions such as protein-to-protein interactions. Compositional epistasis refers to what William Bateson originally defined as biological epistasis. Statistical epistasis was attributed to the definition by Ronald Fisher. Phillips suggested that compositional and statistical epistasis are complementary to one another (Phillips 2008). When two genes interact statistically it is more likely that they also interact physically (VanderWeele 2010). The physical molecular interactions occur between various genes (VanderWeele 2010). Therefore, statistical epistasis can provide useful information in the biological understanding of genetic architectures that underlie complex disease.

While genome scans have helped to unravel and identify the genetic risk factors involved in common and complex human disease (Visscher et al. 2012), association studies have used statistical methods to analyse and explore individual SNPs one at a time. Consequently, they do not consider possible interactions present between genetic markers. GWAS have been unsuccessful at detecting epistasis as they commonly focus on identifying the main genetic associations with additive effect. However, it has been hypothesized that the non-additive effects between genes, particularly epistatic interactions, could contribute to our understanding of the underlying genetic architecture of phenotypic variation (Phillips 2008). Current attempts to study such interactions in complex human disorders have focused on the interactions between pre-identified genes that exist in candidate regions (Rishika De et al. 2015; Rose & Bell 2012). This is an important area of research given that it has been suggested that epistasis might account for the remaining unexplained heritability within many common complex disorders (Manolio et al. 2009). In particular, Maher (Maher 2008) describe such epistatic interactions as “underground networks” in which missing heritability could be concealed.

3.3 Epistasis Challenges

It is difficult to detect epistasis or SNP-SNP interactions in large-scale genome-wide settings, for three fundamental reasons as outlined by Ritchie (Ritchie 2013). In particular, variable selection, model building, and model interpretation in the context of human biology have been the primary focus in many research initiatives.

Identifying appropriate SNPs and evaluating increasingly higher-order combinations from very high-dimensional data (of which GWAS is) is computationally difficult. The International HapMap Consortium reported that to capture most of the relevant genetic markers across the human genome, they needed approximately 300,000 carefully selected SNPs (Olivier 2005). Under this assumption, Gilbert-Diamond and Moore (Gilbert-Diamond & Moore 2011) highlighted that with 300,000 SNPs the generated pairwise combinations of SNPs would be 4.5×10^{10}. This exhaustive evaluation without high performance computing resources would be computationally infeasible. As such, a computational algorithm to filter the genome-wide datasets into smaller subsets is often needed.

The second challenge is model building. This involves the development of robust computational and statistical methods to model the relationship between high-order SNP combinations and disease susceptibility. Traditional parametric-based statistical approaches, such as logistic and linear regression, are ineffective at dealing with the problem of exponentially increased dimensionality associated with multi-locus testing. The epidemiological sample in the study must be exponentially larger to allow for enough subjects to be tested with the generated genotype combination for the genetic effects to be accurately detected. Therefore, non-parametric approaches more specifically data mining and machine learning methods such as Multifactor Dimensionality Reduction, Neural Networks, Random Forest, and Support Vector Machines have proven to be more powerful

approaches than parametric statistical approaches. Nonetheless, they are not without their own limitations.

The third challenge is the interpretation of epistasis models and their biological context. Making biological inferences from computational statistical models can be more challenging than detecting and characterizing epistatic interactions. Cordell (Cordell 2009) suggested that inferring biological mechanisms from statistical model results is complex and limited. Cordell argues that statistical interaction does not necessarily reflect interaction on a cellular level and that it is possible for biological epistasis to arise in the absence of statistical epistasis. The relationship between statistical and biological epistasis has been discussed in detail by Moore and Williams (Moore & Williams 2005). They proposed two significant questions *"First, when does statistical evidence of epistasis in human populations imply underlying biomolecular interactions in the aetiology of disease? Second, when do biomolecular interactions produce patterns of statistical epistasis in human populations?"* They concluded that the relationship between biological and statistical epistasis is difficult to comprehend.

Yet, interpreting statistical epistasis results at a biomolecular level in the context of human health and diseases will help provide a central framework for employing genetic information to improve diagnosis, prevention, and treatment strategies.

3.4 Strategies for Detecting Epistasis in Genome-Wide Association Studies

Despite the spectacular effort in developing statistical methods and computational strategies to detect SNP interactions in large GWAS data, epistasis analysis in GWAS remains in its infancy. Perhaps one of the reasons is the logistical difficulties associated with large combinatorial analysis in high-order SNP interactions. This is in addition to the low statistical power caused by small sample sizes in GWAS cohorts (Cordell 2009). Various statistical methods have been developed to exhaustively search pairwise or high-order interactions between SNPs to detect epistatic effects in genome-wide case-control studies.

The pairwise method of interaction involves two loci while high-order interactions require three or more loci that interact jointly (Cordell 2009; Taylor & Ehrenreich 2015). These methods employ different searching strategies that include exhaustive search (Cordell 2009), search based on probability (Prabhu & Pe'er 2012), candidate region search (Rishika De et al. 2015), and search based on the filtering (Ding 2014) of interesting SNPs selected through a priori knowledge. Moreover, these statistical methods vary in the way they select biomolecule units to test for interactions, such as SNPs, genes, and/or proteins.

3.4.1 Exhaustive Search of Pairwise Interaction

Pairwise interaction is arguably one of the simplest methods to perform when detecting interactions in genome-wide data (Cordell 2009). This method is used to test all possible pairs of loci across the genome and implement interaction tests for each two-locus combination. Although, pairwise search is computationally feasible, it is in practice, an un-scalable and time-consuming process. Given the number of genetic markers routinely generated in genome-wide studies (anything between 500,000 and 1 million SNPs), it is clear to see this approach has limited utility, particularly in complex diseases where many interacting SNPs are the route cause. Therefore, performing such a large number of statistical tests may suffer from low statistical power (Cordell 2009). However, the evolution and availability of parallel processing facilities, i.e. banks of Graphical Processing Units (GPUs), will make such tasks possible within a reasonable time frame (Chatelain et al. 2018; Hemani et al. 2011).

3.4.2 Exhaustive Search of Higher-Order Interaction

In the context of genome-wide data, implementing an exhaustive search over higher-order interactions, i.e. third and fourth-order, poses a significant challenge (Sailer & Harms 2017). There are an enormous number of multiple tests generated and these are proportional to order level interactions. Consequently, the number of comparisons required increases exponentially and thus the time required to perform such analysis (Taylor & Ehrenreich

2015). In addition, the fact that higher-order interaction analysis requires many degrees of freedom will potentially reduce statistical power in studies (Cordell 2009). It has been suggested that in order to mitigate such issues a two-stage procedure should be employed (Faye & Bull 2011; Nguyen et al. 2015). The first stage focuses on a subset of loci identified through single-locus threshold analysis, and the second stage, using this subset to perform the exhaustive search of all possible interactions between these loci. There is a debate for selecting loci at the first stage. This concern is that while some loci are truly associated with the phenotype, they are often discarded due to threshold selection. This is particularly true for loci with no marginal effects. The selection process of loci based of single-locus thresholds could be altered to select loci based on a priori knowledge of biology, genetic impact and pathway information, but this would discount the hypothesis-free nature of genome-wide analysis (Herold et al. 2009).

3.4.3 Computational Statistical Approaches for Epistasis Detection

It is critical to model complex interactions between genetic markers if epistasis is to be detected. The challenge of identifying epistatic interactions in large-scale GWAS case-control data has attracted a great deal of research interest. Up to now, there are almost one hundred computational software tools designed and developed for epistasis detection. The omictools website provides a full list of tools to be used with GWAS data analysis (https://omictools.com/epistasis-detection-category). In this section the focus will be on software methods that have become popular and shown particular promise for identifying epistasis in genome-wide case-control studies using the statistical epistasis definition.

Zhang and Liu (Zhang & Liu 2007) developed Bayesian Epistasis Association Mapping (**BEAM**), which uses a Bayesian partitioning model to model disease-associated markers and their interactions. BEAM computes the posterior probability that each individual marker set is related with the disease via Markov chain Monte Carlo. Zhang *et al*. (Zhang et al. 2010) proposed Tree-Based Epistasis Association Mapping (**TEAM**), using an exhaustive search

pairwise algorithm for fast detection of SNP-SNP interactions in GWAS settings. This program utilizes permutation tests over the common Bonferroni correction adjustment method to control family-wise error and false discovery rates. In addition, TEAM applies minimum spanning tree structures that significantly increase the performance and accelerate the process of epistasis detection in GWAS data. Wan *et al.* (Wan et al. 2010) developed the BOolean Operation-based Screening and Testing (**BOOST**) software tool, which is a computationally and statistically feasible and fast program for the detection of all pairwise epistatic interactions. BOOST was designed based on a Boolean representation of genotype data that uses fast logic operations (bitwise) to generate contingency tables that promote space and CPU efficiency. In addition, this program was developed using a two-stage search method; screening and testing. The selected SNPs in the screening stage are forwarded to the testing stage to measure the interaction effects of SNP pairs by employing the likelihood ratio statistic and log-linear model. To further improve computation time, a GPU-based version of BOOST was introduced called **GBOOST** (Yung et al. 2011). Wang *et al.* (M. H. Wang et al. 2016) introduced a fast and powerful *W*-test for identifying pairwise epistatic interactions. The test is particularly powerful when using low frequency variants, in which MAF is between 1 and 5 percent in GWAS data. The test is advantageous over alternative methods. First, it is model-free such that no assumptions are made about the genetic effect model. Second it incorporates a Chi-Squared distribution that has data-adaptive degrees of freedom, allowing for robust association testing in genome scans. Herold *et al.* (Herold et al. 2009) introduced the **INTERSNP** tool for genome-wide interaction analysis that considers two and three-markers for association tests. INTERSNP selects combinations of SNPs for interaction analysis based on a priori information including statistical evidence for single-marker association, genetic relevance of SNP genomic location, and the biological relevance of SNP function and pathway information. The authors concluded that the proposed tool can help elucidate the actual relevance of gene interactions in complex diseases and demonstrate the potential and feasibility of completing three-marker interaction

analysis. Recently, Fang *et al.* (Fang et al. 2017) developed a technique based on high-dimensional grouped variable selection, called Two-Stage Grouped Sure Independence Screening (**TS-GSIS**) for detecting SNP-SNP interactions with or without marginal effects as well as identifying causal SNP effects within a certain gene and their corresponding SNP-SNP interaction effects. Moreover, Lasso regression is used with the TS-GSIS approach to select important SNPs in candidate genes to reduce the dimension of data by determining the size of candidate genes in models. This is a powerful characteristic to balance model complexity and predictive performance.

Terada *et al.* (Terada et al. 2016) proposed a software tool named **LAMPLINK** for the detection of statistically significant high-order interactions from genome-wide case-control data. The authors incorporate Limitless Arity Multiple-testing Procedure (LAMP) (Terada et al. 2013), a statistical method to list statistically significant combinatorial effects that consist of three or more SNPs in each combination using PLINK (Purcell et al. 2007) to perform association analysis for GWAS. LAMPLINK is limited to dominant and recessive models neglecting the additive genetic model which might provide new insights into the missing heritability problem. In terms of time performance, LAMPLINK outperforms existing traditional techniques such as logistic regression and multifactor dimensionality reduction when performing combinatorial interaction analysis.

Although, these above-mentioned techniques have been widely used for the detection of SNP-SNP interactions they are often criticised for their inability to deal with high-dimensional data. Consequently, these approaches are not scalable and will likely become redundant as the number of SNPs sequenced significantly increases over time.

3.4.4 Data Mining and Machine Learning Approaches for Epistasis Detection

A variety of data mining approaches, including data reduction and data recognition, have been used to detect interactions between genes in large-scale genetic studies. Data reduction approaches involve the transformation of data to a lower dimensional space (Rehman et al.

2016). There are several examples of data reduction approaches including the restricted partition method (RPM) (Culverhouse 2010), Combinatorial partitioning method (CMP) (Nelson 2001), set association (Ott & Hoh 2003), and multifactor dimensionality reduction (MDR) (Ritchie et al. 2001). Advances have also been made in pattern recognition studies, where patterns and regularities in the data can be used to classify and discriminate between groups using high-decisional data sets, such as GWAS (Nandy & Padariya 2016). This has been achieved using several traditional machine learning algorithms that include support vector machines (SVMs), artificial neural networks (ANNs), and random forests (RF).

3.4.4.1 Data Reduction Approach

Multifactor dimensionality reduction (MDR) has been successfully applied to detect common interactions between loci for a wide variety of human diseases including T2D (Barna et al. 2018), Bladder Cancer (Andrew et al. 2008), Bipolar Disorder (Oh et al. 2012), Alzheimer (Martin et al. 2006), Obesity (R De et al. 2015), and Sporadic Amyotrophic Lateral Sclerosis (Greene et al. 2010).

MDR is a feature or attribute constructive induction algorithm (Moore 2007) that performs data reduction by converting high-dimensional data, e.g., multi-loci data, into one-dimension with two levels: high and low risk. The process of attribute construction is performed by pooling a new single attribute from multiple variables, e.g., a single SNP from multiple SNPs so that a new attribute acts as a function of two or more other attributes (Moore 2007). MDR was developed to detect interactions between loci in the absence of marginal effects. In 2001 (Ritchie et al. 2001), MDR became a breakthrough approach and an alternative solution to parametric regression paradigms such as logistic regression where interactions are explored exclusively among loci that exhibit statistically significant effects.

MDR was one of the earliest approaches developed to facilitate the detection, characterization, and interpretation of epistatic interactions in genetic studies of human disease. This approach was evaluated using Sporadic Breast Cancer in population-based

studies (Ritchie et al. 2001). The study revealed statistically significant high-order interactions among four polymorphisms from three different Estrogen-Metabolism genes. This was the first report of such interactions associated with a common complex multifactor human disorder (Ritchie et al. 2001).

The popularity associated with the use of MDR in epistasis analysis was found due to the fact that the model is a non-parametric approach (Moore & Andrews 2015) in which no hypothesis about the value of a statistical parameter is made. It is a genetic model-free approach (Moore & Andrews 2015) that assumes no particular inheritance model. This is particularly useful for complex diseases in which the mode of inheritance is obscure and complex. MDR also uses a highly constructive induction algorithm to detect non-linear interactions among discrete genetic attributes. This is achieved by selecting two or more SNPs and reducing them to a single feature thus permitting interaction effects to be detected. Moreover, the integration of cross-validation resampling with MDR adds additional characteristics to the model. This is specifically important to avoid overfitting and minimize the false-positive rate in GWAS settings.

MDR provides a comprehensive and powerful data mining approach for detecting, characterizing, and interpreting non-linear epistatic interactions by combining attribute selection, attribute construction, classification, cross-validation and visualization, but it does come with its own limitations. The main limitation is scalability (Bush et al. 2006). It does not scale up when a large number of predictor variables are used. In the case of GWAS analysis, the number of genetic markers (predictors) can be between 500,000 and 1 million and in some cases much higher. By performing pairwise search using MDR for GWAS settings this would seem impractical. MDR on more than a few hundred loci will be computationally difficult (Bush et al. 2006). Therefore, to apply MDR on GWAS data, the predictor variables need to be reduced. Variables for MDR analysis are often selected from

candidate gene studies or extracted from a large set of genetic markers using one of the filtering approaches (Ritchie 2013).

Furthermore, MDR does not distinguish between marginal effects from pure interaction effects and this can make it difficult to interpret. More importantly, model power is reduced significantly when 50% genetic heterogeneity is present (Upstill-Goddard et al. 2013).

Over the last two decades numerous extensions to MDR have emerged to improve and overcome some of the limitations evident in the original model. Extended MDR implementations include methods to handle unbalanced datasets (Yang et al. 2013; Velez et al. 2007), missing data (Namkung et al. 2009), covariate adjustment (Calle et al. 2008), and model-based MDR in the presence of noise (Cattaert et al. 2011). Additionally, others were developed to make large-scale analysis of epistasis tractable, i.e. MDR-based solutions that utilise GPUs (Greene et al. 2010) to accelerate epistasis analysis (Sinnott-Armstrong et al. 2009). Unified model-based MDR approaches (UM-MDR) (Yu et al. 2016) have also helped to overcome the limitation of evaluating the significance of multi-locus models. Empirical Fuzzy MDR (EF-MDR) (Leem & Park 2017) was developed to avoid the difficulty in reflecting the uncertainty of high-risk and low-risk in binary classification settings. Although significant advances have been made, MDR and its variants are considered computationally prohibitive.

3.4.4.2 Filtering Approach

Because the search space in multivariate models of genomic datasets is large, due to the large number of features (SNPs) considered, detecting epistatic interactions remains a significant challenge. Thus, filtering approaches have seen widespread use to select important variables prior to epistasis analysis to improve efficiency in data mining and machine learning studies. Various methods have been proposed to perform feature selection. One such approach is ReliefF (Greene et al. 2009; Robnik-ˇSikonja & Kononenko 2003), which is an attribute quality estimator. ReliefF is based on detecting conditional dependencies between attributes

and searching for nearest neighbours. In addition to its natural interpretation, the algorithm is effectively scaled up to include a large number of examples and features. However, the algorithm can be sensitive to a large number of variants that are irrelevant to the classification of the trait. Mckinney *et al.* (McKinney et al. 2007) proposed an alternative approach named Evaporative Cooling (EC) for feature selection that overcomes ReliefF limitations.

Recently, Verma *et al.* (Verma et al. 2018) proposed a collective feature selection approach to select true positive epistatic variables using various parametric, non-parametric, and data mining methods. Using this approach proves to be effective for selecting features with epistatic effects in the presence of incomplete penetrance, and polygenic inheritance.

3.4.4.3 Pattern Recognition Approach

As elucidated previously, pattern recognition is a complex process that deals with real and noisy data and recognizing patterns and regularities that can be used to classify and discriminate between groups using the full dimensionality of the data. One of the popular and appropriate pattern recognition approaches for large-scale genomic data analysis is random forest.

Random forest (RF) is one of the non-parametric machine learning algorithms that is based on a randomized decision tree ensemble (Breiman 2001). It exhibits the potential to capture epistatic interactions through the process of variable selection (Chen & Ishwaran 2012; Kawaguchi 2012). It ranks variables using variable importance measures (Breiman 2001) and detects interactions between features (Strobl et al. 2008). The major limitation with RFs, however, is that the detection of gene–gene interactions depends on the presence of main effects (Kim et al. 2009). Thus, epistatic interactions with no marginal effects are often left undiscovered when RF analysis is performed.

Another limitation with the RF algorithm is that was designed to analyse data with no more than a few thousands features (SNPs) on a standard machine (Qi 2012). Schwarz *et al.*

(Schwarz et al. 2010) developed a Random Jungle (RJ) algorithm based on an extension to RF. The RJ is a computational, and memory efficient method designed to handle large-scale GWAS datasets with hundreds and thousands of samples and SNPs. RJ is based on variable backward elimination while maintaining all other options provided by the original RF particularly the permutation importance measure. In addition, it uses multithreading and a Message Passing Interface (MPI) across processes that can be implemented on multiple CPUs simultaneously. A real GWAS dataset from a Crohn's disease study consisting of 513 cases and 515 controls with 317,503 genotyped SNPs was used to implement RJ. Analysing GWAS data using RJ seems to be feasible with respect to time and memory consumption and the results show that the RJ is a promising method for high-dimensional GWAS data. The application of RJ to GWAS may help to identify interacting SNPs that were not found using traditional parametric statistical approaches.

Another extension to RF is SNPInterForest (Yoshida & Koike 2011). SNPInterForest was built based on a modification to the RF construction framework, which allows for either a combination of SNPs or a single SNP when choosing a split variable at each node. This prevents the important scores of SNPs with no marginal effects from being underestimated. Furthermore, the interaction score measurement is introduced to discover interacting SNP combinations. Thus, if a certain SNP combination appears frequently on a branch, the interaction strength is calculated based on the number of simultaneous appearances of SNP combinations in each branch of each tree. Accordingly, it is more likely that these SNPs for the corresponding SNP combinations can identify interactions between them. SNPInterForest has been evaluated on a real Rheumatoid Arthritis GWAS dataset from The Wellcome Trust Case Control Consortium (WTCCC), which contains 500,000 genotype SNPs and 3499 cases and controls (Yoshida & Koike 2011). The evaluation revealed that SNPInterForest achieved considerable improvements in detecting pure epistatic interactions in comparison to an RF ensemble learning algorithm. Furthermore, SNPInterForest

outperformed existing methods based on exhaustive search strategies, which include BOOST (Wan et al. 2010) and SNPHarvester (Yang et al. 2009). However, computational burden is considered one of the main limitations of SNPInterForest implementation, as is the case with many other similar approaches.

In addition to RF, Neural Networks (NNs) and Support Vector Machines (SVMs) have shown excellent power in identifying epistatic interactions in complex human traits. Neural Networks with a feedforward, and backpropagation architecture are capable of dealing with large datasets (e.g. large-scale GWAS data). NN algorithms with advanced characteristics can sufficiently detect epistatic interactions including genetic heterogeneities, incomplete penetrance (high effect size), and polygenic inheritance.

While SVM algorithms are as robust as NNs and demonstrate significant power when used to detect epistasis, in comparison to MDR, Chen *et al.* (Chen et al. 2008) conducted an experiment in a case-control Prostate Cancer study population employing SVM and MDR. The authors revealed that an SVM outperformed MDR under all the scenarios particularly in the presence of 5% genotyping error, 5% missing data, or a combination of both under different pairwise epistasis models with a variety of allele frequencies. The following sections discuss these approaches in more detail.

3.5 Artificial Neural Networks

An artificial neural network (ANN) is a machine learning technique that is inspired by the way biological nervous systems (human brain) process information (Haykin 1994). The brain is a highly complex, non-linear, parallel computer system composed of millions of highly interconnected neurons organized to perform computation, e.g. pattern recognition, vehicle control, and human vision. Typically, these neurons have the ability to transmit and receive information (signals) and process inputs to produce any number of different outputs. In the human brain, connections between neurons transmit signals between interconnected neurons. The direction of these signals can be unidirectional or bidirectional. The learning

process in the human brain is based on experience (training). Learning in biological systems is achieved by making small compensatory adjustments to connections that exist between the neurons (tuneable weights) as well as changing neuron activation thresholds.

ANN is a crude simulation that attempts to mimic the behaviour of our brain (Haykin 1994). Neural networks constructed from a group of interconnected neurons are organised into layers. Input, hidden, and output layers when combined describe the structure of the network. The input layer is the first layer, and its neurons receive information signals from external sources. The output layer is the last layer in the network, and its neurons present their output to the outside world. The middle layers are referred to as the hidden layers, and they are located between the input and output layers. The hidden neurons receive their inputs and transmit their outputs internally in the network. Every neuron in the network is a processing unit that takes an input signal with its weight and performs a fixed mathematical operation using an activation function. The activation function defines the output of the neuron and the scale based on predefined thresholds. In order for neural networks to learn and produce the desired output, the weights are adjusted during the learning or training process.

Based on the theoretical definition of ANNs in (Anthony & Bartlett 1999), the basic computational units in neural networks are neurons, each neuron takes n input values $x_1, x_2, \dots x_n$, and a bias intercept term represented by $+1$ (not included in the input), which is a constant term used to overcome the problem with input patterns that are zero. The network outputs a hypothesis $h_{W,b}(x)$ where W and b are weight and bias parameters that can be learned from input data, x. The neuron output is defined as:

$$h_{W,b}(x) = f(W^T x) = f(\sum_{i=1}^{n} W_i x_i + b) \tag{3.1}$$

where $f: \mathbb{R} \mapsto \mathbb{R}$ represents the activation function. Figure 3.1 illustrates a basic example of a neuron.

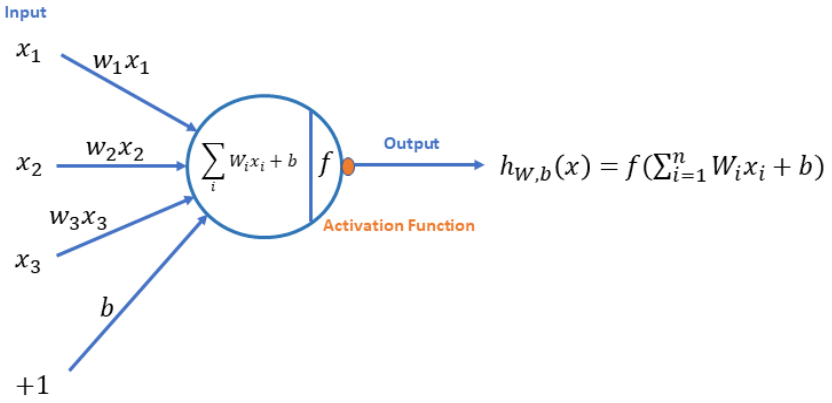


Figure 3.1: Single Neuron

There are various types of activation function; sigmoid function, hyperbolic tangent, rectifier linear unit, and maxout (Anthony & Bartlett 1999; Candel et al. 2018). The formal definition of these activation functions are as follows:

- Sigmoid Activation Function:

  Sigmoid is the non-linear activation function that corresponds to the input-output mapping defined in logistic regression. Sigmoid is used to scale the neuron's output to a range of [0,1]. The sigmoid function is represented as:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.2}$$

  where x denotes the input to the neuron. Figure 3.2 (a) presents the graphical representation of a sigmoid function.

- Hyperbolic Tangent Activation Function:

  The Hyperbolic Tangent (tanh) activation function is another common non-linear activation function and used to scale the output between [-1,1]. Hence, the tanh function is a rescaled version of the sigmoid function mainly used for classification. The tanh function is formulated as:

$$f(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3.3}$$

  where x denotes the input to the neuron. Figure 3.2 (b) presents the graphical representation of a tanh function.

- Rectifier linear unit Activation Function:

  The rectifier linear unit (ReLU) is a very popular non-linear activation function in deep neural networks. The main reason is that not all the neurons are activated at the same time, allowing sparsity to be added to the network, which helps reduce computational overheads. It is used to scale the output between [0 and infinity]. This activation function has a zero threshold and is given as follows:

$$f(x) = \max(0, x) \tag{3.4}$$

  Hence,

$$f(x) = \begin{cases} 0 \ \ for\ x < 0 \\ x\ for\ x \geq 0 \end{cases} \tag{3.5}$$

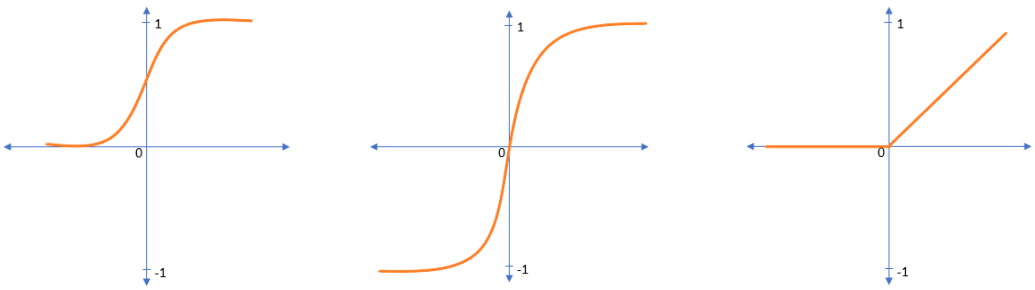
where x denotes the input to the neuron. Figure 3.2 (c) provides a graphical representation of the ReLU function.

- Maxout Activation Function:

  The maxout activation function is a generalized version of ReLU. It is the max of two inputs. Maxout does not suffer from dying neurons (transferring negative inputs to the ReLU function as zero). This means the gradient is zero and the neurons can never be activated in this region. Maxout is used to scale the output between [-infinity and infinity]. The maxout activation function is defined as follows:

$$f(w^T x + b) = max(w_1^T x + b_1, w_2^T x + b_2) \tag{3.6}$$

  where x denotes the input to the neuron.



a. Sigmoid b. Tanh c. ReLU

Figure 3.2: Activation Functions Graphical Representation

It is often difficult to determine which activation function to adopt for your data; as each may outperform the other in different scenarios (Candel et al. 2018). Thus, grid search models are often used to compare activation functions and select the one that is best for your data.

3.5.1 Characteristics of Artificial Neural Networks

Using ANN offers various useful properties and capabilities (Haykin 1994).

- A neural network is a ***non-linear model***, where each neuron in the network is basically a non-linear unit. Neurons are used to construct the network. Non-linearity is a highly important characteristic in neural networks, particularly if underlying datasets are non-linear. Moreover, non-linearity offers additional flexibility to the neural network in modelling real-world complex relationships.
- I***nput-output mappings*** allow neural networks to learn using a supervised learning paradigm (labelled training data that corresponds to target responses). The training of the network is performed iteratively, and the weights modified until the network reaches a steady state.
- Neural networks are considered ***data driven self-adaptive*** algorithms that can be effectively adapted to given datasets. They are designed to reject ambiguous patterns that arise in classification tasks and provide confidence values for decisions made.
- ANNs are described as ***universal approximation functions*** that can approximate any complex non-linear function with arbitrary accuracy.
- Neural networks are ***massively parallel systems***, similar to the parallel distributed structure of the brain, and have the ability to capture truly complex behaviour in a highly hierarchical fashion. This feature makes ANNs appealing for solving large-scale and complex real-world applications.

3.5.2 Structure of Artificial Neural Networks

In ANN the communication links (connections) between neurons are responsible for information propagation (Haykin 1994). There are two common types of ANN architecture widely used for classification and prediction problems. These are feedforward neural networks and recurrent neural networks. In the former type, information is transmitted in a forward direction through the network layers, on a layer-by-layer basis, starting from the input layer through to the output layer. In the later type, the structure of the network integrates feedback loop connections in each neuron in the hidden layer to provide dynamic behaviour in the neurons. In this thesis only feedforward neural networks are considered, in particular multilayer perceptrons.

3.5.3 Multilayer Feedforward Neural Networks

Multilayer feedforward neural networks are also known as multilayer perceptrons (MLP). MLPs are distinguished by the presence of one or more hidden layers in their structure (Haykin 1994). Each hidden layer contains a number of hidden neurons. The function of hidden neurons is to interconnect input and output neurons. These hidden neurons enable the neural network to learn non-linear complex tasks by extracting meaningful features. Extracting higher order features is particularly valuable when the input vector is large (Haykin 1994). The neurons in MLPs exhibit a high degree of connectivity, as the output of one neuron is the input to all other neurons in the adjacent forward layer. The number of hidden layers and the number of the hidden neurons in each layer determines the performance of neural networks (Heaton 2008). Therefore, different neural network structures generate different outcomes. If a limited number of hidden neurons are used, this can lead to underfitting (Heaton 2008), where the model is unable to extract and learn the non-linear structure in complex high-dimensional datasets. On the other hand, using too many hidden neurons can lead to overfitting (Heaton 2008), where the network is tuned to the training data resulting in a network that cannot generalise using unseen data samples (Haykin 1994).

The backpropagation algorithm provides a computationally efficient method for training MLPs for supervised learning. The learning process minimizes a cost function in accordance with an error-correction rule (Haykin 1994). If the response of the network generates output not close to the desired response (target), the weights of the network are adjusted to minimize the error (cost function).

3.5.4 Backpropagation Algorithm

Backpropagation (Werbos 1982; Rumelhart et al. 1986; Werbos 1974; LeCun et al. 1998) is a learning algorithm for implementing gradient descent in weight space for neural networks and is widely used for training multilayer feedforward networks. The intuition behind this technique is to efficiently compute gradient vectors (partial derivatives) of the cost function, to minimize the overall cost function with respect to weights and bias. The backpropagation process contains two passes through the different layers of the network (forward pass and a backward pass) (Haykin 1994). During the forward pass, the input vector is fed forward through the network, layer by layer, to produce a set of outputs. The error term, which is the difference between the actual response from the network and the desired response (target), is calculated. In the backward pass, the error term is propagated backwards to the previous layers through the network to adjust the weights between the units. During this training stage the weights of the network are adjusted iteratively using gradient descent optimization to minimise cost function errors, i.e. the actual response becomes closer to the desired response.

3.6 Deep Learning

Deep learning (DL) is an efficient fast-growing class of machine learning that has its foundation in artificial neural networks. Early deep learning networks were built using ANNs in the 1980s (Fukushima 1980). However, the popularity of DL was not seen till breakthroughs by Hinton began to appear in 2006 (Hinton & Salakhutdinov 2006). Since then, DL has been used across many domains, including image recognition (Krizhevsky et

al. 2012), speech recognition (Hinton et al. 2012), natural language processing (Collobert et al. 2011), and pharmaceutical formulation analysis (Ekins 2016).

The basic structure of DL networks is an ANN with many hidden layers and neurons – typically more than three hidden layers. This offers better capacity for feature learning and extraction. DLs are known as representation learning methods that consume raw data and automatically discover deep abstract representations to learn complex functions (LeCun et al. 2015). A key aspect of deep learning is its ability to automatically learn features from data and the interactions between data points using a representation learning procedure (Min et al. 2017). This characteristic of DL has helped to make major advances in solving big data problems. However, ANNs with many hidden layers can cause gradient-based training of randomly initialised weights in deep neural networks to get stuck in the local minimum.

Consequently, Hinton and Salakhutdinov (Hinton & Salakhutdinov 2006) proposed a greedy layer-wise pre-trained deep autoencoder to initialise the weights of networks layer-by-layer and learn reduced representations from raw data. This algorithm offers a good solution to the local minimum problem. In addition, this algorithm allows non-linear structures between features to be discovered and extracted in complex and large-scale datasets.

3.6.1 Deep Learning Architecture

The basic architecture in deep learning is a neural network architecture with many hidden layers and neurons. Different architectures have been proposed and many have been successfully used in various domains. Convolutional neural networks propose deep learning structures that are inspired by models of the human visual cortex, which have been widely utilised in image recognition (Krizhevsky et al. 2012) and natural language processing (Collobert et al. 2011). While recurrent neural networks, that build dynamic behaviour into the neurons, have become the primary method for time series data (Graves et al. 2013). Other architectures based on restricted Boltzmann machines (RBMs) (Smolensky 1986) i.e. deep belief networks (DBNs) (Hinton et al. 2006), and deep autoencoders specifically stacked

autoencoders, have also gained popularity in dimensionality reduction and for pre-training deep networks (Hinton & Salakhutdinov 2006). Table 3.1 presents some common DL approaches.

The learning process in deep learning is split into three main categories:

-***Networks for supervised learning***: this type of learning process is designed to train networks using labelled data. It is mostly used for classification tasks.

-***Networks for unsupervised learning***: designed to train networks with unlabelled data. This offers an efficient method to automatically learning features and capturing high-order feature interactions.

-***Networks for semi-supervised learning***: designed to train networks using labelled and unlabelled data. The unlabelled data is used to initialise the weights of a fully connected network for classification tasks using labelled data.

Table 3.1: Different Architectures of DL

Architecture	Description
Deep Neural Networks	• Deep framework with fully connected input, output and multiple hidden layers. • Used for classification and regression tasks. • Automatically learn deep non-linear abstract representations from raw data.
Stacked Autoencoder	• Consists of multiple layers of single autoencoders. Aims to reconstruct the input vector. • Used for dimensionality reduction (feature extraction) and pre-training deep networks. • Mainly for unsupervised learning. • Training process based on a greedy layer-wise learning strategy to initialise the weights of fully connected networks. Then fine-tuned using backpropagation for classification tasks.
Deep Belief Networks	• Applied to supervised and unsupervised learning. • Consists of a composition of restricted Boltzmann machines. Each subnetwork hidden layer is connected to the visible layer of the next RBMs. • The top two layers have undirected connections and directed connections to the lower layers. • Training process based on a greedy layer-wise learning strategy to initialise the network using unlabelled data, followed by fine-tuned training for classification tasks.
Convolutional Neural Networks	• Consists of a sequence of convolutional and subsampling layers followed by a fully connected layer for classification. • Used for feature extraction in two and three-dimensional data such as images. • Unsupervised and Supervised learning process.
Recurrent Neural Networks	• Contains cyclic connection in hidden neurons to perform recurrent computation. • Includes two sources of input to hidden neurons: the past information stored in the hidden unit and the present input. • RNN has memory, therefore it is used in sequential applications where outputs depend on previous input computations. • Unlike other DL architectures, RNNs share the same weights in its forward computation.

3.6.2 Autoencoder

An autoencoder (AE) is an artificial neural network that is utilized for unsupervised learning. AEs automatically learn features from unlabelled data (Le 2015; Ng 2011) and their primary application is in data reduction (Hinton & Salakhutdinov 2006). However, researchers have found that autoencoders can be used as a way to pre-train deep neural networks (Bengio et al. 2007; Erhan et al. 2010; Erhan et al. 2009). An Autoencoder consists of three or more layers: an input layer, a number of hidden layers, and an output or reconstruction layer. A shallow or simple structured autoencoder is a single hidden layer neural network that maps the original data (input values) to compressed data (lower dimensionality than the original data) through an ***encoding*** process which is in turn mapped to an output layer to approximate the original data through a ***decoding*** process (Le 2015). Basically, a shallow AE learns a low-dimensional representation similar to principal components analysis. AE computes the principal components of the input data, which is the optimal basis for linear dimensionality reduction. Figure 3.3 presents a single hidden layer AE, illustrating the encoding and decoding steps.

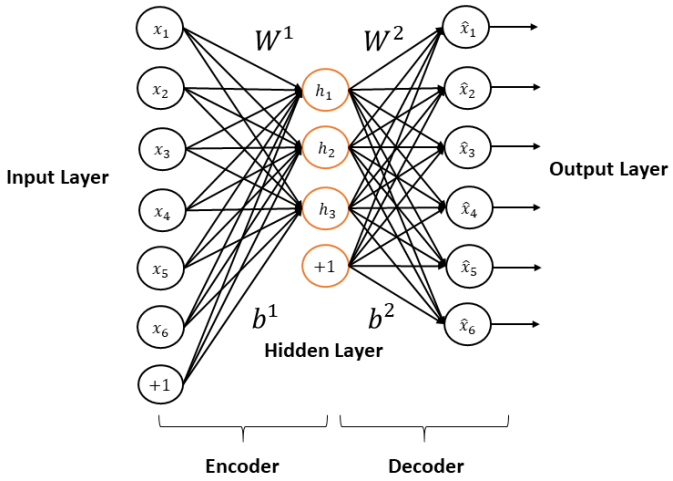


Figure 3.3: Autoencoder

This is a single layer autoencoder procedure. DL architectures have better capabilities when several autoencoders are stacked (Bengio et al. 2013). The following section explains stacked autoencoders in more details.

