A NOVEL VARIABLE SELECTION METHOD FOR CLASSIFICATION WITH APPLICATION TO SINGLE NUCLEOTIDE POLYMORPHISM DATA

Thesis submitted in accordance with the requirements of the University of Liverpool

for the degree of

Doctor of Philosophy

in

Biostatistics

by

Nazatulshima Hassan

September 2018

DECLARATION

I declare that the thesis has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Nazatulshima Hassan

This research was carried out in the Department of Biostatistics, in the Institute of Translational Medicine, at the University of Liverpool, United Kingdom.

TABLE OF CONTENTS

ABSTRACT

A Novel Variable Selection Method for Classification with Application to Single Nucleotide Polymorphism Data

by

Nazatulshima Hassan

Introduction and Aims: In recent years, there has been a growing interest in studying genetic data so as to answer specific medical questions; for example, indicative biomarkers that can accurately predict (classify) outcomes (e.g. healthy and disease or different categories of patients' response to treatment). In genome-wide data analysis, a typical procedure is to use a variable selection approach, often univariable, where the primary aim is to select the most important genetic variants, particularly Single Nucleotide Polymorphisms (SNPs), associated with an outcome of interest. This thesis proposes a novel variable selection method by considering the multivariate nature of the genetic data. The aim of this thesis is threefold: (i) to develop a quantitative variable selection method for classification which can be used in the multivariate setting, computationally inexpensive and easy to understand and to apply, (ii) to propose a multi-step approach that selects SNPs and evaluates the classification performance of the resulting models in a cross-validation framework, and (iii) to jointly model the longitudinal clinical and SNP data for classification using the Standard and New Antiepileptic Drugs (SANAD) dataset.

Methods: A literature search was conducted to study the different approaches of variable selection and their relationship with classification performance. A novel variable selection method, tSNR within a logistic regression framework

was developed to select the most informative SNPs. In addition, a multi-step framework that involved univariable and multivariable selection in a cross-validation setting was proposed. Then, the filter metric tSNR and the multi-step framework were assessed using simulated datasets. The methods were further examined using an epilepsy pharmacogenomics dataset (EpiPGX) in which the phenotype of interest is the remission from seizures status after receiving first well-tolerated antiepileptic drugs (AEDs). A second epilepsy dataset from the SANAD trial was used as the validation dataset. Within the SANAD dataset, the longitudinal clinical and SNP data were jointly modelled using a longitudinal discriminant analysis (LoDA) approach with multivariate generalised linear mixed model (MGLMM). The classification performance was measured by calculating the probability of correct classification (PCC) and area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

Results: The literature review suggested the need for variable selection methods, which could potentially aid better classification accuracy. In the simulation study, the univariable tSNR ranking was able to capture the causal SNPs in the top ten ranked SNPs. In addition, within the proposed framework, the results using simulated datasets suggested that the classification performance using SNPs selected by cumulative tSNR (multivariate) are better than the SNPs based on univariable tSNR ranking. The results were further confirmed using the real clinical datasets. The addition of SNP data to the longitudinal model based on clinical data improved the mean prediction time at which patients who will not achieve remission from seizures within five years of commencing treatment are identified. However, it did not provide an improvement to the classification performance.

Conclusions:

The developed approach using a tSNR filter metric proved to be effective in ranking and selecting subset of SNPs that are associated with the outcome of interest. The SNPs selected by tSNR were shown to give good classification accuracy. Also, by jointly modelling the longitudinal clinical data and SNP data (selected using tSNR) in a longitudinal model the prediction time at which patients can be classified was improved.

MANUSCRIPT AND COMMUNICATIONS

Manuscript in preparation

N Hassan, G Czanner, A Jorgensen and M García-Fiñana (2018). Variable Selection Methods for Classification with Application to Single Nucleotide Polymorphism (SNP) Data: A Review. *(to be submitted to BMC Medical Research Methodology)*

Oral presentations

"A Review on Classifications Approach in SNP Data".Young Statisticians Meeting (YSM 2015), University of Cardiff, UK, 23-24 July 2015.

"tSNR as a Feature Selection Method in SNP Data Analysis". Analytical Methods in Statistics (AMISTAT 2015), Prague, 10-13 November 2015.

"Extension of Simple Logistic Regression as Variable Selection Method for Classification: An Application to SNP Data". 4th Stochastic Modelling Techniques and Data Analysis International Conference (SMTDA2016), Malta, 01-04 June 2016.

"Novel tSNR as Variable Selection Method for Classification: An Application to SNP Data". 37th Annual Conference of the International Society for Clinical Biostatistics (ISCB 2016), Birmingham, 21-25 August 2016.

Poster presentations

"Novel tSNR as Variable Selection Method for Classification: An Application to SNP Data". Faculty Poster Day, Faculty of Health & Life Sciences, University of Liverpool, UK, 10 June 2016.

"Novel tSNR as Variable Selection Method for Classification: An Application to SNP Data". Statistical Methods for Post Genomic Data (SMPGD 2017), London, 12-13 January 2017.

"Combining Clinical and Single Nucleotide Polymorphism Data in Longitudinal Discriminant Analysis". Statistical Analysis of Multi-Outcome Data (SAM 2017), University of Liverpool, UK, 03-04 July 2017.

ACKNOWLEDGEMENTS

I would like to thank various people for their contributions to this thesis. After all these years, there are no proper words to convey my deep gratitude and respect for my supervisors, Dr Marta García-Fiñana, Dr Gabriela Czanner and Dr Andrea Jorgensen for their continuous support and constructive suggestions for this thesis.

My sincere thanks must also go to Dr David Hughes for his help with the longitudinal data analysis, Dr Ben Francis and Dr Ian Smith for their help related to GWAS data analysis. I would also like to thank Dr Graeme Sills and committee for the permission to use the EpiPGX dataset and Professor Tony Marson for his permission with the SANAD dataset. I also thank the members of Multivariate Modelling Group and the Statistical Genetics Group from the department of Biostatistics and my fellow colleagues for advice and encouragement.

I would also like to take this opportunity to thank my viva examiners Professor Mario Cortina Borja and Dr Girvan Burnside for their comments and suggestions to improve the quality of this thesis.

Special thanks to Majlis Amanah Rakyat (MARA) and Universiti Kuala Lumpur (UniKL) as the sponsors and for making me pursuing my PhD possible. I am most grateful to the management and supervisors from UniKL for giving me this opportunity, which there is no way to express how much it meant to me.

I also would like to thank my friends who are together with me along this PhD journey, making it even more meaningful. Finally, I thank with love my supportive family, my parents, Haji Hassan Bin Ismail and Hajjah Habibah Binti Sa'adon and my brother Mohd Khairul Hisham Bin Hassan.

ABBREVIATIONS

AED	Antiepileptic Drug
AIC	Akaike Information Criterion
AUC	Area Under the ROC Curve
BIC	Bayesian Information Criterion
BP	Base Pair
EpiPGX	Epilepsy Pharmacogenomics
FDR	False Discovery Rate
FN	False Negative
FP	False Positive
GLM	Generalised Linear Models
GLMM	Generalised Linear Mixed Models
GWAS	Genome Wide Association Studies
HapMap	Haplotype Map
HWE	Hardy-Weinberg Equilibrium
KNN	k-Nearest Neighbour
LD	Linkage Disequilibrium
LDA	Linear Discriminant Analysis
LoDA	Longitudinal Discriminant Analysis
LR	Logistic Regression
MCMC	Markov Chain Monte Carlo
MGLMM	Multivariate Generalised Linear Mixed Models
MRI	Magnetic Resonance Imaging
NPV	Negative Predictive Value
PCC	Probability of Correct Classification
PLR	Penalised Logistic Regression

PPV	Positive Predictive Value
QC	Quality Control
RF	Random Forest
ROC	Receiver Operating Characteristic
rs	Reference SNP
SANAD	Standard and New Antiepileptic Drugs
SLR	Stepwise Logistic Regression
SNP	Single Nucleotide Polymorphism
SNR	Signal-to-Noise Ratio
SVM	Support Vector Machine

STATISTICAL SYMBOLS AND NOTATIONS

X	SNPs, variables, covariates, features, biomarkers
Y	phenotype, outcome
p	number of SNPs
n	total number of samples, patients, individuals
n_g	number of samples in specific phenotype group
X_i	observation of i-th SNP where $i = 1,2,\dots,p$
X_{ij}	observation of i-th SNP and j-th sample where $j = 1,2,\dots,n$
D'/r^2	LD summary measure
β	coefficient, parameter
σ^2	variance
μ	mean
Ω	set of all samples in SVM
k	the number of splits for cross-validation
d	bias-correction term represents the number of coefficient
G	total number of phenotype groups where $g = 0,1,\dots,G-1$
R	the number of longitudinal biomarkers where $r = 1,2,\dots,R$
T	observation time where $t_r = \{t_{r,1},\dots,t_{r,n_r}\}$
Y_r	longitudinal observations $y_r = \{y_{r,1},\dots,y_{r,n_r}\}$
Y_{SNP_i}	SNP data when jointly model with longitudinal marker (i-th SNP where $i = 1,2,\dots,p$)
X_{SNP_i}	SNP as fixed effects in the longitudinal model (i-th SNP where $i = 1,2,\dots,p$)
b	$b = (b_1,\dots,b_R)$ latent random effects vector
$\emptyset_r$	dispersion parameter
h_r^{-1}	link function in GLMM for r-th marker

$x_{r,j}$	covariates (fixed effects) for r-th marker
$x_{SNP_i,j}$	covariates (fixed effects) for i-th SNP marker
$z_{r,j}$	random effects
α_r	unknown regression coefficients (fixed effects) related to the model for the r-th marker
$\mathcal{MVN}$	multivariate normal distribution
$\mathbb{D}$	covariate matrix
ω_q	mixture distributions which weighted by a factor $q = 1,2,\dots,Q$
φ	density of the multivariate normal distribution
ψ	fixed effects regression coefficients
θ	random effects regression coefficients
f	density of the observed markers
π_g	prior probabilities of belonging to each group
$\mathbb{c}$	Cut off value in LoDA
$\aleph$	total number of samples from a MCMC scheme where $\mathfrak{n} = 1,2,\dots,\aleph$
L_1	lasso penalty
L_2	ridge penalty
λ	positive constant in penalised logistic regression

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

LIST OF FIGURES

Chapter 5

Chapter 6

Chapter 1
Introduction

This thesis seeks to address the need for variable selection in classification (i.e. discriminate samples or individuals between two or more groups) within the context of genetic data analysis using single nucleotide polymorphisms (SNPs). The current chapter provides an introduction to the basic concept of variable selection for classification, discusses the needs for the variable selection and the general view of how this has been so far approached by the research community.

Variable selection for classification can be seen as a problem to choose the variables that carry the most information about the outcome of interest (i.e. signal) while having reasonably high precision (i.e. low variance) so in other words, as the problem of maximising the signal-to-noise ratio (SNR). Therefore,

this chapter will introduce this SNR concept by defining the main terminology mainly taken from engineering and the related concepts from statistics.

Another important concept is classification. This chapter will briefly describe logistic regression as the classification method. In addition, the method of combining SNP and longitudinal clinical data for classification will be discussed.

1.1 Single nucleotide polymorphism data

In this section, the SNP data is defined and introduced. A rich resource of genetic information is provided by single nucleotide polymorphisms (SNPs), which located along the chromosomes where the genetic code tends to vary from one person to another by just a single base [1]. Generally, our DNA sequence is formed from four nucleotide bases namely, Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). The polymorphism appears when for example, at a certain base, the majority of individuals may hold the 'G' nucleotide, whereas some will hold 'T' nucleotide.

The SNPs information are collected for each individual. SNPs represent only one type of genetic data and can be thought of as categorical variables typically showing three levels called genotypes (e.g. AA, AT, TT) [3]. In this thesis, the focus lies on analysing the genetic data consisting of categorical SNPs with binary phenotypes (only two possible outcomes e.g. "Yes" or "No", "0" or "1"). From here on, this type of data will be referred to SNP data.

As an example, an extract of SNP data is shown in Table 1.1. The "PHENOTYPE" column labels the phenotype groups (e.g. cases coded as "1" and controls coded as "0" for binary outcomes). The extended columns to the

right belong to the SNPs. For instance, column "rs2821984_G" is a SNP with G as the minor allele (i.e. less common allele at a SNP). Hence, with C as the major allele (i.e. the most common allele at a SNP), the SNP can have three possible genotypes CC, CG and GG. In this thesis (unless stated otherwise), the data is coded either 0, 1 or 2 for the additive model which represents the count of the minor allele.

Table 1.1: Extract of SNP dataset as an example.

ID	SEX	PHENOTYPE	rs2821984_G	rs2837900_A	...	rs333600_G
1	1	1	2	0		0
2	2	1	1	1		2
3	1	1	1	0		2
4	2	1	1	1		0
5	1	0	2	2		1
6	1	0	0	2		1
7	2	0	0	1	...	1

1.2 Variable selection for classification in genetic studies

Analysis of genetic data using SNPs has become an important area of research due to its association with complex diseases and different reactions to medications and treatments [4]. SNP datasets are often very large consisting of millions or hundreds of thousands of SNPs and are high-dimensional. High-dimensional data refers to the case when the number of variables, p (i.e. the number of SNPs) is much greater than the number of samples (or individuals), n $(p \gg n)$.