3.6.3 Stacked Autoencoders

Stacked autoencoder (SAE) is a neural network comprising multiple layers of sparse autoencoders (Bengio et al. 2013). SAE trains each layer in turn, in a greedy layer-wise unsupervised fashion for pre-training the weights parameters of deep network (Bengio et al. 2007). The greedy layer-wise training approach first involves training the first sparse autoencoder on the raw input vector of the network to obtain initial set of weights parameters for that first layer and to learn the first reduced representation of the raw input, resulting in first order features. This abstract representation of features in first layer acts as an input for the second sparse autoencoder, refers to second layer in SAE. The second layer is trained with similar manner, allowing to initialize weights parameters for that layer and produce reduced representation of features corresponding to the first order features. The process of training the parameters of each layer individually is repeated, using the output (the abstract representation) of each layer as input for the next layer, until the parameters for each layer are initialized (pre-trained). The last layer of SAE represents the deepest layer of the network and it contains the information of interest delivered by the activation vector. This activation vector presents the last and higher order features corresponding to the raw input vector of the network. The output of this last layer of SAE is linked to the softmax classifier. The softmax classifier is a supervised learning algorithm and it is commonly used in conjunction with the unsupervised deep learning for feature learning methods.

Finally, following the unsupervised pre-training stage of SAE, the error term of softmax classifier is computed and then propagated backward into the stacked layers. The whole network is fine-tuned using backpropagation learning algorithm. During backpropagation training stage the weights of the network is adjusted iteratively using the gradient descent optimization until the errors of cost function is minimized so that the actual output vector, which is the network's activation vector, becomes closer to desired output vector through the network. Chapter 4 (section 4.5.4) will explain SAE that is used with T2D GWAS data

in more details. Refer to Figure 4.11 for the entire process of SAE (unsupervised pre-training and supervised fine-tuning stages).

This greedy layer-wise learning algorithm based on training the network layer-by-layer is nonetheless a very efficient way to convert high-dimensional data into low-dimensional data, allowing highly abstract non-linear structure between features to be discovered. Furthermore, this layer-by-layer greedy learning strategy allows weights to be initialised in regions near to a good local minimum, bringing better optimisation and generalisation (Bengio et al. 2007).

In this thesis we use this greedy layer-wise learning algorithm to extract the non-linear epistatic interactions between SNPs in T2D GWAS data and to initialise the weights of a fully connected multilayer perceptron softmax classifier before it is fine-tuned for the binary classification of T2D as either case or control.

3.6.4 Deep Learning Hyperparameters Optimisation

In order to improve and accelerate the performance of neural networks and to achieve better generalization, various studies and investigations have been conducted to optimise the learning process in NNs. One of the primary elements is regularization, which is a strategy used to avoid overfitting and enhance performance. For example, dropout regularization (Srivastava et al. 2014), randomly removes hidden neurons from the network during the training process. Other researchers have developed different techniques to improve the training processes in NNs, such as tuning the learning rate and momentum term (Schmidhuber 2015). A complete description of tuning hyperparameters used to improve the training stage in neural networks in this thesis is provided in Table 3.2.

Table 3.2: Definition of Tuning Parameters used with Neural Networks

Tuning Parameter	Description of Tuning Parameter
Input dropout ratio	A fraction of the features for each training row to be removed from training. It is useful when the feature space is large and noisy. This can improve generalization.
Stopping metric	Is used to determine the metric to use for early stopping.
Stopping tolerance	Is used to set the relative tolerance for metric-based stopping to stop training when improvement is less than tolerance value.
stopping rounds	Is used to stop training if the option selected for the stopping metric doesn't improve for the specified value for training rounds.
Learning rate	Is a function of the difference between the predicted value and the target value (step size of weight to update during training). Backpropagation is used to correct the output at each hidden layer. A large learning rate leads to oscillatory traps in the learning process thus passing the local minimum. While a small learning rate can result in training freezing in a local minimum. Learning rate controls how slowly or quickly an NNs model learns the problem.
Rate annealing	Is used to ensure the learning rate does not freeze into local minimum.
Rate decay	Is used to control the change of the learning rate throughout layers.
Momentum start	Is used to control the amount of momentum at the beginning of training.
Momentum stable	Is used to control the amount of learning for which momentum increases.
Momentum ramp	Is used to control the final momentum value reached after momentum ramp training samples.
Max w2	Is a maximum sum of the squared incoming weights in any single neuron. It is useful when the activation function is set to Rectifier. This helps stability when the Rectifier is used.
L1 (Lasso), L2 (Ridge)	Are regularization techniques to modify the cost function and minimize cost. This helps prevent overfitting and improve generalization.

3.7 Traditional Machine Learning Algorithms

Technically machine learning algorithms are developed to either model linear or non-linear effects. In this thesis, the linear learning algorithm used is the Generalized Linear Model (GLM). The non-linear learning is based on decision trees (i.e. Recursive Partitioning and Regression Trees (RPART)), Random Forest (RF), Stochastic Gradient Boosting (GBM), and Support Vector Machines (SVMs) with Radial Basis Function Kernel.

3.7.1 Generalized Linear Models

Generalized linear models are statistical methods that extend linear modelling frameworks that allow for response variables that are not normally distributed (Nelder & Wedderburn

1972). GLM is commonly used to model binary data and consists of three components: random component, linear predictor, and the link function (Fox 2008). A random component specifies the probability distribution of the response variables. There are several statistical distributions referred to as an exponential family of distributions which include: Gaussian (normal), binomial, Poisson, gamma, and multinomial (Nelder & Wedderburn 1972). In a binary classification model, binomial (Bernoulli distribution) is commonly used where the output is either 1 or 0.

The linear predictor assumes a linear mapping between independent variables and the response variables (outcomes) through a link function. The link function describes the relationship between the linear predictor and the mean (expectation of the response variables) of the probability distribution (Nelder & Wedderburn 1972). In GLM, to fit a dataset the maximum likelihood method is used. This method provides an estimate of the model parameters through an iteratively reweighted least-squares (IRLS) procedure to minimize the loss function (error) with respect to the weights of the independent variables (Fox 2008).

In this thesis, logistic regression which is described as a GLM is utilized for binary classification tasks. A detailed explanation of logistic regression has already been provided in Chapter 2 when discussing association analysis in GWAS (section 2.8.3).

3.7.2 Decision Trees

A decision tree is a recursive partitioning algorithm that can be used in classification and regression tasks (Breiman 1984). The algorithm adopts a tree representation to create a training model to predict target variables (class) by learning decision rules inferred from training data. The decision tree constructs from; a root node, internal nodes, and terminal nodes or leaves. The tree has a single root node assigned to the whole training data, and each internal node corresponds to an attribute. Each terminal node corresponds to a class label (Berk 2016).

Technically, the tree is grown by splitting the source data into subsets (left and right branches) based on the attribute value test following a splitting rule, starting from the root node. This process is repeated in a binary recursive partitioning manner at each node, particularly internal nodes. The tree continues to grow until no additional splits can be created. Figure 3.4 illustrates an example decision tree workflow.

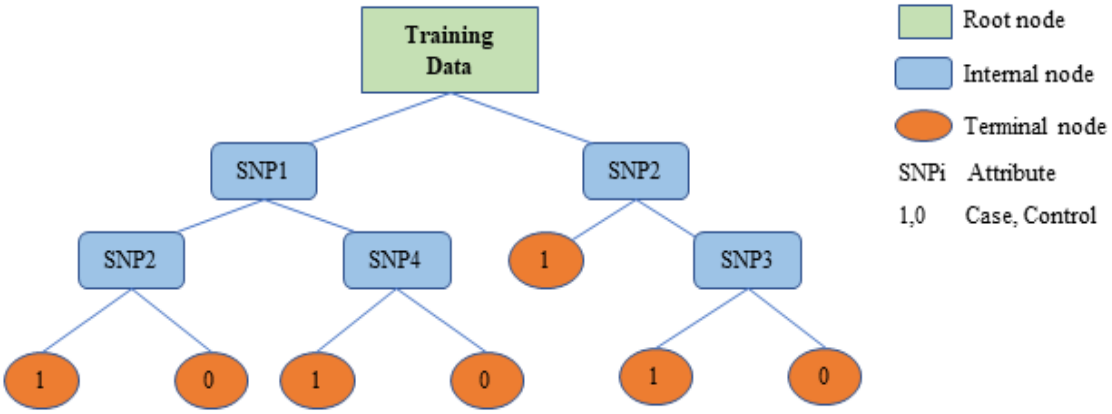


Figure 3.4: Decision Tree Classifier

The primary challenge in the construction of decision trees is to identify which attributes are required during the splitting process at each node and where the split should be imposed. This is defined using splitting rules (Buntine & Niblett 1992) in which node impurity is minimized and homogeneity is maximized using specific criteria. There are several commonly used splitting criteria for classification trees including information gain also known as entropy index, Gini index, and towing (Shih 1999). For example, during the splitting process the attribute with the highest information gain measure is selected.

Another challenge in the construction of decision trees is tree growth (Moisen 2008). Essentially the growing process is stopped when no further splits can be enforced due to a lack of data at a node. This means that the tree continues to go deeper and deeper almost to the point where it fits the training data perfectly, resulting in overfitting and poor accuracy on unseen data. One way to solve this problem and obtain the optimal size of the tree is to use a pruning algorithm (Esposito et al. 1997). Pruning involves reducing the size of a tree

to the optimal size by removing splits that generate two terminal nodes which in turn do not improve the performance of the tree on test data.

Overall, decision trees, offer a simplified method for the interpretation of complex tree results and are capable of handling missing values and outliers in data (Rokach & Maimon 2005).

3.7.3 Random Forests

The random forest algorithm is a randomized decision tree-based ensemble developed for classification and regression tasks (Breiman 2001). RF uses a collection of trees rather than a single tree. These trees are typically grown from thousands of trees and each tree is grown using a bootstrap aggregation or bagging technique. Bagging (Breiman 1996) is one of the ensemble techniques that builds many independent models or learners to allow trees to grow independently. The classification results each tree produces are combined using a voting technique. Bagging is an ideal technique for high-variance data with low-bias (Hastie et al. 2009) where noisy models are averaged, which removes biases and reduces variance.

The random forest is constructed by generating several bootstrap samples using the original data. For each bootstrap sample, the tree is grown, and a random subset of predictor variables is selected to split the tree node. The best split is calculated using these randomly selected candidate variables. This process is continued until the tree is fully grown without pruning, resulting in a forest of decision trees. Each tree is trained on a particular bootstrap sample of observations. Observations not considered in a specific bootstrap are used as out-of-bag (OOB) observations. The OOB samples are used as a test dataset to estimate error and permutation-based variable importance measurement for variable selection. The prediction of unseen data is based on majority voting for classification.

Given an F feature set from the original data, F consist of $fA_1 \dots fn_n$, where n represents the number of predictor features in the given dataset. The random forest starts by selecting

several bootstrap samples from the original data. A random split from the initial data, T, into several decision trees, $T_1, T_2, \ldots, T_t$ using bootstrap samples is performed to construct the forest as illustrated in Figure 3.5. The classification result is obtained using a vote system to identify the most popular classes.

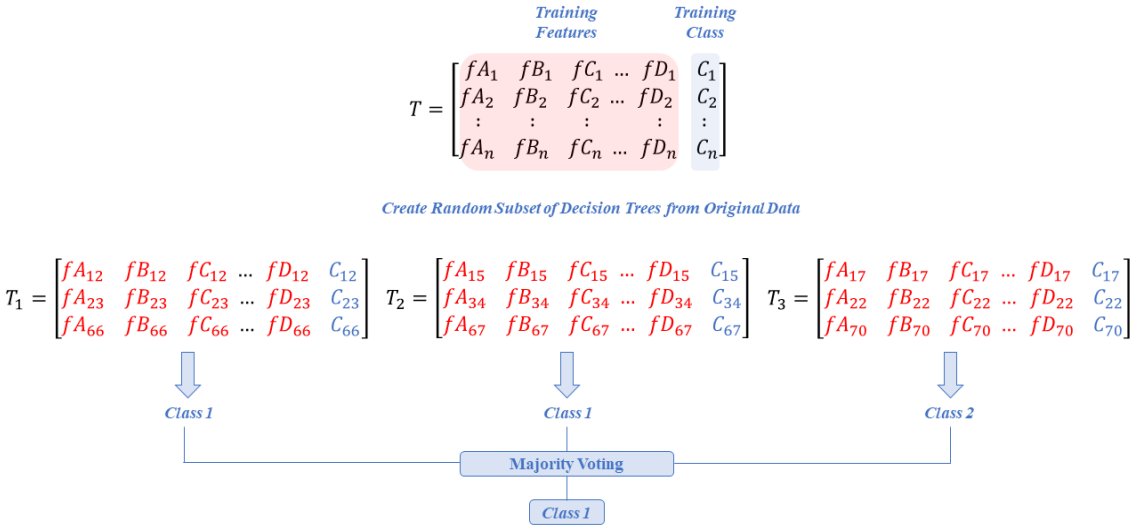


Figure 3.5: Random Forest Workflow

The RF classifier model is a highly recommended algorithm for high-dimensional data such as GWAS (Chen & Ishwaran 2012; Qi 2012). It has been successfully used in many genetic studies (Botta et al. 2014; López et al. 2018; Schwarz et al. 2010; Kursa 2014). This is because the algorithm is highly data adaptive and can handle correlations and interactions between features and can also rank variables using variable importance measures (Chen & Ishwaran 2012). In addition, deep trees promote low bias, while bootstrap aggregation improves the performance of the final model by de-correlating trees and reducing variance (Chen & Ishwaran 2012).

3.7.4 Stochastic Gradient Boosting

Gradient boosting is another ensemble tree-based method based on the combination of two powerful techniques including gradient-based optimization and boosting (Hastie et al. 2009). Gradient-based optimization computes the gradient to minimize a model's loss function in training data. Boosting algorithms (Kearns 1988) sequentially add new weak, base-learner models to the ensemble to create a strong learning system that obtains better performance

than a single model for predictive tasks. A weak learner is a learner whose error rate is only slightly better than random guessing (Hastie et al. 2009). A weak learner is represented by a decision tree model.

The learning procedure of gradient boosting starts by additively fitting weak learners (new models) to obtain a more accurate estimate of the response variable. The results in a new model being trained based on an error from previous models in the ensemble, are trained during each iteration. The algorithm allocates weights to each resulting model and applies a weighted average to produce the final classification result. Figure 3.6 illustrates gradient boosting workflow.

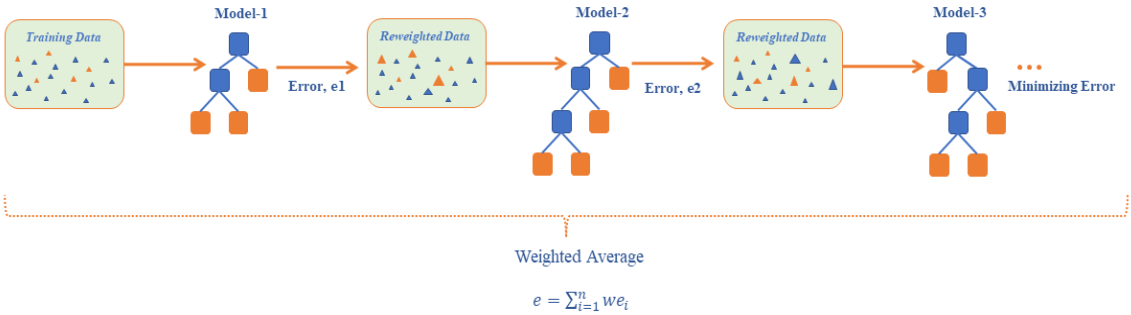


Figure 3.6: Gradient Boosting Workflow

GBM is subjected to overfitting where generalization capabilities are decreased. This is a situation where new decision tree models are added to the ensemble until the training data is completely overfitted. There are a number of different approaches to prevent the GBM model's overfitting. The technique adopted in this thesis is regularization through shrinkage (Natekin & Knoll 2013). Shrinkage also known as the learning rate is used to reduce the impact of each new model added to the ensemble. This means if the model's error is high during one of the boosted iterations, its negative impact on the ensemble model can be corrected in subsequent steps. Setting shrinkage to a small value can improve the model's ability to generalise on new data but at the cost of convergence speed.

The Stochastic gradient boosting (Friedman 2002) algorithm is adopted in the implementation of this research work which is one of the gradient boosting methods

developed to incorporate randomness into the fitting procedure. Specifically, a randomly selected subsample of the training data is used to fit the base-learner model instead of the full training data. The subsampling procedure improves generalization and reduces computational burden (Natekin & Knoll 2013).

3.7.5 Support Vector Machines

Support vector machines (SVMs) are a supervised discriminative classifier formally developed to find decision boundaries represented by hyperplanes in an n-dimensional space (where n refers to the number of features) that classifies the data points samples (attributed to different classes). The original SVM (Vapnik & Lerner 1963) is a non-probabilistic binary linear separation that for a given set of training samples each point in space is marked as belonging to one of two classes. Typically, SVM chooses the optimal hyperplane with the maximum margin distance to the closest training data points (support vectors) of any class instances. In general, the generalisation error of the model improves with larger margin.

For non-linear separation problem (Cortes & Vapnik 1995) SVM uses a technique called kernel. In kernel method (Mercer 1909; Aizerman et al. 1964) the data is transformed into another dimension, mapped into a higher dimensional feature space, that has a clear separating margin between the data points of different instances. This mapping is attained by using one of the kernel functions, i.e. hyperbolic tangent, polynomial, and radial basis function. For SVM algorithm to output the optimal hyperplane that possesses maximum margin, gradient decent optimisation along with the regularisation parameters are used to adjust the weights of the cost function and thus minimise the classification error on unseen data. Although SVM can be used to avoid the difficulties of using linear functions in the high-dimensional feature space by means of the kernel transformation, it lacks the transparency of model results and does not directly provide probability estimates. Figure 3.7 demonstrates an example of a separable problem in features space with the optimal hyperplane and the maximum margin to the nearest support vectors of the two categories.

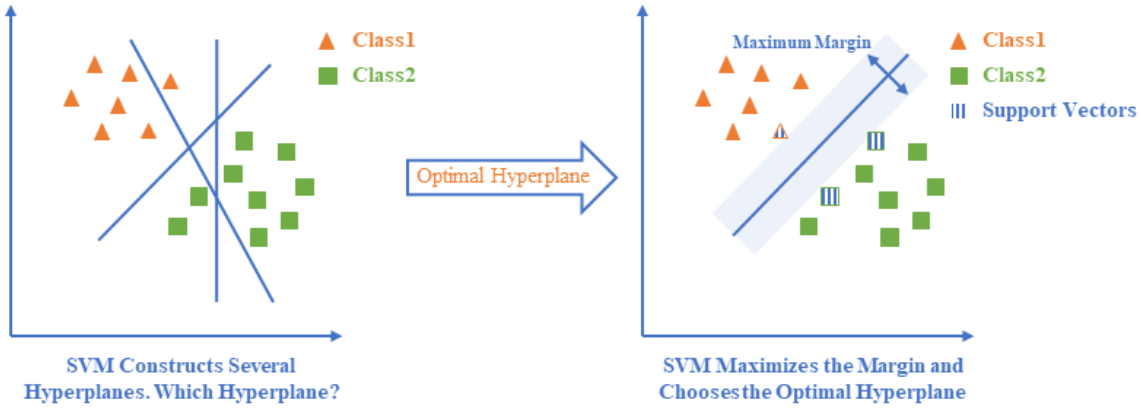


Figure 3.7: Support Vector Machine Example

3.8 Feature Selection

Feature Selection is a technique designed to find an optimal subset of features from the original dataset (Saeys et al. 2007). This is in contrast to other dimensionality reduction and compression techniques where the original representation of the variables is abstracted and altered (Veerabhadrappa & Lalitha 2010). Feature selection has become a necessity in several application domains, offering manifold advantages. These advantages include providing less computationally intensive models, avoiding overfitting, optimized model performance, and model interpretation (De Silva & Leong 2015). There are various types of feature selection techniques and each differs in how they integrate the feature selection search in the model hypothesis space.

Feature selection techniques can be arranged into three categories – filter, wrapper and embedded methods (Chandrashekar & Sahin 2014). Filter methods are based on calculating the feature relevance score and removing the ones that have the lowest scores. The search in the feature subset space is considered separately from the search in the hypothesis space, ignoring the interaction with the model selection and feature dependencies (redundant features may be nominated). In contrast, the wrapper methods incorporate the model hypothesis search within the feature subset search, allowing for the generation and evaluation of various subsets of features in addition to accounting for feature dependencies. Embedded methods are similar to wrapper methods as they are specific to a given learning

model with the advantage of dealing with computational complexity better than wrapper methods.

In this thesis the Recursive Feature Elimination algorithm (RFE) is used (Dong et al. 2015). RFE is a wrapper method that recursively evaluates models by adding or removing features to search for an optimal combination of variables that improve and maximise model performance. The procedure is initiated by fitting an initial model to the training set using all features. During the process of feature selection, each feature is ranked using its importance to the model where the top ranked features are maintained. The model is refitted and its performance reassessed using this subset of top ranked features. To better estimate the performance of the model, a 10-fold cross-validation resampling can be used. Although resampling methods are computationally burdensome, incorporating them with RFE can advance the probabilistic assessment of feature importance and provide better performance estimation than using a single fixed dataset.

3.9 The Application of Machine Learning into T2D

Machine learning has already been successfully applied to a wide range of medical applications to discover SNP interactions and investigate the discriminative capabilities of risk susceptibility to T2D. Zhu *et al.* (Zhu et al. 2013) considered the generalized multifactor dimensionality reduction (GMDR-GPU) approach for detecting gene-gene interactions. The study identified 24 core SNPs that appear to be important for T2D. Wang *et al.* (Wang et al. 2014) investigated gene-gene interaction using the lasso-multiple regression (LMR) approach. Researchers found that the SNPs from genes CDKN2BAS and KCNJ11 are significantly associated to T2D. Random forest and T-Trees (TT) for T2D GWAS have been implemented in (Botta et al. 2014) for exploiting SNP correlations. The investigation suggested that the T-Trees method was able to recover most of the loci already reported in the literature. Furthermore, the T-Trees method outperformed RF with a classification prediction rate of 83.4% and 75.8% respectively.

Ban *et al.* (Ban et al. 2010) conducted a study using an SVM to identify combinations of SNPs for the prediction of individuals' susceptibility to T2D. A subset of the best SNP combinations using 14 SNPs selected from possible candidate SNPs (408 SNPs) was used as features for the classification of disease risk. This work obtained a 65.5% classification prediction rate with 56.7%, and 73.9% for sensitivity and specificity respectively. Furthermore, the authors investigated subpopulation datasets by gender using similar techniques and found different SNP combinations. The results yielded slightly better accuracy rates of 70.9% (Sens = 71.4%, Spec = 70.4%) and 70.6% (Sens = 71.5%, Spec = 69.6%) for men and women datasets respectively. The authors concluded that epidemiological evidence for sex differences exists in T2D. In the study conducted by López *et al.* (López et al. 2018), Random Forest, Support Vector Machine and Logistic regression algorithms were applied for learning predictive models for T2D, first using SNP data only and second using SNP data combined with clinical data. Using SNP data only, the results revealed that the RF produced an AUC = 85.3% which outperformed the LR and SVM methods with AUC = 83.5%, and AUC = 82.5% respectively. However, adding clinical data including sex, Body Mass Index and age the results suggested that the predictive ability of the models improved. The AUC for RF, LR increased to 89%, 84.4% respectively. The authors concluded that the RF is a useful technique for SNP data that can model feature interactions and deal effectively with overfitting and missing value. In the study conducted by Gül *et al.* (GÜL et al. 2014), the authors used binary logistic regression (LR) to investigate missing heritability and early risk prediction for T2D in two separate studies that considered genetic data only followed by genetic and clinical data analysis. The authors revealed that using 798 SNPs, the classification predictive rate of genotype analysis achieved 96.5%. However, the additive contribution of clinical data to the analysis, resulted in 98% classification accuracy.

Kim *et al.* (Kim et al. 2018) tested deep neural network (DNN) using several subsets of SNPs extracted through Fisher's exact test and L1-penalized logistic regression. The results demonstrated that using 678 SNPs in male samples, it was possible to achieve 93.1% and 85.7% predictive accuracy for DNN and LR respectively. Using the female datasets samples, DNN and LR achieved 92.8% and 90.2% respectively. While adding clinical data to the analysis, the results showed improvements in the predictive accuracy of the DNN with 94.8% for males and 94.6% for females. LR produced 84.7% and 83.3% for males and females respectively. Malovini *et al.* (Malovini et al. 2012) proposed a Hierarchical Naïve Bayes (HNB) for the classification of T2D genetic data. The HNB model was designed to account for SNPs in linkage disequilibrium. The results showed that HNB classification performance was higher than those obtained using standard Naïve Bayes (NB) with 92% and 90% respectively. Table 3.3 summarised the previous studies in T2D.

Table 3.3: Previous Works in T2D

Author	Year	Model	AUC		Sens	Spec	Features	Analysis
Ban *et al.*	2010	SVM	0.653 0.709 0.706	-Total -Male -Female	0.567 0.714 0.715	0.739 0.704 0.696	Genetic	Classification
López *et al.*	2018	RF LR SVM RF LR SVM	0.853 0.835 0.825 0.89 0.844 0.825				Genetic Genetic and Clinical 	Classification
Malovini *et al.*	2012	HNB NB	0.92 0.90		0.89 0.89	0.93 0.92	Genetic	Classification
Kim *et al.*	2018	DNN LR DNN LR	0.931 0.857 0.948 0.847	-Male			Genetic Genetic and Clinical 	Classification
		DNN LR DNN LR	0.928 0.902 0.946 0.833	-Female			Genetic Genetic and Clinical 	
Gül *et al.*	2014	LR	0.965 0.980				Genetic Genetic and Clinical	Classification
Zhu *et al.*	2013	GMDR-GPU					Genetic	Gene-Gene Interactions
Wang *et al.*	2014	LMR					Genetic	Gene-Gene Interactions
Botta *et al.*	2014	RF TT	0.758 0.834				Genetic	Gene-Gene Interactions

3.10 Summary

One of the challenges in computational biology is how to explore, understand and interpret complex, large biological data. More specifically, how to extract important information from the raw data, and use it to explain the underlying cause of complex diseases. This chapter presented and discussed deep learning and its effectiveness in converting high-dimensional data to low-dimensional data while maintaining and extracting important information. Deep learning is adopted in this thesis to explore T2D GWAS data. The next chapter will discuss the methodology used in this thesis to explore epistatic interactions in T2D GWAS using deep learning stacked autoencoders.

Chapter 4 Proposed Methodology

4.1 Introduction

This Chapter discusses the proposed methodology. This includes the data authorisation process and a description of the Nurses' Health Study and the Health Professionals Follow-up Study in T2D obtained from the Database of Genotypes and Phenotypes. The data quality control procedures are also discussed, and the results are reported. Following the process of removing unreliable information (individuals and markers), the analysis conducted to perform logistic regression and association analysis for the population-based case-control study design is presented.

Furthermore, this chapter presents the proposed novel framework posited in this thesis. It discusses the use of deep learning stacked autoencoders in large-scale GWAS as a feature extraction mechanism to pre-initialise a multilayer perceptron (MLP) for T2D case-control classification tasks. The classification and evaluation performance for a random forest and multilayer perceptron classifier are presented. This will be utilised in the results chapter to provide a set of baseline results for comparison with the proposed novel framework.

The study investigates genotypic risk factors along with other risk factors that include clinical, environmental exposure, and sociodemographic factors for the classification of T2D in case-control cohorts. The classification and evaluation performance for five traditional supervised machine learning algorithms are presented.

4.2 Data Acquisition

The data utilised in this research was obtained following authorised access to the Database of Genotypes and Phenotypes (dbGaP) (Tryka et al. 2014). The Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS) in T2D (Study Accession: phs000091.v2.p1) are used to demonstrate the applicability of the proposed approach posited in this thesis. The NHS and HPFS cohorts are part of the Gene Environment Association

Studies initiative (GENEVA, http://www.genevastudy.org) funded by the trans-NIH Genes, Environment, and Health Initiative (GEI). The following sections provide an in-depth description of both datasets.

4.2.1 Data Description

The NHS (Nurses' Health Study) was established in 1976, and the HPFS in 1986. NHS participants include 121,700 female registered nurses aged between 30 and 55 of age that reside within 11 U.S states. HPFS participants include 51,529 male health professionals aged between 40 and 75 years from 50 U.S states. The NHS and HPFS participants responded to a questionnaire requesting information related to their medical history and lifestyle characteristics. Since then, on a 2 to 4-year cycle, cohort members have been asked to provide dietary information using validated semi-quantitative food frequency questionnaires. Participants were also asked to provide blood samples, in which 32,826 members of the NHS and 18,225 members of the HPFS responded. The case and control participants were selected from those who provided a blood sample. Case participants were identified as those who reported themselves to be affected by T2D and it was confirmed by a medical record validation questionnaire. Control participants were defined as those without diabetes. The DNA of case and control participants were genotyped at the Broad Centre for Genotyping and Analysis (CGA) using the Affymetrix Genome-Wide Human 6.0 array (Affymetrix is a DNA microarray technology that enables multiplex and parallel analysis of biological systems at the cell, protein, and gene level).

A total of 6041 NHS and HPFS case-control subjects with genotype information across 909622 SNPs successfully passed the initial quality control at the Broad CGA and were used in the final version of the dataset. The NHS subjects consist of 1581 T2D cases and 1854 controls, and HPFS contains 1232 T2D cases and 1374 controls. Participants in the NHS dataset were identified as Hispanic or non-Hispanic and each belongs to one of four racial categories (White, African-American, Asian or Other). Participants were however mainly

White and non-Hispanic representing 97.4% in the NHS dataset. The HPFS participants belong to one of the four racial categories (White, African-American, Asian or Other). They were predominantly White representing 96% in the HPFS dataset. Table 4.1 summarises the NHS and HPFS subjects and their ethnicity.

Table 4.1: NHS and HPFS Subject's Ethnicity

Racial Category	**NHS**		**HPFS**	
	Case	Control	Case	Control
White	1551	1779	1184	1283
African-American	17	13	12	14
Asian	6	11	12	14
Other	7	7	24	27

The NHS and HPFS datasets include clinical and dietary data, along with each participant's age, gender, Body Mass Index (BMI), alcohol intake, smoking status, physical activity, height, weight, family history of diabetes among first degree relatives, high blood pressure, high blood cholesterol, polyunsaturated fat intake, magnesium intake, cereal fibre intake, and glycaemic load as demonstrated in Table 4.2. A comprehensive description for both the GENEVA NHS and HPFS datasets can be found in the quality control report in the GENEVA NHS and HPFS Type 2 Diabetes project (The Nurses' Health Study 2009; The Health Professional Follow-Up Study 2009).

Table 4.2: The Clinical Data for the GENEVA NHS-HPFS Datasets

Variable	**Description**	**Coded values**
idg	GENEVA identification number	
age	Age in years	
bmi	BMI in kg/m2	
hisp	Hispanic ethnicity	1= Hispanic 2= Not-Hispanic
case	Diabetes case status	0= Control 1= Case of T2D 2= Uncertain diabetes type
alcohol	Alcohol intake	
smk	Cigarette smoking	1= Never cigarette smoker 2= Past cigarette smoker 3= Current cigarette smoker
act	Total physical activity	

race	Race variable for NHS	1= White 2= African American 3= American Indian 4= Asian 5= Hawaiian
woman	Sex	1= Woman 2= Man
race2	Race variable for HPFS	1= White 2= Other 3= Asian 4= African American
ht	Height in meters	
Wt	Weight in kg at time of blood draw	
famdb	Family history of diabetes among first degree relatives	1= Yes 0= No
hbp	Reported high blood pressure at/before blood draw	1= With a history of hypertension 0= No history of hypertension
chol	Reported high blood cholesterol at/before blood draw	1= With a history of high cholesterol 0= No history of high cholesterol
pufa	Polyunsaturated fat intake	
magn	Magnesium intake	
ceraf	Cereal fibre intake	
gl	Glycaemic load	

4.2.2 Data Format

Both the NHS and HPFS datasets are in PLINK format. Technically, files in the standard PLINK format are very large and computationally challenging. As such, converting very large files to binary format is recommended and often performed using the PLINK v1.9 toolset. Binary formatted files considerably reduce the file size and significantly enhance computational efficiency.

Standard flat files in the PLINK format include the ped and map files. The ped file contains information about each individual in the study including Family ID (Fam ID), Individual ID (Ind ID), Paternal ID (Pat ID), Maternal ID (Mat ID), Sex, Phenotype (Pheno), and the complete genotyped data. The genotyped data is represented as SNPs. Each SNP is bi-allelic, meaning it contains only two nucleotides coded as A, T, C, or G. The map file contains information about SNPs and associated rsNumbers (SNP), Chromosome (Chr), and the corresponding Base-Pair coordinate *(physical position of SNP to chromosomal position)* as well as Genetic Distance (Gen Dist) *(the measure of genetic difference between species or between populations within a species, zero means no differences).* Table 4.3 and Table 4.4 show examples of standard flat files corresponding to ped and map files.

Table 4.3: Ped File

Fam ID	Ind ID	Pat ID	Mat ID	Sex	Pheno	rs1	rs2	rs3	…
FAM_T2D	60444	0	0	1	-9	CC	GG	TT	…
FAM_T2D	166692	0	0	2	-9	GG	GG	TT	…
FAM_T2D	167773	0	0	1	-9	GC	GG	TT	…
FAM_T2D	167362	0	0	2	-9	CC	GG	TT	…
FAM_T2D	166960	0	0	2	-9	GC	GG	TT	…

Table 4.4: Map File

Chr	SNP	Gen Dist	Base-Pair
2	SNP_A-1820282	0	24049
2	SNP_A-2056638	0	43652
2	SNP_A-1792446	0	49698
2	SNP_A-2063286	0	64387
2	SNP_A-2260913	0	76644

Binary files include bim, bed, and fam files. The bim file contains information similar to that in the map file in addition to Allele1 and Allele2 for each SNP in the ped file. The fam file contains the identification information for each participant. The information in the fam file is similar to the columns described in the ped file excluding the genotype data. The bed file is the largest file of the three in this binary set of files and contains a binary genotype data. This file contains all the SNPs used in the study as well as the genotype for each SNP in each participant. Table 4.5, Table 4.6, and Table 4.7 provide examples of corresponding binary files of fam, bim, and bed files respectively.

Table 4.5: Fam File

Fam ID	Ind ID	Pat ID	Mat ID	Sex	Pheno
FAM_T2D	60444	0	0	1	-9
FAM_T2D	166692	0	0	2	-9
FAM_T2D	167773	0	0	1	-9
FAM_T2D	167362	0	0	2	-9
FAM_T2D	166960	0	0	2	-9

Table 4.6: Bim File

Chr	SNP	Gen Dist	Base-Pair	Allele1	Allele2
2	SNP_A-1820282	0	24049	G	C
2	SNP_A-2056638	0	43652	G	A
2	SNP_A-1792446	0	49698	T	C
2	SNP_A-2063286	0	64387	G	G
2	SNP_A-2260913	0	76644	C	A

Table 4.7: Bed File

01101100	00011011	00000001
00011101	00011100	10010001
11111111	00111110	00011100
11100001	00011000	00101100
11001100	00000001	00110000

4.3 Data Quality Control

Data quality control (QC) and preliminary analysis is performed using PLINK v1.07 and v1.9 (Purcell et al. 2007) for Windows. PLINK is also used to merge the NHS and HPFS datasets (NHS and HPFS participants were genotyped using the Affymetrix Genome-Wide Human 6.0 array) and subsequent filtering procedures. Before QC, the 0 Chromosome was removed, and non-T2D participants, i.e. other types of diabetes (65 NHS, 68 HPFS), the HapMap controls (44 NHS, 29 HPFS), and those belonging to ethnicity other than white (61 NHS, 103 HPFS) were also excluded from the study. This study is restricted to white ancestry to reduce potential bias due to population stratification. The dataset was subjected to pre-established quality control protocols as recommended in (Anderson et al. 2010). In addition, quality control parameters are tuned to meet the requirements of the analysis presented in this study. Quality control assessments for individuals and genetic data are conducted separately.

4.3.1 Individual QC

Individuals that met any of the following criteria were discarded from the analysis. Samples with discordant sex information were identified using the X-chromosome homozygosity rate calculation. The expected homozygosity rate was less than 0.2 for female, more than 0.8 for male resulting in 14 samples being removed from the dataset. Figure 4.1 represents the X-chromosome homozygosity rate for female and male samples.

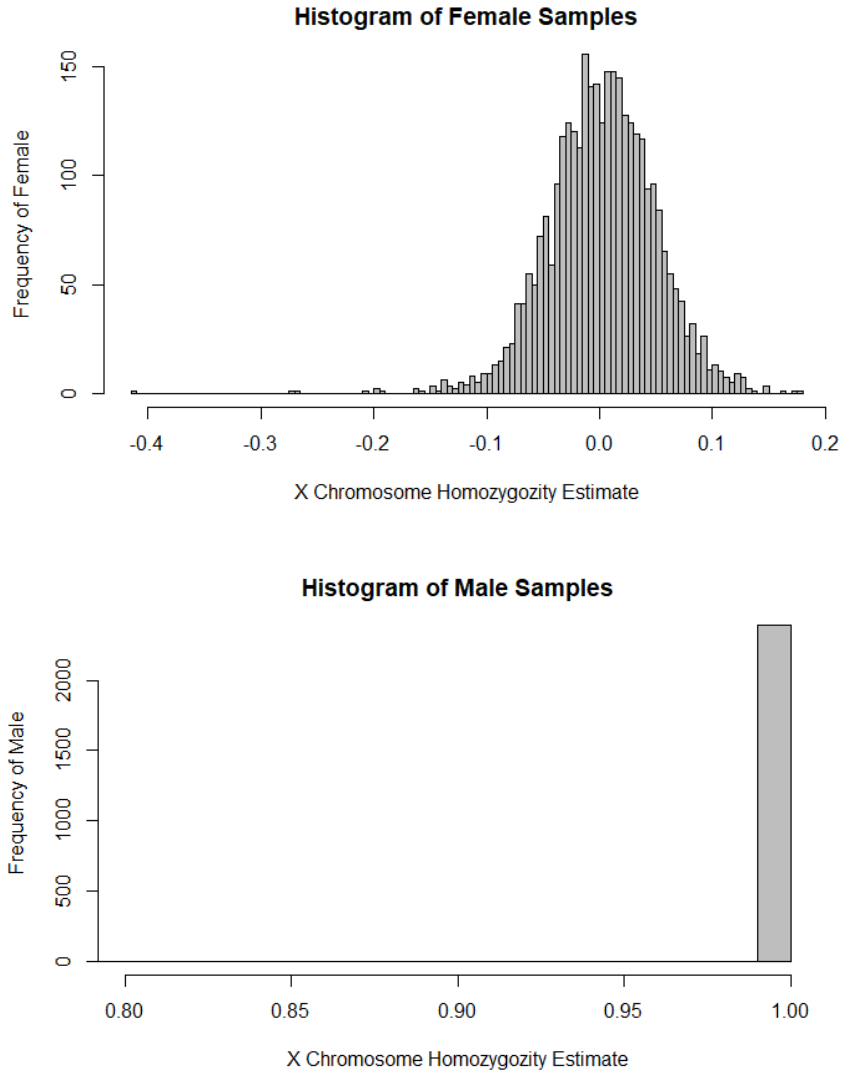


Figure 4.1: X-Chromosome Homozygosity Rate for Female and Male

Individuals with elevated missing data rates (genotype failure rate ≥ 0.05) and outlying heterozygosity rate (heterozygosity rate ±3 standard deviations from the mean) were identified resulting in 131 individuals being discarded from the analysis. Figure 4.2 demonstrates the proportion of missing SNPs with respect to the heterozygosity rate. Dashed lines indicate quality control thresholds and the dots represent the observed samples.

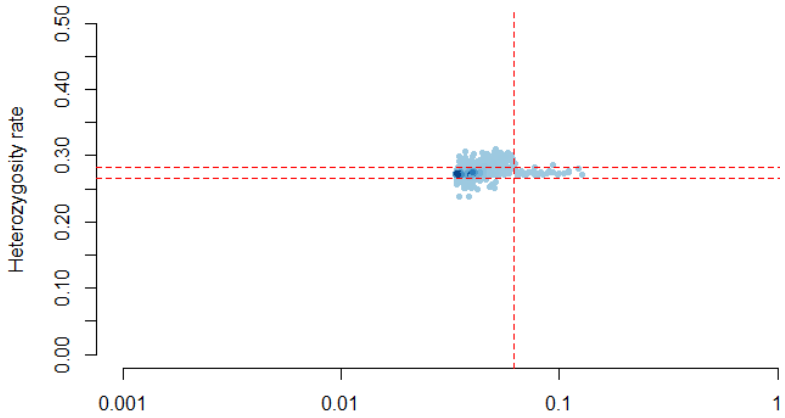


Figure 4.2: Genotype Failure Rate vs. Heterozygosity Rate

Duplicated or related individuals were identified by estimating Identity-by-descent for shared alleles. We have chosen to remove an individual from each pair with an IBD > 0.185, which is halfway between the IBD for third and second-degree relatives. Eight samples were found and excluded from the dataset. Figure 4.3 represents a histogram for the mean pairwise IBD distribution between all pairs of samples in the dataset. Vertical dashed lines indicate quality control thresholds (IBD >0.185).

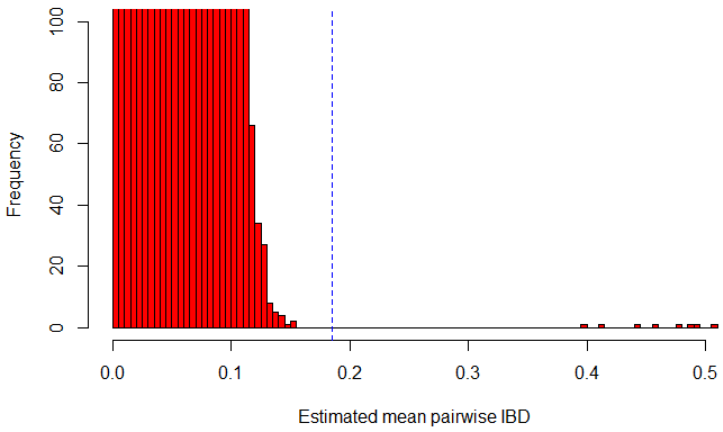


Figure 4.3: Histogram Showing the Distribution of Mean Pairwise IBD

To visualize the degree of relatedness between a pair of individuals the proportion of loci sharing one allele IBD (parent-child pairs), represented by Z1 is compared with the proportion of loci sharing zero allele IBD (unrelated), presented by Z0, in the genome file. Each point on the plot represents the relationship type between a pair of individuals as shown in Figure 4.4. This figure shows that many individuals identified are unrelated (black points).

Blue points describe second, third, fourth, and fifth degree relatives while the green points represent duplicated and first-degree relatives that have subsequently been discarded from the dataset.

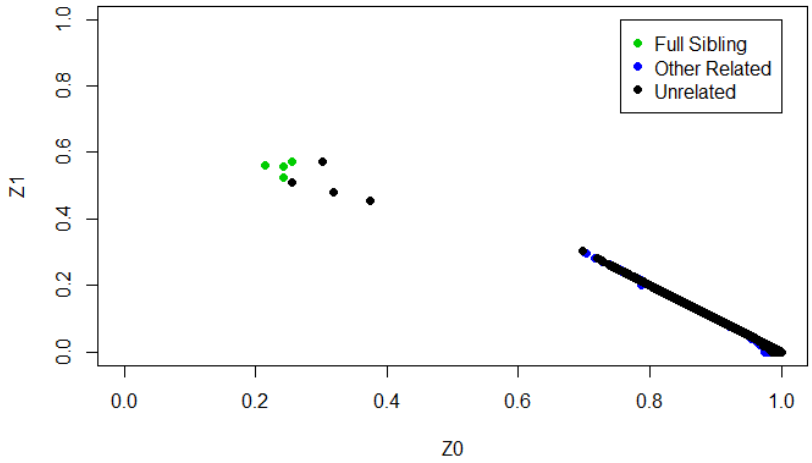


Figure 4.4: Degree of Relatedness

Individuals with divergent ancestry were identified using PCA. PCA is constructed using pruned genome-wide genotype data from a reference panel of HapMap phase III data consisting of four diverse ancestral populations including Europe (CEU), Asia (Chinese (CHB) and Japanese (JPT) populations), and Africa (YRI). The fact that there is large-scale genetic diversity between the four ancestral populations, means it is possible to use the first two principal components to separate and cluster samples from within the four groups. To identify samples with divergent ancestry in our dataset, these samples are clustered alongside the HapMap individuals. Using principal component scores, 51 individuals with a 2nd principal component score of less than 0.061 were removed. Figure 4.5 shows the principal component analysis plot for our dataset using HapMap phase III data for ancestry clustering. The black dots represent our dataset, CEU (red), CHB and JPT (purple) and YRI (green). The grey dashed line is the principal component score for identifying samples for removal. Furthermore, 101 individuals were removed due to missing genotype data rate of 0.05.

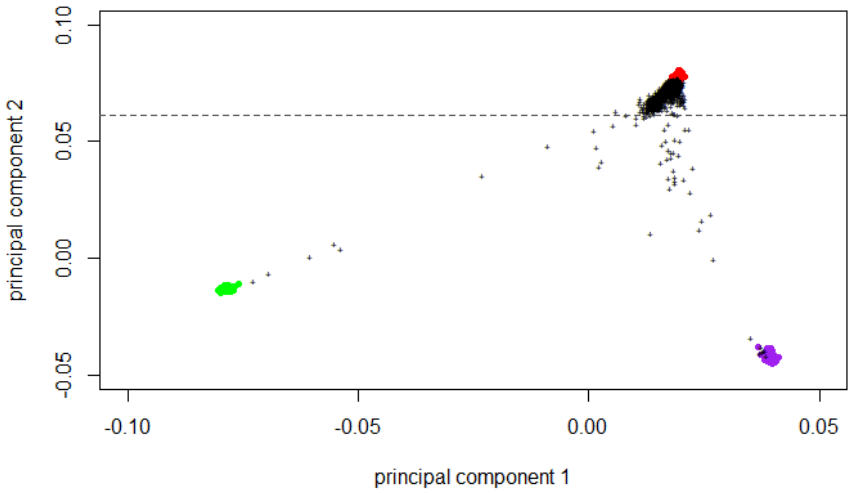


Figure 4.5: Principal Component Analysis

4.3.2 Genetic Marker QC

SNPs with excessively different ($p < 0.00001$) missing data rates between case and control samples were identified and removed, resulting in 29 SNPs being excluded from the analysis. Figure 4.6 shows a histogram of the missing genotype rate to specify the threshold used to elevate genotype failure rates. The dashed line indicates the quality control threshold used for genotype failure rates ≥ 0.05.

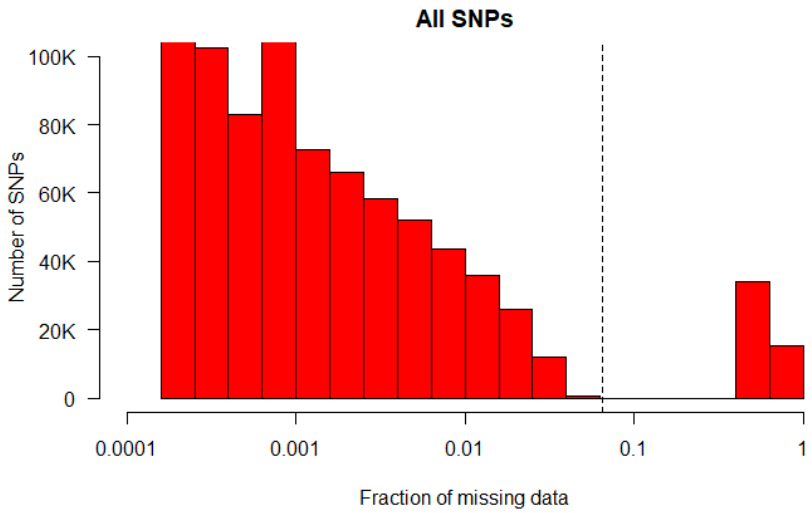


Figure 4.6: Histogram of the Missing Genotype Rate

In this analysis, SNPs that show extensive departure from HWE in control samples were excluded as these can be indicative of genotyping error. The significance threshold for identifying markers in HWE is set to p-value < 0.001. This resulted in 2248 variants being removed from the dataset.

All SNPs with a MAF threshold of <0.05 were identified in the dataset resulting in 178004 variants being excluded. Whereas, markers with low genotyping efficiency (call rate) were identified using a missing genotype rate of 0.01 resulting in 116863 variants being excluded from the dataset.

This concludes QC analysis. The final dataset used for subsequent analysis contained 5393 individuals (2481 cases, 2912 controls) with 608342 markers each. Figure 4.7 summarises the quality control procedures used with the NHS-HPFS dataset to obtain a subset of reliable markers and samples for subsequent association analysis.

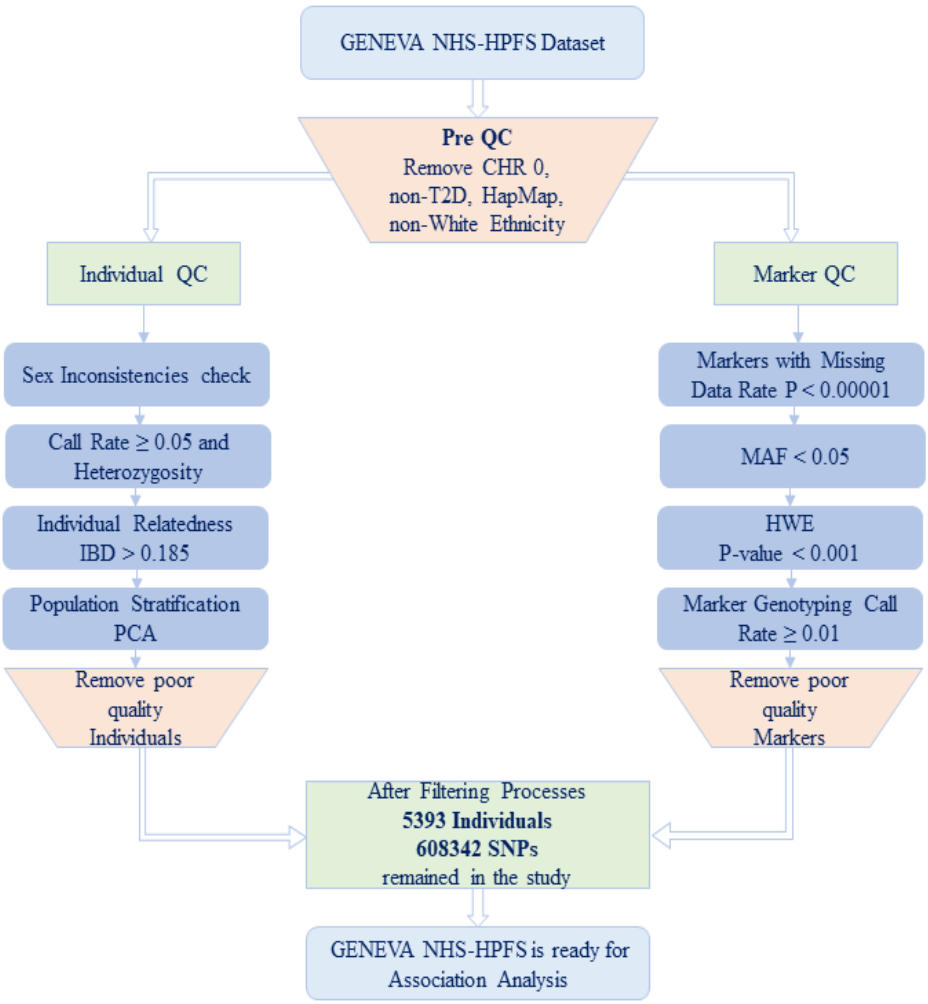


Figure 4.7: Quality Control Workflow for NHS-HPFS Dataset

4.4 Association Analysis using Quality Controlled T2D Dataset

In this section, population-based association mapping is presented. A standard case-control association analysis is conducted in an unrelated, white racial subpopulation to compare the frequency of genotypes at genetic marker loci (SNP) between cases and controls contained in the Geneva NHS and HPFS T2D datasets. Association analysis using logistic regression is performed with PLINK v1.9. This is a widely used approach within GWAS studies, under an additive genetic model to assess the association of all SNPs within the study with disease binary traits (0/1) for case and control subjects. Other models for disease penetrance are available including multiplicative, dominant and recessive models. However, additive models are the most commonly used in genetic association testing when the underlying genetic model is unknown and there are a large number of uncharacterised SNPs and outcomes (Clarke et al. 2011). Disease penetrance associated with a given variant (genotype) is defined as the risk of disease in individuals carrying that variant. In an additive genetic model of disease penetrance, an additive effect indicates that the risk of disease is increased y-fold for genotype Aa and 2y-fold for genotype AA (Clarke et al. 2011).

Furthermore, logistic association testing is adjusted using Genomic Control (GC) to control population structure, and p-values are considered based on a GC inflation factor λ. In addition, to detect statistically significant SNPs the Bonferroni-corrected genome-wide significance threshold $p < 5 \times 10^{-8}$ is used as highlighted in (Dudbridge & Gusnanto 2008). Odds ratio with a 95% confidence interval (95% CI) was measured to evaluate the strength of associations between SNPs and T2D and to determine if there is risk association, no association or protective association between an SNP and the phenotype of interest (in this instance T2D). To report the context of the SNPs identified, the database of single nucleotide polymorphism (dbSNP) was used (Wheeler et al. 2007). This tool is developed and provided by the National Center for Biotechnology Information (NCBI) in collaboration with the

National Human Genome Research Institute (NHGRI) and it contains genetic background information for all identified genetic variations (Wheeler et al. 2007).

4.5 Classification for High-Dimensional T2D GWAS Data

For Classification tasks using T2D GWAS data, state-of-the-art algorithms in machine learning including random forest and multilayer perceptron classifiers are used and benchmarked against a deep learning stacked autoencoder. The performance of these advanced machine learning methods is evaluated to assess their discriminating capabilities when classifying observations with T2D (cases) and without T2D (controls) using the GENEVA NHS-HPFS GWAS dataset. The analyses were conducted using R Studio utilizing the H2O package (Aiello et al. 2018; Candel et al. 2018). Figure 4.8 shows the proposed methodological framework for the approach posited in this thesis.

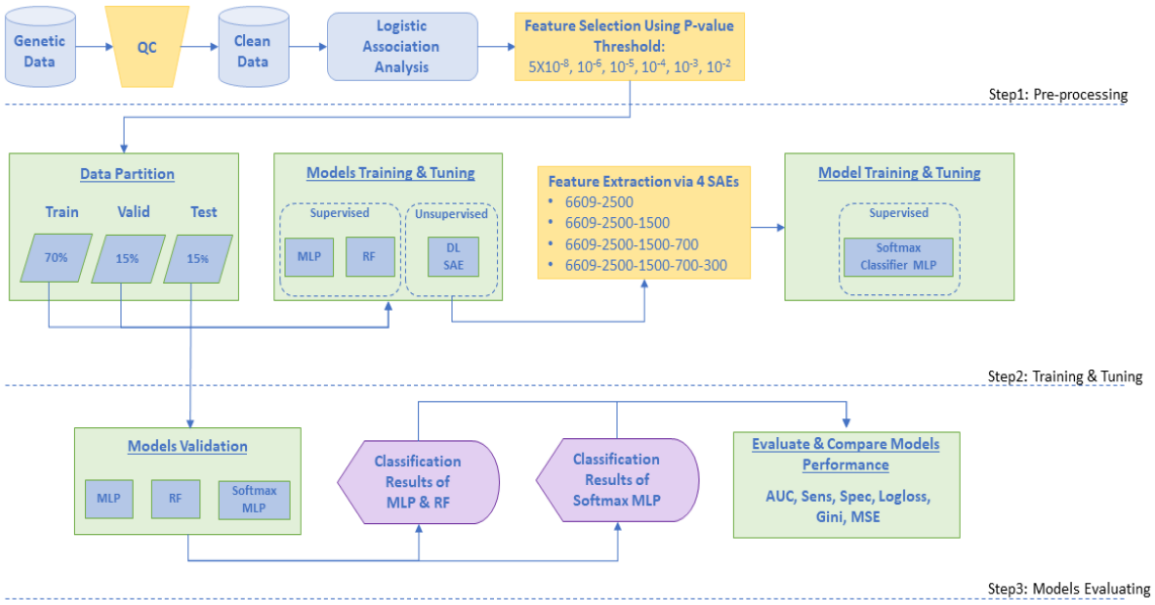


Figure 4.8: Methodology Framework for High-Dimensional Data

The following sections present the experimental configurations and our methodology in more detail.

4.5.1 Extracting Groups of Features from Association Analysis

Logistic regression association analysis is employed to assess the association between all SNPs and the T2D phenotype. Most GWAS genotypes between 500,000 and one million SNPs, and in some studies significantly more. Using such a large number of genetic variables

to train classification models is computationally difficult. One simple and common approach is to filter a subset of genotype SNPs to remove less useful information (Bush & Moore 2012). This can be achieved by selecting a set of SNPs resulting from logistic association for single SNP analysis with different significance thresholds.

In this study several p-value thresholds are considered including 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} resulting in 7, 13, 23, 103, 766, and 6609 SNPs respectively. These subsets of SNPs are used to exhaustively evaluate the non-linear epistatic interactions in each subset and assess the predictive capacity of advanced machine learning in discriminating between cases and controls in T2D GWAS.

4.5.2 Classification using Multilayer Perceptron

A multilayer perceptron (MLP) that is trained with gradient descent optimization using the backpropagation learning algorithm is implemented in this analysis for binary classification tasks, based on the theoretical definitions in (LeCun et al. 2015; Candel et al. 2018; Ng 2011). The MLP is constructed using input, hidden, and output layers containing a pre-defined number of units (neurons) – depending on the evaluation. Let n_l denote the number of layers in the network where l is a layer and L_l is a particular layer. Thus, L_1 is the input layer and L_{n_l} is the output layer in the network. First the input vector is transmitted to the input neurons in the input layer and then the outputs from the input neurons are passed to the hidden neurons in the hidden layer, which is the second layer. This process is continued until the last layer of the hidden layers is reached. Then, the outputs of this last hidden layer are sent to the output neurons in the output layer. In addition to the layers and neurons, the neural network consists of several parameters including the weight and bias. The parameters $(W, b) = \left(W^{(1)}, b^{(1)}, W^{(n)}, b^{(n)}\right)$ where $W_{ij}^{(l)}$ denotes the weight of the connection between unit j in layer l, and unit i in layer $l + 1$. Additionally, the bias unit $b_i^{(l)}$, associated with unit i in layer $l + 1$ is used with output value equal to $+1$. The number of units in layer l is

represented by s_l, and a bias unit $b_i^{(l)}$ which is not counted with s_l. The output value of unit i in layer l is given by an activation vector $a_i^{(l)}$ which is equal to the total weighted sum of inputs (including the bias term), denoted by $z_i^{(l)}$. Thus, $a_i^{(l)} = f\left(z_i^{(l)}\right)$ where $z_i^{(l)}$ is given as:

$$z_i^{(l+1)} = \sum_{j=1}^{n} W_{ij}^{(l)} x_j + b_i^{(l)} \tag{4.1}$$

Given a fixed setting of parameters W, b the neural network hypothesis is defined as $h_{W,b}(x)$ which gives the real number output as:

$$h_{W,b}(x) = a_i^{(l)} = f\left(z_i^{(l)}\right) \tag{4.2}$$

where $f: \mathbb{R} \mapsto \mathbb{R}$ represents the activation function. Basically, there are various types of activation function which include the sigmoid function, hyperbolic tangent, rectifier linear unit, and maxout. It is challenging to choose which activation function to adopt for our dataset, thus we let the network model select which of these activation functions best fits our dataset.

Following the forward pass calculation of all the activations in layer L_2 L_3, and so on up to the output layer L_{n_l} to compute the output value throughout the network, including the output value of the hypothesis $h_{W,b}(x)$, the error term for each unit in previous layers is computed using the backpropagation algorithm. The weights of the network are then adjusted through iterative updates using gradient descent.

Given a fixed training set $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$ of m training samples, parameter x is a vector containing the input features for a sample and y the outcome, the neural network can be trained using gradient descent optimization.

The overall cost function when using the mean squared error cost function is defined as (Ng 2011):

$$J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m} J\left(W,b;x^{(i)},y^{(i)}\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2$$
$$= \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{W,b}\left(x^{(i)}\right) - y^{(i)}\right\|^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2 \quad (4.3)$$

And in case of using the cross-entropy cost function, the overall cost function is defined as:

$$J(W,b) = \left[-\frac{1}{m}\sum_{i=1}^{m} J\left(W,b;x^{(i)},y^{(i)}\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2$$
$$= \left[-\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\log h_{W,b}\left(x^{(i)}\right) + \left(1-y^{(i)}\right)\log\left(1-h_{W,b}\left(x^{(i)}\right)\right)\right] \quad (4.4)$$
$$+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2$$

In cost function $J(W,b)$, The second term is a weight decay term which is a regularization term that penalizes large weights. The weight decay term (L2 regularization penalty) is used to add a penalty to the error function to reduce the magnitude of the weights. This makes the weight values to decay towards zero (but not exactly zero) and thus prevent overfitting. The weight decay parameter, λ, is used to control the relative importance of the first and second terms of the cost function. Typical values of λ range between 0 and 0.1 (Kuhn & Johnson 2013). Too small of a λ can lead to overfitting the data, while too large values of λ can lead to underfitting the data. Therefore, grid search is used to choose the optimised λ value.

To train the neural network model, random initialisation of parameter $W_{ij}^{(l)}$ and each $b_i^{(l)}$ to a value close to zero is applied. This step is essential to stop hidden layer units learning the same function of the input. More specifically, if all the parameters $W_{ij}^{(l)}$ and $b_i^{(l)}$ are initialise using the same values, activations and output values for all units will be the same ($a_1^{(2)} = a_2^{(2)} = a_3^{(2)} = \cdots$) for any input x.

The gradient descent optimization algorithm is used to updates parameters W, b as define below:

$$W_{ij}^{(l)} := W_{ij}^{(l)} - \alpha \frac{\partial}{\partial W_{ij}^{(l)}} J(W, b)$$
$$b_i^{(l)} := b_i^{(l)} - \alpha \frac{\partial}{\partial b_i^{(l)}} J(W, b) \quad (4.5)$$

where α represents the learning rate.

The partial derivatives of the cost function are computed using the backpropagation algorithm (see Algorithm 1). Algorithm 1 describes how backpropagation computes the partial derivatives $\frac{\partial}{\partial W_{ij}^{(l)}} J(W, b; x, y)$ and $\frac{\partial}{\partial b_i^{(l)}} J(W, b; x, y)$ for the cost function $J(W, b; x, y)$ for a single example (x, y).

The backpropagation algorithm first performs a feedforward pass to compute all the activations $a_i^{(l)}$ and the output value of $h_{W,b}(x)$ in the network. An error term $\delta_i^{(l)}$ is calculated for each node i in layer l to measure the contribution of this node to any errors in the output. For hidden nodes, the error term $\delta_i^{(l)}$ is computed using a weighted average $z_i^{(l)}$ of the error terms of the nodes that use $a_i^{(l)}$ as an input. For an output node, the error term $\delta_i^{(n_l)}$ (where n_l is the output layer) signifies the difference between the network's activation and the true target value. Then, the error term $\delta_i^{(l)}$ is propagated backwards to the previous layers through the network to adjust the weights for each node i in layer l.

Algorithm 1 Backpropagation Algorithm

1: Perform a feedforward pass and compute the activations for $L_2, L_3, \ldots, L_{n_l}$ (n_l is the output layer)
2: **for** each output unit i in layer n_l, **do**
3: $\quad \delta_i^{(n_l)} = -\left(\nabla_{a_i^{(n_l)}} J\right) . f'(z_i^{(n_l)})$
4: **end for**
5: **for** $l = n_l - 1, \ldots, 2$, **do**
6: $\quad$ **for** each node i in layer l, **do**
7: $\quad\quad \delta_i^{(l)} = \left(\sum_{j=1}^{s_{l+1}} W_{ji}^{(l)} \delta_j^{(l+1)}\right) f'(z_i^{(l)})$

8: **end for**
9: **end for**
10: Compute the desired partial derivatives:
11: $\frac{\partial}{\partial W_{ij}^{(l)}} J(W, b; x, y) = a_j^{(l)} \delta_i^{(l+1)}$
12: $\frac{\partial}{\partial b_i^{(l)}} J(W, b; x, y) = \delta_i^{(l+1)}$

Once the partial derivatives of the cost function with respect to a single example (x, y) have been computed, the derivative for the overall cost function $J(W, b)$ can be calculated as:

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W, b) = \left[\frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial W_{ij}^{(l)}} J(W, b; x^{(i)}, y^{(i)}) \right] + \lambda W_{ij}^{(l)}$$
$$\frac{\partial}{\partial b_i^{(l)}} J(W, b) = \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial b_i^{(l)}} J(W, b; x^{(i)}, y^{(i)}) \quad (4.6)$$

Thereafter, gradient descent is used to train the neural network as described in Algorithm 2. $\Delta W^{(l)}$ is a matrix with dimensions similar to $W^{(l)}$, and $\Delta b^{(l)}$ is a vector of similar dimension to $b^{(l)}$. Algorithm 2 describes one iteration of gradient descent as follows:

Algorithm 2 Gradient Descent Algorithm

1: Set $\Delta W^{(l)} := 0, \Delta b^{(l)} := 0$ (matrix/vector of zeros) for all l.
2: **for** $i = 1, \ldots, m$, **do**
3: Use backpropagation to compute $\nabla_{W^{(l)}} J(W, b; x, y)$ and $\nabla_{b^{(l)}} J(W, b; x, y)$.
4: Set $\Delta W^{(l)} := \Delta W^{(l)} + \nabla_{W^{(l)}} J(W, b; x, y)$.
5: Set $\Delta b^{(l)} := \Delta b^{(l)} + \nabla_{b^{(l)}} J(W, b; x, y)$.
6: **end for**
7: Update the parameters:
8: $W^{(l)} := W^{(l)} - \alpha \left[\left(\frac{1}{m} \Delta W^{(l)} \right) + \lambda W^{(l)} \right]$
9: $b^{(l)} := b^{(l)} - \alpha \left[\frac{1}{m} \Delta b^{(l)} \right]$

The steps taken in the gradient descent optimization algorithm can be repeatedly applied to minimize the overall cost function $J(W, b)$.

Momentum training and learning rate annealing are advanced optimization tuning parameters that are used to modify backpropagation to allow previous iterations to influence the current version. The velocity vector is defined as follows:

$$v_{t+1} = \mu v_t - \alpha \nabla L(\theta_t)$$
$$\theta_{t+1} = \theta_t + v_{t+1} \quad (4.7)$$

where θ denotes the parameters $W\ and\ b$. The momentum coefficient is represented by μ and the learning rate is α.

Training the Baseline MLP Classifier

The MLP is trained using a training set of labelled observations $(x^{(i)}, y^{(i)})$ where $y^{(i)} \in \mathbb{R}^2$, extracted from the T2D case-control GWAS data and used for supervised learning. The parameter x is a vector of input features obtained from the training samples which are extracted as described in section 4.5.1 (Extracting Groups of Features from Association Analysis). Six feature input vectors consisting of 7, 13, 23, 103, 766, and 6609 SNPs respectively were used to train six separate MLP models. The output y was used for target outcomes (a sample with T2D and a sample without T2D respectively) among observations in the study. The network parameters W and b are randomly initialised close to zero before training is performed. The cost function is set to cross-entropy for binary inputs as defined in equation (4.4).

Hyperparameters Configuration of MLP

All MLP models are trained with several different layer and neuron configurations. In addition, parameters including L1 (Lasso) and L2 (Ridge) regularization penalties, learning rate, rate_annealing, momentum_start, momentum_stable, input_dropout_ratio, and early stopping criteria are configured for model optimisation.

Finding optimal hyperparameters is challenging, yet fundamentally important for model accuracy. Therefore, Random Grid Search (RGS) (Bergstra & Yoshua 2012) is widely used to overcome this issue. The random grid search allows us to test various hyperparameter

combinations and choose configurations that maximise model accuracy (Bergstra & Yoshua 2012). For RGS, a set of hyperparameters and search criteria must be specified. Each hyperparameter is defined with a range of possible values. Search criteria with stopping_metric, stopping_tolerance, and stopping_rounds are specified for early stopping to prevent model overfitting. Based on grid search results, the best model is selected. In some cases, the model with the lowest mean square error or lowest Logloss is considered the best option. While in another case the highest AUC is considered.

In this analysis, RGS with a range of hyperparameter values is implemented to evaluate model accuracy. Figure 4.9 and Figure 4.10 present the R code snipped used to build random and automated search for different network configurations. The activation function coefficient is given several options including Rectifier, Tanh, Maxout, RectifierWithDropout, TanhWithDropout, and MaxoutWithDropout. Two hidden layer configurations are considered - three and four hidden layers. The number of neurons in each layer is set to ten, fifty, and a hundred. The remaining hyperparameters L1 and L2 regularization, dropout, learning rate, and momentum training are given several possible values. To avoid overfitting, early stopping is used to decide when the MLP is optimized and sufficiently accurate. There are several parameters to control early stopping including stopping metric, stopping rounds and stopping tolerance which are set to Logloss, 5, and 1e-2 respectively. The network is finally trained using 100 epochs.

```
# Hyper-parameters for Random Grid Search
hyper_params <- list(
  activation=c("Rectifier","Tanh","Maxout","RectifierWithDropout","TanhWithDropout","MaxoutWithDropout"),
  hidden=list(c(10,10,10),c(50,50,50),c(100,100,100),c(10,10,10,10),c(50,50,50,50),c(100,100,100,100)),
  input_dropout_ratio=c(0,0.1,0.2),
  l1=seq(0,1e-4,1e-6),
  l2=seq(0,1e-4,1e-6),
  rate = c(0, 0.005, 0.001, 0.0001,0.00001),
  rate_annealing = c(1e-8, 1e-7, 1e-6),
  momentum_start = c(0, 0.5),
  momentum_stable = c(0.99, 0.5, 0)
)
hyper_params
```

Figure 4.9: R Code – Hyperparameters Used with RGS in MLP

```
search_criteria = list(strategy = "RandomDiscrete",
                       max_runtime_secs = 900,
                       max_models = 100,
                       stopping_rounds=5,
                       stopping_tolerance=1e-2,
                       stopping_metric="logloss",
                       seed=42
                       )
```

Figure 4.10: R Code – Search Criteria Used with RGS in MLP

4.5.3 Classification using Random Forest Classifier

A Random Forest classifier is implemented for the binary classification of case-control T2D GWAS data. Random Forest classifiers have been used extensively in genetic studies (Botta et al. 2014; López et al. 2018; Schwarz et al. 2010; Kursa 2014) given their ability to deal with high-dimensional data structures, such as GWAS (Qi 2012). Furthermore, it is a useful algorithm for uncovering correlations and interactions within and across a large number of features. In this study we train an RF algorithm using the same data splitting strategy (training, validation, and testing) with the same feature subsets (7, 13, 23, 103, 766, and 6609 SNPs respectively).

The RF algorithms use a randomized decision tree-based ensemble with the number of trees configured to 200 and the depth of each tree set to 20. Increasing the number of trees and their depth will adjust the weakness of each learner. To avoid overfitting, early stopping is used to decide when the RF is optimized and sufficiently accurate. The following parameters stopping metric, stopping rounds and stopping tolerance are set to Logloss, 4, and 1e-2 respectively.

4.5.4 Classification using Deep Learning Stacked Autoencoders

Previous sections have discussed supervised learning for binary classification tasks in T2D GWAS data. However, high-dimensional data can be reduced by removing redundant information and converting the dataset to a lower dimension data using deep learning stacked autoencoders. SAEs use unsupervised feature learning or, more specifically, non-linear dimensionality reduction (Hinton & Salakhutdinov 2006). Unsupervised feature learning with SAE allows information about important features to be captured relating to non-linear

latent representations of these features. Hence, SAE provides a way to learn deep features from original SNP data by capturing information about important SNPs and the cumulative non-linear epistatic interactions between them.

An autoencoder can be used to pre-train the neural network, where the target output values $\hat{x}$ are approximately equal to the input values x using backpropagation. This is achieved by minimizing the discrepancy between the input vector and its reconstruction (output vector). The AE attempts to learn a function $h_{W,b}(x) \approx x$, which means it is trying to learn an approximation to the identity function and consequently to output $\hat{x}$ equal to x. The aim is to discover interesting structure in the data specifically correlations between input features. This is achieved by placing constraints on the network, such as limiting the number of hidden units in the hidden layer. In this case, the network will be enforced to learn a compressed representation of the input, given the vector of hidden unit activations $a^{(2)} \in \mathbb{R}^n$, where n is the number of hidden units, and then try to reconstruct the input x.

An alternative way to limit the number of hidden units is to use the sparsity of hidden units in the network. Sparsity is a useful constraint method particularly with large numbers of hidden units. In sparse autoencoders (Ng 2011), most neurons are inactive. A neuron is considered active if its output value is close to 1, while if it is close to 0 the neuron is considered inactive. Thus, with sparse autoencoders the data is constrained, allowing the network to discover interesting structure in the data, important features, which is then used during reconstruction. To establish the sparsity constraint on the autoencoder, the activation of a hidden unit in the network is denoted as $a_j^{(2)}(x)$ for a given input x. Let

$$\hat{p}_j = \frac{1}{m}\sum_{i=1}^{m}\left[a_j^{(2)}\left(x^{(i)}\right)\right] \tag{4.8}$$

where $\hat{p}_j$ signifies the average activation of hidden unit j, averaged over m training examples. Furthermore, the constraint is imposed such that $\hat{p}_j = p$, where p is a sparsity

parameter representing a small value near to 0. To meet this constraint, the activation of the hidden unit must almost be close to 0 (for example, $p = 0.05$).

To achieve this requirement of $\hat{p}_j = p$, an extra penalty term based on Kullback-Leibler (KL) divergence is added to the cost function $J(W, b)$ that penalizes $\hat{p}_j$ when deviating significantly from p:

$$\sum_{j=1}^{s_2} p \, log \frac{p}{\hat{p}_j} + (1 - p) log \frac{1 - p}{1 - \hat{p}_j} \tag{4.9}$$

where s_2 represents the number of units in the hidden layer, and j is an index for summing the hidden units in the network. KL divergence which is a standard function used to measure how different two different distributions are, is used to impose the penalty term, refer to (4.9), as follows:

$$\sum_{j=1}^{s_2} KL\left(p||\hat{p}_j\right) \tag{4.10}$$

where

$$KL\left(p||\hat{p}_j\right) = p \, log \frac{p}{\hat{p}_j} + (1 - p) log \frac{1 - p}{1 - \hat{p}_j} \tag{4.11}$$

Equation (4.11) represents the Kullback-Leibler divergence between two Bernoulli random variables with mean p and $\hat{p}_j$. This can be equal to 0 if $\hat{p}_j = p$, alternatively it increases monotonically as $\hat{p}_j$ diverges from p. Hence, after adding a sparse penalty term, the overall cost function can now be defined as:

$$J_{sparse}(W, b) = J(W, b) + \beta \sum_{j=1}^{s_2} KL\left(p||\hat{p}_j\right) \tag{4.12}$$

where $J(W, b)$ is the mean squared error cost function defined in equation (4.3), and β is used to control the sparsity penalty term's weight. Typically, the activations of a hidden unit are dependent on W, b parameters, therefore the term $\hat{p}_j$ which is the average activation of hidden unit j depends on W, b.

In order to integrate the KL-divergence term into the derivative calculation during the backpropagation algorithm (see Algorithm 1) the error term can now be computed as:

$$\delta_i^{(l)} = \left(\left(\sum_{j=1}^{s_{l+1}} W_{ji}^{(l)} \delta_j^{(l+1)} \right) + \beta \left(-\frac{p}{\hat{p}_i} + \frac{1-p}{1-\hat{p}_i} \right) \right) f'(z_i^{(l)}) \qquad (4.13)$$

This is a single layer autoencoder procedure. However, to form SAE several AEs are stacked. The concept of SAE is that the outputs of each hidden layer in AE are connected to the inputs of the subsequent AE layer, repeating this process for the next AE layers, and for the classification task the last hidden layer is linked to a softmax classifier. After greedy layer-wise unsupervised pre-training, the resulting deep features can be used as input to a fully connected supervised neural network.