Over the years, many statistical and machine learning methods have been applied to SNP data for classification. Such classification aims to assign each sample correctly to the group it belongs to while using all the SNP data. For

example, in the situation of binary phenotypes, one is either interested to classify the samples into cases (e.g. disease, negative response to treatment) or controls (e.g. healthy, positive response to treatment). The machine learning method however do not do variable selection, instead they find a best approach to use all the SNP data to discriminate patients. Such an approach does not allow to reduce the dimensionality of the data and extra caution should be taken towards validation.

However, there are biological arguments that only a subset of SNPs is linked to the phenotype of interest. Then a subset of SNPs is investigated for association with the phenotype which normally leads to model building for classification. From an analysis point of view, selecting the most informative SNPs and best model is best cast as a statistical problem of variable and model selection [5, 6]. Hence, dealing with the high-dimensionality and very large number of variables in SNP data, raises the needs for feasible, computationally non-complex and flexible variable selection methods or frameworks that can aid in achieving good classification accuracy. Due to the challenge in dealing with high-dimensional data, research in this direction continues.

Variable selection is important in several ways; (i) to reduce the computational time and space required to run specific algorithms, (ii) to improve the performance of classifiers, i.e., by removing the noisy or irrelevant variables and by reducing the likelihood of overfitting to noisy data, and (iii) to identify which variables may be relevant to a specific response (i.e. to identify which SNPs associate with the phenotype) [7]. In this respect, three main variable selection methods for classification, namely, filter, wrapper and embedded are widely discussed in the literature. Many studies have reviewed these approaches in the

context of genetic studies (see for example, Guyon and Elisseeff (2003) [8], Saeys et al. (2007) [9], Schwender et al. (2008) [3] and Hira and Gillies (2015) [10]) and a detailed review of this literature is presented in Chapter 2 (Literature Review).

The importance of variable selection is significant as in most cases it helps to increase classification accuracy. With that in mind, the three methods of variable selection are considered in the development of a novel variable selection method and a modelling framework to analyse SNP data. In this thesis, the hypothesis is that good classification accuracies may be achieved by selecting SNPs that carry high signal as compared to noise (i.e. high signal-to-noise ratio). Furthermore, the main contributions of this thesis in variable selection for classification are threefold: i) development of a novel filter metric tSNR which is based on the signal-to-noise ratio, ii) application of the method in both univariable and multivariate settings, and iii) investigation of the contribution of variables selected by tSNR on classification accuracy when combined with longitudinal clinical data.

1.3 Signal-to-noise ratio

In science and engineering, the signal-to-noise ratio (SNR) is a measure that compares the level of a desired signal to the level of background noise [11]. In the context of magnetic resonance imaging (MRI), SNR is conceptualised by comparing the signal of MRI image to the background noise of the image. Statistically, in the binary outcomes scenario, SNR of a variable is defined as the ratio of the difference in mean between the two groups over the standard deviation from the two groups [12]. In genetic studies, the t-test is a commonly used variable selection method that applies the SNR concept to identify the most informative genes. In their paper, Mishra and Sahu [13] applied the SNR ranking

generated from t-tests to select the most informative genes for cancer classification.

As we know, the t-test is applied to compare two groups of continuous data (e.g. microarray data). However, in their study, Zhou et al. [14-16] have proposed to modify the t-test to fit categorical SNP data. The three values of SNPs are transformed into binary form which allows the calculation of mean and standard deviation in each group. The t-scores are then used to rank the SNPs from the highest to the lowest values.

On the other hand, generalised SNR was recently proposed by Czanner et al. [17] within the context of generalised linear models (GLM) in the recording of single neurons. In the paper, generalised SNR was not introduced as the variable selection method *per se*, but rather as a measure of system fidelity in neural spiking activity. Hence, following this notion, this thesis generalises SNR specifically to work with the binary outcomes and categorical SNPs in SNP datasets. The generalised SNR is applied in the logistic regression framework due to its ability to deal with a categorical outcome. The proposed generalised SNR (from here on will be referred as tSNR) is used as the ranking measure for SNPs. The tSNR is further extended to enable its application as a model selection criterion. A detailed description of the developed method, tSNR is presented in Chapter 3 (Methods).

In this thesis, the direct application of SNR to SNP datasets is challenging for several reasons. First, the number of SNPs can be extremely large (typically half a million) and the effect of SNPs on the phenotype are often small to modest, so the SNR of each SNP is low [18]. Second, it is not possible to consider all SNPs

to develop a model (i.e. filtering method is required). Third, due to a large number of SNPs, a stopping criterion should be developed and imposed on the maximum number of SNPs that can be included in the final model. Fourth, since this method is built outside of the classification algorithm (i.e. it is a filter method), the performance of the variable selection might be independent to the choice of classifier. Hence, it is crucial to select a good classifier that can produce good classification performance. Fifth, misclassification tends to be much higher for binary outcomes than continuous outcomes [18].

1.4 Logistic regression as the framework

Logistic regression (LR) is a classical approach that can be used to test for associations between SNPs and a specific phenotype. This is due to its ability to deal with categorical outcomes or specifically binary outcomes that this thesis is concerned with. The popularity of LR is that it is widely available in most statistical packages and the application of the method is well accepted in many fields [19].

In this thesis, LR is not only used as the framework for the filter metric tSNR, but also as the classification method to assign the samples (or individuals) to their respective phenotype group. In terms of the proposed univariable filter metric tSNR, the deviances (null and residual) from logistic regression are used to calculate the tSNR value. The null deviance from the LR model shows how well the response variable is predicted using only the intercept. Meanwhile, the residual deviance tells how well the response is predicted with inclusion of one SNP.

In the classification problem, LR measures the relationship between the outcome (cases usually coded as '1') and one or more independent variables, by estimating probabilities using its underlying logistic function. These probabilities (which take values between 0 and 1) are then transformed into either 0 or 1 (binary outcomes) according to the probability threshold specified. In this thesis, multiple logistic regression and penalised logistic regression are used with the usual threshold 0.5 is applied.

1.5 Combining SNP and longitudinal clinical data

Apart from variable selection, this thesis also focuses on jointly modelling SNP data and longitudinal clinical data, which could potentially improve the classification performance. Genetic studies have shown that SNPs have become essential variables to consider for predicting an individual's belonging to a particular class of complex diseases and different reactions to medications and treatments [4, 20]. The outcome usually is influenced by both the genetics information and interaction between clinical and environmental variables.

In general, clinical data are often available, and their predictive value is well-validated in the literature [21]. Often, the clinical data are collected once for each patient (cross-sectional). This may be a feasible approach when the diagnosis of a patient is of interest. However, if the effect of therapy or any time-dependent response is of interest, then it is more appropriate to include the longitudinal profiles as well [22].

Predicting the risk of individuals to develop a disease or have a particular response to treatment given their genetic sequence (e.g. SNPs) is a desirable

goal, yet the current ability to make such predictions is relatively poor on its own [23]. One explanation is in the complexity of the data structure; spatial, high-dimensional and categorical. Secondly, the challenge lies on how to best combine the SNP data and longitudinal clinical data. In this thesis, the proposal to jointly model the SNP data (which is cross-sectional) and longitudinal clinical data for the purpose of classification will be discussed in Chapter 6 (Combining SNP Data and Longitudinal Clinical Data).

1.6 Motivations

This thesis is mainly motivated by the fact that variable selection is important when dealing with high-dimensional genetic data. On one hand, variable selection is important to researchers who aim to improve classification performance (or prediction accuracy). On the other hand, variable selection is needed to understand the relationships within the dataset. This is because there are datasets with many variables of which their relationship with the outcome of interest is not known by the experts in the field. Thus, any improvement in this area can represent an important advancement in genetic association studies, with implications to other fields.

In addition, variable selection has been a challenge for researchers of genetic studies due to its computational complexity. As the capacity of the computer and other technologies (e.g. genotyping technologies) increases, the dimensionality of the collected data also becomes higher. However, analysing the large number of variables in the multivariate analysis is challenging. Hence, it is important that a proper framework for variable selection and model selection in analysing the data is proposed.

1.7 Aims of thesis

The aim of this thesis is threefold:

(i) To develop a quantitative variable selection method for classification that satisfies several criteria; (i) it can be easily extended to multivariate applications, (ii) it is computationally inexpensive, and (iii) it is easy to understand and to apply.

(ii) To propose a multi-step approach that selects SNPs and evaluates the classification performance of the corresponding models in a cross-validation framework.

(iii) To jointly model the longitudinal clinical and SNP data for classification using the SANAD dataset.

1.8 Structure of the thesis

This section describes the thesis structure. A detailed literature search that has been undertaken is discussed in *Chapter 2.* In the literature review, three main variable selection approaches that are used in genetic data analysis namely, filter, wrapper and embedded are described. The gaps in the current literature are discussed which leads to the proposal of a novel variable selection method in *Chapter 3.*

The methods chapter *(Chapter 3)* is divided into two main areas; i) variable selection method for classification, and ii) assessment of the classification performance in longitudinal discrimination problems. The existing filter or ranking measure is first described. The concept of signal-to-noise ratio, which functions to capture the signal of a single variable, is discussed. An extension of the metric called tSNR, which can be used in the multivariate setting, is

explained. The proposal to jointly model the longitudinal clinical and SNP data using Longitudinal Discriminant Analysis (LoDA) is described at the end of the chapter. For each application, the classification performance is measured to evaluate the efficacy of the model.

The utility of tSNR is shown using simulated datasets in *Chapter 4*. Ten replicates of SNP datasets are simulated using HAPGEN v2.0 [24] software by setting different objectives. The performance of the filter metric tSNR is measured based on its ability to capture causal SNPs. Furthermore, the classification performance is compared using different classification methods (e.g. penalised logistic regression and stepwise logistic regression).

Chapter 5 focuses on the application of tSNR to a real clinical scenario, the Epilepsy Pharmacogenomics (EpiPGX) study which comprises of two cohorts. In this chapter, the phenotype of interest is defined as the remission status of patients after receiving first well-tolerated antiepileptic drug (AED). The analysis includes the selection of the most informative SNPs by applying univariable and multivariate approaches to the first cohort from the EPiPGX dataset (development set). The results gathered from the development set are then evaluated on the second cohort of the dataset (validation set).

The application using tSNR is further continued in *Chapter 6* on the SANAD dataset. In this study, the interest is to identify patients who will not achieve remission from seizures within five years of commencing treatment diagnosis. Patients who achieve a continuous 12-month period free from seizures within five years of diagnosis are regarded as being in "remission," whereas patients who do not are referred to as "refractory" [25]. In this chapter, the SNPs selected

by tSNR are jointly modelled with longitudinal clinical data as explained in the methods chapter.

The thesis is concluded in *Chapter 7* which highlights the conclusions from the analyses together with recommendations for its use. Also, the limitations are discussed and further work is suggested.

Chapter 2

Literature Review: Variable Selection Methods for Classification with Application to SNP Data

2.1 Introduction

Analysis of genetic data has become increasingly popular for studying complex human disease. The studies examine the associations between single nucleotide polymorphisms (SNPs) and the complex human diseases and different reactions to medications and treatments [4, 26]. The information from hundreds or thousands of individuals with their health status (e.g. healthy or affected) or treatment response (e.g. positive or negative response to treatment) are collected. Each individual is then genotyped at millions of SNPs. Although, there are also continuous or time-to-event outcomes, in this thesis the focus lies on the genetic data with binary phenotypes.

The availability of SNP datasets attracted the interest of researchers due to their ability to identify patterns of data that vary systematically between individuals with different phenotype groups [27]. For instance, in a breast cancer study [28] of a series of individuals who were diagnosed with breast cancer, a higher frequency of a particular SNP allele or genotype can be observed as compared to healthy people. In disease specific studies, the SNP which correlates with the outcome is called "disease SNP". However, in this study, the pharmacogenetics data is also utilised of which the outcome of interest is person's response to medication. Hence, for easier explanation, from here onwards this particular SNP will be referred to as the causal SNP (i.e. the SNP that increases the risk of a specific outcome of interest).

Classification is an important problem and it often relates to a situation where the aim is to predict whether an individual belongs to a certain phenotype group or class. The process of classification can be based on a supervised learning process where, in order to assign the individual to a group, information from individuals who already belong to a specific group is used. However, the analysis of SNP datasets is complicated due to the large variable space, which poses computational time complexity and low accuracy [4]. The challenge of using SNP datasets is that SNPs have only three possible levels, commonly denoted as AA, AB and BB, which gives a SNP dataset a different structure from one containing continuous variables [29]. This extreme sparsity of coverage for discrete variables is a challenge when assessing the relationships between SNPs and the phenotype.

When applied to SNP datasets, variable selection methods can play a crucial role in classification. Not only do they aim to solve the problem of high dimensionality, they also aim to reduce computation time due to data complexity

and to achieve good levels in classification accuracy [4]. There is research on variable selection and classification methods with application to SNP datasets [3, 4, 30, 31]. These publications have highlighted that identifying the variables that offer a good level of discrimination and selecting an appropriate classification method is equally important. However, it is important to note that there is no universal, optimal classification method that fits every dataset [4].

This chapter provides an overview of variable selection methods in the context of classification with specific applications to SNP datasets. The objective of this review is twofold:

(i) To describe the background of existing variable selection methods for classification in this area.

(ii) To explore the applications of several classification methods to SNP datasets (focusing on binary phenotype).

This chapter is organised as follows. Section 2.2 gives the overview of genetic data and the statistical challenges which arise in the analysis of high-dimensional data. Then, description of the simulated dataset used for this chapter is given in Section 2.3. In Section 2.4, the variable selection methods are summarised and the application of each method is shown using the simulated dataset. Afterwards, several classification methods that are well-known in dealing with SNP data are explained in Section 2.5. The chapter is concluded in Section 2.6.

2.2 Overview of SNP data

In this thesis, the focus lies on SNP data from Genome-wide Association Studies (GWAS) not on data with targeted gene (i.e. candidate genes association studies)

which is often smaller. Therefore, in this section, the concept of GWAS is introduced as well as the statistical challenges related to it.

2.2.1 Genome-wide association studies

The main objective of Genome-wide Association Studies (GWAS) is to test the relationship between the genetic variants along the genome and a specific phenotype [32]. GWAS are widely used to investigate genetic association with complex diseases such as cancer, type II diabetes and epilepsy. Nowadays, the studies are also extended to the area of pharmacogenetics where the main aim is to find the association between genetic variants and response to treatment, in terms of either efficacy or toxicity [33].

Specifically, GWAS use SNPs as genetic markers, as they are easy to type and abundant in the human genome [34]. With the currently applied technologies, a large number of SNPs, sometimes in excess of one million are investigated within a single study.

2.2.2 Single nucleotide polymorphisms

Whilst the majority of the human DNA sequence is identical between individuals, some locations along the genome differ from one individual to the next in terms of the nucleotide held at that location. As introduced in Chapter 1 (Introduction), our DNA sequence is formed from four nucleotide bases namely; Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). The polymorphism appears when for example, at a certain location, the majority of individuals may hold the 'T' nucleotide, whereas some will hold the 'G' nucleotide (as illustrated in Figure 2.1). The different possible nucleotides at the location are known as 'alleles'. If the variant allele occurs in at least 1% of the population it is known

as a Single Nucleotide Polymorphism (SNP)[35, 36]. SNPs occur every 100 to 300 bases along the 3-billion-base human genome.

Typically, there can be three categories of genotype at a SNP. For example, SNP with major allele A and minor allele T can have three possible genotypes AA, AT and TT. The genotypes can be summarised in a 2×3 contingency table of the genotype counts for each phenotype (see Table 2.1). Essentially, in an association study, under the null hypothesis of no association with the disease, the genotype frequencies are expected to be approximately the same across phenotype groups (e.g. cases and controls groups) [37]. As a result, a test of association is given by a simple χ^2 test for independence of the rows and columns of the contingency table.

Table 2.1: Contingency table of genotype counts.

Genotypes / Phenotype	AA	AT	TT
Group 1	m_{11}	m_{12}	m_{13}
Group 2	m_{21}	m_{22}	m_{23}

Some diseases (Mendelian diseases) are caused by variation within a single gene, meanwhile for complex diseases and traits, several genetic variants, environmental factors and their interactions are often at play [38]. The goals of the analysis of the association between genetic data and phenotype can therefore be summarised as follows [38]:

(i) To identify interaction between two or more SNPs (i.e. SNP-SNP interactions) or interaction between SNP and environmental factors whose distribution differ substantially between phenotype groups.

(ii) To find SNPs that show a coherent pattern.

(iii) To classify new observations (i.e. patients) into their respective phenotype group based on the SNPs information.

2.2.3 Statistical challenges

Here, the statistical challenges that arise from the perspective of variable selection for classification are discussed. The statistical challenges for analysing SNP data are described in detail in Liang and Kelemen (2008) [39]. In GWAS, researchers often have to analyse high-dimensional datasets, where the number of variables, p is much greater than samples, $(p \gg n)$. Specifically, SNP datasets may consist of hundreds of thousands or even over a million SNPs that are assessed per individual [34]. From there, models are developed to classify samples (or individuals) to their specific phenotype. Normally, a modelling method (e.g. least squares regression) tries to fit a complex model with a large number of variables as perfectly as it can. However, this situation leads to overfitting which essentially means the model follows the error, or noise too closely [40]. Overfitting has become the main issue when discussing high-dimensional datasets due to the concern of inaccurate estimates of outcome on new observations that were not part of the modelling process [40-42].

Typically in GWAS, many of the SNPs will be highly correlated which can reduce the power of the identification of small to moderate genetic effects for complex phenotypes. The condition known as Linkage Disequilibrium (LD) occurs when SNPs are dependent on each other [34]. Certain statistical measures (e.g. D' or R^2) are widely applied to investigate the correlation between SNPs.

As we know, the SNP data is categorical with only three possible levels, AA, AT and TT. The interchangeable values from one level to the other brings a different

meaning entirely (i.e. each level represents 0, 1 or 2 copies of minor allele for additive model). This gives a SNP dataset an entirely different structure than the one containing continuous variables. If plotted, the categorical values will lie on the edges and vertices of a high-dimensional hypercube rather than on a subset of a continuous space [29]. Hence, the extreme sparsity of coverage for discrete variable suggests intuitively that it may be more difficult to discover relationships.

Reproducibility has become a major issue in genetic association studies for complex phenotypes [8, 39, 43]. This problem is observed in a situation when a set of SNPs show highly significant associations with a phenotype group of interest through one method, while not showing these associations when using a different method. In a different situation, the SNPs may be significant in one dataset with a number of samples but show totally different results when tested on different samples with similar SNPs (external validation).

Moreover, due to the low prior probability of causality for each SNP in the genome and the large number of SNPs being tested, rigorous thresholds of statistical significance are needed for genetics association studies in order to ward off a deluge of false positive outcomes. Reducing the number of false positives, while maintaining acceptable power, is often required in biological or biomarker discovery applications since follow-up experiments can be costly and laborious [41]. In statistical hypothesis testing, the aim is to reject the null hypothesis when it is very unlikely (to a certain degree) that the null hypothesis is true. The threshold varies by study, but the conventional threshold in GWAS relates to p-values less than 5×10^{-8} to be statistically significant (i.e. genome-wide significance) [44]. The threshold is based on the estimated number of

independent tests in the genome if all common SNPs in HapMap are tested with direct genotyping or imputation [120,155].

Realising the need to overcome these challenges, the development of statistical and computational algorithms for variable selection of large genetic datasets has become a key area of research. Thus, in the next section the discussion includes available variable selection methods for classification with specific application to SNP data and how these methods could address some of the said statistical challenges, if not all.

2.3 Methods

This section provides the descriptions of methods used in this chapter. A literature search is done in order to provide detailed explanations of existing variable selection and classification methods in the genetic studies. To provide better understanding of each method, the method is applied on a simulated dataset.

2.3.1 Literature search

The literature search for relevant studies is undertaken through the MEDLINE (Ovid) database. The literature search applies the following terms: "classification" combined with "variable selection" and "single nucleotide polymorphism". The search is limited to English language publications between 2005 to 2015. In the initial search, 795 relevant titles for variable selection and classification with specific applications to SNP data were identified. From the 795 articles, 108 were included based on their title and abstract. However, further studies are identified and included if related by examining the reference lists of all the included articles.

2.3.2 Simulated dataset

The methods are applied to a simulated dataset to give better overview of each method. The data simulation is done using genome-wide simulation software, HAPGEN v2.0 [24] with HapMap3 CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) as the reference panel. Essentially, the data are simulated with similar allele frequencies and linkage disequilibrium structure to the reference panel. The simulated dataset consists of 1,000 cases and 1,000 controls with 116,415 SNPs on chromosome 1. In order to see each method's ability to capture the most important SNPs, five causal SNPs are assumed under a log-additive model with high odds ratio [45] as follows:

Table 2.2: The details of causal SNPs which are simulated under a log-additive model with high odds ratio.

SNP	Base-pair Position	Risk allele	Heterozygote disease risk	Homozygote disease risk
rs966321	4215064	1	3.00	5.25
rs914717	156952983	1	2.20	3.00
rs12046196	228881820	1	5.25	10.25
rs1130193	200252354	1	10.25	28.50
rs10888878	55015907	1	15.25	28.50

2.3.3 Sample and genotyping QC

Sample and genotyping quality control (QC) is undertaken as standard data pre-processing procedure using an open-source whole genome association analysis toolset, PLINK 1.9 [46]. Applying standard QC procedures to each SNP, this number is first reduced to 96,697 after applying GWAS thresholds based on minor allele frequency (MAF), SNP genotyping rate and test Hardy-Weinberg Equilibrium (HWE). The screening on MAF only includes the SNPs with MAF

>0.01. Low MAF SNPs could be more susceptible to genotyping errors and their association signals are less robust [47]. For SNP genotyping rate only SNPs with <10% missing genotypes are included. Further, SNPs that are extremely deviated from HWE (p-value $<10^{-6}$) are removed. In principle, a population is in HWE when there is a fixed relationship between allele and genotype frequencies over generations [34]. Hence, deviations from this relationship suggests that there may be quality problems in the genotyping procedure. At the same time, all 2,000 samples passed the standard QC procedure (based on rate of missingness, duplication of samples, relatedness and heterozygosity).

2.3.4 Data pruning

Linkage Disequilibrium (LD) pruning is an important quality assurance step for GWAS analysis. Some tests for association will obtain better results if the markers used are not in LD with each other [48]. LD is defined as an association in the alleles present at each of two sites on a genome [35]. Therefore, the pruning option is undertaken to reduce or eliminate the SNPs that is in approximate LD with each other which intuitively can help minimising the computational complexity. Also, it may help focusing on more signals and allow more region of potential interest.

The pruning option (150 50 0.90) is implemented using PLINK 1.9 [46] software. For this method, all pairs of SNPs within a window of 150 SNPs, 50 SNPs are compared with each other to measure their pairwise LD. If any pair of SNPs within the window are in LD greater than R^2 threshold of 0.9, the first SNP in the pair will be inactivated (pruned). The window and step options (150 and 50 accordingly) are chosen to speed up the pruning process. Meanwhile, the threshold of 0.9 (higher threshold) is chosen to avoid reducing too many SNPs

during the pruning process. In general, the lower the selected threshold, the more SNPs will be pruned [48]. After applying the LD pruning, the simulated data consists of 50,178 SNPs.

2.4 Summary of variable selection methods for SNP data

In this section, the variable selection methods for SNP data are summarised. To provide a clearer view of each method, the discussion includes some examples by using the simulated dataset. Variable selection methods are commonly used to select relevant variables for model construction. In terms of classification, variable selection is an important process to identify variables that can be used to accurately classify individuals (e.g. assigning individuals to their respective phenotype group). In recent years, data can offer a comprehensive picture of the complexity of biological systems at different levels and there has been a growing interest to answer specific biological questions (e.g. which SNPs caused the disease and which SNPs relate to the individuals' response to treatment) [49]. In genetic association studies, there is an emerging need to develop strategies for selecting sets of SNPs likely to be relevant to phenotype group of interest so that poor performance of classifiers can be avoided [50, 51].

In the context of classification, three types of variable selection methods which are frequently discussed are filter, wrapper and embedded [3, 7-10, 52]. These papers on bioinformatics discussed the variable selection methods in detail with a few selected algorithms listed in each method. The discussions are broad which include applications that are not limited to SNP data. In this review, the discussion is limited to the variable selection methods for classification using

SNP data and whether the methods are addressing the challenges that we have discussed in the previous section (Section 2.2.3). Figure 2.1 illustrates the three categories of variable selection methods.

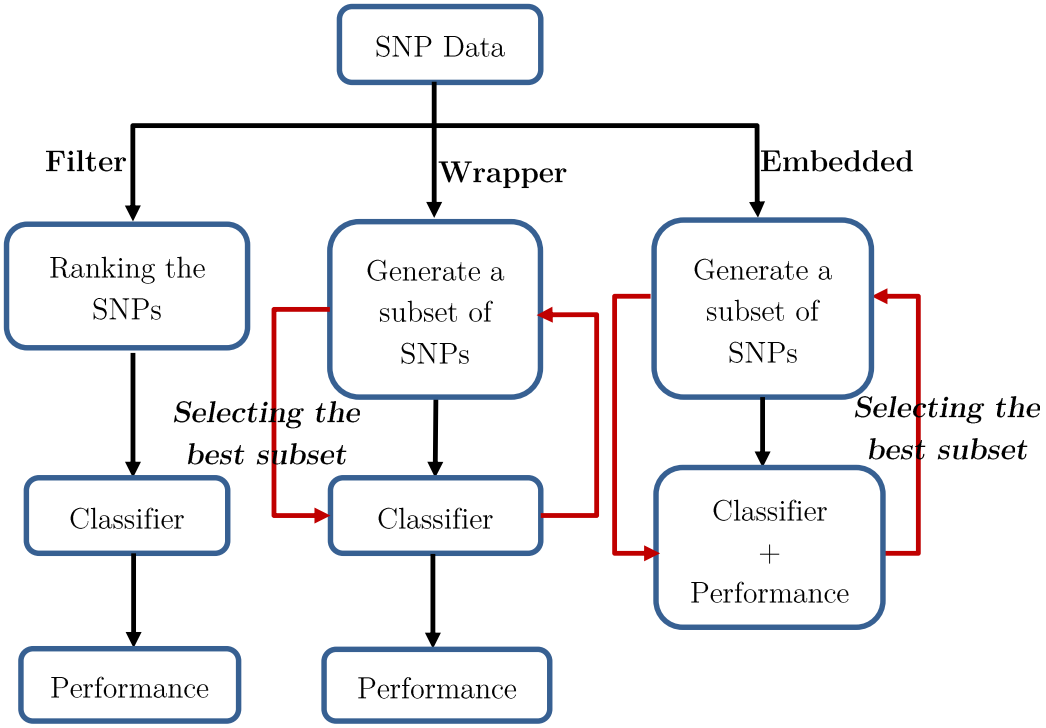


Figure 2.1: Workflow of variable selection methods for classification; filter, wrapper and embedded [53].