Pre-Training the Stacked Autoencoders

In this analysis, to train SAEs, unlabelled training samples are used. These were extracted from the T2D case-control GWAS dataset. A subset of input features consisting of 6609 SNPs was generated using a p-value threshold of 10^{-2}. The SAE is based on the pre-training of weights for fully connected networks in a greedy layer-wise fashion, instead of using random weight values. For each single autoencoder the cost function is set to mean squared error, refers to equation (4.3), and the activation function coefficient is set to the hyperbolic tangent function (tanh) refers to equation (3.3). Epochs set to 1 and L1 regularization penalty is configured to 1e-5. The configuration of the SAE consists of four single autoencoders each containing a single hidden layer with 2500, 1500, 700, and 300 hidden neurons respectively. The configuration is described as follows:

- The first SAE contains one autoencoder with 2500 hidden neurons to form a 6609-2500 neural network.
- The second SAE consists of two autoencoders with 2500 and 1500 hidden neurons and these are connected to form a 6609-2500-1500 neural network.

- The third SAE contains three autoencoders with 2500, 1500, and 700 hidden neurons and these are connected to form a 6609-2500-1500-700 neural network.
- Finally, the fourth SAE includes four autoencoders with 2500, 1500, 700, and 300 hidden neurons and these are connected to form a 6609-2500-1500-700-300 neural network.

This SAE is used to learn the deep features contained within 6609 SNPs by gradually reducing the dimensionality of features to 300 units while only retaining the salient information from each layer. The input vector for the first autoencoder consists of 6609 features and the number of hidden units is 2500. This network is trained to compress the 6609 features using a single hidden layer with 2500 units to learn the salient features in the data and remove redundant information therefore reducing dimensionality. The trained hidden layer containing the 2500 neurons is then used as an input vector to the second autoencoder – again the data is compressed into 1500 hidden units and redundant information is once again removed. This process continues until a layer containing 300 neurons is reached, the last hidden layer in this SAE.

Training the MLP Softmax Classifier

Following the training process of each layer of the network on unlabelled data for feature reduction, the weights parameters are now initialised at a better location in parameter space than if they were randomly initialised. The final hidden layer in SAE (300 deep features) is used to feed into an MLP softmax classifier for supervised binary classification of T2D GWAS data.

In this analysis, to train an MLP softmax classifier, labelled training samples are utilised. The cost function associated with binary classification is set to cross-entropy as defined in equation (4.4). To optimize the predictive capacity of the MLP classifier and to prevent overfitting and improve generalization, a number of hyperparameters to control the configuration topology of the network were experimentally detected and tuned. The network architecture including the number of hidden layers and units were specified in addition to

the regularization parameters L1, L2, and dropout, were experimentally tuned. The activation function coefficient is given several options to determine which of these to adopt for our data. The adaptive learning rate was disabled to allow for momentum training and learning rate annealing to be experimentally detected. Moreover, to find the optimal combination of hyperparameters RGS was employed. Figure 4.9 and Figure 4.10 shown above lists the tuning parameters used with the MLP classifier along with their range of configuration values. Early stopping was also adopted to avoid overfitting. The optimal number of epochs was experimentally set to 100.

Fine-Tuning the Entire Network

SAEs train the parameters of each layer independently. To improve the performance results of SAEs fine-tuning strategy on labelled training samples (supervised fashion) can be used. To perform fine-tuning for the whole network, all layers of SAE are treated as a single model. Hence, in one iteration of fine-tuning all the weights of SAEs can be improved. Fine-tuning can be implemented using backpropagation algorithm with the aim of minimizing the error between the actual output and the expected output of the MLP softmax model. Cross-entropy cost function and gradient descent optimization defined in equations (4.4) and (4.5) are utilized respectively to update the parameters (W, and b) for classification tasks of case-control T2D GWAS data. Figure 4.11 illustrates the configuration used in the pre-training of the final stacked autoencoder network and the subsequent fine-tuning processes.

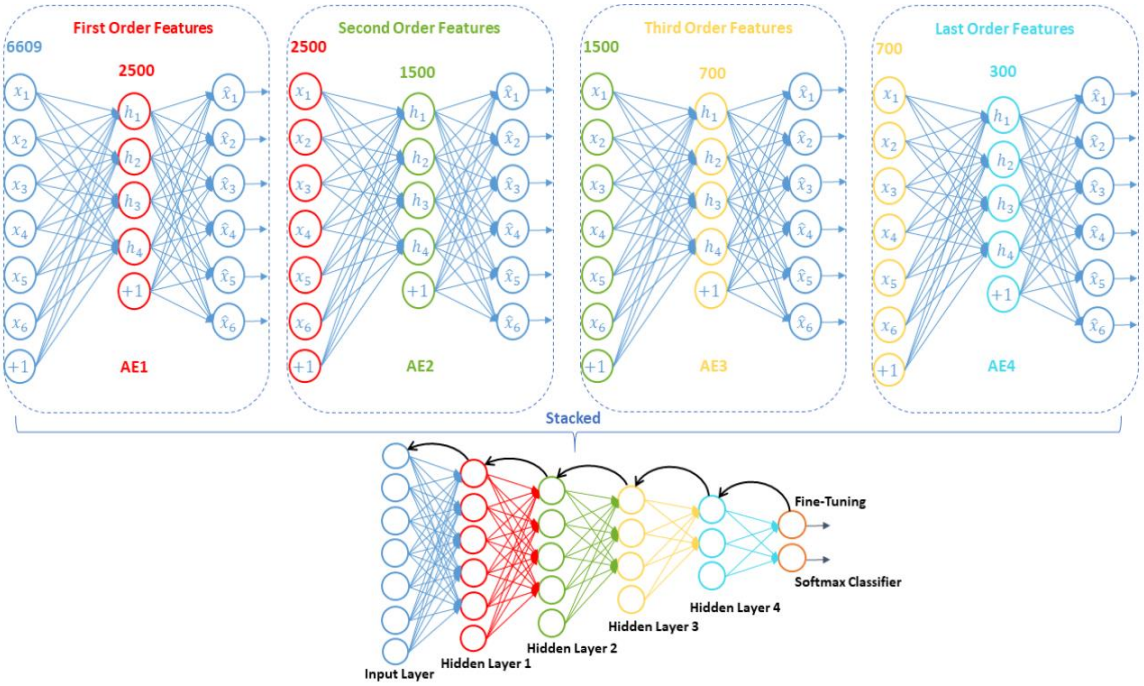


Figure 4.11: Configuration of Stacked Autoencoder for Feature Extraction and the Process of Fine-Tuning

4.6 Classification for Genetic and Clinical Data

Classification has been used to evaluate the performance of various machine learning methods and their ability to discriminate between T2D case-control observations in the GENEVA NHS-HPFS GWAS dataset. In this analysis, genetic and clinical factors are considered to model T2D using five traditional supervised machine learning algorithms. These machine learning algorithms are used to model non-linear and linear effects. The non-linear models include Stochastic Gradient Boosting (GBM), Support Vector Machines with Radial Basis Function Kernel (SVMs), Recursive Partitioning and Regression Trees (RPART), and Neural Networks (NNET). The linear model includes a Generalized Linear Model (GLM). For the evaluation of the predictive abilities of the classification models we used the caret package (Kuhn 2008) in R Studio. Figure 4.12 demonstrates the workflow for the classification of genetic and clinical data using traditional machine learning.

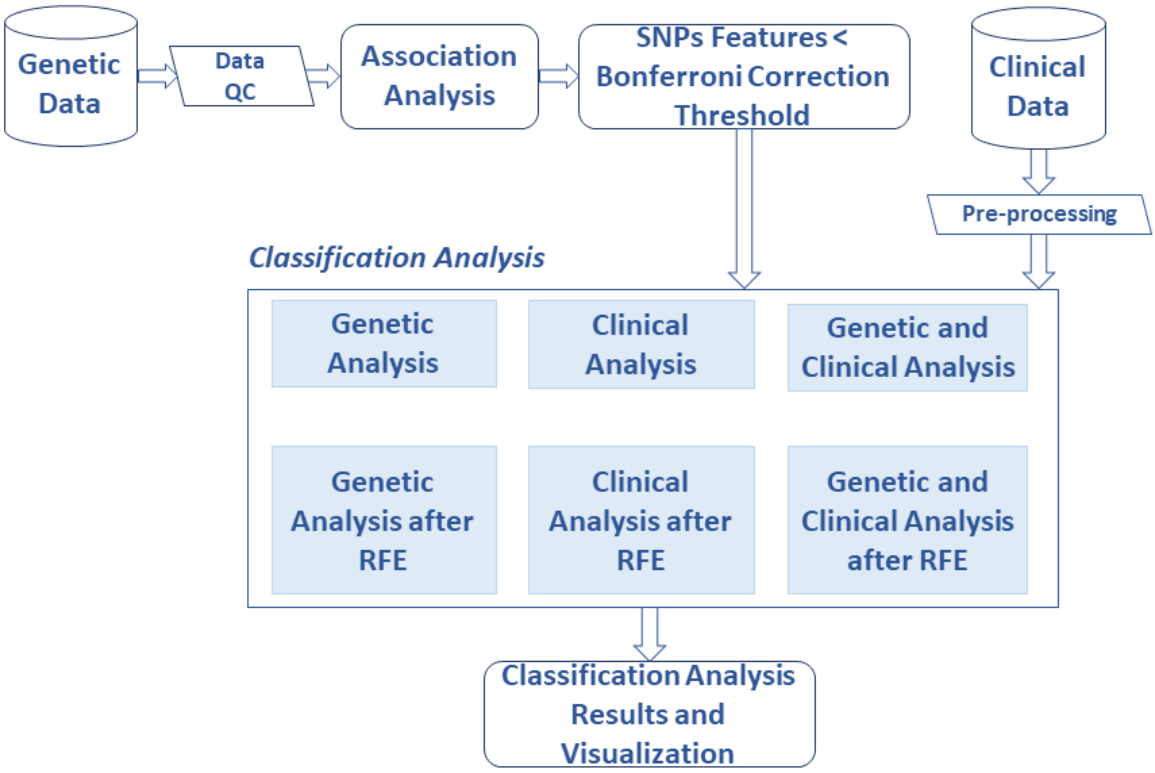


Figure 4.12: The Workflow for the Classification of Genetic and Clinical Data

4.6.1 Genetic Data Analysis

In the genomic data investigations, following association analysis, six common SNPs from logistic association analysis reached the Bonferroni correction genome-wide significance threshold (rs4132670, rs12243326, rs12255372, rs7901695, rs4506565, and rs2371765). These are considered as a set of features for the binary classification of T2D.

4.6.2 Clinical Data Analysis

In clinical analysis nine features were considered strong candidates in this investigation. These include Body Mass Index (BMI), alcohol intake (Alcohol), smoking status (SMK), physical activity (ACT), family history of diabetes (Famdb), high blood pressure (Hbp), high blood cholesterol (Chol), Age and Sex. These are consistent with those found in (Lyssenko et al. 2008; Wilson 2007), who suggest that BMI, SMK, Famdb, Hbp, Age, and Sex are stronge predictors of T2D. BMI, Age, ACT, and Alcohol are continuous variables in our study, and have been converted to a binary representation using the median of the variables (coded as 0, 1). The remaining variables are binary.

4.6.3 Genetic and Clinical Data Analysis

In the joint effect of genetic and clinical data analysis, a combination of six genetic variables and the nine clinical variables are used as input features in the third analysis. The genetic and clinical features are similar to those identified and mentioned previously. The classification performance of the three experiments are evaluated and compared. In addition, variable importance for each specific model is also considered to measure the scale of importance and model performance. This approach is useful to determine which of the predictors (genetic and clinical) contribute more to better model performance.

4.6.4 Feature Selection for Genetic and Clinical Data Analysis

In this analysis, feature selection is conducted using the Recursive Feature Elimination algorithm (RFE) (Dong et al. 2015). The subset of top ranked features selected by RFE is presented as input to conduct three distinct evaluations using genomic data only, clinical data only, and lastly genetic and clinical data combined. These evaluations determine whether a reduced subset of features can improve on or maintain the previous performance results obtained using the original dataset.

4.7 Performance Evaluation Measurement

There are various performance metrics that have been used for evaluating classifier performance. Each metric offers a different perspective on classifier model performance. Consequently, there is no agreed definition on which approach is best. One classifier can produce better results on one performance metric but not on others. Therefore, using several performance metrics to evaluate classifier performance is recommended (Seliya et al. 2009).

In this study, the performance of each classifier is measured using sensitivity (true positives), also called the recall rate, specificity (true negatives), the Area Under the Curve (AUC), Gini, Logarithmic Loss, and Mean Square Error (MSE). Sensitivity and specificity are used to represent the number of correctly identified case and control observations (Trevethan 2017). Sensitivity refers to the proportion of observations who have the disease and give positive

test outcomes (true positive rate). It describes the ability of the test to correctly classify people with T2D. Whereas specificity refers to the proportion of observations without the disease and give negative test outcomes (true negative rate). It describes the ability of a model to correctly classify people without T2D. Generally, sensitivity and specificity metrics of the test are inversely related. This means that both metrics often trade-off with each other. Choosing the optimal trade-off between sensitivity and specificity depends on the purpose of the test. Another approach uses the Youden's index (Youden 1950) to find a balance between sensitivity and specificity. A third possible approach uses the ROC curve to find the closest trade-off value to the (0, 1) top-left corner (Perkins & Schisterman 2006), which defines the optimal cut-off point based on the lower distance to the (0, 1) corner. In the context of this thesis, the true positive rate (sensitivity) is considered a higher priority than the true negative rate (specificity). It is more important not to miss a potential case of T2D rather than misclassify a healthy individual as having T2D, as for the latter, further tests would clarify whether there is any concern or not.

Alternatively, other performance indicators such as positive predictive value (PPV) and negative predictive value (NPV) can be considered to provide healthcare system (i.e. physicians) with information that can be more relevant to patients, if the test result is positive or negative (Trevethan 2017; Florkowski 2008). PPV is the probability that individuals with positive results certainly do have the disease. Whereas NPV is the probability that individuals with negative results indeed do not have the disease. The predictive values (PPV and NPV) depend on the prevalence of the disease in the population under investigation. This considers as one of their limitations as they are influenced by how common the disease is in the population being studied. For example, if a disease is uncommon in a tested population, a large amount of positive test results will be false positives thus the PPV will be low. Whereas if the disease is common in a tested population, the PPV will be high.

Each classifier produces a class prediction which is either mapped to case or control instances. As such, given a classifier and an instance there are four possible prediction outcomes; true positive (TP), true negative (TN), false positive (FP), and false negative (FN) (Okeh & Okoro 2012). These four outcomes are employed to construct a two-by-two confusion matrix that summaries the prediction results of a classifier when a test set is used. Figure 4.13 shows a confusion matrix of classification results with its predictive accuracy terms. In the context of our study, if an observation is T2D (case) and it is classified as T2D, it is counted as a true positive; while if it is classified as non-T2D (control) it is counted as a false negative. The other possible instance is non-T2D - if it is classified as non-T2D, it is counted as a true negative; if it is classified as T2D it is counted as a false positive.

		Actual Class	
		Case	Control
Predicted Class	Case	TP	FN
	Control	FP	TN

Figure 4.13: Confusion Matrix Table

The confusion matrix is used to calculate the performance metrics of binary classification models. Performance metrics including sensitivity, specificity, accuracy, PPV, and NPV are calculated as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.14}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.15}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + TN} \tag{4.16}$$

$$PPV = \frac{TP}{TP + FP} \tag{4.17}$$

$$NPV = \frac{TN}{TN + FN} \tag{4.18}$$

In order to estimate the accuracy of machine learning models, *k*-fold cross-validation (Jung 2018) is used. *K*-fold cross-validation also called rotation estimation is a statistical method used to compare and select a model for a given classification model to estimate the predictive capability of the model on unseen data (data not used during the training or validation phases) (Kohavi 1995). *K*-fold cross-validation involves randomly partitioning the dataset into *k* folds or groups (Refaeilzadeh et al. 2009). These groups are approximately equal size. *K* iterations of training and validation are performed. Subsequently, in the first iteration the first fold is considered as a validation set while the remaining *k*-1 folds are used to fit the model. In the second iteration the second fold is held-out for validation and the remaining *k*-1 folds for training. This procedure is repeated until all the folds are trained and tested. The average error obtained from cross-validation is an estimate of the error obtained for the classifier. The common way to obtain a large number of estimates is to run *k*-fold cross-validation multiple times and this is known as repeated *k*-fold cross validation. This method is performed in our analysis considering 5 folds with 30 repetitions.

The area under the curve (AUC) and the receiver operating characteristic curve (ROC curve) are used in this study to assess and compare classifier performance. Both of these measures are widely used to evaluate binary classifiers (Hand 2009) particularly, in medical decision making (Kumar & Indrayan 2011). The AUC value describes the probability of a correct classification using both positive and negative instances; values range between 0 and 1. A classifer that produces a large area under the curve is preferable. This is because a higher AUC means better classification (Hand 2009). The ROC curve is a graphical plot to visualize and organize the performance of a binary classification model (Hand 2009). It is created by calculating the trade-off between the true positive rate (also known as sensitivity) against the false positive rate which can be represented as (1-specificity) (Hand 2009). ROC graphs are

a preferred approach rather than simple classification accuracy as the latter is generally considered a poor metric for measuring performance (Seliya et al. 2009).

The Gini coefficient value can be derived from the AUC. It represents the area between the ROC curve and the diagonal. The Gini coefficient is usually used in binary classification problems. A Gini value above 60% is considered a good model. The Gini is defined as:

$$Gini = 2 * AUC - 1 \tag{4.19}$$

Logarithmic Loss (Logloss) is a classification loss function often used to measure the performance of a classification model where the predicted input is a probability value between 0 and 1. Logloss increases as the predicted probability (accuracy) decreases (Vovk 2015; Ferri et al. 2009). A Logloss value of 0 indicates a perfect model where the model correctly classifies all class instances. The Logloss is defined as:

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log p_{ij} \tag{4.20}$$

where N denotes the number of observations, and M the number of possible outcomes (actual labels). While, y_{ij} represents a binary indicator to specify if the actual label j is the correct classification for observation i, and p_{ij} is the model probability of assigning label j to observation i. In the case of binary classification where only two classes are specified the mathematical expression of Logloss can be simplified to:

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1 - y_i)\log(1 - p_i)] \tag{4.21}$$

The Mean Squared Error (MSE) performance metric is used to measure the average of the square of the error between the actual values and the predicted values. MSE values closer to 0 mean that the model correctly classifies all class instances (Ferri et al. 2009). The mathematical definition of MSE is expressed as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (4.22)$$

MSE is a common measurement metric for evaluating the predictive ability of neural networks (Ferri et al. 2009).

4.8 Summary

This chapter presented details about the methodology utilized in this thesis. The framework architecture to fulfil the outline of the proposed methodology including the data description and data quality control procedure, along with genetic association analysis, feature engineering, classification and the performance measurements used were described.

Chapter 5 Results

5.1 Introduction

This chapter presents the results of the experiments performed in this thesis. Following the quality control procedures conducted to produce subsets of reliable samples and genetic markers for logistic association analysis, the results and their visualizations are presented. These results highlight statistically significant SNPs that suggest potential disease-associations in T2D. In addition, these findings are biologically discussed and linked to previous studies in the literature.

Furthermore, the results for the binary classification of T2D genetic data using deep learning stacked autoencoders are evaluated and benchmarked against a multilayer perceptron neural networks and random forest classifier.

Five algorithms including GBM, SVM, RPART, NNET, and GLM are employed to conduct three distinct evaluations using genomic data only, clinical data only, and lastly the joint effect of genetic and clinical data. The results generated by the three evaluations models are compared.

5.2 Logistic Association Analysis Results

A case-control study design is used for association analysis tests to find statistically significant SNPs associated with T2D. Logistic regression association analysis under an additive genetic model adjusted for GC inflation factor shows that six genotyped SNPs passed the Bonferroni-corrected genome-wide significance threshold. In addition, 22 (16+6 GWS) SNPs were found to be above the suggestive association threshold. The Manhattan plot in Figure 5.1 illustrates the logistic association analysis results. Each point in the plot represents an individual SNP and the chromosome number along the x-axis and the negative log of the corresponding p-value on the y-axis. The plot shows the SNPs that reached the Bonferroni level of significance, and those that reached the suggestive threshold.

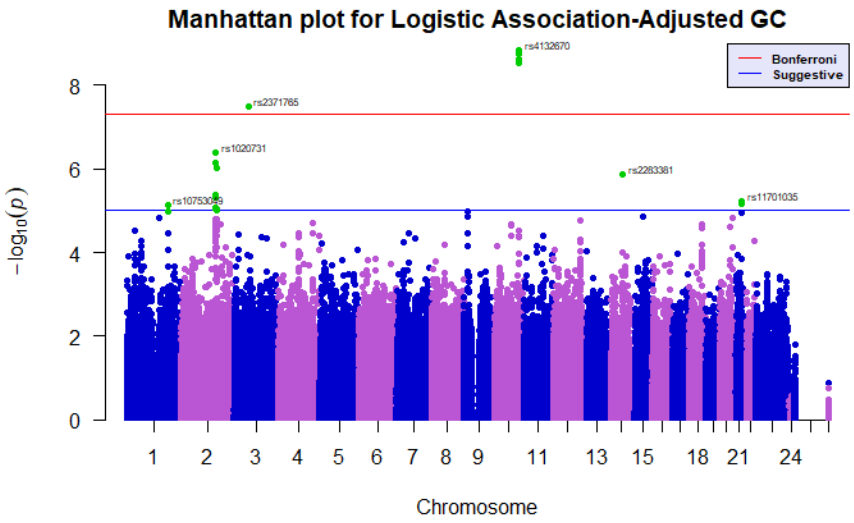


Figure 5.1: Manhattan Plot for Logistic Regression Analysis Adjusted GC

Logistic analysis uncovered the SNPs shown in Table 5.1. The OR for these SNPs is > 1 indicating that these SNPs are more likely to appear in cases and thus signifying an association with T2D.

Table 5.1: SNPs from Logistic Regression Test of Association

Chr	Nearest Gene	SNP	P-Value	OR	Association Type
10	TCF7L2	rs4132670	1.412×10^{-9}	1.288	Risk Association
10	TCF7L2	rs12243326	1.767×10^{-9}	1.295	Risk Association
10	TCF7L2	rs12255372	2.400×10^{-9}	1.290	Risk Association
10	TCF7L2	rs7901695	2.561×10^{-9}	1.283	Risk Association
10	TCF7L2	rs4506565	2.991×10^{-9}	1.281	Risk Association
3	ADAMTS9	rs2371765	3.146×10^{-8}	1.240	Risk Association

The analysis shows that there are 5 significant associative SNPs located in chromosome 10 as listed in Table 5.1. These SNPs have an association with T2D and a OR > 1. These SNPs can be found in the Transcription Factor-7-Like-2 (*TCF7L2*) Gene region. The *TCF7L2* gene is known as the susceptibility gene with the largest effect on T2D predisposition (Gloyn et al. 2009). In line with (Chuan-zhen et al. 2008; Cauchi et al. 2006; Scott et al. 2006; Sale et al. 2007; Ibrahim et al. 2016) studies that suggested that (*TCF7L2*) genetic variants have been associated with T2D in several ethnic groups, our findings confirm this and thus could serve as a starting point for future investigations. In addition, one of the reported genes found

in chromosome 3 has been previously associated with T2D related traits. Particularly, ADAM metallopeptidase with thrombospondin type 1 motif 9 (*ADAMTS9*) gene which has previously been reported by Voight et.al (Voight et al. 2010).

Of the list of SNPs obtained from logistic association tests, we found 10 SNPs (rs1020731, rs10181181, rs6718526, rs2925757, rs7572970, rs7593730, rs4589705, rs4077463, rs11693602, and rs9287795) above the suggestive threshold. These were located in chromosome 2 as listed in Table 5.2. Chromosome 2 is one of the largest chromosomes in the human genome and gene abnormalities have been linked to several important diseases particularly T2D (Hanis et al. 1996). These SNPs are found in the RNA binding motif single stranded interacting protein 1 (*RBMS1*) gene and show protective association to T2D with OR < 1. Even though the odds ratio for these SNPs indicate that the condition is more likely to be in the control group, the large-scale meta-analysis study conducted by Qi *et al.* (Qi et al. 2010) proved that the *RBMS1* gene was a T2D associated variant represented by SNP rs7593730 at 2q24 locus. Variants in this region were usually linked to lower fasting glucose suggesting that the 2q24 locus may influence T2D risk by affecting glucose metabolism and insulin resistance. Thus, these SNPs identified by our study should be considered for further investigation. Figure 5.2 shows the Manhattan plot for chromosome 2 and highlights the SNPs that reached the suggestive line.

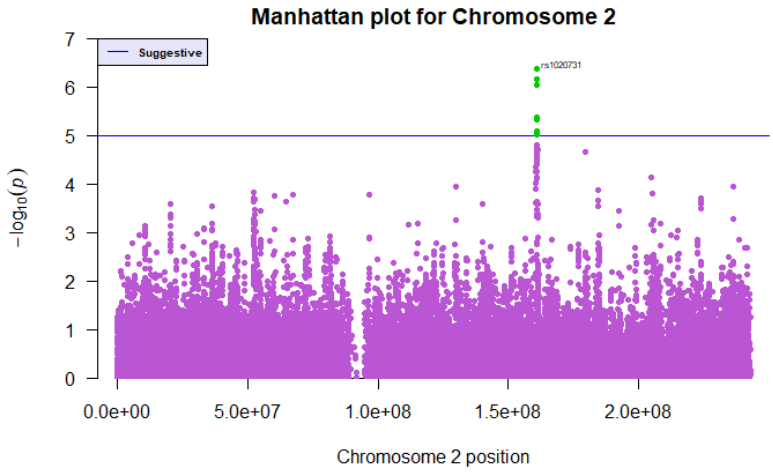


Figure 5.2: Manhattan Plot for Chromosome 2

Two of the genes reported in our analysis were found to be associated with different disease traits in other studies. Particularly, the regulator of G-protein signaling 6 (RGS6) gene that is associated with Parkinson's disease (Ahlers et al. 2016), Cancer (Ahlers et al. 2016), and Heart Rate (HR) variability (Verweij et al. 2018). Additionally, the holocarboxylase synthetase (HLCS) gene has been linked with the pathogenesis of Chronic Rhinosinusitis with Nasal Polyps (Bohman et al. 2017).

Table 5.2: SNPs above Suggestive Threshold

Chr	Nearest Gene	SNP	P-Value	OR	Association Type
2	RBMS1	rs1020731	4.135×10^{-7}	0.8022	Protective
2	-	rs10181181	7.014×10^{-7}	0.8061	Protective
2	RBMS1	rs6718526	9.192×10^{-7}	0.7836	Protective
14	RGS6	rs2283381	1.362×10^{-6}	0.8071	Protective
2	-	rs2925757	4.251×10^{-6}	0.7923	Protective
2	RBMS1	rs7572970	4.652×10^{-6}	0.8179	Protective
21	HLCS	rs11701035	6.026×10^{-6}	1.2560	Risk Association
21	HLCS	rs2835530	6.899×10^{-6}	1.2550	Risk Association
1	-	rs10753049	7.403×10^{-6}	1.2670	Risk Association
1	-	rs6425178	7.508×10^{-6}	1.2670	Risk Association
1	-	rs6667131	7.567×10^{-6}	1.2670	Risk Association
2	RBMS1	rs7593730	8.343×10^{-6}	0.8095	Protective
2	RBMS1	rs4589705	8.813×10^{-6}	0.8100	Protective
2	RBMS1	rs4077463	9.101×10^{-6}	0.8102	Protective
2	RBMS1	rs11693602	9.118×10^{-6}	0.8102	Protective
2	RBMS1	rs9287795	9.631×10^{-6}	0.8105	Protective

In genetic association analysis Q-Q plots are used to check for possible systematic bias in the study dataset. In Figure 5.3 the Q-Q plot shows the p-values for SNPs against expected p-values. The figure shows the relationship between the expected distribution of p-values (null) and the observed distribution of p-values for association tests. The logistic test shows that systematic bias is not possible due to population stratification as there is no early

deviation from the diagonal. Logistic regression under the additive genetic model, adjusted for GC, shows that there is moderate deviation in the upper right tail from the y=x line and also that there is a severe deviation at the top. This suggests the existence of some form of association between SNPs and the disease of interest. The red line represents the null hypothesis of no association and the blue dot refers to the observed $-log_{10}(p)$.

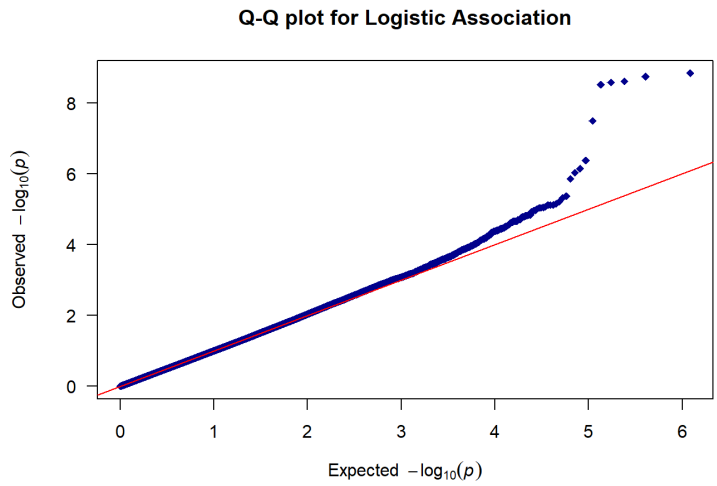


Figure 5.3: Q-Q Plot for Logistic Test Adjusted GC

5.3 Results for the Classification of High-Dimensional Genetic Data

Following the identification of feature subsets obtained from the original GWAS data, deep learning stacked autoencoders for unsupervised feature learning is adopted to find latent representations in SNPs. Learned features are utilised to fine-tune an MLP for T2D binary classification of case-control observations. The results obtained are benchmarked against two supervised machine learning algorithms (multilayer perceptron neural network and random forest classifiers), using the original SNP sequences obtained from GWAS analysis.

In this analysis, the performance of the MLP, deep learning SAE and RF are measured using the Area Under the Curve (AUC), Sensitivity, Specificity, Gini, Logarithmic Loss, and MSE values.

5.3.1 Data Splitting

The dataset used in this analysis is split randomly into training (70%), validation (15%), and testing (15%). The training set is used to train the model. The validation set is used to optimize

the hyperparameters. Following hyperparameter optimisation, the test set is utilised to evaluate model performance on unseen data.

5.3.2 Baseline Multilayer Feedforward Neural Network

This section presents the classification results for the MLP using the T2D dataset. Several association analysis thresholds are considered including 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} resulting in 7, 13, 23, 103, 766, and 6609 SNPs respectively. The architecture of the MLP network for each subset of features with its tuning parameters including activation function, input_dropout_ratio, L1 and L2 penalties, momentum_stable, momentum_start, learning rate, and rate_annealing is presented in Table 5.3. Several neuron and hidden layers were tested to specify the optimal network topology to obtain the best results. For input neurons with 103, 766, and 6609 SNPs the best performance was obtained using three hidden layers with one hundred neurons in each and one output neuron. For 23 SNPs (input nodes), three hidden layers with ten nodes in each is specified. For 13 and 7 SNPs configurations four hidden layers with fifty nodes in each is used. Based on empirical analysis, these configurations produced the best results.

Table 5.3: Configuration of the Network for MLP for Different Subsets of Features

Input neurons	Activation function	Hidden	Input dropout	L1	λ	Momentum stable	Momentum start	Learning rate	Rate annealing
6609	RectifierWithDropout	[100,100,100]	0.2	9.9e-5	3.7e-5	0.99	0.5	0.001	1.0e-6
766	RectifierWithDropout	[100,100,100]	0.2	9.9e-5	3.7e-5	0.99	0.5	0.001	1.0e-6
103	RectifierWithDropout	[100,100,100]	0.2	9.9e-5	3.7e-5	0.99	0.5	0.001	1.0e-6
23	TanhWithDropout	[10,10,10]	0.0	4.6e-5	8.1e-5	0.0	0.0	1.0e-4	1.0e-7
13	MaxoutWithDropout	[50,50,50,50]	0.0	1.2e-5	2.2e-5	0.0	0.0	0.001	1.0e-8
7	MaxoutWithDropout	[50,50,50,50]	0.0	1.2e-5	2.2e-5	0.0	0.0	0.001	1.0e-8

Table 5.4 provides the performance metrics for the MLP using the validation set. Metric values for p-value thresholds 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} were obtained using an optimized F1 threshold with values 0.4822, 0.4879, 0.4686, 0.5155, 0.3914, and 0.3005 respectively.

Table 5.4: Performance Metrics of MLP for Validation Set

p-value	AUC	Sens	Spec	Logloss	Gini	MSE
10^{-2}	0.9479	0.9133	0.8319	0.3166	0.8959	0.0966
10^{-3}	0.8682	0.8430	0.7338	0.4545	0.7365	0.1482
10^{-4}	0.6938	0.6814	0.6246	0.6330	0.3813	0.2211
10^{-5}	0.6176	0.8477	0.3305	0.6720	0.2117	0.2395
10^{-6}	0.6052	0.7142	0.4257	0.6722	0.2110	0.2397
5×10^{-8}	0.5778	0.8501	0.2745	0.6811	0.1537	0.2440

Table 5.5 provides the performance metrics for the MLP using the test set. Metric values for 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} were gained using an optimized F1 threshold with values 0.4762, 0.4564, 0.4712, 0.5789, 0.3919, and 0.3502 respectively. Comparatively, the results are lower than those obtained by the validation set, which is expected.

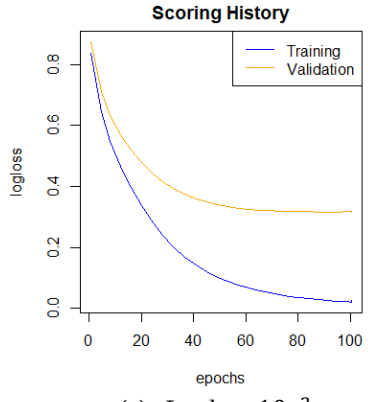
Table 5.5: Performance Metrics of MLP for Test Set

p-value	AUC	Sens	Spec	Logloss	Gini	MSE
10^{-2}	0.9534	0.9439	0.8086	0.2849	0.9069	0.0885
10^{-3}	0.8375	0.8528	0.6709	0.5031	0.6751	0.1643
10^{-4}	0.6810	0.5397	0.7346	0.6397	0.3660	0.2245
10^{-5}	0.5774	0.8060	0.3035	0.6773	0.2009	0.2421
10^{-6}	0.5699	0.8878	0.1836	0.6806	0.1634	0.2438
5×10^{-8}	0.5434	0.8528	0.2117	0.6899	0.0848	0.2484

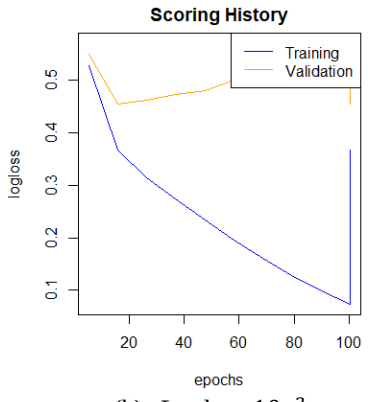
The classification accuracy for the MLP classifier model shows significant improvement with values ranging between 57.78% for 5×10^{-8} and 94.79% for 10^{-2} in the validation set. This is also the case for the test set with values of 54.34% and 95.34% for 5×10^{-8} and 10^{-2} p-value thresholds respectively. Sensitivity and specificity metrics for the MLP using

the validation and test sets are imbalanced for lower p-value thresholds, with signs of bias in sensitivities. However, this is not the case with higher thresholds 10^{-4}, 10^{-3}, and 10^{-2} where the number of SNPs increases.

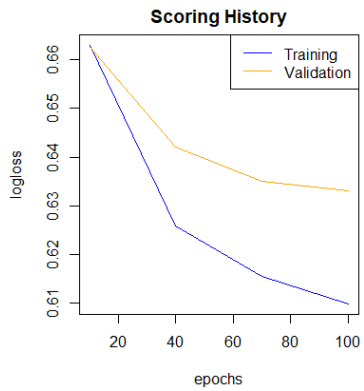
Figure 5.4 shows the model learning curve for the training and validation sets and is useful to model overfitting. Early stopping was adopted to avoid overfitting during model learning. Logloss stopping metrics was used to stop model learning on the validation set when the model's Logloss value did not improve by 1e-2 (stopping_tolerance) after reaching 5 scoring epochs (stopping_rounds). An overfitted model can be diagnosed from the Logloss plot where the train loss curve slopes down while the validation loss curve slopes down, hits the inflection point (the epochs points) and begins to slope up. As can be seen in Figure 5.4 (a) and (b), there are signs of overfitting in both cases. In same figure, a small amount of overfitting can be observed in (c), (d), (e), and (f) which in the future can be addressed by imposing heavier regularisations. Furthermore, the AUC plots are visualized to highlight if overfitting occurs. AUC plots provide useful information about early divergence between the training and validation curves, Figure 5.5 shows that there is a small amount of overfitting, again this can be addressed using heavier regularisations.