2.4.1 Filter methods

Filter methods identify the relevant SNPs by eliminating the uninformative SNPs. Filter methods are known to be fast, scalable (i.e. computationally efficient since it may involve one SNP at a time) and independent of the classifier [9, 39, 54]. A typical procedure for a filter method is using a conventional p-value approach (e.g. Pearson Chi-squared and Fisher Exact tests). The SNPs are ranked by the lowest to the highest p-value to represent the importance of each SNP in relation to the phenotype group. Filter methods are considered as pre-processing steps since they are independent of the choice of classifiers [8]. This

is a particular advantage of filter methods because the chosen SNPs might perform differently from one classifier to another.

Another advantage of filter methods is that they are computationally inexpensive and easy to implement. However, most proposed filter methods are univariable. This means that each variable is considered separately, thereby ignoring correlation structures among SNPs, which may lead to worse classification and prediction performance when compared to other types of variable selection methods [39]. SNPs might be non-significant when analysed on their own, but may become significant when analysed in combination with other SNPs (due to their interaction). In order to overcome the problem of ignoring variable dependencies, a number of multivariate filter methods [4, 5, 23, 55] have been proposed, aimed to some degree, at the incorporation of variable dependencies.

Logistic regression is a classical approach to test for association between SNPs and phenotype in genetic association studies due to its ability to deal with categorical outcome. The general model of logistic regression that an individual has a particular disease can be written as,

$$Pr(Y = 1|X) = \frac{e^{\beta_0+\beta_i X_i}}{1 + e^{\beta_0+\beta_i X_i}} \tag{2.1}$$

where β_0 is the intercept, meanwhile β_1 represents the coefficient for each of the SNP, X_i fitted in the logistic regression model. The coefficient is defined as the additive increase in the log of the odd ratios resulting from a one-unit increase in X_i. The values of $X_i, i = 1,2, \dots, p$ can be represented by different values depending on the genetic mode of inheritance assumed. The different possible values are specified in Table 2.3.

Table 2.3: Genotype coding (values of X_i) according to genetic model.

	AA	AT	TT
Recessive model	0	0	1
Dominant model	0	1	1
Additive model	0	1	2
Codominant model			
Variable 1	0	1	0
Variable 2	0	0	1

A nominal p-value can be calculated for all SNPs $i,\ i = 1,2,\dots,p$. Although usually applied as a univariable filter method, equation (2.1) can also be easily extended to a multivariate setting. The SNPs will go through a ranking procedure that allow the selection of a subset of SNPs to be used in the following step (i.e. classification). One may select some limited number of SNPs regardless of their overall significance [56]. For instance, one may select the SNPs with the top 1% or 5% of p-values.

For illustration, the logistic regression analysis is performed on the 50,178 SNPs using the --logistic function in PLINK 1.9 [46]. The output mainly contains the odds ratio, coefficient t-statistics and asymptotic p-value for t-statistics of each SNP. The associated p-values of the SNPs are shown in the Manhattan plot (Figure 2.2). The plot shows $-\log_{10}$ of the p-value for each SNP against its position in chromosome 1. The five simulated causal SNPs which reached genome-wide significance, are highlighted in green.

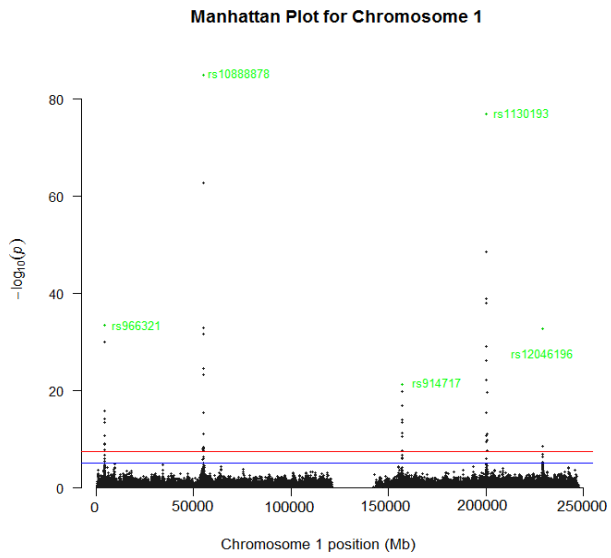


Figure 2.2: Manhattan plot for 50,178 SNPs for chromosome 1.

The horizontal red line and blue line represent the genome-wide significance threshold of p-value $< 5 \times 10^{-8}$ and suggestive line significance threshold of p-value $< 5 \times 10^{-5}$ accordingly. Next, the corresponding p-values are sorted increasingly and regard SNPs at the beginning of the list as the most significant ones.

The ranking of the top 20 SNPs are shown in Table 2.4. Indeed, the simulated causal SNPs are listed among the top 20 SNPs (those highlighted in bold font). Subsequently, the top SNPs may be used for developing the model for classification. The new sample will then be assigned to a specific phenotype group given the genotype values of the SNPs used in the classification model. The results for classification performance using logistic regression will be presented in Section 2.5.

Table 2.4: Top 20 SNPs based on p-values from logistic regression model. The simulated causal SNPs are highlighted in bold font.

1	2	3	4	5	6
rs10888878	**rs1130193**	rs2286202	rs10920304	rs2819362	rs3820439
1.18×10^{-85}	1.31×10^{-77}	2.08×10^{-63}	2.90×10^{-49}	1.24×10^{-39}	1.19×10^{-38}
7	8	9	10	11	12
rs966321	rs2270001	**rs12046196**	rs1180964	rs2411738	rs2819365
4.40×10^{-34}	1.57×10^{-33}	2.01×10^{-33}	2.16×10^{-32}	1.30×10^{-30}	7.65×10^{-30}
13	14	15	16	17	18
rs2735784	rs2270003	rs2527848	rs2819360	rs9427715	**rs914717**
6.10×10^{-27}	2.93×10^{-25}	5.93×10^{-24}	6.14×10^{-23}	8.95×10^{-23}	6.40×10^{-22}
19	20				
rs2511200	rs4950760				
1.92×10^{-20}	2.61×10^{-20}				

In contrast to logistic regression which can be applied directly to the categorical SNP data, *Modified* $\boldsymbol{t}$*-test* and $\boldsymbol{F}$*-statistics* apply mean approximation of the SNP data. The nominal values of the genotypes are transformed into a set of vectors to calculate the mean. These methods were introduced by Zhou and Wang (2007) [14]. Similar to other filter methods, the study suggests that before proceeding with a classification procedure, it would be best to undergo a variable selection procedure in order to produce good classification results. The study first ranks each SNP using a ranking measure. Then from the ranking list, different SNP subsets are formed by sequentially choosing different numbers of SNPs (e.g. 5, 10, 50, 100, 500) with top ranking values.

Each of the subsets is then used in a classifier. The study chooses Support Vector Machine (SVM) which apply the information from the subset of SNPs to classify the samples (i.e. patients) into their respective groups. SVM is chosen due to its attractive features, such as effectively avoiding overfitting and can accommodate large feature spaces and fast [14]. Further usage of SVM is explained in Section

2.5 Classification. The classification performance is measured using sensitivity, specificity and probability of correct classification. The best subset of SNPs is later determined based on the highest values of classification measures.

The t-score is defined for SNP i to be the greatest t-score for all phenotype groups, $g = 0,1$ (0 for controls and 1 for cases) for SNP i:

$$t_i = max\left\{\frac{\left|\vec{\bar{X}}_{ig} - \vec{\bar{X}}_i\right|}{M_g S_i}, \qquad g = 0,1,\dots,G-1\right\} \tag{2.2}$$

$\vec{\bar{X}}_{ig}$ and $\vec{\bar{X}}_i$ are two row vectors indicating the mean of i-th SNP in the g-th phenotype group and the mean for all groups. $\left|\vec{\bar{X}}_{ig} - \vec{\bar{X}}_i\right|$ denotes the Euclidean distance of the two vectors. In this equation, the categorical values of 0, 1 and 2 of each SNP are transformed into a vector of three dimensions, i.e., $0 \Rightarrow \vec{X}_i^{(1)} = \{1,0,0\}$, $1 \Rightarrow \vec{X}_i^{(2)} = \{0,1,0\}, 0 \Rightarrow \vec{X}_i^{(3)} = \{0,0,1\}$.

$$M_g = \sqrt{\frac{1}{n_g} + \frac{1}{N}} \tag{2.3}$$

$$S_i^2 = \frac{1}{N-G}\sum_{g=0}^{G-1}\sum_{j\in g}\left(\vec{X}_{ij} - \vec{\bar{X}}_{ig}\right)\left(\vec{X}_{ij} - \vec{\bar{X}}_{ig}\right)^T \tag{2.4}$$

Here, $\vec{X}_{ij}$ refers to the vector of the i-th SNP of the j-th sample; N is the total number of samples in all phenotype groups; n_g is the number of samples in the particular phenotype group, g and S_i is the within-group standard deviation. The mean difference in the formula indicates the dispersion of the mean of phenotype group of interest from the mean of all phenotype groups. The t-score is calculated for each SNP which indicates a signal that a specific SNP carries.

Among the total of p SNPs, the higher the value of the t-score means the more informative the SNP is.

Using one of the simulated causal SNPs, rs10888878 as an example, Table 2.5 shows the transformation done for each nominal value of the SNP which allows t-score to be calculated.

Table 2.5: Calculation of t-score for SNP i, rs10888878 with 2,000 samples.

				Transformation of nominal SNP value into vector		
j	PHENOTYPE	Genotype	Additive	AA	AT	TT
1	1	AT	1	0	1	0
2	1	AT	1	0	1	0
3	1	AA	0	1	0	0
4	1	AT	1	0	1	0
5	1	AA	0	1	0	0
:	:	:	:	:	:	:
1000	1	TT	2	0	0	1
$n_1 = 1000$			$\bar{\vec{X}}_{i1}$	0.457	0.529	0.014
1001	0	AA	0	1	0	0
1002	0	AA	0	1	0	0
1003	0	AA	0	1	0	0
1004	0	AT	1	0	1	0
1005	0	AT	1	0	1	0
:	:	:	:	:	:	:
2000	0	AA	0	1	0	0
$n_0 = 1000$			$\bar{\vec{X}}_{i0}$	0.933	0.065	0.002
$N = 2000$			$\bar{\vec{X}}_i$	0.695	0.297	0.008

From here, the within group standard deviation, S_i is calculated for each SNP. Then all the values are plugged into equation (2.2) which will produce two values of t-scores (one for each phenotype group). Table 2.6 shows the t-scores of 20 SNPs in descending order. Similar with the p-value ranking, the t-score ranking managed to capture the five causal SNPs within the top 20 SNPs ranking.

Table 2.6: Top 20 SNPs based on t-scores from modified t-test. The simulated causal SNPs are highlighted in bold font.

1	2	3	4	5	6
rs10888878	rs2286202	**rs1130193**	**rs12046196**	rs2270001	rs1180964
15.21	11.30	8.52	7.71	7.70	6.71
7	8	9	10	11	12
rs3820439	rs10920304	rs2270003	**rs914717**	rs2819362	rs2819365
5.92	5.87	5.43	5.21	5.09	4.70
13	14	15	16	17	18
rs966321	rs2511200	rs2411738	rs2527848	rs12023371	rs16849483
4.69	4.62	4.47	4.46	4.31	4.12
19	20				
rs2735784	rs10489842				
4.08	3.97				

F-statistics applies the similar concept of ranking measure as the modified t-test. Assuming there are two phenotype groups for a given dataset and each SNP contains two alleles, the F-statistics (F_{st}) value is calculated as

$$F_{st} = \frac{Var_a}{\bar{a} \cdot \bar{t}} \tag{2.5}$$

$$\bar{a} = \sum_{g=0}^{1} a_g \tag{2.6}$$

$$Var_a = \sum_{g=0}^{1} \frac{(a_g - \bar{a})^2}{2} \tag{2.7}$$

where a and t are the two alleles' frequencies, respectively, in each group; $\bar{a}$ and $\bar{t}$ are the mean frequencies of the two alleles across groups; Var_a refers to the variance of one allele and a_g is designated as the frequency of one allele for the g-th group. Principally, SNPs with larger F_{st} values are more significant and will be on top of the rank. The F-statistic involves calculating two single alleles A

and T which can be represented as 0 and 1 respectively. Table 2.7 shows the transformation and the calculation done by using SNP rs10888878 as an example.

Table 2.7: Calculation of F_{st} values for SNP i, rs10888878 with 2,000 samples.

				Transformation of nominal SNP value into vector	
j	PHENOTYPE	Genotype	SNP i	Allele 1	Allele 2
1	1	AT	1	0	1
2	1	AT	1	0	1
3	1	AA	0	0	0
4	1	AT	1	0	1
5	1	AA	0	0	0
:	:	:	:	:	:
1000	1	TT	2	1	1
				$a_1 = 14$	$t_1 = 543$
1001	0	AA	0	0	0
1002	0	AA	0	0	0
1003	0	AA	0	0	0
1004	0	AT	1	0	1
1005	0	AT	1	0	1
:	:	:	:	:	:
2000	0	AA	0	0	0
				$a_0 = 2$	$t_0 = 67$
$N = 2000$				$\bar{a} = 16$	$\bar{t} = 610$

From here, Var_a is calculated for each SNP. Then, all the values gathered are plugged into equation (2.7). Similar to the modified t-test the SNPs are ranked in descending order. Table 2.8 shows the results of F-statistics for top 20 SNPs. Compared to logistic regression and modified t-test ranking, F-statistics can only capture one simulated causal SNP within the top 20 ranking.

Table 2.8: Top 20 SNPs based on F_{st} from F-statistics. The simulated causal SNPs are highlighted in bold font.

1	2	3	4	5	6
rs1130193	rs7540530	rs2795275	rs11117808	rs832521	rs17556883
306.30	294.55	294.31	294.29	293.44	292.46
7	8	9	10	11	12
rs1326005	rs11590608	rs338466	rs7522034	rs11576909	rs3942955
292.14	291.73	291.52	290.89	290.81	290.46
13	14	15	16	17	18
rs4614251	rs560426	rs17348602	rs10799593	rs11163752	rs696859
290.44	290.25	290.08	290.05	290.01	289.97
19	20				
rs2093765	rs4951338				
289.96	289.96				

The filter methods that have been discussed earlier are usually done for each SNP (univariable) i.e. the statistic is calculated for each SNP. However, they are easily extended to multivariable (multiple independent variables) selection by considering a subset of SNPs at a time. Unlike the univariable approach which analyse a SNP at a time, the multivariable approach takes into account the correlation between SNPs. This advantage of the multivariable approach may contribute to an increment in the classification accuracy [23, 57].

2.4.2 Wrapper methods

The classical examples for wrapping in which the classification method itself is used to select the predictors are backward elimination, forward and stepwise selection in linear regression [3]. The wrapper methods enhance the filter methods by wrapping around a particular learning algorithm that can assess the selected subsets of variables in terms of the estimated classification errors and then build the final classifier [9].

The *forward selection* starts with one variable and incrementally adds more variables in the model [7]. Every time a variable is added, the contribution of the variable to the model is evaluated (e.g. classification measures). The optimum subset of variables may be determined by observing any increment of the classification accuracy.

The *backward elimination* method begins with a model that includes all variables [58]. The method then eliminates variables one by one until a subset of variables show a significant contribution to the model. For example, the elimination rule can be based on a classification measure with specific cut-off value.

Similar to forward selection, *stepwise selection* starts with one variable. However, the variables that are already in the model do not necessarily remain [59]. Variables are added and removed depending on their contribution to the classification performance. However, it is important to note that both backward elimination and stepwise selection methods are only computationally feasible when the number of variables are small [7]. For example, Park and Hastie (2007) [60] applied a stepwise selection only after penalised logistic regression was carried out. Here, the classification method is the logistic regression itself. Stepwise selection is applied to further reduce the number of SNPs selected by the penalised logistic regression.

In the wrapper approach, every time a subset of SNPs is introduced to the classifier, classification measures are calculated. This process will help to reduce the number of variables which may improve the classification accuracy. The evaluation of a specific subset of variables is obtained by training and testing a specific model, tailoring this approach to a specific classification algorithm.

Wrapper methods tend to perform better in this setup. However, it also leads to a big disadvantage of the wrapper method, that is the computational inefficiency which is more apparent as the number of variables grows [10].

As an example, we apply stepwise logistic regression (SLR) for wrapper method. Since it is computationally expensive to include all the SNPs at one time, a subset of top 100 SNPs is selected based on the ranking gathered by logistic regression earlier. Akaike Information Criterion (AIC) is implemented as the decision measure to either retain or remove each SNP during the modelling process.

Cross-validation is applied by dividing the samples into training and test sets. 80% of the samples belong to the training data and the remaining 20% to the test data. The process is repeated 100 times from which the values of mean and standard deviation are calculated. Further explanation on classification performance will be described in Section 2.5. The analysis is done using the '`step`' function in statistical software, R [61]. The final model, which consists of 26 SNPs from the subset of 100 SNPs, is determined based on the highest accuracy of classifying between cases and controls patients. The list of the 26 SNPs selected by SLR is shown in Table 2.9.

Table 2.9: The 26 SNPs selected by stepwise logistic regression (SLR). The simulated causal SNPs are highlighted in bold font.

No	SNPs	No	SNPs	No	SNPs
1	**rs10888878**	11	rs926247	21	rs2095769
2	**rs1130193**	12	rs927888	22	rs1028543
3	**rs12046196**	13	rs4915919	23	rs12023371
4	**rs966321**	14	rs4927134	24	rs6656470
5	**rs914717**	15	rs2286202	25	rs7534558
6	rs6429449	16	rs12118215	26	rs3737599
7	rs4661077	17	rs1180964		
8	rs12731187	18	rs2527848		
9	rs771132	19	rs11206502		
10	rs10915476	20	rs2564856		

2.4.3 Embedded methods

Embedded methods work by embedding a variable selection method inside the classifier. For example, penalisation is the most common approach in this technique [51, 60, 62, 63]. Embedded methods have the advantage that they include the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods.

Penalised logistic regression (PLR) is known as an extension of simple logistic regression which applies penalisation to reduce the number of SNPs and later classifies the phenotype groups [3, 4]. Penalisation methods allow the building of more powerful models with a large number of variables in the model. In high-dimensional data analysis, it is possible to fit all p variables in one model using a technique that constrains or regularises the coefficient estimates towards zero [40]. The two renowned methods for shrinking the regression coefficients towards zero are ridge regression and the least absolute shrinkage and selection operator (lasso) methods [64].

Park and Hastie (2007) [60] proposed to maximise the log-likelihood of the simple logistic regression model (2.1). The log-likelihood is maximised using ridge regression which uses L_2 norm as the penalty of the coefficients which minimise the following equation:

$$L(\beta_0, \beta, \lambda) = -l(\beta_0, \beta) + \frac{\lambda}{2} \|\beta\|_2^2 \tag{2.8}$$

where l indicates the binomial log-likelihood and λ is a positive constant. The tuning parameter λ controls the strength of the penalty which will shrink the β towards zero. The `glmnet` function in R applies cross-validation to select the value of λ. The cross-validation error is computed over a grid of λ values [40]. The tuning parameter is selected based on the smallest cross-validation error. In ridge regression, $\|\beta\|_2$ denotes L_2 norm which is defined as $\|\beta\|_2^2 = \sum_{i=1}^{p} \beta_i^2$. In penalisation, the norm measures the distance of β from zero. The study highlights that with L_2 penalisation, none of the coefficients is set to zero which will include all SNPs in the final model. Therefore, stepwise selection is implemented to reduce the number of SNPs in the model which will then improve the classification accuracy and interpretability.

Similar to the wrapper approach, the top 100 SNPs based on the p-value ranking are utilised for PLR. The method is able to reduce the number of SNPs used in the model to five with all simulated causal SNPs included (rs10888878, rs1130193, rs12046196, rs966321 and rs914717).

Wu et al. (2009) [65] overcome the disadvantage of L_2 penalisation, and reduce further the number of SNPs by implementing a lasso based on the L_1 norm in their PLR method. Lasso is a shrinkage technique which will shrink the

coefficient estimates towards zero or exactly to zero. The L_1 norm is defined as $\|\beta\|_1$ where $\|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$ which can replace the penalty in equation (2.8). The study concludes that lasso penalised regression is easily capable of identifying important SNPs in highly correlated data.

2.5 Classification

This section summarises the classification methods which are usually done after the variable selection procedure. Supervised learning is now a well-received approach in statistics, where a specific algorithm is trained on a training samples (training data) with a set of variables and known outcome from which we can build a classifier. The primary aim is for the classifier to perform well not only on the training data but also on test data that are not used to train the classifier [40].

The quality of classification can be assessed using the well-known receiver operating characteristics (ROC) methodology and calculating the area under the ROC curve (AUC) [23, 31, 66]. The AUC can be thought of as the probability that a classifier will correctly predict the phenotype group the sample belongs to. The greater the AUC, the better is the performance of the classifier. Besides AUC, Probability of Correct Classification (PCC), sensitivity, specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are also usually presented to evaluate the performance of specific classifiers. The six classification measures will be reported in this thesis. A detailed description of each measure will be discussed in the next chapter.

One of the possible setups to undertake classification is the following. Assuming we are using the SNPs selected using different variable selection methods

presented earlier. The main objective is to explore the effect of variable selection towards classification accuracy. We assume that there is a relationship between SNPs selected and the performance of classification accuracy. Cross-validation is applied by dividing the samples into training and test sets. 80% of the samples belong to the training data and the remaining 20% to the test data. The procedure is repeated 100 times from which the values of mean and standard deviation of the classification measures will be calculated. Figure 2.3 illustrates the classification strategy implemented on different subset of SNPs selected by filter, wrapper and embedded methods using the simulated dataset.

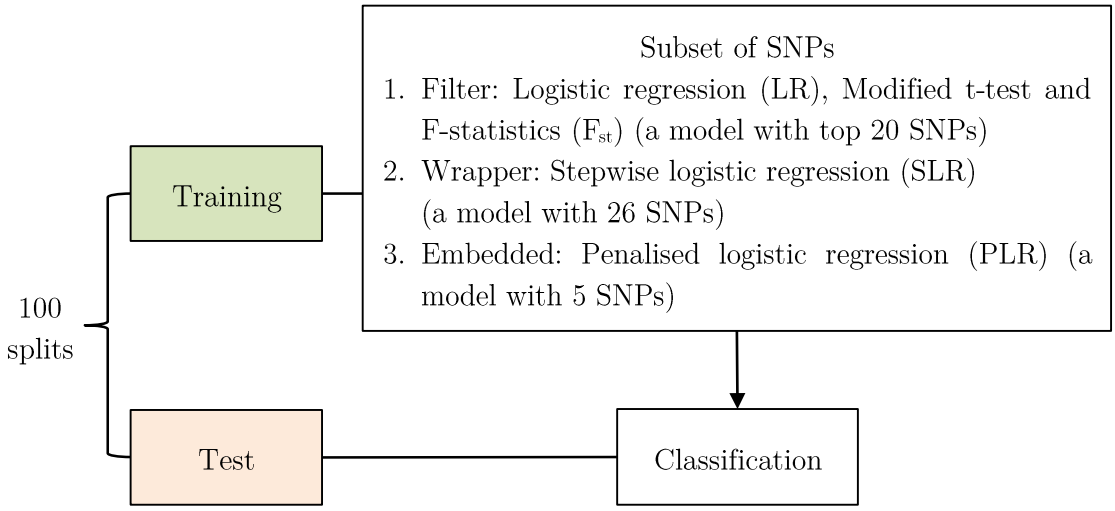


Figure 2.3: Classification strategy with SNPs selected by filter, wrapper and embedded methods.

In GWAS, the few classification methods mainly follow two methodological frameworks; *logistic regression* and *Bayesian principle* [67]. In the first framework, SNPs are modelled as discrete variables (e.g. additive, recessive). Meanwhile in the second framework, SNPs are modelled as ternary categorical variables and no assumptions are usually made on pre-specified genetic models. In what follows, the results of classification performance using logistic regression are compared with those using the Bayesian principle with a Naïve Bayes

classifier. Later, other classification methods that usually apply the Bayesian principle for example KNN, SVM and CART are discussed.

For categorical outcome, Naïve Bayes classification is often used. The word "naïve" means that the variables are independent of each other and conditional on the same outcome [68]. The Bayes' theorem states that the probability of $Y = g$, given the observed SNP, $X = x$:

$$Pr(Y = g|X = x) = \frac{Pr(Y = g) \times Pr(X = x|Y = g)}{Pr(X = x)} \qquad (2.9)$$

where $Pr(Y = g)$ represents the overall or prior probability that a given sample, j is associated with the g-th category of the phenotype group and $Pr(X = x)$ denotes the probability of the sample to have genotype x (e.g. 0,1 or 2 for additive model). $Pr(X = x|Y = g)$ is the density function of X for a sample that comes from the g-th group. Essentially, the learning algorithm in the classifier builds a probabilistic model of the variables and uses that model to predict the classification of a new sample.

Other than classification, the Bayes approach can also be utilised to rank variables from which the top-ranked will most likely contain the most informative variables for prediction of the underlying phenotype group [49]. Several studies [31, 69-78] have suggested the usage of Bayes classifier as a classification method with specific application to SNP datasets. Table 2.10 shows the classification performance of both logistic regression and Naïve Bayes classifiers within the simulated dataset, following different variable selection methods shown in the previous section (Section 2.4). The Naïve Bayes classification is performed using the '`naïvebayes`' package in R [79].

Table 2.10: Classification performance of logistic regression and Naïve Bayes.

Variable selection methods	Logistic Regression				Naïve Bayes			
	PCC	AUC	Sens	Spec	PCC	AUC	Sens	Spec
LR	0.82	0.82	0.80	0.84	0.80	0.80	0.78	0.81
t-test	0.82	0.82	0.81	0.84	0.79	0.79	0.77	0.81
F_{st}	0.68	0.68	0.72	0.63	0.68	0.69	0.77	0.59
SLR	0.84	0.84	0.82	0.86	0.78	0.79	0.73	0.83
PLR	0.83	0.83	0.83	0.82	0.80	0.80	0.74	0.86

Note: Sens = Sensitivity; Spec = Specificity

In real data, the conditional probability of the Y given X is unknown, therefore, computing Bayes classifier is impossible. Many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given sample to the phenotype group with highest estimated probability [40]. *k-Nearest Neighbour* (KNN) is one such method and is famous for its simplicity and effectiveness. Schwender et al. (2004) [30] describe in detail the application of KNN to SNP dataset and compares it with other classification methods (e.g. bagging, CART and Random Forest). It shows KNN performs slightly better, with a smaller misclassification rate.