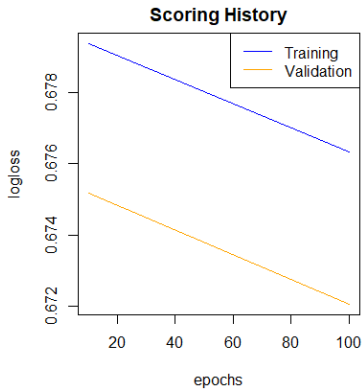
(a) Logloss 10^{-2}



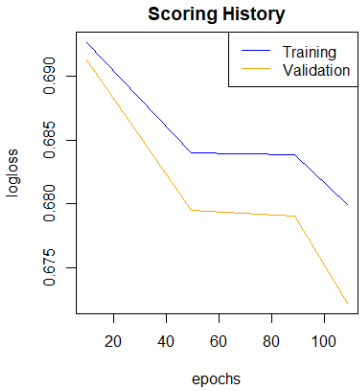
(b) Logloss 10^{-3}



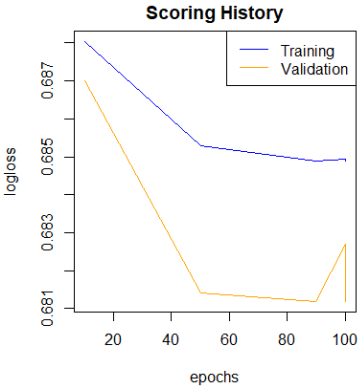
(c) Logloss 10^{-4}



(d) Logloss 10^{-5}

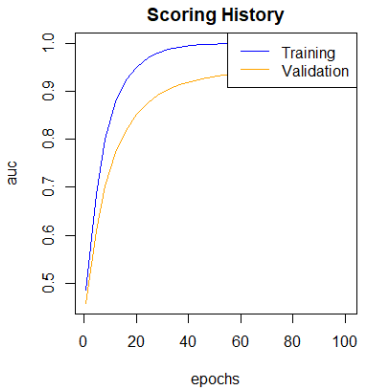


(e) Logloss 10^{-6}

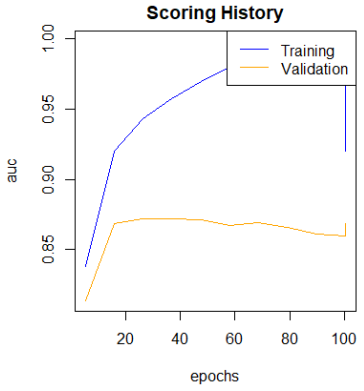


(f) Logloss 5×10^{-8}

Figure 5.4: (a) to (f) Logloss Plots against Epochs for p-value 10^{-2} to 5×10^{-8}

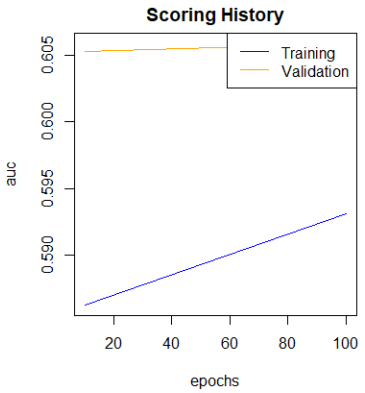


(a) AUC 10^{-2}

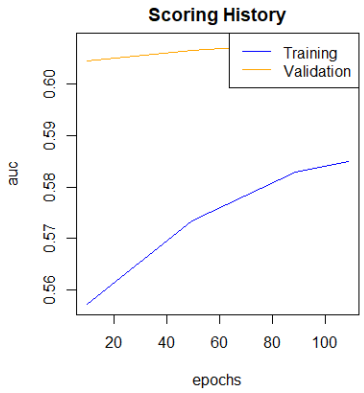


(b) AUC 10^{-3}

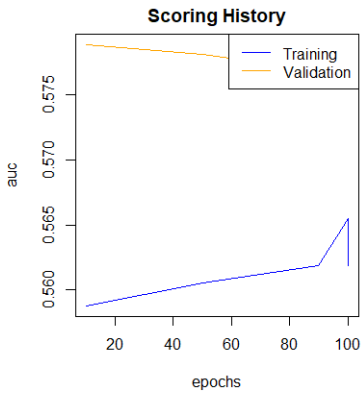
(c) AUC 10^{-4}



(d) AUC 10^{-5}



(e) AUC 10^{-6}



(f) AUC 5×10^{-8}

Figure 5.5: (a) to (f) AUC Plots against Epochs for p-value 10^{-2} to 5×10^{-8}

5.3.3 Baseline Random Forest Ensemble Method

An RF classifier is utilised which is a randomized decision tree-based ensemble for the classification of T2D case-control observations and to benchmark the performance of the MLP classifier. In this evaluation, using the same subset of features obtained from association analysis p-value thresholds (5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2}), with the number of trees set to 200 and a maximum depth of 20, the early stopping criterion is adopted to build an optimized RF model and to avoid overfitting. The early stopping criterion is determined using stopping metrics, stopping rounds, and stopping tolerance values set to Logloss, 4, and 1e-2 respectively. Model learning stops fitting new trees, when the model's Logloss value, fails to increase more than 1e-2 after 4 scoring intervals.

Table 5.6 shows the performance metrics for the RF using the validation set. Metric values for 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} were obtained using optimized F1 threshold values 0.4952, 0.4963, 0.2858, 0.5474, 0.5287, and 0.5500 respectively.

Table 5.6: Performance Metrics of RF for Validation Set

p-value	AUC	Sens	Spec	Logloss	Gini	MSE
10^{-2}	0.7471	0.6416	0.7507	0.6517	0.4943	0.2295
10^{-3}	0.7366	0.7213	0.6610	0.6422	0.4733	0.2250
10^{-4}	0.6610	0.6112	0.6302	0.6518	0.3221	0.2298
10^{-5}	0.5473	0.9203	0.1092	0.7520	0.0947	0.2686
10^{-6}	0.5620	0.7423	0.3697	0.7209	0.1241	0.2578
5×10^{-8}	0.5552	0.8103	0.2857	0.6936	0.1104	0.2489

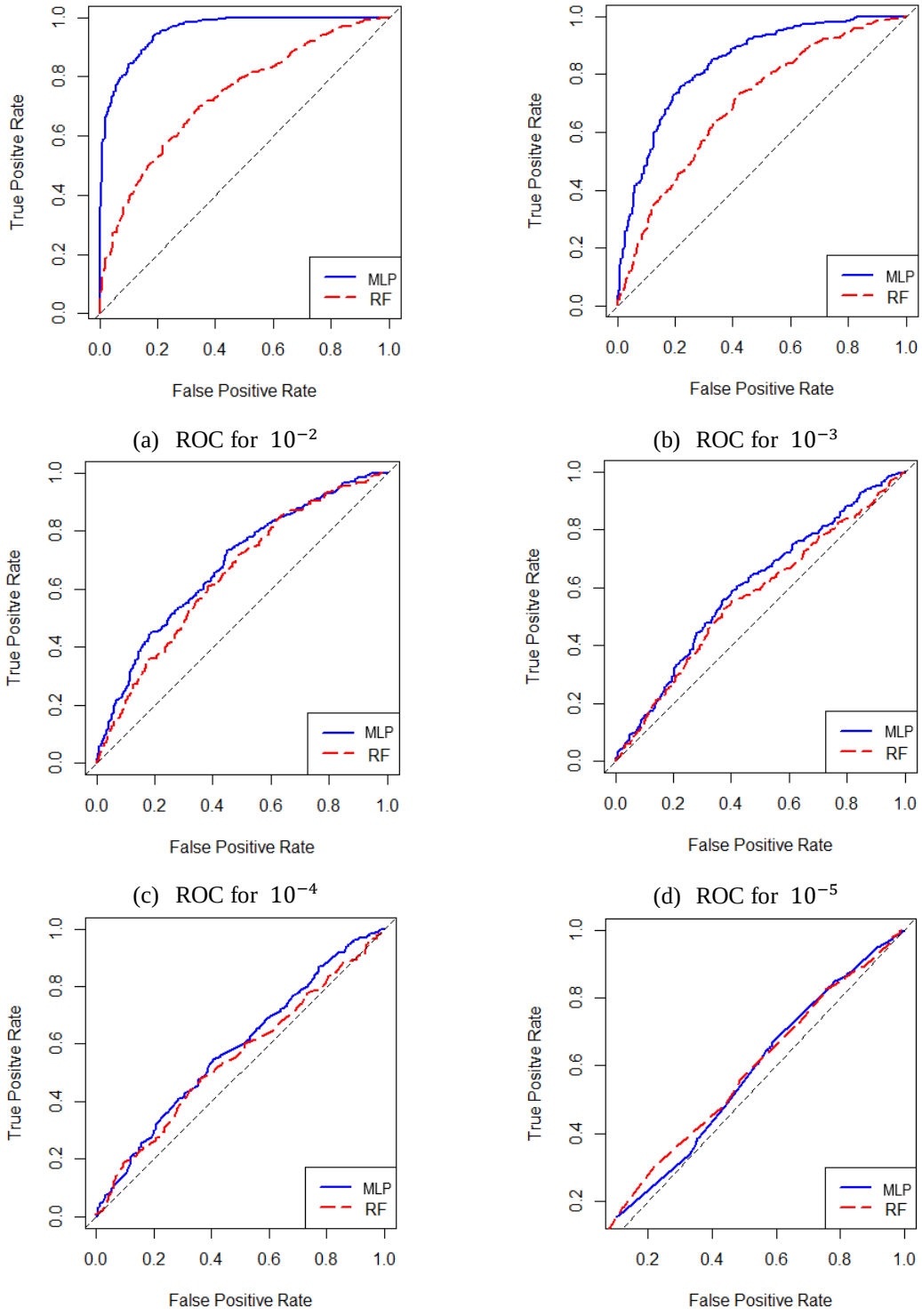
Table 5.7 presents the performance metrics for the RF using the test set. Metric values for 5×10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} were obtained using optimized F1 threshold values 0.4944, 0.4872, 0.5677, 0.5118, 0.5350, and 0.5508 respectively. The classification accuracy metric partially shows, in some cases, worse results than those obtained using the validation set, which is to be expected.

Table 5.7: Performance Metrics of RF for Test Set

p-value	AUC	Sens	Spec	Logloss	Gini	MSE
10^{-2}	0.7353	0.5654	0.7831	0.6566	0.4706	0.2320
10^{-3}	0.6966	0.7313	0.5841	0.6551	0.3932	0.2313
10^{-4}	0.6498	0.7196	0.5102	0.6595	0.2996	0.2335
10^{-5}	0.5661	0.5560	0.5969	0.7444	0.1322	0.2663
10^{-6}	0.5517	0.7245	0.3346	0.7300	0.1034	0.2615
5×10^{-8}	0.5518	0.8294	0.2423	0.6924	0.1036	0.2496

The RF classifier model shows 19.19% improvement for the validation set and 18.35% improvement for the test set. Sensitivities and specificities for the RF using the validation and test sets are imbalanced for lower p-value thresholds, indicating that RF is not able to distinguish between cases and controls. However, for higher thresholds 10^{-4}, 10^{-3}, and 10^{-2} where the number of SNPs increases a much higher improvement can be observed.

Figure 5.6 presents the ROC curves generated for the MLP and RF classifiers. The performance of the MLP and RF are similar when using 5×10^{-8}, 10^{-6}, 10^{-5}, and 10^{-4}. However, the MLP classifier model outperforms the RF model when using 10^{-3} and 10^{-2} thresholds.



(a) ROC for 10^{-2}

(b) ROC for 10^{-3}

(c) ROC for 10^{-4}

(d) ROC for 10^{-5}

(e) ROC for 10^{-6}

(f) ROC for 5×10^{-8}

Figure 5.6: (a) to (f) Performance ROC Curves for MLP and RF Test Sets using p-value Threshold 10^{-2} to 5×10^{-8}

5.3.4 Deep Learning Stacked Autoencoder Results

This section presents the classification results for T2D that were obtained using a Deep Learning SAE. This evaluation considers SNPs generated with a p-value threshold 10^{-2} which filters the dataset to 6609 SNPs. Deep learning stacked autoencoders use these SNPs

to extract latent information and non-linear epistatic interactions between SNPs. The results are based on four stacked autoencoders. The first SAE consists of 2500 hidden neurons while the second, third and fourth SAE use (2500, 1500), (2500, 1500, 700), and (2500, 1500, 700, 300) hidden neurons respectively.

The network architecture for each SAE layer along with its tuning parameters is presented in Table 5.8. Several neuron and hidden layer configurations are tested to determine the optimal network topology for softmax classification tasks. For input neurons with 2500 and 1500 compressed SNPs the best performance was obtained using four hidden layers with ten neurons in each, and one output neuron. Using 700 SNPs, three hidden layers and one hundred nodes in each are used. While, 300 SNPs, with three hidden layers and ten nodes in each, produced sufficient results.

Table 5.8: Configuration of the Network for MLP Softmax Classifier for Four SAEs

Input neurons	Activation function	Hidden	Input dropout	L1	λ	Momentum stable	Momentum start	Learning rate	Rate annealing
2500	MaxoutWithDropout	[10,10,10,10]	0.2	3.5e-5	9.6e-5	0.99	0.0	1.0e-4	1.0e-8
1500	MaxoutWithDropout	[10,10,10,10]	0.2	9.6e-5	0.99	0.0	1.0e-4	1.0e-8	9.6e-5
700	RectifierWithDropout	[100,100,100]	0.2	9.9e-5	3.7e-5	0.99	0.5	0.001	1.0e-6
300	RectifierWithDropout	[10,10,10]	0.2	3.3e-5	9.3e-5	0.5	0.0	0.001	1.0e-6

Table 5.9 illustrates the performance metrics for the deep learning SAE using the validation set. The metric values for the first SAE (2500 hidden units), second SAE (2500, 1500), third SAE (2500, 1500, 700), and fourth SAE (2500, 1500, 700, 300) were obtained using an optimized F1 threshold with values 0.2904, 0.3619, 0.3388, and 0.4662 respectively.

Table 5.9: Performance Metrics of DL SAE for Validation Set

SAE	AUC	Sens	Spec	Logloss	Gini	MSE
2500	0.9400	0.9297	0.7815	0.3484	0.8800	0.1024
1500	0.9039	0.9039	0.7226	0.3910	0.8079	0.1244
700	0.8909	0.9320	0.6834	0.4406	0.7818	0.1344
300	0.8064	0.8477	0.6526	0.5416	0.6128	0.1792

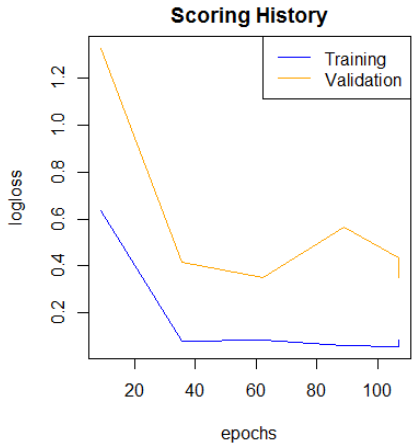
Table 5.10 presents the performance metrics obtained using the test set. The metric values for the first SAE (2500 hidden units), second SAE (2500, 1500), third SAE (2500, 1500, 700), and fourth SAE (2500, 1500, 700, 300) were obtained using an optimized F1 threshold with values 0.6173, 0.4149, 0.3514 and 0.3889 respectively. The results are lower than those produced using the validation set but not for all cases, for SAE (2500 hidden units) a 0.25% improvement was observed.

Table 5.10: Performance Metrics of DL SAE for Test Set

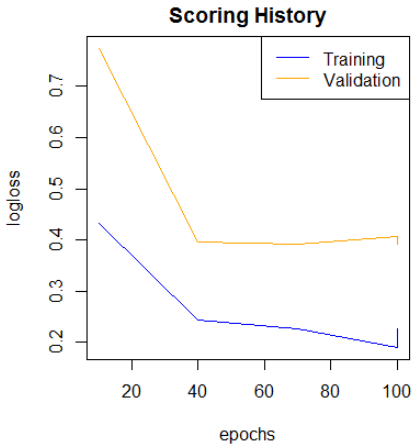
SAE	AUC	Sens	Spec	Logloss	Gini	MSE
2500	0.9425	0.8714	0.8954	0.3405	0.8851	0.0948
1500	0.8947	0.8761	0.7448	0.4226	0.7895	0.1330
700	0.8689	0.9228	0.6326	0.4969	0.7379	0.1536
300	0.8078	0.8785	0.5306	0.5511	0.6157	0.1843

The classification accuracy using the SAE approach shows a progressive deterioration as the input features are steadily compressed to 300 hidden neurons using both the validation and test sets. Despite the gradual deterioration in performance, satisfactory results were achieved with 1500 hidden units. Figure 5.7 presents the model learning curves for both the training and validation sets to detect overfitting. Early stopping was adopted to ensure this does not happen and the Logloss stopping metric was used to stop the model learning when the model's Logloss value does not improve by 1e-2 after reaching 5 scoring epochs. Again, as

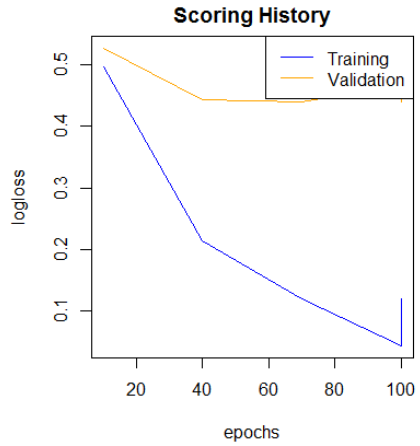
can be seen in Figure 5.7 and Figure 5.8 (a), (b), and (c) there are signs of overfitting which can also be addressed with heavier regularisations. But the last figure with 300 hidden units (d), shows that overfitting is appropriately managed.



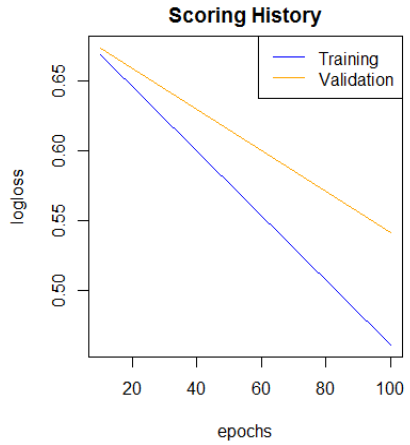
(a) Logloss for hidden=2500



(b) Logloss for hidden=1500

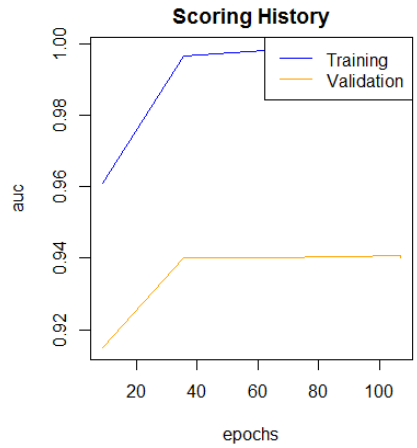


(c) Logloss for hidden=700

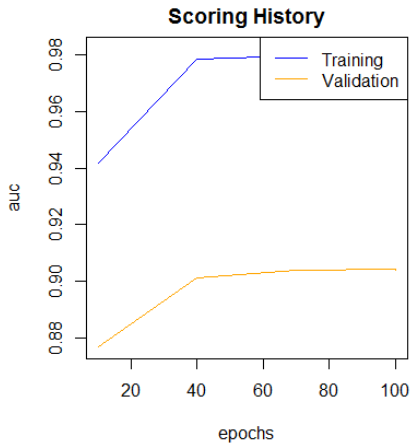


(d) Logloss for hidden=300

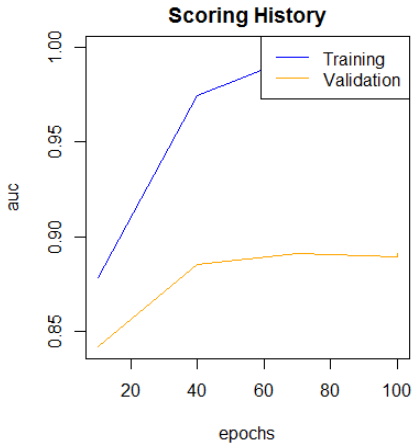
Figure 5.7: (a) to (d) Logloss Plots against Epochs for 2500, 1500, 700, and 300 Compressed Units



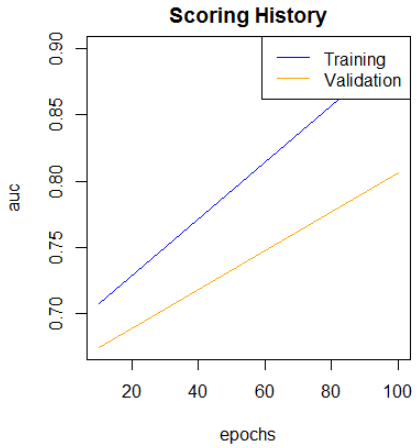
(a) AUC for hidden=2500



(b) AUC for hidden=1500



(c) AUC for hidden=700



(d) AUC for hidden=300

Figure 5.8: AUC Plots against Epochs for 2500, 1500, 700, and 300 Compressed Units

The ROC curve in Figure 5.9 illustrates the cut-off values for the false and true positive rates using the test set. In this evaluation, the ROC curve shows a gradual deterioration in the performance of the softmax classifiers as the initial 6609 features (SNPs) are gradually compressed down to 300 hidden units in the stacked autoencoder configurations. Although the predictive accuracy degraded from 94.25% for 2500 hidden units and to 80.78% for 300 hidden units, the results still remain acceptable.

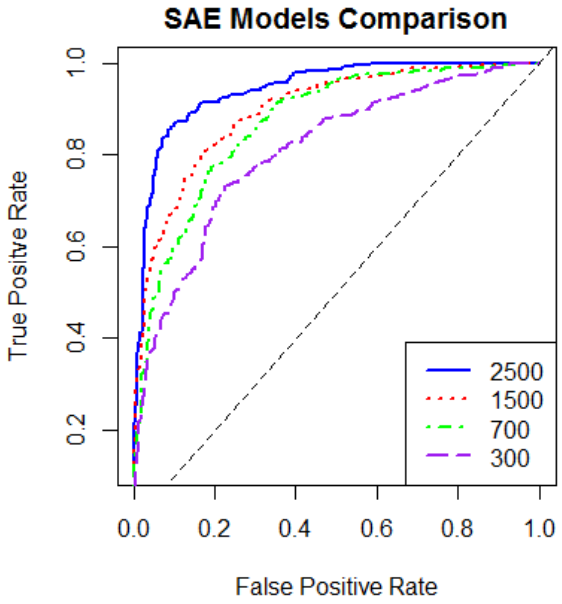


Figure 5.9: Performance ROC Curves of DL SAE for Test Set for 2500, 1500, 700, and 300 Hidden Units

5.4 Results for the Classification of Genetic and Clinical Data using Traditional Machine Learning

Several experiments are considered in this section to investigate and evaluate the predictive capacity of traditional machine learning models using genetic data only, clinical data only, and lastly when genetic and clinical data are combined. Five algorithms including GBM, SVM, RPART, NNET, and GLM are employed to conduct three distinct experiments with genomic data only, clinical data only, and lastly genetic and clinical data combined. The performance of each model is measured using AUC, Sensitivity, and Specificity.

5.4.1 Data Splitting

The dataset is split randomly into training (80%) and testing (20%) to evaluate model performance. For resampling, 5-fold cross-validation with 30 repetitions is employed and the average performance is calculated.

5.4.2 Genetic Analysis Results

The first experiment was conducted using genomic features only; these include SNPs (rs4132670, rs12243326, rs12255372, rs7901695, rs4506565, and rs2371765) extracted from logistic association analysis using Bonferroni correction threshold. Five machine learning algorithms including GBM, SVM, RPART, NNET, and GLM were designed and evaluated using training and testing sets as specified previously. The hyperparameters for these models were experimentally detected. GBM, a boosted tree model, is used for classification with the number of trees (n.trees) in the range 50 to 150, and the number of leaves in each tree between 1 and 3 (represents the complexity of tree (interaction.depth)). The minimum number of samples in tree nodes (n. minobsinnode) is set to 10, with the learning rate (shrinkage) set to 0.1. The best tuning values used to fit the model include n.trees set to 50 and interaction.depth set to 1. An SVM, a classifier model based on radial kernel function, is used for classification with sigma set to 0.2731565 and the cost (C) parameter range between 0.25 and 1. RPART is also used, which is a decision tree model,

for classification; the size and the splits in the decision tree are controlled with the complexity parameter configured to 0.001526718. In addition, NNET, a neural network model with single hidden layer, is employed for classification with 1, 3, and 5 hidden layer units. The regularization parameter (decay) is configured to 0.1, which is useful to avoid overfitting. Table 5.11 highlights the optimal values used along with the selected machine learning algorithms used in classification tasks.

Table 5.11: Tuning Parameters for Models using Genetic Data

Classifier	**Best Tuning Parameters**
GBM	number of trees = 50 number of leaves in a tree = 1 learning rate = 0.1 minimum number of training set samples in a node = 10
SVM	sigma = 0.2731565 cost = 0.25
RPART	complexity parameter = 0.001526718
NNET	size = 1 decay = 0.1

The results presented in Table 5.12 show that sensitivities and specificities are imbalanced for all the models, sensitivities are lower than specificities. This indicates that the selected features for these models are inadequate at distinguishing between cases and controls. This analysis also reveals that the performance using the AUC for linear and non-linear classifiers are almost the same ranging between 57.09% for SVM and NNET classifiers and 57.46% for the GLM. Figure 5.10 illustrates the ROC curve for the chosen models.

Table 5.12: Predictive Results for Genetic Analysis

Classifier	Accuracy	Sensitivity	Specificity
GLM	0.5746	0.2668	0.8348
GBM	0.5718	0.2546	0.8399
SVM	0.5709	0.2424	0.8485
RPART	0.5737	0.2607	0.8382
NNET	0.5709	0.2587	0.8348

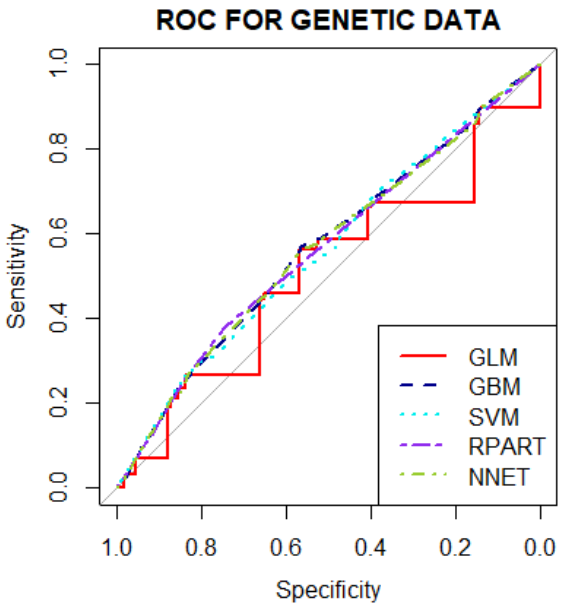


Figure 5.10: ROC Curve for Five Models using Six SNPs Reached Bonferroni Correction Threshold

5.4.3 Genetic Analysis Results using Feature Selection

Using the RFE algorithm for feature selection, individual features are assessed to determine their rank within all features considered. Figure 5.11 shows the results for various feature combinations. The results highlight that the optimal number of features is three with an AUC=0.558. The three ranked features are rs2371765, rs12255372, and rs4132670 and these are used as input features for the five models to investigate whether this reduced feature set can enhance or maintain the previous result when the whole set of features are used.

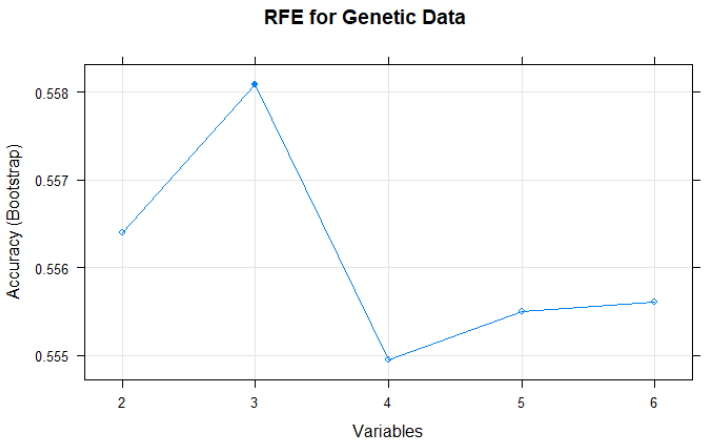


Figure 5.11: Recursive Feature Elimination Plot for Genetic Data

5.4.3.1 Classifier Performance of Genetic Data using Features Extracted from RFE

This experiment was conducted using the three features selected by RFE (rs2371765, rs12255372, and rs4132670). The hyperparameters are experimentally detected. For the GBM, the number of trees is configured to 150 with the number of leaves in a tree set to 1. The minimum number of samples in tree nodes is set to 10 with the learning rate set to 0.1. Whereas for the SVM, sigma is set to 0.2950747 and the cost parameter is configured to 0.25. For the RPART classifier model, the size and the split of the decision tree is controlled with the complexity parameter configured to 0.001526718. Finally, for the NNET model, the number of units in the hidden layer is set to 3 and the regularization parameter configured to 1e-04. Table 5.13 describes the tuning hyperparameters for these chosen models.

Table 5.13: Tuning Parameters for Models using Genetic Data Selected by RFE

Classifier	Best Tuning Parameters
GBM	number of trees = 150 number of leaves in a tree = 1 learning rate = 0.1 minimum number of training set samples in a node = 10
SVM	sigma = 0.2950747 cost = 0.25
RPART	complexity parameter = 0.001526718
NNET	size = 3 decay = 1e-04

Table 5.14 describes the predictive results obtained from each of the models using the reduced feature set from the genomic dataset. It is evident that the GLM, GBM, and RPART classifiers perform slightly worse in terms of accuracy measurement. Sensitivity for the GBM improved by 1.83% indicating that this classifier can classify case instances much better using these selected features. However, the SVM and NNET can improve on the previous results demonstrated using the full features set.

Figure 5.12 illustrates the ROC curves for the five models which shows no significant improvements on the previous set of results for the five models when the full set of features are used.

Table 5.14: Predictive Results for Genetic Analysis using Features Selected by RFE

Classifier	Accuracy	Sensitivity	Specificity
GLM	0.5709	0.2587	0.8348
GBM	0.5709	0.2729	0.8227
SVM	0.5728	0.2587	0.8382
RPART	0.5737	0.2607	0.8382
NNET	0.5718	0.2546	0.8399

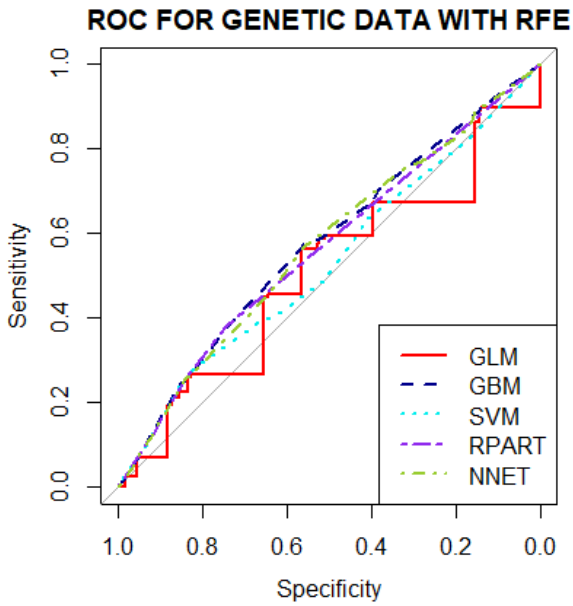


Figure 5.12: ROC Curve for Five Models using Three SNPs Chosen using RFE

5.4.4 Clinical Analysis Results

A separate analysis is conducted using clinical variables only. These include Body Mass Index (BMI), alcohol intake (Alcohol), smoking status (SMK), physical activity (ACT), family history of diabetes (Famdb), high blood pressure (Hbp), high blood cholesterol (Chol), AGE and SEX. The same five machine learning algorithms used in the previous experiment are used in this experiment. Table 5.15 presents the optimal values tested for these models.

Table 5.15: Tuning Parameters for Models using Clinical Data

Classifier	Best Tuning Parameters
GBM	number of trees = 150 number of leaves in a tree = 1 learning rate = 0.1 minimum number of training set samples in a node = 10
SVM	sigma = 0.07629681 cost = 0.25
RPART	complexity parameter = 0.007097792
NNET	size = 1 decay = 0.1

The results in Table 5.16 show that the GBM classifier yields the best accuracy with 71.06%. Although GBM produced the best AUC performance, the model classifies control instances better than cases with 64.63% and 76.52% for sensitivity and specificity, respectively. The AUC values for the SVM and RPART are lower than other classifiers. However, RPART is the only classifier with sensitivity higher than specificity (73.89%, 61.47%) which means that RPART model can recognize cases better than controls. Figure 5.13 shows the ROC curve for the selected models. In fact, in general the NNET actually performs better than all others given that the balance between sensitivity and specificity is much closer with an AUC of 71%. Although the classification performance of clinical data shows satisfactory results, the value of clinical information is limited as it is only useful when individuals in the study have already developed the disease (in this case T2D).

Table 5.16: Predictive Results for Clinical Analysis

Classifier	Accuracy	Sensitivity	Specificity
GLM	0.7086	0.6505	0.7581
GBM	0.7106	0.6463	0.7652
SVM	0.6893	0.6379	0.7330
RPART	0.6718	0.7389	0.6147
NNET	0.7096	0.6800	0.7348

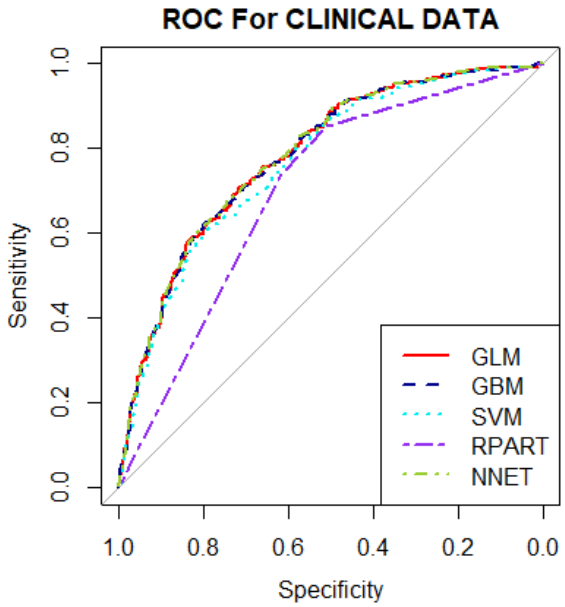


Figure 5.13: ROC Curve for Five Models using Clinical Data

5.4.5 Clinical Analysis Results using Feature Selection

In this experiment the RFE algorithm is utilised to determine the optimal number of clinical features for model training. Figure 5.14 illustrates the accuracy results for various feature combinations. The optimal number of features is eight which produces an AUC=0.7024. The eight ranked features are BMI, Famdb, Hbp, Chol, SMK, Alcohol, SEX, and ACT.

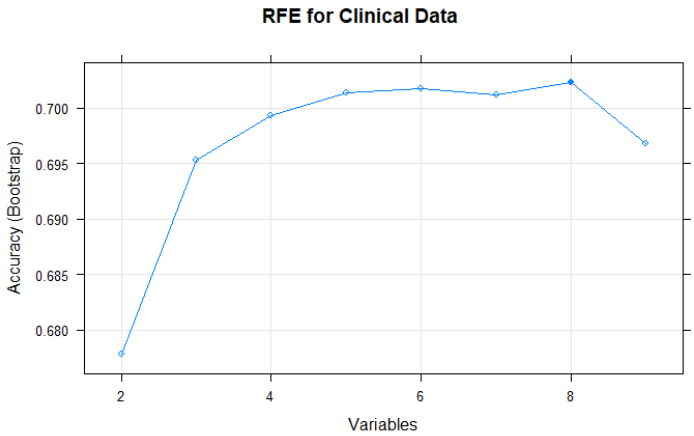


Figure 5.14: Recursive Feature Elimination Plot for Clinical Data

5.4.5.1 Classifier Performance of Clinical Data using Features Extracted from RFE

The eight clinical variables extracted using RFE are used as input for the selected models. Again, hyperparameter tuning is performed experimentally with the coefficient values shown in Table 5.17.

Table 5.17: Tuning Parameters for Models using Clinical Data Selected by RFE

Classifier	Best Tuning Parameters
GBM	number of trees = 100 number of leaves in a tree = 2 learning rate = 0.1 minimum number of training set samples in a node = 10
SVM	sigma = 0.08207512 cost = 0.25
RPART	complexity parameter = 0. 007097792
NNET	size = 1 decay = 0.1

The accuracy for all classifiers deteriorates slightly using the reduced feature set as presented in Table 5.18. The sensitivity for the GBM model shows a 4% improvement indicating that this classifier can classify cases much better using the reduced feature set. The ROC curve in Figure 5.15 shows that the five classifiers obtain similar results to the previous experiment.

Table 5.18: Predictive Results for Clinical Analysis using Features Selected by RFE

Classifier	Accuracy	Sensitivity	Specificity
GLM	0.7076	0.6421	0.7634
GBM	0.7096	0.6863	0.7294
SVM	0.6912	0.6505	0.7258
RPART	0.6718	0.7389	0.6201
NNET	0.7076	0.6821	0.7294

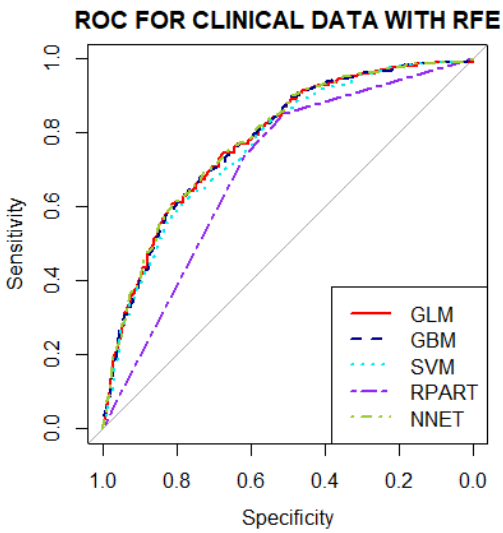


Figure 5.15: ROC Curve for Five Models using Clinical Data Selected by RFE

5.4.6 Genetic and Clinical Analysis Results

A combination of six SNPs along with nine clinical variables is used as input features in the third experiment. Again, the same five machine learning algorithms are used. The best tuning hyperparameters for those models are shown in Table 5.19.