The principle of the KNN algorithm is to classify a given sample to the phenotype class based on highest estimate probability [4]. In GWAS, the classifier uses training samples with known outcomes and SNP genotypes to predict responses in an independent dataset of samples (test dataset) [80]. Given a positive integer K and genotype from test data, x_0, the KNN classifier first identifies the K points in the training dataset that are closest to x_0, represented by N_0. It then estimates the conditional probability for phenotype group g as the fraction of points in whose response values equal to g:

$$Pr(Y = g|X = x_0) = \frac{1}{K}\sum_{j\in N_0} I(y_j = g). \tag{2.10}$$

Then, KNN classifies the test observation, x_0 to the phenotype group with the largest probability. The performance of KNN crucially depends on how the distance between the test observation and its closest neighbours is measured. Normally, the closest neighbours are measured by a distance function, for example, Euclidean, Manhattan or Minkowski for continuous variables, or Hamming for categorical variables. To visualise the three different scenarios ($K = 1, 3, 5$) KNN is applied on the simulated dataset using '`class`' package in R [82]. Afterwards, classification measures are calculated. The results improve slightly when increasing the $K = 1$ to $K = 3$ and later $K = 5$. But, increasing the value of K further results in no further improvement. Table 2.11 shows the classification performance using KNN with highest recorded value by $K = 5$, for different variable selection methods.

Table 2.11: Classification performance of KNN using $K = 3, 5$.

Variable selection methods	$K = 3$				$K = 5$			
	PCC	AUC	Sens	Spec	PCC	AUC	Sens	Spec
LR	0.77	0.78	0.74	0.81	0.79	0.79	0.75	0.83
t-test	0.77	0.77	0.72	0.82	0.78	0.78	0.73	0.83
F_{st}	0.60	0.60	0.58	0.62	0.61	0.61	0.59	0.63
SLR	0.76	0.77	0.73	0.80	0.78	0.78	0.73	0.83
PLR	0.82	0.82	0.79	0.85	0.82	0.82	0.79	0.85

Note: Sens = Sensitivity; Spec = Specificity

Another well-known classifier which could be applied to SNP data is *Support Vector Machine* (SVM). The idea behind SVM is to construct an optimal separating hyperplane (or simply a straight line in a two-dimensional setting) between two groups. Here, optimal means that the distance of the hyperplane to the closest point of either group is maximised [30].

SVM can be defined as a concept in statistics and computer science for a set of related supervised learning methods that analyse data and recognise patterns, used for linear classification and regression analysis [84]. Often, the classification analysis through SVM is done after the variable selection procedure takes place [30, 85]. When SVM classifies, it separates a given set of binary-labeled training data with a hyperplane that is maximally distant from the point of each set [86]. For cases in which no linear separation is possible, SVM can work using kernel functions (radial or polynomial), which automatically follows a non-linear decision boundary in the input space.

In the simplest SVM case of two-dimensional setting, a hyperplane is defined by the equation,

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \tag{2.11}$$

for parameters β_0, β_1 and β_2 at any $X = (X_1, X_2)^T$. As mentioned earlier, a two-dimensional setting is simply a straight line as shown in Figure 2.6. Therefore, in a p-dimensional setting the hyperplane can be easily extended to

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0 \tag{2.12}$$

Like most classifiers, SVM is developed based on training data that will correctly classify the test observation. Suppose a hyperplane (2.12) that separates the training data perfectly according to their phenotype groups exists, in two phenotype groups $y_j \in \{0,1\}$, a hyperplane has the following properties

$$\beta_0 + \beta_1 X_{j1} + \beta_2 X_{j2} + \cdots + \beta_p X_p > M \text{ if } y_j = 1 \tag{2.13}$$

and

$$\beta_0 + \beta_1 X_{j1} + \beta_2 X_{j2} + \cdots + \beta_p X_p < M \text{ if } y_j = 0 \quad (2.14)$$

for all samples $j = 1,2,\ldots,n$. M represents the margin of the hyperplane. Naturally, a test observation is assigned to a phenotype group depending on which side of the hyperplane it lies, either (2.13) or (2.14). We now apply the SVM on the simulated dataset using the '`e1071`' package in R [87]. Firstly, the SVM is fitted to the training dataset with the subset of the SNPs according to the variable selection methods. Here, we assume the data are following the non-linear decision boundaries (kernel=radial) and determine the value M through cross-validation (cost=1). Then, SVM will compute the scores based on equations (2.13) and (2.14) to classify the new samples (test data) into one group or the other. Table 2.12 shows the classification performance of SVM on the simulated dataset, for different variable selection methods.

Table 2.12: Classification performance of SVM.

Variable selection methods	PCC	AUC	Sensitivity	Specificity
LR	0.82	0.82	0.78	0.86
t-test	0.82	0.82	0.78	0.86
F_{st}	0.69	0.70	0.81	0.56
Stepwise	0.82	0.82	0.81	0.84
PLR	0.82	0.82	0.79	0.85

Tree-based methods algorithm was first published by Morgan and Sonquist in 1963 [156]. Since then, with applications in machine learning and engineering fields, tree-based methods have become attractive tools for classification. Several studies [88-91] apply tree-based methods to SNP data to identify informative SNPs, detect the interaction between SNPs or improve the classification accuracy of phenotype groups. The tree-based method involves stratifying and

segmenting the variables space into a number of simple regions [40]. The splitting method in the analysis mirrors the tree, hence the name tree-based method.

In mid 1980s, Breiman, Friedman, Olshen and Stoned introduced one of the well-known tree-based methods is *Classification and Regression Tree* (CART)[40]. The decision trees in CART include; (1) Classification Trees which are used when the target variable is categorical and the tree is used to identify the class within which a target variable would likely fall into; meanwhile (2) Regression Trees which are applicable when the target variable is continuous and the tree is used to predict its value. A few recursive algorithms of CART are previously discussed [35, 92].

Given binary phenotype groups g where $g = 0,1$ and a p-dimensional vector X_i containing the values of the p SNPs $X_1, X_2, \dots, X_p$ for each sample $j,\ j = 1,2, \dots, n.$ For simplicity, let X_1 be the most predictive SNP and the set of all samples denoted as Ω. The samples are first divided into Ω_L and Ω_R where both are the left and right subset groups of Ω respectively. The samples are divided based on the value of SNP, X_1 which in additive model represented is by a three-level factor variable 0, 1 and 2. The three possible splits are given by,

$$\begin{aligned}(1) &= \begin{cases}\Omega_L\ if\ X_i \in (0)\\ \Omega_R\ if\ X_i \in (1,2)\end{cases}\\ (2) &= \begin{cases}\Omega_L\ if\ X_i \in (0,1)\\ \Omega_R\ if\ X_i \in (2)\end{cases}\\ (3) &= \begin{cases}\Omega_L\ if\ X_i \in (0,2)\\ \Omega_R\ if\ X_i \in (1)\end{cases}\end{aligned} \tag{2.15}$$

Similarly, in the next step the most predictive SNP of $Y = g$ in each subsets is determined. For instance, X_2 is the most predictive SNP of Ω_L and X_3 is the

most predictive SNP of Ω_R. Again the split will follow the three possible ways as given in (2.15).

One important note to be considered is the measure to split the samples into their respective groups. This measure is commonly referred to as Bayes error, minimum error or misclassification cost [35]. Another commonly used measure is the Gini index [35], also called the nearest neighbour error which is defined as,

$$i(\Omega) = 2Pr_{\Omega}(1 - Pr_{\Omega}) \tag{2.16}$$

where Pr_{Ω} is the probability of being case, conditional on belonging to Ω. For illustration, we apply the classification tree model to the top SNPs using package '`rpart`' in R [93]. Figure 2.4 shows the tree structure constructed using the top five SNPs in the model. It is shown that only two SNPs (rs10888878 and rs1130193) are important in classifying the phenotype groups.

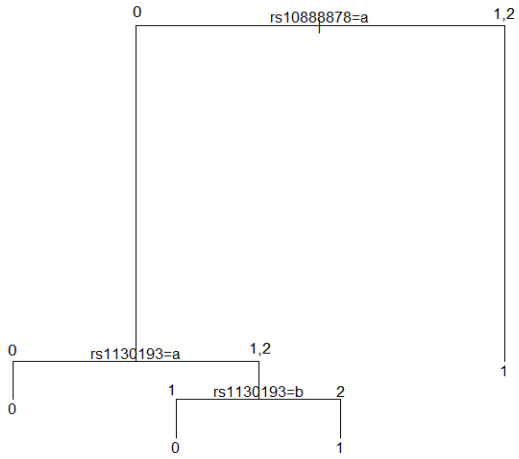


Figure 2.4: Tree structure with top five SNPs rs10888878, rs1130193, rs2286202, rs10920304 and rs2819362.

Based on the output, the splitting starts from rs1088878. The samples with genotype equal to 1 or 2, will be classified as cases $(Y = 1)$. Meanwhile, those

with genotype 0 will go to the left side of the split. Then, further splitting will follow the second important SNP, rs1130193. The samples with genotype 0 will be classified as controls $(Y = 0)$, otherwise the samples are assigned to the split on the right side. Finally, the samples that have genotype 1 for rs1130193 will be assigned in controls group and genotype 2 into cases group.

Similar to our previous analysis, we repeat the analysis for the other subset of SNPs according to the variable selection methods and classify the samples on the test set. The following table shows the results of the classification using CART.

Table 2.13: Classification performance of CART.

Variable selection methods	PCC	AUC	Sensitivity	Specificity
LR	0.80	0.81	0.79	0.82
t-test	0.80	0.81	0.78	0.82
F_{st}	0.70	0.75	0.93	0.47
Stepwise	0.80	0.80	0.78	0.82
PLR	0.80	0.81	0.78	0.83

The advantages of CART can be summarised in threefold; i) the method is not constrained by distributional assumptions, ii) CART has reasonable precision to find complex interactions, and iii) the method allows inclusion of all potential predictors into the model (including continuous variables) [90]. However, CART generally does not have the same level of predictive accuracy as some of the other regression and classification approaches (e.g. SVM and logistic regression). A few tree-based methods have been introduced to improve on this, such as boosting, bagging and random forests [40]. All these methods have the same basic concept as CART with extension in certain parts of the algorithms.

To summarise, Table 2.14 show the overall performance of each classifier according to different variable selection methods and classification measures.

Table 2.14: Summary of classification performance for different variable selection methods and classifiers.

(a) Probability of Correct Classification (PCC)

Variable selection methods	**Logistic Regression**	**Naïve Bayes**	**KNN ($K = 5$)**	**SVM**	**CART**
LR	0.82	0.80	0.79	0.82	0.80
t-test	0.82	0.79	0.78	0.82	0.80
F_{st}	0.68	0.68	0.61	0.69	0.70
SLR	0.84	0.78	0.78	0.82	0.80
PLR	0.83	0.80	0.82	0.82	0.80

(b) Area Under the (ROC) Curve (AUC)

Variable selection methods	**Logistic Regression**	**Naïve Bayes**	**KNN ($K = 5$)**	**SVM**	**CART**
LR	0.82	0.80	0.79	0.82	0.81
t-test	0.82	0.79	0.78	0.82	0.81
F_{st}	0.68	0.69	0.61	0.70	0.75
SLR	0.84	0.79	0.78	0.82	0.80
PLR	0.83	0.80	0.82	0.82	0.81

(c) Sensitivity

Variable selection methods	**Logistic Regression**	**Naïve Bayes**	**KNN ($K = 5$)**	**SVM**	**CART**
LR	0.80	0.78	0.75	0.78	0.79
t-test	0.81	0.77	0.73	0.78	0.78
F_{st}	0.72	0.77	0.59	0.81	0.93
SLR	0.82	0.73	0.73	0.81	0.78
PLR	0.83	0.74	0.79	0.79	0.78

(d) Specificity

Variable selection methods	**Logistic Regression**	**Naïve Bayes**	**KNN ($K = 5$)**	**SVM**	**CART**
LR	0.84	0.81	0.83	0.86	0.82
t-test	0.84	0.81	0.83	0.86	0.82
F_{st}	0.63	0.59	0.63	0.56	0.47
SLR	0.86	0.83	0.83	0.84	0.82
PLR	0.82	0.86	0.85	0.85	0.83

2.6 Concluding remarks

In this chapter, the possible ways of using variable selection methods and several classifiers were reviewed by showing examples using a simulated SNP dataset. The review was focusing on categorical SNP data with binary outcomes. The main challenge in dealing with SNP data is its high dimensionality feature. Hence, in order to reduce the dimensionality of the SNP data, variable selection step is undertaken. The variable selection is not only important to reduce the dimensionality of the data but is believed to aid in improving the classification performance as reported by other studies.

In this review, three variable selection methods that are common in genetic data analysis, namely filter, wrapper and embedded were summarised. Each method was applied to a simulated SNP dataset with five causal SNPs. By using the simulated dataset, each variable selection method was investigated on its ability to capture the causal SNPs. Also, different subsets of SNPs were analysed with different classifiers and their classification accuracies were compared.