Table 5.19: Tuning Parameters for Models using Genetic and Clinical Data

Classifier	Best Tuning Parameters
GBM	number of trees = 100 number of leaves in a tree = 2 learning rate = 0.1 minimum number of training set samples in a node = 10
SVM	sigma = 0.04330685 cost = 0.25
RPART	complexity parameter = 0.007360673
NNET	size = 1 decay = 0.1

The results in Table 5.20 show that the best classification accuracy was 72.99% which was obtained using the NNET algorithm. The AUC values for the GLM, GBM, SVM, RPART, NNET did yield better results than experiments that used clinical or genomic data only. In addition, sensitivities and their corresponding specificities for the GBM and NNET classifiers are balanced with values ~70%.

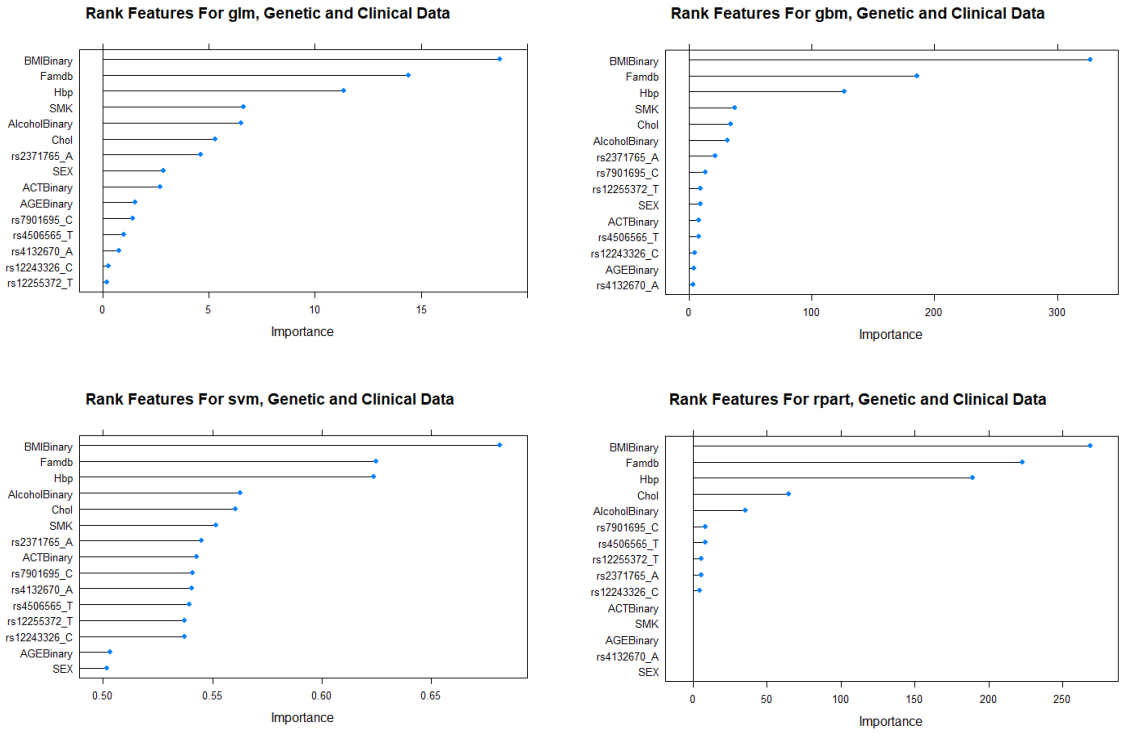
Table 5.20: Predictive Results for Genetic and Clinical Analysis

Classifier	Accuracy	Sensitivity	Specificity
GLM	0.7289	0.6758	0.7742
GBM	0.7212	0.7011	0.7384
SVM	0.7067	0.6695	0.7384
RPART	0.6718	0.7389	0.6147
NNET	0.7299	0.7011	0.7545

As can be seen in Figure 5.16 from the variable important plots the predictive values of the machine learning models used in this investigation are due to clinical data, with slight evidence arising from genetic data. BMI and Famdb were significantly important. Moreover, the importance of other clinical variables including Hbp, Chol, SMK, Sex, Alcohol, ACT,

and AGE appeared to vary among these five models. For the RPART model, the ranked features for ACT, SMK, AGE, and SEX seemed to have a trivial effect.

The importance of genetic variables, in relation to the predictive values for these five algorithms, varied and proved to be less relevant than clinical variables. Although all six genetic variables were used to model the SVM, and NNET, their rank measurement is low. For the GLM, GBM and RPART, however, not all genetic variables were considered, given that they showed minor to no influence on the predictive results. Again, the importance of clinical information to the prediction of developing T2D is only useful when the observations in the study have already been diagnosed with it. Using genetic data gives much better indication to the early prediction of the risk of developing the disease (T2D). Figure 5.17 presents the ROC curves for the selected models.



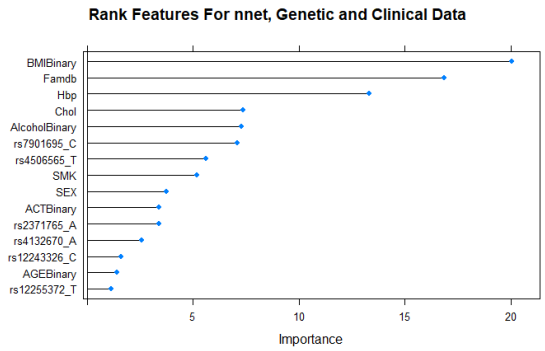


Figure 5.16: Variable Important Plots for Each Model

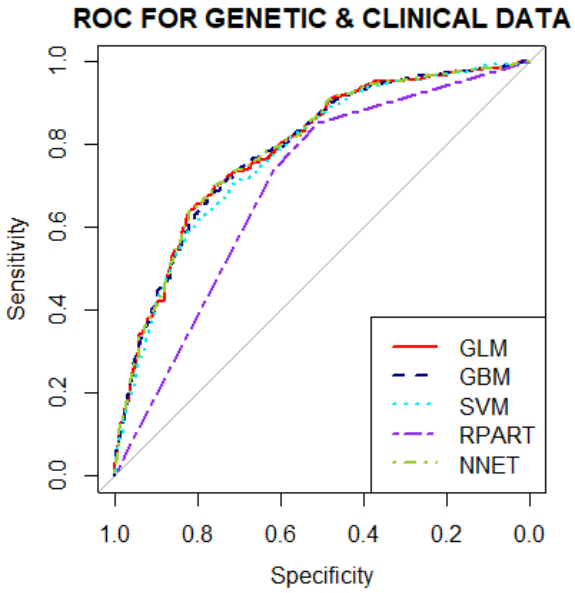


Figure 5.17: ROC Curve for Five Models using Genetic and Clinical Data

5.4.7 Genetic and Clinical Analysis Results using Feature Selection

Of the fifteen features containing genetic and clinical data, six optimal features are selected using the RFE algorithm. This reduced feature set is utilized to determine whether or not the performance capacity of the five models can improve on or maintain the previously reported set of results. Figure 5.18 illustrates the accuracy results for various feature combinations. The results show that six features achieved the highest performance with AUC=0.7061. The six ranked features are BMI, Famdb, Hbp, Chol, Alcohol, and rs4132670.

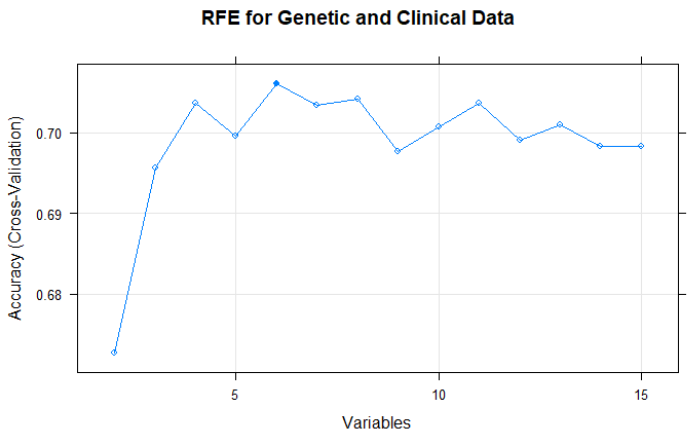


Figure 5.18: Recursive Feature Elimination Plot for Genetic and Clinical Data

5.4.7.1 Classifier Performance of Genetic and Clinical Features Extracted from RFE

The feature set contains one SNP and five clinical features. Again, Table 5.21 presents the optimal tuning hyperparameters for the chosen models.

Table 5.21: Tuning Parameters for Models using Genetic and Clinical Data Selected using RFE

Classifier	Best Tuning Parameters
GBM	number of trees = 150 number of leaves in a tree = 1 learning rate = 0.1 minimum number of training set samples in a node = 10
SVM	sigma = 0.1351249 cost = 0.25
RPART	complexity parameter = 0.007097792
NNET	size = 1 decay = 0.1

Table 5.22 shows the results of the five models using the six features. The performance of the five machine learning models deteriorates in this experiment in comparison to the previously reported results. Although sensitivities and their corresponding specificities for all models are relatively balanced, they do not improve on the previous results. Again, the ROC curve in Figure 5.19, demonstrates that the five classifiers do not improve on the previous set of results.

Table 5.22: Predictive Results for Genetic and Clinical Analysis using RFE

Classifier	Accuracy	Sensitivity	Specificity
GLM	0.7086	0.6526	0.7563
GBM	0.7115	0.6211	0.7885
SVM	0.7096	0.6232	0.7832
RPART	0.6718	0.7389	0.6147
NNET	0.7135	0.6842	0.7384

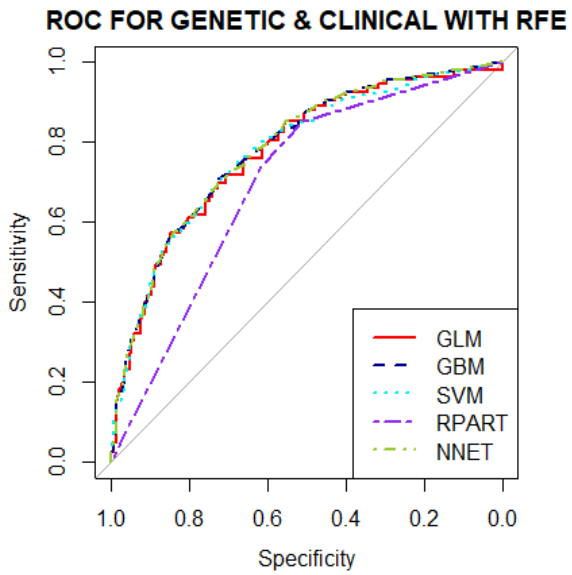


Figure 5.19: ROC Curve for Five Models using Genetic and Clinical Data Selected using RFE

As can be seen in Figure 5.20 the variable importance plots, performance is largely due to clinical data, with only one SNP from the genetic data being used. BMI, Famdb, and Chol were significantly important in all models and their rank was always at the top. In the case of the genetic feature rs4132670 its rank was always low.

In comparison to the full features set, the results confirm that the importance of genetic variables for the five algorithms using traditional machine learning algorithms, appears to be less relevant in comparison to clinical variables.

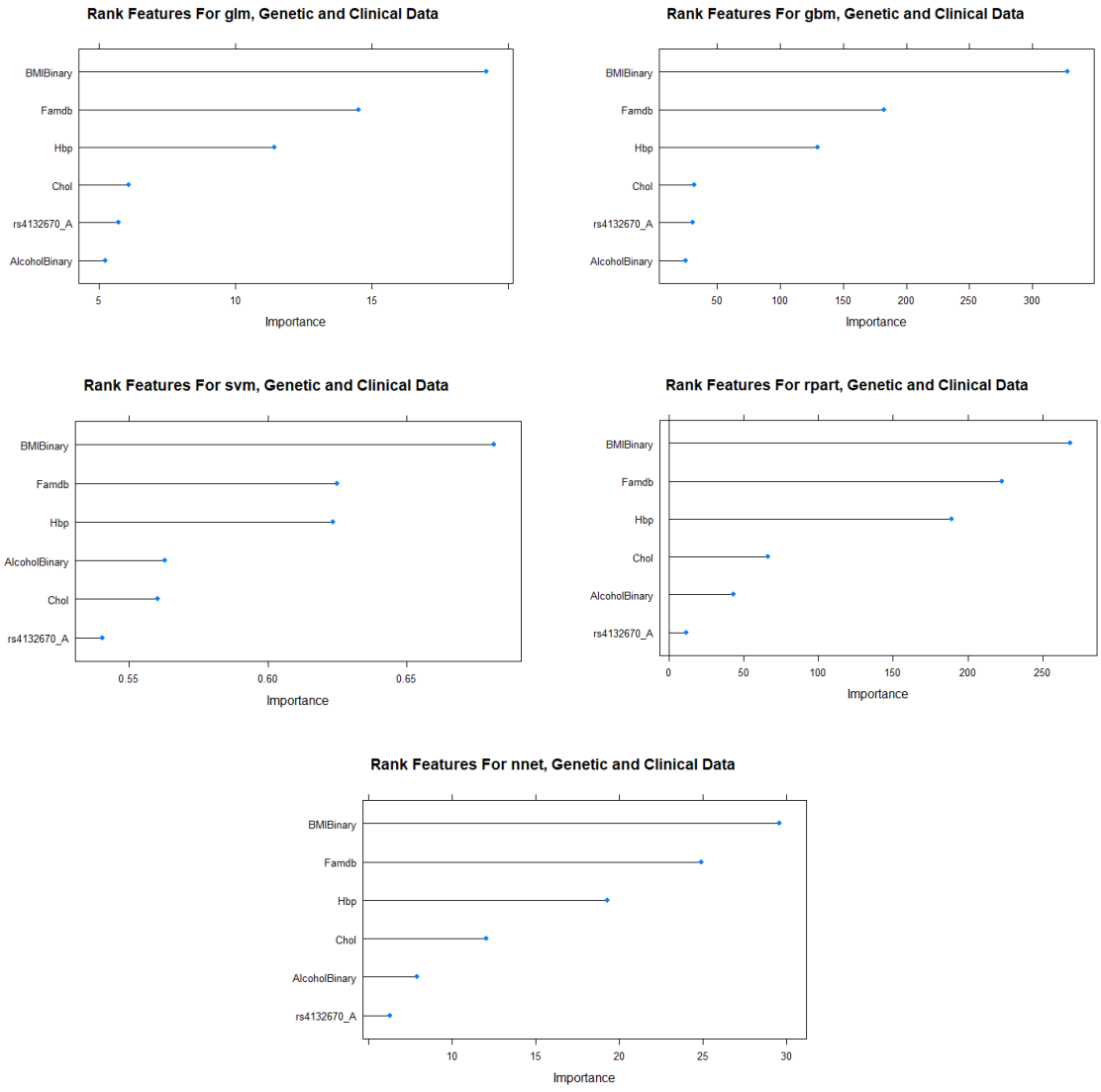


Figure 5.20: Variable Important Plots for Each Model with Features Selected using RFE

5.5 Summary

This chapter presented the results obtained from several experiments conducted in this thesis. Different performance evaluation measurements were presented and compared. In addition, the obtainable results for the binary classification of T2D high-dimensional genetic data using advanced machine learning algorithms were presented and benchmarked against simpler and less computationally expensive machine learning algorithms. These results support the arguments and novel claims made in this thesis that high-dimensional GWAS data and deep learning stacked autoencoders for unsupervised feature learning are sufficient to find epistatic interactions between SNPs and subsequently improve model accuracy in classification tasks.

Chapter 6 Discussion

Genetic association studies (GWAS) have significantly expanded our understanding of the genetic variants that predispose us to complex human diseases. Using a standard statistical test for single-SNP analysis has proven ineffective in complex disease given that single genetic loci (SNP) do not act independently to increase disease risk. The occurrence of complex diseases results from interactions between multiple genetic loci (Morris et al. 2012; Robinson et al. 2014; Lee et al. 2012). Modelling the complexity of these genotype-phenotype interactions in complex disorders is considered a significant challenge. Therefore, GWAS is more suitable for capturing linear interactions in diseases such as Cystic Fibrosis where a single SNP mutation is the cause (Cutting 2015).

In the field of bioinformatics where large and complex biological data structure exists, researchers have focused on the use of traditional machine learning algorithms to perform multi-SNP (epistatic) analysis. Random forests and support vector machine have been implemented in (Botta et al. 2014; Nguyen et al. 2015; Ban et al. 2010; Tello et al. 2013) to discover SNP correlations. Generalized multifactor dimensionality reduction (GMDR) (Zhu et al. 2013) have been successfully applied in the analysis of gene-gene interactions. However, investigating all possible SNP combinations in GWAS is computationally complex and expensive and as such has seen little success outside of large data centres equipped to provide such facilities. Therefore, in this thesis we proposed an alternative approach for studying epistatic interactions using a universal approximator to handle the complex non-linear correlations and interactions between features. This has the benefit of extracting the salient information that exists in genetic data and provides an interesting machine learning methodology for the classification of high-dimensional T2D. Using a deep learning stacked autoencoder we were able to detect non-linear epistatic interactions between features and use the weights of these learned features to initialise a multilayer perceptron softmax classifier for the binary classification tasks in hand.

This approach contributes to the body of knowledge in the area as genetic risk alleles associated with T2D identified by GWAS only answer 10 and 20 percent of the missing heritability of this complex disorder (Prasad & Groop 2015). Therefore the prediction of disease risk based on highly ranked SNPs demonstrates little predictive power (Mittag et al. 2012) as confirmed in the results generated in this thesis. Wei *et al.* (Wei et al. 2009) and Gül *et al.* (GÜL et al. 2014) found that much higher predictive accuracy is obtained when increasing the number of SNPs, while comparatively poorer performance is attained when including only SNPs above genome-wide significance thresholds.

6.1 Advanced Machine Learning with p-value < 10^{-2}

Initial baseline results were obtained using an MLP classifier model to investigate and evaluate its ability to distinguish between cases and controls in T2D genomic data with different feature size combinations. The best result was obtained using a p-value threshold of 10^{-2} (6609 SNPs) (AUC=95.34%, Sens=94.39%, Spec=80.86%, Logloss= 28.49%, Gini=90.69%, and MSE=8.85%). The results show that there was a clear deterioration in performance as the p-value threshold is increased. Using the Bonferroni genome-wide significance threshold 5×10^{-8} (7 SNPs) attained the worst results (AUC=54.34%, Sens=85.28%, Spec=21.17%, Logloss=68.99%, Gini=8.48%, and MSE=24.84%). It is clear a much higher predictive accuracy is obtained by increasing the number of SNPs. Sensitivities and specificities for the MLP model are imbalanced for lower p-value thresholds (10^{-5} (23 SNPs), 10^{-6} (13 SNPs), and 5×10^{-8} (7 SNPs)) with (Sens=80.60%, Spec=30.35%), (Sens=88.78%, Spec=18.36%), and (Sens=85.28%, Spec=21.17%) respectively. It is reasonable to conclude that the MLP cannot learn specificities when the number of SNPs was reduced using the dataset we had; given that MLP cannot learn sufficient relationships between SNPs to be able to classify cases and controls in a balanced way, this is hardly surprising. Therefore, the MLP classifier required a high number of SNPs (6609 SNPs) to learn the non-linear relationships between SNPs and to find the latent

representations of the relevant information in the dataset and thus to be able to classify case and control observations sufficiently. Although, MLP can learn and capture the latent representations between SNPs formed by epistatic interactions when using high number of features yet the proportion of data that represents noise cannot be controlled.

RF machine learning models have been successfully used in genetic studies (Botta et al. 2014; López et al. 2018; Schwarz et al. 2010; Kursa 2014). In this thesis, the results show that the best classification performance of RF model was obtained using a p-value threshold of 10^{-2} (6609 SNPs) (AUC=73.53%, Sens=56.54%, Spec=78.31%, Logloss= 65.66%, Gini=47.06%, and MSE=23.20%). In general, the sensitivities and specificities are unstable for lower p-value thresholds indicating that the RF classifier has low discriminatory capacity to separate case and control observations when using low number of SNPs (13 SNPs, and 7 SNPs). Fitting an RF model using genome-wide significance threshold (7 SNPs), the result dropped to (AUC=55.18%, Sens=82.94%, Spec=24.23%, Logloss= 69.24%, Gini=10.36%, and MSE=24.96%) classification performance as these SNPs are often false positives.

The RF classifier model is a highly recommended algorithm for data such as GWAS (Qi 2012) because the algorithm is a randomized decision tree-based ensemble that is highly data adaptive and can handle correlations and interactions among features while at the same time ranking those variables that are important (Chen & Ishwaran 2012). In our data set using a p-value threshold of 10^{-2} to extract 6609 SNPs (features) the results show that the MLP outperformed the RF. For the MLP (AUC=95.34%, Sens=94.39%, Spec=80.86%, Logloss= 28.49%, Gini=90.69%, and MSE=8.85%), while for the RF (AUC=73.53%, Sens=56.54%, Spec=78.31%, Logloss= 65.66%, Gini=47.06%, and MSE=23.20%). Both Models (MLP, and RF) are used to fit non-linear data however in our case the capacity of RF was in most cases lower than the MLP model as shown in Table 5.5 and Table 5.7. This is probably because the RF is not as efficient as the MLP to learn the non-linear relationships between SNPs.

Using our proposed deep learning stacked autoencoder approach to extract the latent representations from the 6609 SNPs through progressively smaller hidden layer units (2500, 1500, 700, and 300), the results using the validation and test sets demonstrate a gradual deterioration. The classification accuracy value of the 300 compressed neurons is reasonably high (80.78% in the test set). The best result was obtained when using 2500 compressed units (AUC=94.25%, Sens=87.14%, Spec=89.54%, Logloss= 34.05%, Gini=88.51%, MSE=9.48%). Moreover, the Logloss scoring history plot shown in Figure 5.7 (a-c) suggest that overfitting was exist between the training and validation datasets when using 2500, 1500, and 700 compressed units. The validation Logloss deviates from the training Logloss after reaching 40 epochs. Whereas in Figure 5.7 (d) with 300 hidden unites overfitting was appropriately managed.

Although SAE with 2500 compressed units (initially 6609 SNPs) achieved less predictive accuracy than MLP using (6609 SNPs) the results are still comparable and significant for both models with an AUC=94.25% for the SAE and an AUC=95.34% for the MLP. Furthermore, it is noticeable that sensitivity and specificity for the SAE with 2500 hidden layer units attained better stability (Sens=87.14%, and Spec=89.54%) than the comparable MLP using 6609 SNPs (Sens=94.39%, and Spec=80.86%). This indicates that learning the deep features within the 6609 SNPs by reducing the dimensionality to 2500 better represents the data (removes data considered to be noise). This allowed us to train the softmax classifier (MLP) to better discriminate between case and control observations. The results show that we obtained high results even when the original data is compressed to 300 SNPs with a predictive accuracy of 80.78%. This is encouraging and demonstrates the potential for applying DL to high-dimensional GWAS data to extract features for classification modelling.

The SAE with 300 hidden layer units (AUC=80.78%) showed significantly higher results than those produced by the RF (AUC=73.53%). This is because the multiple hidden layers

compress the input features into abstract representations to model the complexity of the non-linearity of SNP interactions generally observed in genetic data, while removing less important information. This automated feature extraction algorithm outperforms the traditional supervised classification models presented in this thesis and offers a powerful way to enhance GWAS data analysis.

Our T2D NHS-HPFS GWAS data is used in Kim's work (Kim et al. 2018). The authors utilized different genetic association mappings (Fisher's exact test and L1-penalized logistic regression) to our approach to extract different subsets of SNPs (96, 214, 399, and 678 SNPs). Deep neural network with 2 hidden layers of 50 neurons was used in their work to classify T2D case-control observations. In comparison to Kim's work, the results presented in this thesis using DL SAE with 300 compressed units achieved ~81% predictive accuracy while they obtained (~79% for male, and ~82% for female) for 214 SNPs which is a comparable result to our classification performance result. For 399 SNPs, they achieved (~87% for male, ~86% for female). Although their results using 399 SNPs are comparably higher than our result, yet our 300 compressed features extracted using DL SAE represents the reduced, non-linear and latent information from the initial features (6609 SNPs). These, we consider, are a better representation of features than using direct features as in Kim's work from statistical logistic approach.

Deep learning is used in DeepWAS (Arloth et al. 2016) to identify individual regulatory SNPs by investigating genomic location and sequence alterations before association analysis is conducted. This approach differs to the approach presented in this thesis, in that, QC and association analysis are conducted using all of the SNPs genotyped in the T2D NHS-HPFS study data. Pre-SNP selection, based on functional regulatory effects, is not applied since our aim is to find epistatic interactions between SNPs. While DeepWAS concentrates more on biological outcomes (i.e. regulatory mechanisms in GWAS), this thesis focuses on testing DL SAE for epistatic interactions and classification analysis.

Using DL SAE for feature extraction provides a more effective approach than using direct features from statistical approaches such as logistic regression in association analysis for classification tasks. This automated feature extraction algorithm outperforms the traditional supervised classification models and offers a powerful way to enhance overall model performance while reducing the dimensional space and managing overfitting in GWAS data analysis.

6.2 Traditional Machine Learning with Statistically Significant SNPs and Clinical Data

Genetic variables obtained from logistic regression association analysis, mainly SNP variables, and clinical/sociodemographic variables are also considered to investigate and evaluate the predictive capacity of five traditional machine learning models when distinguishing between case and control T2D observations. Three experiments are performed which includes genomic data only, clinical data only, and genetic and clinical data combined.

In the first experiment, six of the most statistically significant SNPs extracted from logistic association analysis are utilized as inputs to model the traditional machine learning algorithms. In general, as shown in Table 5.12, the classification accuracy for all five machine learning models is low with 57.09% for the SVM and NNET classifiers and 57.46% for the GLM. The low accuracy values indicate that genomic data, particularly SNPs, fail to classify case and control observations. This is likely to be caused by the fact that statistically significant SNPs are often false positives. Consequently, highly significant association SNPs demonstrate little predictive power (Mittag et al. 2012). This can be explained due to limited heritability (Dudbridge 2013), which means how much of the phenotypic variance (combines the genotype variance with the environmental variance) is due to genetic variance (Moore et al. 2010).

A much higher predictive accuracy is obtained using clinical variables only. Among the selected models, the GBM achieved the best accuracy 71.06% with 64.63% for sensitivity

and 76.52% for specificity. Moreover, the predictive accuracy when employing both genomic and clinical data as input features showed satisfactory results with the NNET classifier achieving the best result 72.99%. Comparatively, the GLM, GBM, SVM, RPART yielded better results than when using clinical or genomic data separately. The results suggest that the improvement of the accuracy for all classifiers is entirely due to clinical variables, with no predictive value emerging from genotype variables alone. This is confirmed through the use of variable importance as illustrated in Figure 5.16. Although the variables for each model showed the disparity in relation to their rank measurement, variable importance in the tested models showed that clinical data, specifically BMI, was the most important compared to other features. Although the predictive power is mainly due to the clinical variables, combining genetic and clinical information showed that the GBM classification accuracy values improved dramatically from 57.18% for genetic variables to 71.06% and 72.12% for clinical variables and the joint effects of genetic and clinical variables respectively.

Additional evaluations using the features selected by RFE were also considered. For genetic data only with six features, three features were ranked important and used to fit the models. The results showed that the GLM, GBM, and RPART classifiers performed slightly worse in terms of accuracy which was not the case for the SVM and NNET. For clinical data only with nine features, eight features were considered important and utilized to fit the models. The evaluation showed that the accuracy of all classifiers deteriorates slightly using the reduced feature set. For the joint effects of genetic and clinical data, the RFE algorithm eliminates nine from the original 15 features that are considered to be unimportant or have no influence on the model performance. The six features were employed to fit the models and the results demonstrate that there is no improvement in the performance capacity of the five models in comparison to previous results employing the full feature set. More importantly, this experiment illustrates that using the RFE algorithm to select the most important features among six genetic and nine clinical features yielded a single genetic feature and five clinical

features. The clinical data has significantly higher discriminatory capacity than only using the six statistically significant SNPs. Again, this proves that statistically significant SNPs extracted from GWAS analysis is not important for the classification of disease outcomes; given that of the six statistically significant SNPs only one was selected.

Although the classification performance of clinical data in this thesis gives higher predictive results than using genetic data (six statistically significant SNPs), the advantage of clinical information is limited as it is only useful when individuals in the study have already developed the disease (in this case T2D). The prediction of disease risk based on highly significant SNPs demonstrates little predictive power while increasing the number of SNPs gives much higher performance in comparison to clinical data as showed in the results generated in this thesis.

6.3 Summary

This chapter discussed the results obtained from several experiments conducted in this thesis. Binary classification of T2D high-dimensional genetic data using advanced machine learning algorithms was discussed and compared against simpler and less computationally expensive machine learning algorithms. In addition, the evaluation of genetic and clinical data individually and combined was also discussed. This thesis presents a novel framework for the classification of T2D case-control GWAS data. The combination of unsupervised learning, using a DL SAE to extract latent representation from large scale biological data structures, and the use of these subsets of SNPs to initialise a multilayer feedforward softmax classifier for the classification tasks, form the fundamental components in our proposed framework. Although unsupervised deep learning stacked autoencoders are widely used to learn the compressed representation of the data input in many domains (Vařeka & Mautner 2017; Deng et al. 2017). This thesis claims that this is the first study of its kind that introduces the idea of using unsupervised learning with a deep learning algorithm based on

stacked autoencoder to extract the epistatic interaction between SNPs in GWAS for the classification of T2D.

The proposed framework could provide a starting point for researchers and professionals investigating the aetiology of T2D that has the potential to help better understand the missing heritability that the traditional statistical approaches fail to explain. This could lead to an improvement in diagnostic testing for early intervention to minimise the risk of disease onset and may support the future direction of personalized medicine.

Chapter 7 Conclusion and Future Work

7.1 Conclusion

This thesis presented a novel framework based on unsupervised machine learning using deep learning stacked autoencoders to extract complex interactions between SNPs to model a fully connected MLP to classify between cases and controls in T2D GWAS data. The fact that T2D is a polygenic condition means; it is no longer possible to consider a single SNP or gene to investigate the aetiology of such a complex disorder. Considering the interaction of a SNP including SNP-SNP and SNP-environment interactions is increasingly important particularly in complex human diseases like T2D. When considering large-scale GWAS data, investigating the interactions between SNPs formed by epistasis, is known to be complicated. The research conducted in this study focused specifically on finding a computationally and statistically efficient way to identify the epistatic interactions that exist between SNPs.

This study utilized the NHS and HPFS cohorts in T2D provided by the Genotypes and Phenotypes (dbGap) database. Various GWAS tools and techniques were considered to perform stringent quality control assessment steps followed by logistic regression and single-SNP association analysis. For high-dimensional T2D GWAS data, deep learning stacked autoencoders were employed to extract latent information and reduce the features space. In particular, to learn the non-linear epistatic interactions that exist between SNPs. These features were then applied to a fully connected MLP to initialise the weights before it was fine-tuned for binary classification tasks.

The findings using the proposed methodology demonstrate promising and encouraging results. Reducing the feature space from 6609 SNPs to a smaller number of neurons in each layer (2500, 1500, 700, and 300), showed that it was possible to obtain (AUC=94.25%, Sens=87.14%, Spec=89.54%, Logloss= 34.05%, Gini=88.51%, MSE=9.48%) using 2500 hidden layer units. The classification accuracy value of 300 compressed neurons remains

satisfactory (AUC=80.78%, Sens=87.85%, Spec=53.06%, Logloss= 55.11%, Gini=61.57%, MSE=18.43%). This provides a very efficient way to convert high-dimensional GWAS data into low-dimensional data, while maintaining good overall model performance. The greedy layer-wise machine learning solution performed using stacked autoencoders is based on training the network layer-by-layer using unlabelled data. This method allows us to extract latent representation of SNPs in each layer that are formed from the non-linear interaction between them. The produced features represent the reduced compressed features from the original data and only contain information deemed important to the classification tasks in hand. Therefore, we believe that this approach will enhance the quality of further biomedical experiment investigations.

Despite the suitability of deep learning and its potential application to biological data, the adoption of deep learning in biomedical research has been slow. Deep learning is still in its infancy. However, the success presented in this thesis will contribute to the bioinformatics and computational biology research fields as well as other non-genetic domains that comprise complex, large-scale data. Furthermore, we believe that this work provides new insights into the potential use of unsupervised deep learning as an automated feature extraction tool for use in supervised machine learning systems.

7.2 Future Research Directions

Despite the encouraging results, there are still many areas for improvement to further enhance classification results. These are discussed below.

7.2.1 Remove GWAS Stage

The emergence of GWAS has undoubtedly helped to better understand genetics. However, single-locus GWAS is best placed to test an individual SNP independently from SNPs and associations with traits in case-control dataset. In other words, GWAS focuses on the identification of SNPs with main effects. Genetic risk alleles identified by GWAS only answer between 10 and 20 percent of the missing heritability in T2D (Billings & Florez 2010;

Prasad & Groop 2015). Using statistically significant SNPs limits the number of SNPs used in the analysis. Consequently, it would be interesting if this step could be skipped, allowing for more SNPs to be included in the analysis of T2D. Investigating all SNPs will explain a much larger proportion of missing heritability and may support the discovery of unidentified loci with smaller or no main effects. Furthermore, this would allow us to reveal more combinations of SNPs and assess their influence on disease risk. This will present a number of challenges, the most difficult being the significant increase in memory and computer resources required as more increasingly complex feature combinations are considered.

7.2.2 Filter by Biological Plausibility

Another interesting area of research would be to select SNPs based on the strength of the independent main effects. SNPs selection can be established using biological plausibility (prior biological knowledge). The potential advantage of this approach would be to only consider SNPs that interact biologically through their biochemical reactions, functions, pathways, and networks, as some of these SNPs may be missed when using simpler statistical filters. A number of approaches that integrate expert biological knowledge into epistatic interaction analysis to identify important SNPs have been proposed. Bush *et al.* (Bush et al. 2009) developed the Biofilter algorithm to reduce the search space for assessing specific combinations of SNPs based on prior statistical and biological knowledge. The Biofilter uses biological information about SNP-SNP relationships and disease-related SNPS to produce multi-SNP models prior to any statistical analysis being conducted. This knowledge-driven approach offers a way to reduce both the computational and statistical burden inherent in testing for SNP-SNP interactions analysis while simultaneously providing biological information for corresponding statistically significant results that are identified. The utility of this approach has been proven in a number of studies and would warrant further study (Kim et al. 2016; Hohman et al. 2016; S. D. Turner et al. 2011; Pendergrass et al. 2015).

7.2.3 Interpretation of Deep Learning Models

One of the major limitations of deep learning models, and indeed the approach in this thesis, is the inability to interpret the model outcomes. While deep learning is a powerful mechanism for data representation, it is difficult to interpret the results. In a biological context, uncovering complex causal and structural relationships is important if we are to provide better biological insights about complex diseases. From the biomedical point of view, obtaining good results simply is not enough. It is important to provide logical reasoning about genetic components. Several approaches have been developed to remedy this pitfall; however, it is a relatively new research direction. With regards to microarray, DNA and RNA binding sequence input, Alipanahi *et al.* (Alipanahi et al. 2015) proposed the DeepBind tool, a visualisation method and mutation map that illustrates the effect of genetic markers on binding scores that are experimentally detected using deep convolutional neural networks. This work however is still in the early stages of development but would be worth considering in further studies.

Within other application domains, for example in image classification, Zeiler and Fergus (Zeiler & Fergus 2014) proposed a visualization technique through a deconvolution network to reconstruct and visualize intermediate feature layers by mapping these intermediate layers back to the input space. Other research utilized gradient optimization through backpropagation to visualize the response of deep hidden unit architectures in the input space (Mahendran & Vedaldi 2015). Again, all these ideas would be worth exploring in future work.

In our own study, the 300 deep features obtained from the 6609 input SNPs features demonstrate reasonable predictive results. However, it is difficult to know how such results are derived internally. In other words, which of the 6609 SNPs contribute to those 300 features? Consequently, it would be beneficial to integrate an interpretable approach such as Interpretable Decision Boundaries with deep learning stacked autoencoder to facilitate the

transparency and the interpretability of the model. Interpretable Decision Boundaries (Wu et al. 2018) is based on linking the numerical values in a prediction made by the model to the training data points associated with the prediction. These training data points are organized using an Explicable Boundary Tree (EB-tree), which is based on the distances in the deep learning transformed space. The data structure of the EB-tree represents deep learning decision boundaries and the data points in the tree are efficiently able to approximate the predictions of the model via these points. Again, interesting research directions to consider in future work.

7.2.4 Computational and Hyperparameter Optimization of Deep Learning

The training process in deep learning and stacked autoencoders in particular is usually computationally intensive and time-consuming. Parallelizing the training of deep learning can dramatically increase speed and improve efficiency in addition to adding feasibility. The extension into parallel implementations requires access to graphical processing units (GPUs). A recent open sourced framework for high performance computational programming is TensorFlow™ (www.tensorflow.org). It was originally developed by Google and provides huge support for Artificial Intelligence development. In fact, it is being broadly used to implement deep learning and develop solutions with deep learning architectures (Ramsundar & Zadeh 2018).

As a future direction for our work, it would be advantageous to implement deep learning stacked autoencoders on a highly parallel processing framework such as the aforementioned TensorFlow. This would allow us to increase the speed of the model's implementation and optimization of the hyperparameters using much larger SNP combinations.

References

Abdulaimma, B. et al., 2017. Association Mapping Approach into Type 2 Diabetes Using Biomarkers and Clinical Data. In D.-S. Huang, K.-H. Jo, & J. C. Figueroa-García, eds. *Lecture Notes in Computer Science, Springer, Cham*. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 325–336. Available at: http://link.springer.com/10.1007/978-3-319-63312-1_29.

Abdulaimma, B. et al., 2018. Improving Type 2 Diabetes Phenotypic Classification by Combining Genetics and Conventional Risk Factors. In *2018 IEEE Congress on Evolutionary Computation (CEC)*. Rio de Janeiro, Brazil: IEEE. Available at: https://ieeexplore.ieee.org/document/8477647/.

Abdurakhmonov, I., 2016. Bioinformatics: Basics, Development, and Future. In I. Abdurakhmonov, ed. *Bioinformatics - Updated Features and Applications*. IntechOpen. Available at: https://www.intechopen.com/books/bioinformatics-updated-features-and-applications.

Adam, A., 2015. A global reference for human genetic variation. *Nature*, 526, pp.68–74. Available at: http://www.nature.com/nature/journal/v526/n7571/fig_tab/nature15393_SF1.html%5Cnhttp://dx.doi.org/10.1038/nature15393%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26432245.