In the filter method, logistic regression, modified t-test and F-statistics were discussed. Essentially, the filter method selects the most informative SNPs, which later serve as the attributes for the classifier. Intuitively, a good set of SNPs with a specific genotype pattern will allow us to predict the phenotype group to which a sample belongs [94]. Logistic regression and modified t-test successfully selected five simulated causal SNPs. Meanwhile, F-statistics was not able to capture all causal SNPs. In terms of classification performance, the results when using the subsets of SNPs selected by logistic regression and modified t-test were comparable with wrapper and embedded methods. However, it is important to note that filter method is not only restricted to univariable approaches. Hence, it

was mentioned by Saeys et al. [9] that the utility of the filter method as a multivariable selection algorithm is promising, and should be further explored.

Unlike the filter method, the wrapper method can be applied to analyse data in the multivariable selection approach. The wrapper method enhances the filter method through use of variable selection procedures by implementing forward, backward or stepwise selection. These methods are commonly used in GWAS because one SNP might not be significant by itself but shows different result when added with other SNPs. However, as we know with a large number of SNPs to be considered at one time, computational complexity issue will arise. Hence, the wrapper approach is normally applied on a smaller subset of SNPs to evaluate the classification performance. In this review, we applied stepwise logistic regression on a subset of top 100 SNPs. The method successfully selected the five causal SNPs and the highest classification accuracy of 84% was recorded with logistic regression as the classifier.

Another route for variable selection for classification within a SNP dataset is by using an embedded method. This method requires an insertion of a penalty into the classifier. Essentially, the penalty is expected to turn all the coefficients of unimportant SNPs to zero [60]. Similar to the wrapper method, we feel that this method would be computationally burdensome due to the consideration of all SNPs at the initial stage of analysis. Hence, penalised logistic regression was applied to a subset of 100 SNPs from the overall dataset. In the final model, only the five causal SNPs were included.

To conclude, high hopes are associated with the concept of personalised medicine in which the patients are stratified according to disease status or response to

treatment, leading to patient-specific diagnosis and therapy [3]. This chapter helps to emphasize the importance of variable selection for classification. The results using the simulated dataset suggested that it is important to select the most informative SNPs as it can aid to classification performance. Statistically, an ideal variable selection method would make use the SNP data as much as possible, easy to apply and interpret. It also would require less computational burden.

Chapter 3
Methods

3.1 Introduction

In recent years, the advent of technologies such as microarray, proteomics, and next-generation sequencing have enabled the collection of large datasets [49]. Specifically, in GWAS, thousands or millions of SNPs are being measured on individuals of which the number is much smaller. The situation is referred as high-dimensional problem, also called the $p \gg n$ problem [95]. One solution to tackle this problem is to reduce the dimensionality by doing variable selection, which allows the optimal subset of SNPs associated with the phenotype being selected. In the context of classification, implementing a variable selection procedure could aid better classification performance, i.e., assigning samples or individuals to their respective phenotype groups accurately.

As previously discussed in Chapter 2 (Literature Review), three main methods exist, namely, filter, wrapper and embedded methods, which vary with respect to computational complexity or contribution to classification performance. Typically, SNPs are identified individually using univariable analysis methods. However, in reality, the SNPs may be correlated with one another or other environmental variables. Therefore, there is an argument to employ a multivariable approach to variable selection (i.e. to take into account all variables in a single analysis rather than sequentially investigating one variable at a time). The common approach in statistical model-building is to include or exclude the variables based on certain criteria (e.g. AIC, Bayesian Information Criteria (BIC) or R^2) [59]. This procedure is done until the most parsimonious model, which can describe the outcome well, is produced.

There is no universally optimal variable selection method that fits well with every type of data [4]. However, this thesis is concerned with developing a variable selection method that can reduce computational burden and help in improving the classification performance with specific focus on SNP datasets with a binary outcome. With this purpose in mind, the focus lies on filter-type variable selection methods that employ simple yet multivariable metrics of a signal. Specifically, the aim is to study and develop further the generalised signal-to-noise ratio (SNR) within the framework of logistic regression with specific application to SNP data.

Generalised SNR was recently proposed by Czanner et al. [17] within the context of generalised linear models (GLM) in the recording of single neurons. In the paper, generalised SNR was not introduced as the variable selection method *per se*, but rather as a measure of system fidelity in neural spiking activity. This thesis

generalises SNR specifically to SNP datasets by using logistic regression as the framework.

In this thesis, the direct application of SNR to SNP datasets is challenging for several reasons. First, the number of SNPs can be extremely large, the effect of SNPs on the phenotype tend to be small to modest, so the SNR of each SNP is low [18]. Second, it is not possible to apply SNR directly to all variables (SNP data is often large with more than 100,000 SNPs for each chromosome). Third, due to a large number of SNPs, a stopping criterion should be developed and imposed on the maximum number of SNPs that should be included in the final model. Fourth, since this method is built outside of the classification algorithm (i.e. it is a filter method), the performance of the variable selection might be independent with the choice of classifier. Hence, it is crucial to select a good classifier that can produce good classification performance. Fifth, misclassification tends to be much higher for binary outcomes than continuous outcomes [18].

Apart from focusing on variable selection, this thesis also focuses on jointly modelling SNP data and clinical information, which could potentially improve the classification performance. As we know, the clinical data are often available, and their predictive value is well-validated in the literature [21]. Often, the clinical data is collected once for each patient (cross-sectional). This may be a feasible approach for the most common case where the diagnosis of a patient is of interest. However, if the effect of therapy or any time-dependent response is of interest, then it is more appropriate to include the longitudinal profiles as well [22]. Hence, the proposal is to jointly model the longitudinal as well as the cross-sectional data. This chapter describes the existing methods of variable selection as well as the methods proposed in this thesis. Section 3.2 presents the proposed variable

selection method, tSNR. The motivation and how the method works in both univariable and multivariable scenarios are discussed, followed by the existing statistical concepts related to tSNR. Then, the proposal to jointly model the longitudinal clinical and SNP data is elaborated on in Section 3.3. Table 3.1 summarises the structure of the chapter.

Table 3.1: Structure of the chapter.

<table>
<tr><th>Objectives</th><th>Type of data</th><th>Methods of variable (SNPs) selection</th><th>Classification</th></tr>
<tr><td>To select the most informative SNPs</td><td>• SNP data (cross-sectional; categorical)</td><td rowspan="2">• Filter-metric tSNR and the algorithm (Section 3.2.1, 3.2.2, 3.2.3)
• Penalised logistic regression (lasso)
• Stepwise logistic regression
• Other statistical concepts (BIC, AIC, R^2) (Section 3.2.4)</td><td>• Logistic regression</td></tr>
<tr><td>Combining clinical and SNP data</td><td>• Clinical data (longitudinal; count, binary, continuous)
• SNP data (cross-sectional; categorical)</td><td>• Joint modelling of longitudinal clinical and SNP data using MGLMM (Section 3.3)
• Longitudinal Discriminant Analysis (LoDA) (Section 3.3)</td></tr>
</table>

Note: Multivariate Generalised Linear Mixed Model (MGLMM)

3.2 Selection of the most informative SNPs

3.2.1 Proposed tSNR and the variable selection algorithm

In science and technology, signal-to-noise ratio (SNR) is a measure that compares the level of the desired signal to the level of background noise [11]. A higher value of SNR means a better model, since there is more useful information (the signal) than there is unwanted data (the noise) [96]. Generally, SNR is defined as

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}} \tag{3.1}$$

where σ^2_{signal} is the variance of a signal that describes how one variable (i.e. independent) is related to another variable (i.e. dependent) [97]. Meanwhile, σ^2_{noise} is the variance of noise which is essentially the standard error (of the difference), which quantifies the sampling variability and thereby the statistical uncertainty of the comparison measure.

In the context of variable selection in GWAS, the t-test is a well-known statistical test that applies the SNR concept. In other words, it is a signal-to-noise ratio. The t-test is often applied in a situation in which the maximisation of distance or difference between two population means is needed. The variable is chosen or considered important if there is significant evidence (e.g. based on p-value) that the variable is associated with the outcome of interest. These studies [14-16, 98, 99] apply the t-test approach as a variable selection method to GWAS data. For example, Hotelling's T^2 and modified t-test are applied as a ranking measure to microarray (continuous) and SNP (categorical) data accordingly.

Since the focus of this thesis is categorical SNP data, it is essential to discuss the modified t-test at a greater length in this chapter. As discussed in Chapter 2 (Literature Review), Zhou et al. [14-16] have proposed to modify the t-test to fit categorical SNP data. The proposal extended the t-statistic algorithm to rank the p SNPs for all phenotype groups, G. That is, the t-score of SNP X_i $(i = 1,2, \dots, p)$ is calculated as the greatest t-score for all phenotype groups:

$$t_i = max\left\{\frac{\left|\vec{\bar{X}}_{ig} - \vec{\bar{X}}_i\right|}{M_g S_i}, \qquad g = 0,1,\ldots,G-1\right\} \tag{3.2}$$

$\vec{\bar{X}}_{ig}$ and $\vec{\bar{X}}_i$ are row vectors and represent the mean of the i-th SNP in the g-th phenotype group and the mean of the i-th SNP for all groups, respectively. $\left|\vec{\bar{X}}_{ig} - \vec{\bar{X}}_i\right|$ denotes the Euclidean distance of the two vectors. In this equation, the categorical variable with three levels (0, 1 and 2) for each SNP are transformed into a 3-dimensional variable vector $(0 \Rightarrow \vec{X}_i^{(1)} = \{1,0,0\},\ 1 \Rightarrow \vec{X}_i^{(2)} = \{0,1,0\}, 0 \Rightarrow \vec{X}_i^{(3)} = \{0,0,1\})$. Also,

$$M_g = \sqrt{\frac{1}{n_g} + \frac{1}{N}} \tag{3.3}$$

$$S_i^2 = \frac{1}{N-G}\sum_{g=0}^{G-1}\sum_{j\in g}\left(\vec{X}_{ij} - \vec{\bar{X}}_{ig}\right)\left(\vec{X}_{ij} - \vec{\bar{X}}_{ig}\right)^T \tag{3.4}$$

where $\vec{X}_{ij}$ refers to the vector of the i-th SNP of the j-th sample (or individual); N is the total number of samples in all phenotype groups; n_g is the number of samples in the particular phenotype group, g and S_i is the within-group standard deviation. The difference in means in equation (3.2) indicates the dispersion of the mean of the phenotype group of interest from the mean of all phenotype groups. The t-score is calculated for each SNP, which relates to the signal that a specific SNP carries. Among the total of p SNPs, the higher the value of the t-score means the more informative the SNP is.

The t-score (ranking) is limited in the way it considers the SNPs one at a time (univariable). As we know, the need to consider the SNPs in a multivariable setting is important (i.e. presented in a model). Hence, the SNR concept proposed

by Czanner et al. [17], (which is similar to the concept of the t-statistic) is more approriate since it was extended to the generalised linear model (GLM) systems. GLM is an established statistical framework for performing regression analyses, which uses deviance to measure the lack of fit between model and data. GLM makes it possible to perform regression analyses to relate observations from any model in the exponential family to a set of variables. This family includes well-known probability models such as the Gaussian, Bernoulli, Binomial, Poisson, Gamma and inverse Gaussian.

To introduce the definition of SNR for GLM models, let us assume that for a given link function $g()$ the expected value of Y given X can be expressed as $g\big(E(Y|X)\big) = X_1\beta_1 + X_2\beta_2$, where $X_1\beta_1$ is the component of mean unrelated to the signal and $X_2\beta_2$ is the mean that is related to the signal. The partitions $X = [X_1, X_2]$, $\beta = (\beta_1, \beta_2)^T$, where X_1 is a $n \times p_1$ matrix of non-signal variables, X_2 is a $n \times p_2$ matrix of signal variables, β_1 is a $p_1 \times 1$ vector of non-signal coefficients and β_2 is a $p_2 \times 1$ vector of signal coefficients. Now, $p_1 + p_2 = p$.

Further, the deviance in GLM, also called the log-likelihood (ratio) statistic, provides a way of assessing the goodness of fit for a proposed model (or model of interest) by comparing it with a more general model with the maximum number of parameters that can be estimated, which is called the saturated model [100]. Suppose $\hat{\beta}_{max}$ denote the maximum likelihood of the saturated model. The likelihood function for the saturated model evaluated at $\hat{\beta}_{max}$ is given by $\mathcal{L}(\hat{\beta}_{max}|y)$. Let $\hat{\beta}$ denote the maximum likelihood estimate of β which represents the proposed model and $\mathcal{L}(\hat{\beta}|y)$ be the likelihood corresponding to the proposed model. Hence, the deviance is a measure of a distance between the proposed model and the saturated model, given by,

$$Dev(y, X, \hat{\beta}) = -2\, log \frac{\mathcal{L}(\hat{\beta}|y)}{\mathcal{L}(\hat{\beta}_{max}|y)} \tag{3.5}$$

Then, the SNR estimates for GLM is the ratio of deviance defined as,

$$\widehat{SNR} = \frac{Dev(y, X_1, \hat{\beta}_1) - Dev(y, X, \hat{\beta})}{Dev(y, X, \hat{\beta})} \tag{3.6}$$

where the numerator gives the reduction in the deviance due to the signal $X_2 \widehat{\beta}_2$ when controlling for the effect of the non-signal component $X_1 \widehat{\beta}_1$. Here, $\widehat{\beta}_1$ and $\hat{\beta}$ are the maximum likelihood estimates obtained from the two separate fits of the models respectively to Y. Meanwhile, the denominator is the variability in Y due to noise.