Ahlers, K.E., Chakravarti, B. & Fisher, R.A., 2016. RGS6 as a Novel Therapeutic Target in CNS Diseases and Cancer. *The AAPS Journal*, 18(3), pp.560–572. Available at: http://link.springer.com/10.1208/s12248-016-9899-9.

Aiello, S. et al., 2018. Machine Learning with R and H2O. Available at: http://h2o.ai/resources/.

Aizerman, M., Braverman, E. & Rozonoer, L., 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25, pp.821–837.

Al-Sinani, S. et al., 2014. Familial clustering of type 2 diabetes among Omanis. *Oman Medical Journal*, 29(1), pp.51–54. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3910414/pdf/OMJ-D-13-00245.pdf.

Alberts, B. et al., 2014. *Essential Cell Biology* fourth., United States of America: Garland

Science, Taylor & Francis Group.

Alberts, B. et al., 2015. *Molecular Biology of The Cell* Sixth., United States of America: Garland Science, Taylor & Francis Group.

Ali, O., 2013. Genetics of type 2 diabetes. *World Journal of Diabetes*, 4(4), pp.114–123. Available at: http://www.wjgnet.com/1948-9358/full/v4/i4/114.htm.

Alipanahi, B. et al., 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8), pp.831–838. Available at: http://www.nature.com/articles/nbt.3300.

American Diabetes Association, 2018. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes. *Diabetes Care*, 41(Supplement 1), pp.S13–S27. Available at: http://care.diabetesjournals.org/lookup/doi/10.2337/dc18-S002.

Amos, W., Driscoll, E. & Hoffman, J.I., 2011. Candidate genes versus genome-wide associations: which are better for detecting genetic susceptibility to infectious disease? *Proceedings of the Royal Society B: Biological Sciences*, 278, pp.1183–1188. Available at: http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2010.1920.

Anderson, C. a et al., 2010. Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9), pp.1564–1573. Available at: http://www.nature.com/articles/nprot.2010.116.

Andrew, A.S. et al., 2008. DNA Repair Polymorphisms Modify Bladder Cancer Risk: A Multi-factor Analytic Strategy. *Human Heredity*, 65(2), pp.105–118. Available at: https://www.karger.com/Article/FullText/108942.

Anthony, M. & Bartlett, P.L., 1999. *Neural Networks: Theoretical Foundations*, Cambridge: Cambridge University Press.

Arloth, J. et al., 2016. DeepWAS: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. *bioRxiv*. Available at: https://www.biorxiv.org/content/early/2018/12/05/069096.

Balding, D.J., 2006. A tutorial on statistical methods for population association studies. *Nature reviews. Genetics*, 7(10), pp.781–91. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16983374.

Ban, H.-J. et al., 2010. Identification of Type 2 Diabetes-associated combination of SNPs

using Support Vector Machine. *BMC Genetics*, 11(26). Available at: http://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-11-26.

Barna, B. et al., 2018. A multifactor dimensionality reduction model of gene polymorphisms and an environmental interaction analysis in type 2 diabetes mellitus study among Punjabi, a North India population. *Meta Gene*, 16, pp.39–49. Available at: https://linkinghub.elsevier.com/retrieve/pii/S2214540018300100.

Bartlett, A., Penders, B. & Lewis, J., 2017. Bioinformatics : indispensable , yet hidden in plain sight ? *BMC Bioinformatics*, 18(311). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5480157/.

Bateson, W., 1909. *Mendel's principle of Heredity*, New York: Cambridge University Press.

Behjati, S. & Tarpey, P.S., 2013. What is next generation sequencing? *Archives of Disease in Childhood. Education and Practice Edition*, 98(6), pp.236–238. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3841808/.

Bengio, Y. et al., 2007. Greedy Layer-Wise Training of Deep Networks. In *Proceeding NIPS'06 Proceedings of the 19th International Conference on Neural Information Processing Systems*. Canada: MIT Press Cambridge, MA, USA, pp. 153–160. Available at: https://dl.acm.org/citation.cfm?id=2976476.

Bengio, Y., Courville, A. & Vincent, P., 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), pp.1798–1828. Available at: http://ieeexplore.ieee.org/document/6472238/.

Bergstra, J. & Yoshua, B., 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, pp.281–305. Available at: https://dl.acm.org/citation.cfm?id=2188395.

Berk, R.A., 2016. Classification and Regression Trees (CART). In *Statistical Learning from a Regression Perspective. Springer Texts in Statistics*. Springer, Cham, pp. 129–186. Available at: http://link.springer.com/10.1007/978-3-319-44048-4_3.

Billings, L.K. & Florez, J.C., 2010. The genetics of type 2 diabetes: what have we learned from GWAS? *Annals of the New York Academy of Sciences*, 1212(1), pp.59–77. Available at: http://doi.wiley.com/10.1111/j.1749-6632.2010.05838.x.

Blanco-Gómez, A. et al., 2016. Missing heritability of complex diseases: Enlightenment by genetic variants from intermediate phenotypes. *BioEssays*, 38(7), pp.664–673.

Available at: http://doi.wiley.com/10.1002/bies.201600084.

Bland, M., 2015. *An Introduction to medical statistics* 4th ed., Oxford University Press.

Bohman, A. et al., 2017. A family-based genome-wide association study of chronic rhinosinusitis with nasal polyps implicates several genes in the disease pathogenesis L. Zhang, ed. *PLOS ONE*, 12(12), p.e0185244. Available at: https://dx.plos.org/10.1371/journal.pone.0185244.

Botta, V. et al., 2014. Exploiting SNP Correlations within Random Forest for Genome-Wide Association Studies L. Chen, ed. *PLoS ONE*, 9(4), p.e93379. Available at: https://dx.plos.org/10.1371/journal.pone.0093379.

Breiman, L., 1996. Bagging predictors. *Machine Learning*, 26, pp.123–140. Available at: http://link.springer.com/10.1007/BF00058655.

Breiman, L., 1984. *Classification and Regression Trees*, Wadsworth International Group.

Breiman, L., 2001. Random Forests. *Machine Learning*, 45, pp.5–32. Available at: https://link.springer.com/article/10.1023/A:1010933404324.

Browning, S.R. & Browning, B.L., 2012. Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*, 46(1), pp.617–633. Available at: http://www.annualreviews.org/doi/10.1146/annurev-genet-110711-155534.

Buntine, W. & Niblett, T., 1992. A Further Comparison of Splitting Rules for Decision-Tree Induction. *Machine Learning*, 85(1), pp.75–85. Available at: https://link.springer.com/article/10.1023/A:1022686419106.

Burton, P.R. et al., 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), pp.661–678. Available at: http://www.nature.com/doifinder/10.1038/nature05911.

Bush, W. s & Moore, J.H., 2012. Chapter 11: Genome-Wide Association Studies F. Lewitter & M. Kann, eds. *PLoS Computational Biology*, 8(12), p.e1002822. Available at: https://dx.plos.org/10.1371/journal.pcbi.1002822.

Bush, W.S., Dudek, S.M. & Ritchie, M.D., 2009. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp.368–79. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19209715.

Bush, W.S., Dudek, S.M. & Ritchie, M.D., 2006. Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. *Bioinformatics*, 22(17), pp.2173–2174. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl347.

Calle, M.L. et al., 2008. Improving strategies for detecting genetic patterns of disease susceptibility in association studies. *Statistics in Medicine*, 27(30), pp.6532–6546. Available at: https://www.ncbi.nlm.nih.gov/pubmed/18837071.

Can, T., 2014. Introduction to Bioinformatics. In Springer, pp. 51–71. Available at: http://link.springer.com/10.1007/978-1-62703-748-8_4.

Candel, A. et al., 2018. Deep Learning With H2O. In A. Bartz, ed. United States of America: H2O.ai, Inc. Available at: http://h2o.ai/resources.

Cardon, L.R. & Palmer, L.J., 2003. Population stratification and spurious allelic association. *The Lancet*, 361(9357), pp.598–604. Available at: https://www.ncbi.nlm.nih.gov/pubmed/12598158.

Cattaert, T. et al., 2011. Model-Based Multifactor Dimensionality Reduction for detecting epistasis in case-control data in the presence of noise. *Annals of Human Genetics*, 75(1), pp.78–89. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059142/.

Cauchi, S. et al., 2006. Transcription factor TCF7L2 genetic study in the French population: expression in human beta-cells and adipose tissue and strong association with type 2 diabetes. *Diabetes*, 55(10), pp.2903–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17003360.

Chandrashekar, G. & Sahin, F., 2014. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), pp.16–28. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0045790613003066.

Chatelain, C. et al., 2018. Performance of epistasis detection methods in semi-simulated GWAS. *BMC Bioinformatics*, 19(231). Available at: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2229-8.

Cheema, A.K. et al., 2015. Genetic Associations of PPARGC1A with Type 2 Diabetes: Differences among Populations with African Origins. *Journal of diabetes research*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25977930.

Chen, S.-H. et al., 2008. A support vector machine approach for detecting gene-gene

interaction. *Genetic Epidemiology*, 32(2), pp.152–167. Available at: https://www.ncbi.nlm.nih.gov/pubmed/17968988.

Chen, X. & Ishwaran, H., 2012. Random forests for genomic data analysis. *Genomics*, 99(6), pp.323–329. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22546560.

Chen, Z. et al., 2014. Joint Effects of Known Type 2 Diabetes Susceptibility Loci in Genome-Wide Association Study of Singapore Chinese: The Singapore Chinese Health Study Q. Huang, ed. *PLoS ONE*, 9(2), p.e87762. Available at: https://www.ncbi.nlm.nih.gov/pubmed/24520337.

Chen, Z., Huang, H. & Ng, H.K.T., 2014. An improved robust association test for GWAS with multiple diseases. *Statistics and Probability Letters*, 91, pp.153–161.

Chuan-zhen, J. et al., 2008. Sequence Analysis in Vicinity of Type 2 Diabetes Related SNP rs7903146. In *2008 IEEE International Conference on Bioinformatics and Biomedical Engineering*. Shanghai, China, pp. 50–53. Available at: https://ieeexplore.ieee.org/document/4534899.

Cirulli, E.T. & Goldstein, D.B., 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reniews GENETICS*, 11, pp.415–425. Available at: https://www.nature.com/articles/nrg2779.

Clarke, G.M. et al., 2011. Basic statistical analysis in genetic case-control studies. *Nature Protocols*, 6(2), pp.121–133. Available at: http://www.nature.com/articles/nprot.2010.182.

Collobert, R. et al., 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12, pp.2493–2537. Available at: https://dl.acm.org/citation.cfm?id=2078186.

Cooper, A.J. et al., 2012. A prospective study of the association between quantity and variety of fruit and vegetable intake and incident type 2 diabetes. *Diabetes Care*, 35(6), pp.1293–1300. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22474042.

Cordell, H.J., 2009. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6), pp.392–404. Available at: http://www.nature.com/articles/nrg2579.

Cordell, H.J., 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20), pp.2463–2468. Available

at: https://www.ncbi.nlm.nih.gov/pubmed/12351582.

Cortes, A., Medland, S.E. & Renterı, M.E., 2013. Using PLINK for Genome-Wide Association Studies (GWAS) and Data Analysis. In C. Gondro, J. van der Werf, & B. Hayes, eds. *Genome-Wide Association Studies and Genomic Prediction*. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 193–213. Available at: http://link.springer.com/10.1007/978-1-62703-447-0.

Cortes, C. & Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20(3), pp.273–297. Available at: http://link.springer.com/10.1007/BF00994018.

Cox, D.R., 1958. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), pp.215–232. Available at: http://doi.wiley.com/10.1111/j.2517-6161.1958.tb00292.x.

Culverhouse, R., 2010. The Restricted Partition Method. In *Advances in Genetics*. Elsevier Inc., pp. 117–139. Available at: https://www.ncbi.nlm.nih.gov/pubmed/21029851.

Cutting, G.R., 2015. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nature Reviews Genetics*, 16(1), pp.45–56. Available at: http://www.nature.com/articles/nrg3849.

De, R. et al., 2015. Characterizing gene-gene interactions in a statistical epistasis network of twelve candidate genes for obesity. *BioData Mining*, 8(45), pp.1–16. Available at: http://biodatamining.biomedcentral.com/articles/10.1186/s13040-015-0077-x.

De, R. et al., 2015. Identifying gene-gene interactions that are highly associated with Body Mass Index using Quantitative Multifactor Dimensionality Reduction (QMDR). *BioData Mining*, 8(41). Available at: http://biodatamining.biomedcentral.com/articles/10.1186/s13040-015-0074-0.

Deng, L., Fan, C. & Zeng, Z., 2017. A sparse autoencoder-based deep neural network for protein solvent accessibility and contact number prediction. *BMC Bioinformatics*, 18(569). Available at: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1971-7.

Ding, B., 2014. High-Throughput Genetic Interaction Study. In *Between the Lines of Genetic Code*. Elsevier, pp. 55–80. Available at: https://linkinghub.elsevier.com/retrieve/pii/B9780123970176000040.

Donaldson, P. et al., 2016. *Genetics of Complex Disease*, New York: Garland Science,

Taylor & Francis Group.

Dong, Z. et al., 2015. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC Cancer*, 15(489). Available at: http://bmccancer.biomedcentral.com/articles/10.1186/s12885-015-1492-6.

Dudbridge, F., 2013. Power and Predictive Accuracy of Polygenic Risk Scores N. R. Wray, ed. *PLoS Genetics*, 9(3), p.e1003348. Available at: https://dx.plos.org/10.1371/journal.pgen.1003348.

Dudbridge, F. & Gusnanto, A., 2008. Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, 32(3), pp.227–234. Available at: https://www.ncbi.nlm.nih.gov/pubmed/18300295.

Duggal, P. et al., 2008. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics*, 9(516). Available at: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-9-516.

Duggirala, R. et al., 2015. *Genome Mapping and Genomics in Human and Non-Human Primates* 1st ed. R. Duggirala et al., eds., Berlin, Heidelberg: Springer Berlin Heidelberg. Available at: http://link.springer.com/10.1007/978-3-662-46306-2.

Durbin, R.M. et al., 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), pp.1061–1073. Available at: http://www.nature.com/doifinder/10.1038/nature09534.

Ekins, S., 2016. The Next Era: Deep Learning in Pharmaceutical Research. *Pharmaceutical Research*, 33(11), pp.2594–2603. Available at: https://www.ncbi.nlm.nih.gov/pubmed/27599991.

Erhan, D. et al., 2009. The Difficulty of Training Deep Architectures and the Effect of Unsupervised Pre-Training. In *Proceedings of the 12th International Confe-rence on Artificial Intelligence and Statistics (AISTATS)*. Florida, USA, pp. 153–160.

Erhan, D. et al., 2010. Why Does Unsupervised Pre-training Help Deep Learning ? *The Journal of Machine Learning Research*, 11, pp.625–660.

Esposito, F. et al., 1997. A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5), pp.476–493. Available at: http://ieeexplore.ieee.org/document/589207/.

Evans, D.M., 2011. Gene-Gene Interaction and Epistasis. In E. Zegginin & A. Morris, eds. *Analysis of Complex Disease Association Studies*. Elsevier, pp. 197–213. Available at: http://linkinghub.elsevier.com/retrieve/pii/B9780123751423100124.

Fadista, J. et al., 2016. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, 24(8), pp.1202–1205. Available at: http://www.nature.com/doifinder/10.1038/ejhg.2015.269.

Fang, Y.-H., Wang, J.-H. & Hsiung, C.A., 2017. TSGSIS: a high-dimensional grouped variable selection approach for detection of whole-genome SNP–SNP interactions O. Stegle, ed. *Bioinformatics*, 33(22), pp.3595–3602. Available at: https://academic.oup.com/bioinformatics/article/33/22/3595/3884655.

Farsani, F.S. et al., 2013. Global trends in the incidence and prevalence of type 2 diabetes in children and adolescents: a systematic review and evaluation of methodological approaches. *Diabetologia*, 56(7), pp.1471–1488. Available at: https://www.ncbi.nlm.nih.gov/pubmed/23677041.

Faye, L.L. & Bull, S.B., 2011. Two-stage study designs combining genome-wide association studies, tag single-nucleotide polymorphisms, and exome sequencing: accuracy of genetic effect estimates. *BMC Proceedings*, 5(S9). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3287903/.

Fergus, P. et al., 2018. Utilising Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Available at: https://ieeexplore.ieee.org/document/8454302/.

Ferri, C., Hernández-Orallo, J. & Modroiu, R., 2009. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1), pp.27–38. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0167865508002687.

Fisher, R.A., 1918. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2), pp.399–433. Available at: http://www.journals.cambridge.org/abstract_S0080456800012163.

Florkowski, C.M., 2008. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical biochemist. Reviews*, 29 Suppl 1, pp.S83-7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18852864.

Foulkes, A.S., 2009. *Applied Statistical Genetics with R*, New York, NY: Springer New York. Available at: http://link.springer.com/10.1007/978-0-387-89554-3.

Fox, J., 2008. Generalized Linear Models. In *Applied Regression Analysis and Generalized Linear Models*. United States of America: SAGE Puplications, pp. 379–424.

Friedman, J.H., 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), pp.367–378. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167947301000652.

Fukushima, K., 1980. Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics*, 36(4), pp.193–202. Available at: https://link.springer.com/article/10.1007/BF00344251.

Gauthier, J. et al., 2018. A brief history of bioinformatics. *Briefings in Bioinformatics*. Available at: https://www.ncbi.nlm.nih.gov/pubmed/30084940.

Gibbs, R.A. et al., 2003. The International HapMap Project. *Nature*, 426(6968), pp.789–796. Available at: http://www.nature.com/articles/nature02168.

Gilbert-Diamond, D. & Moore, J.H., 2011. Analysis of Gene-Gene Interactions. *Current Protocols in Human Genetics*, 32(3), pp.506–509. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4086055/.

Gloyn, A.L., Braun, M. & Rorsman, P., 2009. Type 2 Diabetes Susceptibility Gene TCF7L2 and Its Role in B-Cell Function. *Diabetes*, 58(4), pp.800–802. Available at: http://diabetes.diabetesjournals.org/cgi/doi/10.2337/db09-0099.

Graffelman, J. & Weir, B.S., 2016. Testing for Hardy–Weinberg equilibrium at biallelic genetic markers on the X chromosome. *Heredity*, 116(6), pp.558–568. Available at: http://www.nature.com/articles/hdy201620.

Graves, A., Mohamed, A. & Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, BC, Canada: IEEE, pp. 6645–6649. Available at: http://ieeexplore.ieee.org/document/6638947/.

Green, E.D., Watson, J.D. & Collins, F.S., 2015. Human Genome Project: Twenty-five years of big biology. *Nature*, 526, pp.29–31. Available at: http://www.nature.com/doifinder/10.1038/526029a.

Greene, C.S. et al., 2010. Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. *Bioinformatics*, 26(5), pp.694–695. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq009.

Greene, C.S. et al., 2009. Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Mining*, 2(5). Available at: http://biodatamining.biomedcentral.com/articles/10.1186/1756-0381-2-5.

GÜL, H., AYDIN SON, Y. & AÇIKEL, C., 2014. Discovering missing heritability and early risk prediction for type 2 diabetes: a new perspective for genome-wide association study analysis with the Nurses' Health Study and the Health Professionals' Follow-Up Study. *Turkish Journal of Medical Sciences*, 44(6), pp.946–954. Available at: http://journals.tubitak.gov.tr/medical/issues/sag-14-44-6/sag-44-6-7-1310-77.pdf.

Guo, X. et al., 2014. Genome-Wide Interaction-Based Association of human diseases - A survey. *Tsinghua Science and Technology*, 19(6), pp.596–616. Available at: http://ieeexplore.ieee.org/document/6961029/.

Hand, D.J., 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), pp.103–123. Available at: http://link.springer.com/10.1007/s10994-009-5119-5.

Hanis, C.L. et al., 1996. A genome–wide search for human non–insulin–dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nature Genetics*, 13(2), pp.161–166. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8640221.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning* 2nd ed., New York, NY: Springer New York. Available at: http://www.springerlink.com/index/10.1007/b94608.

Haykin, S., 1994. *Neural Networks A Comprehensive Foundation* J. Griffin, ed., United States of America: Prentice-Hall.

Heaton, J., 2008. *Introduction to Neural Networks with Java* 2 edition., Heaton Research, Inc.

Heller, R. & Yekutieli, D., 2014. Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics*, 8(1), pp.481–498. Available at: http://projecteuclid.org/euclid.aoas/1396966295.

Hemani, G. et al., 2011. EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics*, 27(11), pp.1462–1465. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr172.

Herman, W.H. et al., 2015. Early Detection and Treatment of Type 2 Diabetes Reduce Cardiovascular Morbidity and Mortality: A Simulation of the Results of the Anglo-Danish-Dutch Study of Intensive Treatment in People With Screen-Detected Diabetes in Primary Care (ADDITION-Europe). *Diabetes Care*, 38(8), pp.1449–1455. Available at: http://care.diabetesjournals.org/lookup/doi/10.2337/dc14-2459.

Herold, C. et al., 2009. INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics*, 25(24), pp.3275–3281. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp596.

Hex, N. et al., 2012. Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabetic Medicine*, 29(7), pp.855–862. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22537247.

Hind, J. et al., 2017. A robust method for the interpretation of genomic data. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 3385–3390. Available at: http://ieeexplore.ieee.org/document/7966281/.

Hinton, G. et al., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6), pp.82–97. Available at: http://ieeexplore.ieee.org/document/6296526/.

Hinton, G.E., Osindero, S. & Teh, Y.-W., 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7), pp.1527–1554. Available at: https://dl.acm.org/citation.cfm?id=1161605.

Hinton, G.E. & Salakhutdinov, R.R., 2006. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786), pp.504–507. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16873662.

Hohman, T.J. et al., 2016. Discovery of gene-gene interactions across multiple independent data sets of late onset Alzheimer disease from the Alzheimer Disease Genetics

Consortium. *Neurobiology of Aging*, 38, pp.141–150. Available at: https://www.ncbi.nlm.nih.gov/pubmed/26827652.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), pp.417–441. Available at: https://psycnet.apa.org/record/1934-00645-001.

Howie, B. et al., 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44(8), pp.955–959. Available at: http://www.nature.com/articles/ng.2354.

Ibrahim, A.T. et al., 2016. Candidate gene analysis supports a role for polymorphisms at TCF7L2 as risk factors for type 2 diabetes in Sudan. *Journal of Diabetes & Metabolic Disorders*, 15(4). Available at: https://www.ncbi.nlm.nih.gov/pubmed/26937418.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–945. Available at: http://www.nature.com/doifinder/10.1038/nature03001.

Inzucchi, S.E. et al., 2012. Management of Hyperglycemia in Type 2 Diabetes: A Patient-Centered Approach: Position Statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care*, 35(6), pp.1364–1379. Available at: http://care.diabetesjournals.org/cgi/doi/10.2337/dc12-0413.

Jung, Y., 2018. Multiple predicting K-fold cross-validation for model selection. *Journal of Nonparametric Statistics*, 30(1), pp.197–215. Available at: https://www.tandfonline.com/doi/full/10.1080/10485252.2017.1404598.

Kawaguchi, A., 2012. Variable Ranking by Random Forests Model for Genome-Wide Association Study. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Hong Kong.

Kearns, M., 1988. Thoughts on Hypothesis Boosting. *Machine Learning class project*.

Kim, D. et al., 2016. BIOFILTER AS A FUNCTIONAL ANNOTATION PIPELINE FOR COMMON AND RARE COPY NUMBER BURDEN. *Pacific Symposium on Biocomputing.*, 21, pp.357–68. Available at: https://www.ncbi.nlm.nih.gov/pubmed/26776200.

Kim, J. et al., 2018. Genetic prediction of type 2 diabetes using deep neural network. *Clinical*

Genetics, 93(4), pp.822–829. Available at: https://www.ncbi.nlm.nih.gov/pubmed/29136281.

Kim, Y. et al., 2009. Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. *BMC Proceedings*, 3(Suppl 7). Available at: http://www.ncbi.nlm.nih.gov/pubmed/20018058%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2795965.

Köbberling, J. & Tillil, H., 1982. Empirical risk figures for first-degree relatives of non-insulin dependent diabetics. *The Genetics of Diabetes Mellitus*, pp.201–209.

Kohavi, R., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *IIJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Canada, pp. 1137–1143. Available at: https://dl.acm.org/citation.cfm?id=1643047.

Koo, C.L. et al., 2013. A Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology. *BioMed Research International*, 2013(432375). Available at: https://www.ncbi.nlm.nih.gov/pubmed/24228248.

Korkiakangas, E.E., Alahuhta, M.A. & Laitinen, J.H., 2009. Barriers to regular exercise among adults at high risk or diagnosed with type 2 diabetes: a systematic review. *Health Promotion International*, 24(4), pp.416–427. Available at: https://www.ncbi.nlm.nih.gov/pubmed/19793763.

Kotnik, P. et al., 2018. Identification of novel alleles associated with insulin resistance in childhood obesity using pooled-DNA genome-wide association study approach. *International Journal of Obesity*, 42, pp.686–695. Available at: http://www.nature.com/doifinder/10.1038/ijo.2017.293.

Krizhevsky, A., Sutskever, I. & Geoffrey E., H., 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Lake Tahoe, Nevada, pp. 1097–1105. Available at: https://dl.acm.org/citation.cfm?id=2999257.

Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software*, 28(5), pp.1–26. Available at: http://www.jstatsoft.org/v28/i05/paper.

Kuhn, M. & Johnson, K., 2013. *Applied Predictive Modeling*, Springer New York. Available at: http://link.springer.com/10.1007/978-1-4614-6849-3.

Kumar, R. & Indrayan, A., 2011. Receiver operating characteristic (ROC) curve for medical researchers. *Indian pediatrics*, 48(4), pp.277–87. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21532099.

Kursa, M.B., 2014. Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics*, 15(8). Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-8.

Laurie, C.C. et al., 2010. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, 34(6), pp.591–602. Available at: https://www.ncbi.nlm.nih.gov/pubmed/20718045.

Le, Q. V, 2015. A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks.

LeBlanc, M. et al., 2016. Identifying Novel Gene Variants in Coronary Artery Disease and Shared Genes With Several Cardiovascular Risk Factors. *Circulation Research*, 118(1), pp.83–94. Available at: https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.115.306629.

LeCun, Y. et al., 1998. Efficient BackProp. *Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science*, 1524, pp.9–50. Available at: http://link.springer.com/10.1007/978-3-642-35289-8_3.

LeCun, Y., Bengio, Y. & Hinton, G., 2015. Deep learning. *Nature*, 521, pp.436–444. Available at: http://www.nature.com/articles/nature14539.

Lee, S.H. et al., 2012. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*, 44(3), pp.247–250. Available at: http://www.nature.com/articles/ng.1108.

Leem, S. & Park, T., 2017. An empirical fuzzy multifactor dimensionality reduction method for detecting gene-gene interactions. *BMC Genomics*, 18(115). Available at: https://www.ncbi.nlm.nih.gov/pubmed/28361694.

Lewis, C.M. & Knight, J., 2012. Introduction to Genetic Association Studies. *Cold Spring Harbor Protocols*, 2012(3), pp.297–306. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22383645.

Lewis, J.P. et al., 2010. Analysis of candidate genes on chromosome 20q12-13.1 reveals evidence for BMI mediated association of PREX1 with type 2 diabetes in European Americans. *Genomics*, 96(4), pp.211–219. Available at: https://www.ncbi.nlm.nih.gov/pubmed/20650312.

Li, H. et al., 2013. A Genome-Wide Association Study Identifies GRK5 and RASGRP1 as Type 2 Diabetes Loci in Chinese Hans. *Diabetes*, 62(1), pp.291–298. Available at: http://diabetes.diabetesjournals.org/lookup/doi/10.2337/db12-0454.

Li, W., 2007. Three lectures on case control genetic association analysis. *Briefings in Bioinformatics*, 9(1), pp.1–13. Available at: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbm058.

López, B. et al., 2018. Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction. *Artificial Intelligence in Medicine*, 85, pp.43–49. Available at: https://www.ncbi.nlm.nih.gov/pubmed/28943335.

Lyssenko, V. et al., 2008. Clinical Risk Factors, DNA Variants, and the Development of Type 2 Diabetes. *New England Journal of Medicine*, 359(21), pp.2220–2232. Available at: https://www.ncbi.nlm.nih.gov/pubmed/19020324.

Mahajan, A., 2014. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics*, 46(3), pp.234–244. Available at: https://www.nature.com/articles/ng.2897 [Accessed March 12, 2016].

Mahendran, A. & Vedaldi, A., 2015. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Boston, MA, USA: IEEE, pp. 5188–5196. Available at: https://ieeexplore.ieee.org/document/7299155.

Maher, B., 2008. Personal genomes: The case of the missing heritability. *Nature*, 456(6), pp.18–21. Available at: http://www.nature.com/doifinder/10.1038/456018a.

Malovini, A. et al., 2012. Hierarchical Naive Bayes for genetic association studies. *BMC Bioinformatics*, 13(S14). Available at: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S14-S6.

Manolio, T.A. et al., 2009. Finding the missing heritability of complex diseases. *Nature*, 461(8), pp.747–53. Available at:

http://www.nature.com/doifinder/10.1038/nature08494.

Manolio, T.A. & Collins, F.S., 2009. The HapMap and Genome-Wide Association Studies in Diagnosis and Therapy. *Annual Review of Medicine*, 60(1), pp.443–456. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2717504/.

Martin, E.R. et al., 2006. A novel method to identify gene–gene effects in nuclear families: the MDR-PDT. *Genetic Epidemiology*, 30(2), pp.111–123. Available at: https://www.ncbi.nlm.nih.gov/pubmed/16374833.

Mathers, C.D. & Loncar, D., 2006. Projections of Global Mortality and Burden of Disease from 2002 to 2030 J. Samet, ed. *PLoS Medicine*, 3(11), p.e442. Available at: https://dx.plos.org/10.1371/journal.pmed.0030442.

McCaughan, J.A. et al., 2013. Comprehensive Investigation of the Caveolin 2 Gene: Resequencing and Association for Kidney Transplant Outcomes U. Sen, ed. *PLoS ONE*, 8(5), p.e63358. Available at: http://dx.plos.org/10.1371/journal.pone.0063358.

McKinney, B.A. et al., 2007. Evaporative cooling feature selection for genotypic data involving interactions. *Bioinformatics*, 23(16), pp.2113–2120. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm317.

Medici, F. et al., 1999. Concordance rate for Type II diabetes mellitus in monozygotic twins: actuarial analysis. *Diabetologia*, 42(2), pp.146–150. Available at: http://link.springer.com/10.1007/s001250051132.

Meigs, J.B., Cupples, L.A. & Wilson, P.W., 2000. Parental transmission of type 2 diabetes: the Framingham Offspring Study. *Diabetes*, 49(12), pp.2201–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11118026.

Mercer, J., 1909. Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 209(441–458), pp.415–446. Available at: http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.1909.0016.

Min, S., Lee, B. & Yoon, S., 2017. Deep learning in bioinformatics. *Briefings in Bioinformatics*, 18(5), pp.851–869. Available at: https://www.ncbi.nlm.nih.gov/pubmed/27473064.

Mitchell, K.J., 2012. What is complex about complex disorders? *Genome Biology*, 13(237).

Available at: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-1-237.

Mittag, F. et al., 2012. Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities. *Human Mutation*, 33(12), pp.1708–1718. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22777693.

Moisen, G., 2008. *Classification and Regression Trees*, Oxford, UK: Elsevier.

Montana, G., 2006. Statistical methods in genetics. *Briefings in Bioinformatics*, 7(3), pp.297–308. Available at: https://www.ncbi.nlm.nih.gov/pubmed/16963464.

Moore, J.H., 2007. Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics. In X. Zhu & I. Davidson, eds. *Knowledge discovery and data mining: challenges and realities*. Heyersh: IGI Global, pp. 17–30.

Moore, J.H. & Andrews, P.C., 2015. Epistasis Analysis Using Multifactor Dimensionality Reduction. In S. M. Williams, ed. *Epistasis: Methods and Protocols*. New York: Humana Press, pp. 301–314. Available at: http://link.springer.com/10.1007/978-1-4939-2155-3_16.

Moore, J.H., Asselbergs, F.W. & Williams, S.M., 2010. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4), pp.445–455. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2820680/.

Moore, J.H. & Williams, S.M., 2005. Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. *BioEssays*, 27(6), pp.637–646. Available at: https://www.ncbi.nlm.nih.gov/pubmed/15892116.

Morris, A., Voight, B. & Teslovich, T., 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9), pp.981–990. Available at: http://www.nature.com/articles/ng.2383.

Motsinger-Reif, A.A. et al., 2008. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genetic Epidemiology*, 32(4), pp.325–340. Available at: https://www.ncbi.nlm.nih.gov/pubmed/18265411.

Namkung, J. et al., 2009. Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method. *Genetic Epidemiology*,

33(7), pp.646–656. Available at: https://www.ncbi.nlm.nih.gov/pubmed/19241410.

Nandy, D. & Padariya, R., 2016. An overview of Pattern Recognition. *International Journal for Innovative Research in Science & Technology*, 2(9).

Natekin, A. & Knoll, A., 2013. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7(21). Available at: http://journal.frontiersin.org/article/10.3389/fnbot.2013.00021/abstract.

Nelder, J.A. & Wedderburn, R.W.M., 1972. Generalized Linear Models. *Journal of the Royal Statistical Society*, 135(3), pp.370–384.

Nelson, M.R., 2001. A Combinatorial Partitioning Method to Identify Multilocus Genotypic Partitions That Predict Quantitative Trait Variation. *Genome Research*, 11(3), pp.458–470. Available at: https://www.ncbi.nlm.nih.gov/pubmed/11230170.

Ng, A., 2011. Sparse autoencoder. *CS294A Lecture notes*, pp.1–19. Available at: http://www.stanford.edu/class/cs294a/sae/sparseAutoencoderNotes.pdf.

Nguyen, T.-T. et al., 2015. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics*, 16(Suppl 2). Available at: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-16-S2-S5.

Oh, S. et al., 2012. A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR. *BMC Bioinformatics*, 13(Suppl 9). Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S9-S5.

Okeh, U. & Okoro, C., 2012. Evaluating Measures of Indicators of Diagnostic Test Performance: Fundamental Meanings and Formulars. *Journal of Biometrics & Biostatistics*, 3(1). Available at: https://www.omicsonline.org/evaluating-measures-of-indicators-of-diagnostic-test-performance-fundamental-meanings-and-formulars-2155-6180.1000132.php?aid=4054.

Olafsdottir, E. et al., 2016. Early detection of type 2 diabetes mellitus and screening for retinopathy are associated with reduced prevalence and severity of retinopathy. *Acta Ophthalmologica*, 94(3), pp.232–239. Available at: https://www.ncbi.nlm.nih.gov/pubmed/26855250.

Olivier, M., 2005. A haplotype map of the human genome. *Nature*, 437(7063), pp.1299–

1320. Available at: http://www.nature.com/articles/nature04226.

Ott, J. & Hoh, J., 2003. Set Association Analysis of SNP Case-Control and Microarray Data. *Journal of Computational Biology*, 10, pp.569–574. Available at: https://www.ncbi.nlm.nih.gov/pubmed/12935345.

Ott, J., Wang, J. & Leal, S.M., 2015. Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, 16(5), pp.275–284. Available at: http://www.nature.com/articles/nrg3908.

Panagiotou, O.A. & Ioannidis, J.P.A., 2012. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology*, 41(1), pp.273–286. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22253303.

Pareek, C.S., Smoczynski, R. & Tretyn, A., 2011. Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, 52(4), pp.413–435. Available at: http://link.springer.com/10.1007/s13353-011-0057-x.

Patnala, R., Clements, J. & Batra, J., 2013. Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC Genetics*, 14(39). Available at: https://www.ncbi.nlm.nih.gov/pubmed/23656885.

Pendergrass, S.A. et al., 2015. Next-generation analysis of cataracts: determining knowledge driven gene-gene interactions using biofilter, and gene-environment interactions using the Phenx Toolkit*. *Pacific Symposium on Biocomputing.*, pp.495–505. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25741542.

Penrod, N.M. & Moore, J.H., 2014. Data Science Approaches to Pharmacogenetics. *Current Molecular Medicine*, 14(7), pp.805–813. Available at: https://www.ncbi.nlm.nih.gov/pubmed/25109795.

Perkins, N.J. & Schisterman, E.F., 2006. The inconsistency of "optimal" cut-points using two ROC based criteria. *American journal of epidemiology*, 163(7), pp.670–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16410346.

Perreault, L.L.-P. et al., 2013. pyGenClean: efficient tool for genetic data clean up before association testing. *Bioinformatics*, 29(13), pp.1704–1705. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt261.

Phani, N.M. et al., 2014. Population Specific Impact of Genetic Variants in KCNJ11 Gene to Type 2 Diabetes: A Case-Control and Meta-Analysis Study M. E. Saez, ed. *PLoS ONE*, 9(9), p.e107021. Available at: https://www.ncbi.nlm.nih.gov/pubmed/25247988.

Philippe, J. et al., 2015. What Is the Best NGS Enrichment Method for the Molecular Diagnosis of Monogenic Diabetes and Obesity? K. Brusgaard, ed. *PLOS ONE*, 10(11), p.e0143373. Available at: https://www.ncbi.nlm.nih.gov/pubmed/26599467.

Phillips, P.C., 2008. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11), pp.855–867. Available at: http://www.nature.com/articles/nrg2452.

Prabhu, S. & Pe'er, I., 2012. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Research*, 22(11), pp.2230–2240. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22767386.

Prasad, R. & Groop, L., 2015. Genetics of Type 2 Diabetes—Pitfalls and Possibilities. *Genes*, 6(1), pp.87–123. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4377835/.

Purcell, S. et al., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3), pp.559–575. Available at: https://www.ncbi.nlm.nih.gov/pubmed/17701901.

Qi, L. et al., 2010. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Human Molecular Genetics*, 19(13), pp.2706–2715. Available at: https://www.ncbi.nlm.nih.gov/pubmed/20418489.

Qi, Y., 2012. Random Forest for Bioinformatics. In C. Zhang & Y. Ma, eds. *Ensemble Machine Learning*. Boston, MA: Springer US, pp. 307–323. Available at: http://link.springer.com/10.1007/978-1-4419-9326-7_11.

Qiu, L. et al., 2014. Quantitative Assessment of the Effect of KCNJ11 Gene Polymorphism on the Risk of Type 2 Diabetes J. Devaney, ed. *PLoS ONE*, 9(4), p.e93961. Available at: https://www.ncbi.nlm.nih.gov/pubmed/24710510.

Ramsundar, B. & Zadeh, R.B., 2018. *TensorFlow for Deep Learning* 1st ed. R. Roumeliotis & A. Young, eds., Beijing: O'Reilly Media.

Refaeilzadeh, P., Tang, L. & Liu, H., 2009. *Encyclopedia of Database Systems* L. LIU & M. T. ÖZSU, eds., Boston, MA: Springer US. Available at:

http://link.springer.com/10.1007/978-0-387-39940-9.

Rehman, M.H. ur et al., 2016. Big Data Reduction Methods: A Survey. *Data Science and Engineering*, 1(4), pp.265–284. Available at: http://link.springer.com/10.1007/s41019-016-0022-0.

Ritchie, M.D. et al., 2001. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *Am. J. Hum. Genet*, 69(1), pp.138–147. Available at: https://www.ncbi.nlm.nih.gov/pubmed/11404819.

Ritchie, M.D., 2013. Reducing Dimensionality in the Search for Gene-Gene Interactions. In L. Padyukov, ed. *Between the Lines of Genetic Code: Genetic Interactions in Understanding Disease and Complex Phenotypes*. Amsterdam: Elsevier Inc., pp. 25–37. Available at: http://dx.doi.org/10.1016/B978-0-12-397017-6.00002-7.

Robinson, M.R., Wray, N.R. & Visscher, P.M., 2014. Explaining additional genetic variation in complex traits. *Trends in Genetics*, 30(4), pp.124–132. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4639398/.

Robnik-ˇSikonja, M. & Kononenko, I., 2003. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning Journal*, 53, pp.23–69. Available at: https://link.springer.com/article/10.1023/A:1025667309714.

Rokach, L. & Maimon, O., 2005. Decision Trees. In *Data Mining and Knowledge Discovery Handbook*. Springer, Boston, MA, pp. 165–192. Available at: http://link.springer.com/10.1007/0-387-25465-X_9.

Rose, A.M. & Bell, L.C.K., 2012. Epistasis and immunity: the role of genetic interactions in autoimmune diseases. *Immunology*, 137(2), pp.131–138. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22804709.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J., 1986. learning representations by back-propagating errors. *Nature*, 323, pp.533–536. Available at: https://www.nature.com/articles/323533a0.

Saeys, Y., Inza, I. & Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), pp.2507–2517. Available at: https://www.ncbi.nlm.nih.gov/pubmed/17720704.

Sailer, Z.R. & Harms, M.J., 2017. High-order epistasis shapes evolutionary trajectories J.

Krug, ed. *PLOS Computational Biology*, 13(5), p.e1005541. Available at: https://www.ncbi.nlm.nih.gov/pubmed/28505183.

Sale, M.M. et al., 2007. Variants of the Transcription Factor 7-Like 2 (TCF7L2) Gene Are Associated With Type 2 Diabetes in an African-American Population Enriched for Nephropathy. *Diabetes*, 56(10), pp.2638–2642. Available at: http://diabetes.diabetesjournals.org/cgi/doi/10.2337/db07-0012.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. *Neural Networks*, 61, pp.85–117. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0893608014002135.

Schwarz, D.F., König, I.R. & Ziegler, A., 2010. On safari to random Jungle: A fast implementation of random forests for high-dimensional data. *Bioinformatics*, 26(14), pp.1752–1758. Available at: https://www.ncbi.nlm.nih.gov/pubmed/20505004.

Scott, L.J. et al., 2006. Association of Transcription Factor 7-Like 2 (TCF7L2) Variants With Type 2 Diabetes in a Finnish Sample. *Diabetes*, 55(9), pp.2649–2653. Available at: http://diabetes.diabetesjournals.org/lookup/doi/10.2337/db06-0341.

Searls, D.B., 2010. The Roots of Bioinformatics. *PLoS Computational Biology*, 6(6), p.e1000809. Available at: https://www.ncbi.nlm.nih.gov/pubmed/20589079.

Sebastiani, P. & Solovieff, N., 2011. *Problem Solving Handbook in Computational Biology and Bioinformatics* L. S. Heath & N. Ramakrishnan, eds., Boston, MA: Springer US. Available at: http://link.springer.com/10.1007/978-0-387-09760-2.

Seliya, N., Khoshgoftaar, T.M. & Van Hulse, J., 2009. Aggregating performance metrics for classifier evaluation. In *2009 IEEE International Conference on Information Reuse & Integration*. Las Vegas, Nevada, USA: IEEE, pp. 35–40. Available at: http://ieeexplore.ieee.org/document/5211611/.

Sharma, N. & Saroha, K., 2015. Study of dimension reduction methodologies in data mining. In *International Conference on Computing, Communication & Automation*. IEEE, pp. 133–137. Available at: http://ieeexplore.ieee.org/document/7148359/.

Shields, R., 2011. Common Disease: Are Causative Alleles Common or Rare? *PLoS Biology*, 9(1), p.e1001009. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3022533/.

Shih, Y.S., 1999. Families of splitting criteria for classification trees. *Statistics and*

Computing, 9(4), pp.309–315. Available at: https://link.springer.com/article/10.1023%2FA%3A1008920224518.

De Silva, A.M. & Leong, P.H.W., 2015. *Feature Selection*, Singapore: SpringerBriefs in Computational Intelligence. Available at: http://link.springer.com/10.1007/978-981-287-411-5.

Sinnott-Armstrong, N.A. et al., 2009. Accelerating epistasis analysis in human genetics with consumer graphics hardware. *BMC Research Notes*, 2(149). Available at: https://www.ncbi.nlm.nih.gov/pubmed/19630950.

Smolensky, P., 1986. Information Processing in Dynamical Systems : Foundations of Harmony Theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. Cambridge, MA: MIT Press, pp. 149–281.

Soric, B., 1989. Statistical "Discoveries" and Effect-Size Estimation. *Journal of the American Statistical Association*, 84(406), pp.608–610. Available at: https://www.jstor.org/stable/2289950?origin=crossref.

Srivastava, N. et al., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1), pp.1929–1958. Available at: https://dl.acm.org/citation.cfm?id=2670313.

Storey, J.D. & Tibshirani, R., 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16), pp.9440–5. Available at: https://www.ncbi.nlm.nih.gov/pubmed/12883005.

Strobl, C. et al., 2008. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307). Available at: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-307.

Tabangin, M.E., Woo, J.G. & Martin, L.J., 2009. The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proceedings*, 3(S7). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2795940/.

Taeger, D. & Kuhnt, S., 2014. *Statistical Hypothesis Testing with SAS and R*, Chichester, UK: John Wiley & Sons, Ltd. Available at: http://doi.wiley.com/10.1002/9781118762585.

Taylor, M.B. & Ehrenreich, I.M., 2015. Higher-order genetic interactions and their contribution to complex traits. *Trends in Genetics*, 31(1), pp.34–40. Available at:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4281285/.

Teare, M.D. & Koref, M.F.S., 2014. Linkage analysis and the study of mendelian disease in the era of whole exome and genome sequencing. *Briefings in Functional Genomics*, 13(5), pp.378–383. Available at: https://www.ncbi.nlm.nih.gov/pubmed/25024279.

Tello, C.J.C., Hernández-Ramírez, D. & García-Sepúlveda, C.A., 2013. Support vector machine algorithms in the search of KIR gene associations with disease. *Computers in Biology and Medicine*, 43(12), pp.2053–2062. Available at: https://www.ncbi.nlm.nih.gov/pubmed/24290921.

Terada, A. et al., 2016. LAMPLINK: Detection of statistically significant SNP combinations from GWAS data. *Bioinformatics*, 32(22), pp.3513–3515. Available at: https://www.ncbi.nlm.nih.gov/pubmed/27412093.

Terada, A. et al., 2013. Statistical significance of combinatorial regulations. *Proceedings of the National Academy of Sciences*, 110(32), pp.12996–13001. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1302233110.

The Health Professional Follow-Up Study, G., 2009. *GENEVA Type 2 Diabetes Project Quality Control Report*,

The HEALTHY Study Group, 2010. A School-Based Intervention for Diabetes Risk Reduction. *New England Journal of Medicine*, 363(5), pp.443–453. Available at: https://www.ncbi.nlm.nih.gov/pubmed/21038513.

The Nurses' Health Study, G., 2009. *GENEVA Type 2 Diabetes Project Quality Control Report*,

Thompson, E.A., 2013. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics*, 194(2), pp.301–326. Available at: https://www.ncbi.nlm.nih.gov/pubmed/23733848.

Trevethan, R., 2017. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health*, 5(307). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5701930/.

Tryka, K.A. et al., 2014. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Research*, 42, pp.D975–D979. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965052/.

Tsai, F.-J. et al., 2010. A Genome-Wide Association Study Identifies Susceptibility Variants for Type 2 Diabetes in Han Chinese. *PLoS Genetics*, 6(2), p.e1000847. Available at: https://www.ncbi.nlm.nih.gov/pubmed/20174558.

Tudies, S. et al., 2012. Genetic and environmental factors associated With type 2 diabetes and diabetic vascular complications. *The Review of Diabetic Studies*, 9(1), pp.6–22. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22972441.

Turner, S., 2018. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software*, 3(25). Available at: http://joss.theoj.org/papers/10.21105/joss.00731.

Turner, S. et al., 2011. Quality Control Procedures for Genome-Wide Association Studies. *Current Protocols in Human Genetics*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21234875.

Turner, S.D. et al., 2011. Knowledge-Driven Multi-Locus Analysis Reveals Gene-Gene Interactions Influencing HDL Cholesterol Level in Two Independent EMR-Linked Biobanks M. B. Gravenor, ed. *PLoS ONE*, 6(5), p.e19586. Available at: https://www.ncbi.nlm.nih.gov/pubmed/21589926.

Upstill-Goddard, R. et al., 2013. Machine learning approaches for the discovery of gene-gene interactions in disease data. *Briefings in Bioinformatics*, 14(2), pp.251–260. Available at: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs024.

VanderWeele, T.J., 2010. Epistatic Interactions. *Statistical Applications in Genetics and Molecular Biology*, 9(1). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2861312/.

Vanitha, C.D.A., Devaraj, D. & Venkatesulu, M., 2015. Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection. *Procedia Computer Science*, 47, pp.13–21. Available at: http://www.sciencedirect.com/science/article/pii/S1877050915004469.

Vapnik, V. & Lerner, A.Y., 1963. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24(6), pp.774–780.

Vařeka, L. & Mautner, P., 2017. Stacked Autoencoders for the P300 Component Detection. *Frontiers in Neuroscience*, 11(302). Available at: http://journal.frontiersin.org/article/10.3389/fnins.2017.00302/full.

Veerabhadrappa & Lalitha, R., 2010. Multi-Level Dimensionality Reduction Methods Using Feature Selection and Feature Extraction. *International Journal of Artificial Intelligence & Applications*, 1(4), pp.54–68. Available at: http://www.airccse.org/journal/ijaia/papers/1010ijaia05.pdf.

Velez, D.R. et al., 2007. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic Epidemiology*, 31(4), pp.306–315. Available at: https://www.ncbi.nlm.nih.gov/pubmed/17323372.

Verma, S.S. et al., 2018. Collective feature selection to identify crucial epistatic variants. *BioData Mining*, 11(5). Available at: https://www.ncbi.nlm.nih.gov/pubmed/29713383.

Verweij, N., van de Vegte, Y.J. & van der Harst, P., 2018. Genetic study links components of the autonomous nervous system to heart-rate profile during exercise. *Nature Communications*, 9(898). Available at: http://www.nature.com/articles/s41467-018-03395-6.

Visscher, P.M. et al., 2012. Five years of GWAS discovery. *American Journal of Human Genetics*, 90(1), pp.7–24. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22243964.

Voight, B.F. et al., 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics*, 42(7), pp.579–589. Available at: http://www.nature.com/articles/ng.609.

Vovk, V., 2015. The Fundamental Nature of the Log Loss Function. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 307–318. Available at: http://link.springer.com/10.1007/978-3-319-23534-9_20.

Waaijenborg, S. & Zwinderman, A.H., 2009. Correlating multiple SNPs and multiple disease phenotypes: penalized non-linear canonical correlation analysis. *Bioinformatics*, 25(21), pp.2764–2771. Available at: https://www.ncbi.nlm.nih.gov/pubmed/19689958.

Wan, X. et al., 2010. BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *The American Journal of Human Genetics*, 87(3), pp.325–340. Available at: https://www.ncbi.nlm.nih.gov/pubmed/20817139.

Wang et al., 2016. Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction. *Journal of Diabetes*, 8(1), pp.24–35. Available at: https://www.ncbi.nlm.nih.gov/pubmed/26119161.

Wang, M.H. et al., 2016. A fast and powerful W-test for pairwise epistasis testing. *Nucleic Acids Research*, 44(12), p.e115. Available at: https://www.ncbi.nlm.nih.gov/pubmed/27112568.

Wang, M.H. et al., 2014. Four pairs of gene–gene interactions associated with increased risk for type 2 diabetes (CDKN2BAS–KCNJ11), obesity (SLC2A9–IGF2BP2, FTO–APOA5), and hypertension (MC4R–IGF2BP2) in Chinese women. *Meta Gene*, 2(1), pp.384–391. Available at: https://www.ncbi.nlm.nih.gov/pubmed/25606423.

Wang, X. et al., 2016. *Application of Clinical Bioinformatics* X. Wang et al., eds., Dordrecht: Springer Netherlands. Available at: http://link.springer.com/10.1007/978-94-017-7543-4.

Weale, M.E., 2010. Quality Control for Genome Wide Association Studies. In M. R. Barnes & G. Breen, eds. *Genetic Variation*. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 341–372. Available at: http://link.springer.com/10.1007/978-1-60327-367-1.

Wei, W.-H., Hemani, G. & Haley, C.S., 2014. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11), pp.722–733. Available at: http://www.nature.com/articles/nrg3747.

Wei, Z. et al., 2009. From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes P. M. Visscher, ed. *PLoS Genetics*, 5(10), p.e1000678. Available at: https://www.ncbi.nlm.nih.gov/pubmed/19816555.

Werbos, P.J., 1982. Applications of advances in nonlinear sensitivity analysis. *System Modeling and Optimization. Lecture Notes in Control and Information Sciences*, 38, pp.762–770. Available at: https://link.springer.com/chapter/10.1007/BFb0006203.

Werbos, P.J., 1974. *Beyond regression : new tools for prediction and analysis in the behavior sciences*. Harvard University, Cambridge, MA.

Wheeler, D.L. et al., 2007. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35, pp.D5–D12. Available at:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1781113/.

Whiting, D.R. et al., 2011. IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes Research and Clinical Practice*, 94(3), pp.311–321. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0168822711005912.

Wigginton, J.E., Cutler, D.J. & Abecasis, G.R., 2005. A Note on Exact Tests of Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics*, 76(5), pp.887–893. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0002929707607356.

Wilson, P.W.F., 2007. Prediction of Incident Diabetes Mellitus in Middle-aged Adults:the Framingham Offspring Study. *Archives of Internal Medicine*, 167(10), pp.1068–1074. Available at: https://www.ncbi.nlm.nih.gov/pubmed/17533210.

Winkler, T.W. et al., 2014. Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols*, 9(5), pp.1192–1212. Available at: http://www.nature.com/doifinder/10.1038/nprot.2014.071.

Wise, J., 2018. NHS diabetes prevention programme helps weight loss, analysis shows. *BMJ*, 360. Available at: http://www.bmj.com/lookup/doi/10.1136/bmj.k1196.

World Health Organization, 2016. *Global Report on Diabetes*, Available at: https://www.who.int/diabetes/global-report/en.

Wu, H. et al., 2018. Sharing Deep Neural Network Models with Interpretation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. Lyon, France, pp. 177–186. Available at: http://dl.acm.org/citation.cfm?doid=3178876.3185995.

Xuan, J. et al., 2013. Next-generation sequencing in the clinic: Promises and challenges. *Cancer Letters*, 340(2), pp.284–295. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0304383512006726.

Yang, C.-H. et al., 2013. MDR-ER: Balancing Functions for Adjusting the Ratio in Risk Classes and Classification Errors for Imbalanced Cases and Controls Using Multifactor-Dimensionality Reduction M. M. Abad-Grau, ed. *PLoS ONE*, 8(11), p.e79387. Available at: https://www.ncbi.nlm.nih.gov/pubmed/24236125.

Yang, C. et al., 2009. SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*, 25(4), pp.504–511. Available at: https://www.ncbi.nlm.nih.gov/pubmed/19098029.

Yoshida, M. & Koike, A., 2011. SNPInterForest: A new method for detecting epistatic interactions. *BMC Bioinformatics*, 12(469). Available at: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-469.

Youden, W.J., 1950. Index for rating diagnostic tests. *Cancer*, 3(1), pp.32–35. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15405679.

Yu, W., Lee, S. & Park, T., 2016. A unified model based multifactor dimensionality reduction framework for detecting gene–gene interactions. *Bioinformatics*, 32(17), pp.i605–i610. Available at: https://www.ncbi.nlm.nih.gov/pubmed/27587680.

Yung, L.S. et al., 2011. GBOOST: a GPU-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics*, 27(9), pp.1309–1310. Available at: https://www.ncbi.nlm.nih.gov/pubmed/21372087.

Zeiler, M.D. & Fergus, R., 2014. *Visualizing and Understanding Convolutional Networks*, Springer, Cham. Available at: http://link.springer.com/10.1007/978-3-319-10590-1_53.

Zeng, P. et al., 2015. Statistical analysis for genome-wide association study. *Journal of Biomedical Research*, 29(4), pp.285–297. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4547377/.

Zhang, X. et al., 2012. Chapter 10: Mining Genome-Wide Genetic Markers F. Lewitter & M. Kann, eds. *PLoS Computational Biology*, 8(12), p.e1002828. Available at: https://www.ncbi.nlm.nih.gov/pubmed/23300418.

Zhang, X. et al., 2010. TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, 26(12), pp.i217–i227. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2881371/.

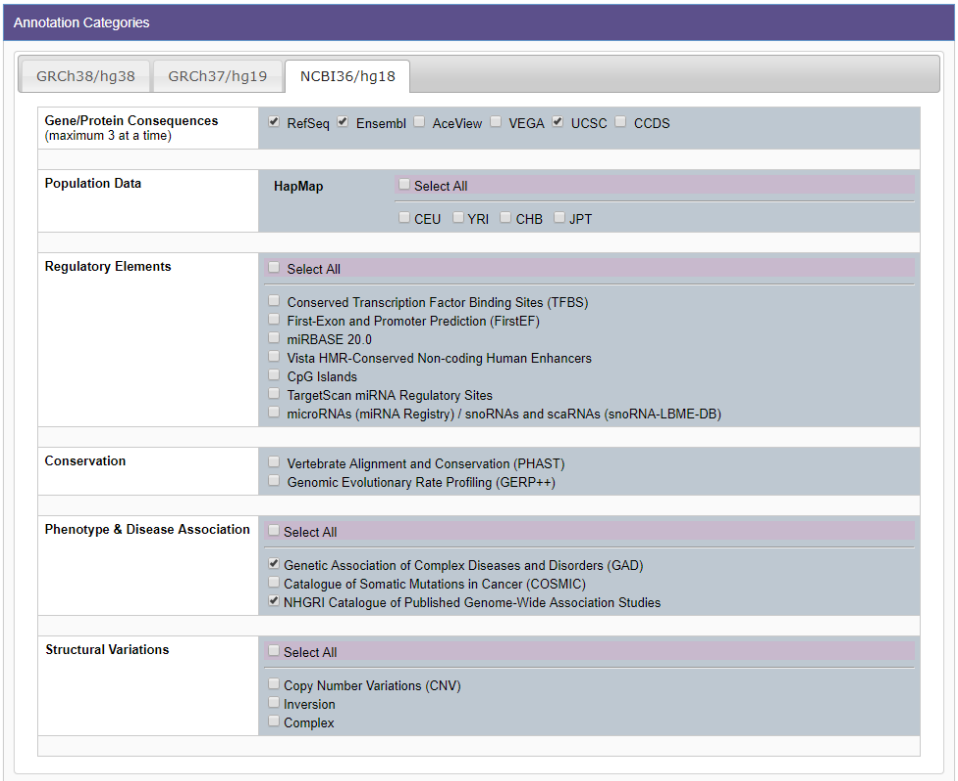
Zhang, Y. & Liu, J.S., 2007. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39(9), pp.1167–1173. Available at: http://www.nature.com/doifinder/10.1038/ng2110.

Zhu, Z. et al., 2013. Development of GMDR-GPU for Gene-Gene Interaction Analysis and Its Application to WTCCC GWAS Data for Type 2 Diabetes H. Zhang, ed. *PLoS ONE*, 8(4), p.e61943. Available at: https://www.ncbi.nlm.nih.gov/pubmed/23626757.

Appendix

The 6609 SNPs extracted from association analysis was used as an input to SNP nexus annotation tool to extract comprehensive annotation of query variants (SNPs) and report the overlapped or closest genes according to the NCBI36/hg18 assembly. To find the relationship between the queried SNPs and disease association (phenotype) we utilized two disease association datasets: ***The Genetic Association Database (GAD)*** and ***The Catalogue of Published Genome-Wide Association Studies (GWAS Catalogue)***. GAD provides information about published scientific papers on human genetic association studies of complex disorders. GWAS Catalogue provides information about SNPs identified by published GWAS. Appendix A is a screenshot of SNPnexus annotation categories selections used to query the 6609 SNPs. A list of phenotype and disease association is presented in Appendix B. The list was limited to include SNPs associated to T2D and other disorders that may influence T2D predisposition. Of the 6609 SNPs, there are a number of SNPs have been reported by previous GWAS studies and were presented in Appendix C. These candidate SNPs can be investigated further to help better understand epistasis in T2D using GWAS.

Appendix A: SNP Nexus Annotation Categories



Appendix B: List of SNPs Associated to T2D and other Diseases Related to T2D Reported via GAD

SNP	Assoc	Phenotype	Disease Class	Gene
rs10885409	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides, birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs11196205	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides, birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs11196208	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides, birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs12243326	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides, birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs12255372	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides, birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs4132670	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides, birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs4506565	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides, birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs7901695	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides, birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs10787472	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides, birth weight	metabolic	*TCF7L2*

		glucose small for gestational age		
rs4074720	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides, birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs6585201	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides, birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs2516478		cardiomyopathy	cardiovascular	*BAT1*
rs6103716	Y	diabetes, type 2; kidney failure, chronic, cholesterol metabolic syndrome triglycerides, type 2 diabetes glucose insulin, gestational	metabolic	*HNF4A*
rs6866823		coronary artery bypass grafting; platelet hyperreactivity	cardiovascular	*ITGA1*
rs7731949		coronary artery bypass grafting; platelet hyperreactivity	cardiovascular	*ITGA1*
rs10518694		diabetes, type 2	metabolic	*ONECUT1*
rs7735277	Y	vascular disease; coronary artery disease; stroke	cardiovascular	*ITGA2*
rs7180600		diabetes, type 2	metabolic	*ONECUT1*
rs7735277	Y	diabetes, type 2	metabolic	*ITGA2*
rs17061580		diabetes, type 2	metabolic	*KLF12*
rs2325583		diabetes, type 2	metabolic	*KLF12*
rs10805519		stroke, ischemic; atherosclerosis, carotid	cardiovascular	*PDE4D*
rs6138948		coronary artery disease; diabetes, type 2; lipoproteins; longevity	cardiovascular	*PTPRA*
rs6886001		coronary artery bypass grafting; platelet hyperreactivity	cardiovascular	*ITGA1*
rs9318218		diabetes, type 2	metabolic	*KLF12*
rs10466028	N	obesity	metabolic	*PRKG1*
rs13037313		coronary artery disease; diabetes, type 2; lipoproteins; longevity	cardiovascular	*PTPRA*
rs9530249		diabetes, type 2	metabolic	*KLF12*

rs16925235	N	obesity	metabolic	*PRKG1*
rs6133002		coronary artery disease; diabetes, type 2; lipoproteins; longevity	cardiovascular	*PTPRA*
rs4620533	Y	diabetes, type 2	metabolic	*PKLR*
rs1125392		diabetes, type 2	metabolic	*FTO*
rs1125392	Y	obesity	metabolic	*FTO*
rs11153664		atherosclerosis, coronary; myocardial infarct	cardiovascular	*ROS1*
rs1979398		coronary artery bypass grafting; platelet hyperreactivity	cardiovascular	*ITGA1*
rs755886	N	diabetes, type 2	metabolic	*CACNA1D*
rs3746619	Y	obesity	metabolic	*MC3R*
rs3746619	N	diabetes, type 2	metabolic	*MC3R*
rs3746619	N	physical activity	normalvariation	*MC3R*
rs1780365	N	diabetes, type 2; insulin	metabolic	*PBX1*
rs1763908		diabetes, type 2 glucose tolerance; insulin	metabolic	*ARHGEF11*
rs1324392		coronary artery disease; diabetes, type 2; lipoproteins; longevity	cardiovascular	*PTPRA*
rs7767391		diabetes, type 2 triglycerides	metabolic	*CDKAL1*
rs17875671		diabetes, type 2	metabolic	*IKBKB*
rs17383719		diabetes, type 2; insulin	metabolic	*PBX1*
rs10940659		stroke, ischemic; atherosclerosis, carotid	cardiovascular	*PDE4D*
rs1923882	Y	cholesterol, HDL; cardiovascular disease; obesity; hyperuricemia	metabolic	*HTR2A*
rs6138948		coronary artery disease; diabetes, type 2; lipoproteins; longevity	cardiovascular	*PTPRA*
rs651821	Y	triglycerides; atherosclerosis, coronary; diabetes, type 2; obesity; cholesterol, HDL; lipids; lipoproteins; body mass; type 2 diabetic nephropathy triglycerides	metabolic	*APOA5*
rs8056879		diabetes, type 2	metabolic	*PRKCB1*
rs10492795		diabetes, type 2	metabolic	*PRKCB1*
rs6480638	N	obesity	metabolic	*PRKCB1*

rs13274396		vascular response; blood pressure, arterial; heart rate; adrenaline; coronary flow velocity; ECG; noradrenaline	cardiovascular	*ADRA1A*
rs6859355	Y	diabetes, type 2; metabolism disorders; myocardial infarction; stroke, ischemic; cholesterol, HDL	metabolic	*ITGA2*
rs3138139		diabetes, type 2; liver disease	metabolic	*RDH5*
rs1040558		diabetes, type 2; triglycerides	metabolic	*CDKAL1*
rs12136288	Y	diabetes, type 2 insulin	metabolic	*CHRM3*
rs2612026	N	diabetes, type 2	metabolic	*CACNA1D*
rs9940128		diabetes, type 2; obesity	metabolic	*FTO*
rs1904013	N	obesity	metabolic	*PRKG1*
rs2245407		diabetes, type 2; MODY; cholesterol, HDL; glucose tolerance; metabolic syndrome; gestational	metabolic	*TCF1*
rs179250	Y	insulin	metabolic	*TSHR*
rs1212595		obesity, localized	metabolic	*NCOA3*
rs4952404	Y	hypertension	cardiovascular	*SLC8A1*
rs1212595		insulin-like growth factor-1; estrogen metabolism	normalvariation	*NCOA3*
rs4073288	Y	body mass; cholesterol HDL; diabetes, type 2; glucose insulin metabolic syndrome triglycerides; birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs9295495	Y	diabetes, type 2; diabetes, type 2 triglycerides	metabolic	*CDKAL1*
rs9465871	Y	diabetes, type 2; diabetes, type 2 triglycerides	metabolic	*CDKAL1*
rs12883673	Y	insulin	metabolic	*TSHR*
rs10935840	Y	peripheral arterial disease; heart disease, ischemic; peripheral vascular disease	cardiovascular	*P2RY12*
rs4732958		blood pressure, arterial; heart rate; adrenaline; coronary flow velocity; ECG; noradrenaline; hypertension	cardiovascular	*ADRA1A*
rs2209726		diabetes, type 2	metabolic	*KLF12*
rs12152938	Y	stroke, ischemic	cardiovascular	*PDE4D*
rs2680649	N	diabetes, type 2	metabolic	*CACNA1D*

rs8080702		diabetes, type 2	metabolic	*GCGR*
rs8080702		diabetes, type 2	metabolic	*GAPD*
rs9565045		diabetes, type 2	metabolic	*KLF12*
rs11000404	N	obesity	metabolic	*PRKG1*
rs9573312		diabetes, type 2	metabolic	*KLF12*
rs9939973	Y	diabetes, type 2; obesity	metabolic	*FTO*
rs1980445		diabetes, type 2; lipids; glucose	metabolic	*PLA2G4A*
rs10426094		body mass; cholesterol; triglycerides; insulin; glucose; blood pressure, arterial; diabetes, type 2; hypertension; insulin; obesity	metabolic	*INSR*
rs10426094		atherosclerosis, coronary	cardiovascular	*INSR*
rs1544791	Y	stroke, ischemic	cardiovascular	*PDE4D*
rs2328549		diabetes, type 2 triglycerides	metabolic	*CDKAL1*
rs3820700		diabetes, type 2	metabolic	*ALMS1*
rs9460598		diabetes, type 2 triglycerides	metabolic	*CDKAL1*
rs10757270		diabetes, type 2	metabolic	*MTAP*
rs9465970		diabetes, type 2 triglycerides	metabolic	*CDKAL1*
rs1121980	Y	diabetes, type 2; obesity	metabolic	*FTO*
rs11789818		coronary artery disease; atherosclerosis, coronary	cardiovascular	*ABCA1*
rs11789818		diabetes, type 2; body mass; cholesterol; cholesterol, HDL; lipoprotein, LDL; triglycerides	metabolic	*ABCA1*
rs1206883		obesity, localized	metabolic	*NCOA3*
rs780390	N	diabetes, type 2	metabolic	*ALMS1*
rs2164660	Y	stroke, ischemic	cardiovascular	*PDE4D*
rs3811942		diabetes, type 2; obesity; hypertension; insulin	metabolic	*PCSK1*
rs4809639		obesity, localized	metabolic	*NCOA3*
rs16978425	Y	blood pressure	cardiovascular	*SLC14A2*
rs2689249		diabetes, type 2; obesity	metabolic	*FTO*
rs3827103		diabetes, type 2; obesity; insulin; leptin; glucose tolerance insulin	metabolic	*MC3R*
rs6554137	Y	heart transplant complications	cardiovascular	*PDGFRA*
rs9460546	Y	diabetes, type 2	metabolic	*CDKAL1*

rs2053086	Y	myocardial infarct	cardiovascular	*HNRPUL1*
rs621060	Y	diabetes, type 2 insulin	metabolic	*CHRM3*
rs17742120		stroke, ischemic; atherosclerosis, carotid	cardiovascular	*PDE4D*
rs16880453		coronary artery bypass grafting; platelet hyperreactivity	cardiovascular	*ITGA1*
rs9530244		diabetes, type 2	metabolic	*KLF12*
rs6084231		coronary artery disease; diabetes, type 2; lipoproteins; longevity	cardiovascular	*PTPRA*
rs1537306		obesity, localized	metabolic	*NCOA3*
rs6456397		diabetes, type 2 triglycerides	metabolic	*CDKAL1*
rs3799559		heart disease, ischemic; peripheral arterial disease; coronary heart disease; stroke, ischemic	cardiovascular	*F13A1*
rs11000467	N	obesity	metabolic	*PRKG1*
rs17816224	N	diabetes, type 2 glucose tolerance obesity	metabolic	*SCG5*
rs7901275	Y	body mass cholesterol, HDL diabetes, type 2 glucose insulin metabolic syndrome triglycerides; birth weight glucose small for gestational age	metabolic	*TCF7L2*
rs9888532		diabetes, type 2	metabolic	*KLF12*
rs10490053	Y	hypertension	cardiovascular	*SLC8A1*

Appendix C: List of SNPs Reported via GWAS Catalogue

SNP	**Region**	**Genes**	**Allele_frequency**	**Trait**
rs10400419	12q14.3	HMGA2	0.3837	Axial length
rs10906115	10p13	CDC123, CAMK1D	0.57	Type 2 diabetes
rs10911902	1q31.1	Intergenic	0.17	Schizophrenia
rs1106766	12q13.3	R3HDM2, INHBC	0.23	Urate levels
rs11097407	4q22.3	NR	NR	Bipolar disorder and schizophrenia
rs11118346	1q41	LYPLAL1	0.47	Height
rs1121980	16q12.2	FTO	NR	Body mass index

rs1121980	16q12.2	FTO	0.41	Obesity (early onset extreme)
rs11259933	15q25.2	ADAMTSL3	0.51	Height
rs11259936	15q25.2	ADAMTSL3	0.48	Height
rs11777747	8q24.3	FLJ43860	0.03	Coronary artery calcification
rs12000445	9p21.3	HuB	NR	Response to platinum-based chemotherapy in non-small-cell lung cancer
rs12148477	15q21.2	NR	0.21	Follicule stimulating hormone
rs12243326	10q25.2	TCF7L2	NR	Two-hour glucose challenge
rs12619788	2p16.3	Intergenic	NR	Economic and political preferences (immigration/crime)
rs12970134	18q21.32	MC4R	0.30	Weight
rs12970134	18q21.32	MC4R	0.30	Body mass index
rs12970134	18q21.32	MC4R	0.36	Waist circumference and related phenotypes
rs1436953	15q22.2	C2CD4A, C2CD4B	0.64	Type 2 diabetes
rs1436955	15q22.2	C2CD4B	0.73	Type 2 diabetes
rs16887552	4p15.33	Intergenic	NR	Response to mTOR inhibitor (everolimus)
rs17042171	4q25	PITX2	0.12	Atrial fibrillation
rs17112901	10q24.31	PKD2L1	0.163	Obesity-related traits
rs17157663	7p21.3	Intergenic	0.32	Quantitative traits
rs17382202	5q12.1	PDE4D	0.153	Response to antipsychotic treatment
rs17465637	1q41	MIA3	0.74	Coronary heart disease
rs17465637	1q41	MIA3	0.72	Myocardial infarction (early onset)
rs17465637	1q41	MIA3	0.71	Coronary heart disease
rs17782313	18q21.32	MC4R	0.76	Height
rs17782313	18q21.32	MC4R	0.18	Obesity
rs17782313	18q21.32	MC4R	0.21	Body mass index
rs17782313	18q21.32	MC4R	0.24	Body mass index
rs1997111	12q24.23	Intergenic	NR	T-tau

rs2200733	4q25	Intergenic	0.12	Atrial fibrillation
rs2200733	4q25	NR	0.11	Stroke (ischemic)
rs2200733	4q25	PITX2, ENPEP	0.11	Atrial fibrillation/atrial flutter
rs2271293	16q22.1	CTCF, PRMT8	0.87	HDL cholesterol
rs2271293	16q22.1	LCAT	0.11	HDL cholesterol
rs2546890	5q33.3	IL12B	0.52	Multiple sclerosis
rs2546890	5q33.3	IL12B	NR	Multiple sclerosis
rs2546890	5q33.3	IL12B	0.56	Psoriasis
rs255052	16q22.1	LCAT	0.17	HDL cholesterol
rs2785980	1q41	LYPLAL1	NR	Fasting insulin-related traits (interaction with BMI)
rs2791553	1q41	LYPLAL1	0.60	Adiponectin levels
rs2820464	1q41	LYPLAL1	0.66	Waist-hip ratio
rs2847281	18p11.21	PTPN2	0.16	Esophageal cancer (squamous cell)
rs2847281	18p11.21	PTPN2	NR	C-reactive protein
rs339331	6q22.1	GPRC6A, RFX6	0.37	Prostate cancer
rs354033	7q36.1	ZNF767, ZNF746	NR	Multiple sclerosis
rs4301033	3q25.1	TSC22D2	NR	Adiponectin levels
rs4506565	10q25.2	TCF7L2	0.31	Fasting glucose-related traits
rs4506565	10q25.2	TCF7L2	0.32	Type 2 diabetes
rs4660293	1p34.3	MACF1, PABPC4	0.23	HDL cholesterol
rs4739466	8p11.23	NR	NR	Bipolar disorder
rs476828	18q21.32	MC4R	0.24	Obesity (early onset extreme)
rs4842838	15q25.2	ADAMTSL3	0.29	Height
rs4842838	15q25.2	ADAMTSL3	0.32	Height
rs651821	11q23.3	APOA1, APOC3, APOA4, APOA5	NR	Lipid metabolism phenotypes
rs651821	11q23.3	APOA5	NR	Triglycerides

rs6589566	11q23.3	APOA, APOC	0.0176	Triglycerides
rs6589566	11q23.3	APOA1, APOC3, APOA5	0.06	LDL cholesterol
rs6681460	1p31.3	SGIP1	0.3755	Presence of antiphospholipid antibodies
rs6736587	2p12	CTNNA2	0.16	Orthostatic hypotension
rs6887695	5q33.3	IL12B	0.32	Crohn's disease
rs7234864	18q21.32	PMAIP1, MC4R	0.26	Body mass index
rs7341475	7q22.1	RELN	0.62	Schizophrenia
rs742134	22q13.2	BIK	NR	Prostate cancer
rs7562790	2p22.2	CRIM1	0.40	Ventricular conduction
rs7593730	2q24.2	RBMS1, ITGB6	0.78	Type 2 diabetes
rs765855	7p21.3	NR	NR	Breast Cancer in BRCA1 mutation carriers
rs7667	1p36.13	CAPZB	NR	Crohn's disease and psoriasis
rs7901695	10q25.2	TCF7L2	0.45	Coronary heart disease
rs7901695	10q25.2	TCF7L2	NR	Type 2 diabetes
rs9268877	6p21.32	MHC	NR	Ulcerative colitis
rs9268877	6p21.32	HLA-DRA, BTNL2	0.45	Ulcerative colitis
rs9465871	6p22.3	CDKAL1	0.18	Type 2 diabetes
rs9813712	3q22.1	Intergenic	NR	Response to amphetamines
rs9940128	16q12.2	FTO	0.44	Body mass index
rs9940128	16q12.2	FTO	0.42	Metabolic syndrome

NR: Not Reported