To define the SNR for the SNP data it can be noted that, it is a common approach to fit the logistic regression model to test for association between SNPs and binary phenotype. Similar to the concept explained earlier, to measure the signal or strength of information a SNP carries, the proposal is to generalise the definition (3.5) in a logistic regression framework. Assume there is a null model which only contains an intercept, β_0 and a saturated model with only one SNP, X_1 with coefficient β_1. By using the analogy in (3.6), here the numerator gives the reduction in the deviance due to the signal of $X_1\beta_1$.

As we know, with logistic regression, more than one SNP, X_i $(i = 1,2, \dots, p)$ can be added into the model with each of the SNPs coded as 0, 1 and 2 (additive model which represents the number of minor alleles held). Hence, the following SNR measure, referred to as tSNR is proposed as the generalisation of (3.6) in the logistic regression framework. The estimation is given by,

$$\widehat{tSNR}_i = \frac{Dev\left(y, \hat{\beta}_0\right) - Dev\left(y, X_i, \hat{\beta}_i\right)}{Dev\left(y, X_i, \hat{\beta}_i\right)} \tag{3.7}$$

where $Dev\left(y, \hat{\beta}_0\right)$ is the null deviance of the null model showing how well the response variable is predicted by a model that includes only the intercept, $\hat{\beta}_0$. Meanwhile, $Dev\left(y, X_i, \hat{\beta}_i\right)$ refers to the deviance of the fitted model which includes the intercept, $\hat{\beta}_0$ and the estimated coefficient, $\hat{\beta}_i$ associated with each SNP, X_i $(i = 1,2, \ldots, p)$.

The adjusted tSNR

As discussed in their paper, Czanner et al. [17] the GLM SNR estimator (3.6), is biased. The bias arises because the SNR numerator always gives positive estimates. The situation is still acceptable when the model considers one variable at a time (univariable approach). However, the value of the numerator will keep increasing with inclusion of more variables (multivariable) causing a positive bias.

Therefore, the proposal was to use *Mittlböck* and Waldhör to obtain a bias-corrected SNR estimator [17]. By adapting the same notion, in (3.7) the difference in the deviance in the numerator and the deviance in the denominator is corrected by the number of the coefficients in the corresponding models,

$$\widehat{tSNR}\ adjusted = \frac{Dev\left(y, \hat{\beta}_0\right) - Dev\left(y, X_i, \hat{\beta}_i\right) + d_0 - d_i}{Dev\left(y, X_i, \hat{\beta}_i\right) + d_i} \tag{3.8}$$

where d_0 is the number of coefficients in the null model, and therefore is equal to one, meanwhile d_i is the number of coefficients in the fitted model. From now on,

equation (3.7) will be used as a univariable filter metric, meanwhile equation (3.8) will be utilised as the criterion for multivariable model selection.

3.2.2 The workflow

By using tSNR as the main filter metric for the variable or model selection, there is a need to propose a workflow on how to select the most informative SNPs. There are a few criteria to be considered when deciding on the workflow of variable selection; i) univariable or multivariable selection, ii) cross-validation, and iii) evaluation of classification performance. Hence, in this section the methods involved in the workflow (Figure 3.1) are described.

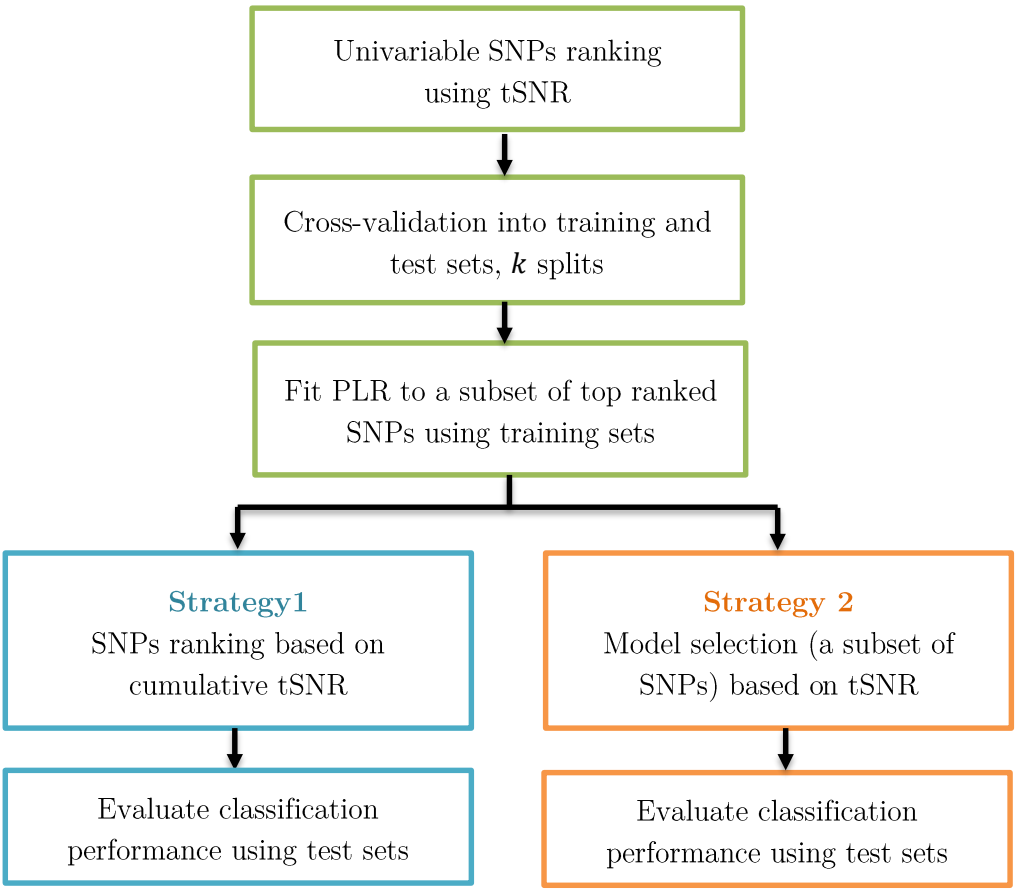


Figure 3.1: Diagram of model building pipeline including (i) univariable tSNR as preselection process; (ii) splits of sample into training and test sets; (iii) model building using penalised logistic regression (PLR); (iv) strategies to select a small subset of SNPs; and (v) model evaluation using test sets.

3.2.2.1 Univariable variable selection

The most common summary measure of inference regarding a single SNP in GWAS is the p-value [101]. The measure is obtained based on univariable statistical tests (e.g. logistic regression, chi-squared). This univariable analysis can be seen as a preprocessing step to infer the association between a single variable and the outcome of interest. In the filter method approach, the variables are often ranked from the most important to the least important variable, where importance is measured by the value of the statistical test (e.g. chi-squared's p-value, Pearson's correlation coefficient). Dealing with a large number of SNPs, the univariable analysis is necessary to filter out the unimportant SNPs before further analysis.

In this thesis, the proposal is to apply tSNR as the univariable selection as a preprocessing step. The larger the tSNR value means that the SNP carries a higher signal in explaining the outcome, and therefore, the more important the SNP is. For this univariable ranking, it can be shown that the ranking produced by tSNR is identical to the ranking produced by p-value from the chi-squared test. The equality is due to the numerator of equation (3.6) being simply the test statistics with a chi-squared distribution with degrees of freedom equal to the difference between the intercept and the number of coefficients estimated. The chi-squared test is widely applied to measure the association between categorical variables [102].

Table 3.2 shows the ranking results produced by tSNR and chi-squared test using a simple simulated dataset. The dataset was simulated with one binary outcome with equal proportion and three SNPs with three different proportions for the genotypes 0, 1 and 2. The specified proportions for each SNP are given by (0.5,

0.3, 0.2), (0.35, 0.35, 0.3) and (0.6, 0.3, 0.1) accordingly. The proportions for first and third SNPs are chosen to represent the causal SNPs, meanwhile no effect for the second SNP. The analysis is done using the '`glm`' function in R [61].

Table 3.2: Ranking comparison between tSNR and p-value of chi-squared statistics.

SNPs	Null deviance	Residual deviance	tSNR	SNR ranking	df	Pr (>chi)	Chi-squared ranking
SNP 1	138.63	123.93	0.1186	2	1	8.101e-05	2
SNP 2	138.63	137.82	0.0059	3	1	0.658	3
SNP 3	138.63	107.83	0.2856	1	1	1.493e-06	1

While in some scenarios the tSNR and chi-squared test give identical results (Table 3.2), there are scenarios where the chi-squared test cannot be applied. The chi-squared test can only be applied to nested models and on the same set of patients. On the other hand, tSNR can be used even if set of patients differ between two models or when there are missing values in the observations for some patients. Furthermore, tSNR can be used in the scenario to compare two non-nested models. The standardisation imposed from the ratio of the deviance making it unrestricted from the number of patients used. Therefore, in this situation tSNR is more useful as it can be interpreted in both nested and non-nested scenarios. The definition and explanation of nested models will be discussed further in the following section.

3.2.2.2 Multivariate variable selection

Here, the multivariable selection methods are discussed. The methods can be used for the implementation of the multivariable part of the algorithm in Figure 3.1 above. It assumes that a set of variables have been already pre-selected.

Penalised logistic regression (PLR)

After going through the filtering process, a subset of SNPs will be used in the modelling process in which the importance of the SNPs will be evaluated once again. The difference between variable selection at this stage and univariable selection discussed earlier, is how the SNPs are assessed. By using multivariable selection, a subset of SNPs are evaluated together (i.e. the correlations between SNPs are considered). In this thesis, penalised logistic regression (PLR) with lasso penalty is applied for the multivariable selection analysis. PLR is chosen due to its ability to deal with large number of SNPs and works well with categorical variables [60, 65, 103].

Let Y represent the outcome variable and assume that there are p SNPs, X_i $(i = 1, 2, \dots, p)$ coded as 0, 1 or 2 (additive). Each sample may belong to one of G phenotype groups represented by $g = 0,1, \dots, G-1$. For a binary phenotype where $G = 2,\ g = 1$ denotes the cases group (e.g. disease, not achieved remission), meanhwile $g = 0$ denotes the controls group (e.g. healthy, achieved remission). The probability that the j–th sample $(j = 1,2, \dots, n)$ belong to group $g = 1$ with i-th SNP is given by $Pr(Y_j = 1|X_i)$ and can be written as,

$$Pr(Y_j = 1|X_i) = \pi_j = \frac{e^{\beta_0 + X_{ij}^T \beta_i}}{1 + e^{\beta_0 + X_{ij}^T \beta_i}}$$

or (3.9)

$$Pr(Y_j = 0|X_i) = 1 - \pi_j = \frac{1}{1 + e^{\beta_0 + X_{ij}^T \beta_i}}$$

where X_{ij} is the observation for i-th SNP for j-th patient. The coefficient vector $\theta = (\beta_0, \beta_1, \dots, \beta_p)^T$ is usually estimated by maximising the log-likelihood [65]

$$L(\theta) = \sum_{j=1}^{n} \left[Y_j \, log \, \pi_j + (1 - Y_j) log(1 - \pi_j) \right] \tag{3.10}$$

Lasso penalty is subtracted from the log-likelihood as follows:

$$g(\theta) = L(\theta) - \lambda \sum_{i=1}^{p} |\beta_i| \tag{3.11}$$

Note that the intercept β_0 is ignored in the lasso penalty, $\lambda \sum_{i=1}^{p} |\beta_i|$. The tuning parameter λ controls the strength of the penalty which will shrink the β_i towards zero. Hence, this procedure will help to further reduce the number of SNPs used during the modelling and later classification process. The PLR with lasso penalty is implemented using '`glmnet`' package in R [104].

Stepwise logistic regression (SLR)

In addition to PLR, stepwise logistic regression is used as comparison in Chapter 4 (Simulation Study). The stepwise selection applies both forward selection and backward elimination of the variables. The addition and deletion of each variable is considered using either AIC or BIC. The stepwise selection is attractive since it keeps evaluating the model each time variable is added or eliminated.

Cross-validation

To explore the problem of over fitting, *cross-validation* with k splits is applied. For this thesis, the split is set for 100. By training and testing the model containing selected variables on different subsets of the data the strength of prediction for classification can be evaluated [105]. As shown in Figure 3.1, the data is split into a training set (80% or 70% of the data) and a test set (20% or 30% of the data) with which the classification is done. The data is split k times

which will produce k different models for the training set and k classification values (e.g. AUC, sensitivity, specificity) using the test set. From here, two strategies will be followed to choose the best model among the k models produced by the split.

Strategy 1: Multivariable ranking using tSNR (cumulative tSNR ranking)

Strategy 1 mainly ranks the SNPs based on the cumulative tSNR for each SNP. Firstly, the tSNR is calculated for each model which is fitted with cross-validation procedure. Then, the tSNR value for each model is multiplied by each of the SNPs present within the model. The tSNR which is considered as the weight for each SNP, ω_i is the summation of the tSNR across the k models,

$$\omega_i = \sum_{k=1}^{100} \widehat{tSNR} \tag{3.12}$$

The SNPs are ranked from the highest cumulative tSNR to the lowest. The total number of SNPs to be selected is determined by the adjusted tSNR. The inclusion of SNPs in the model is halted when there is no increment in adjusted tSNR. Intuitively, the decreasing of the value of tSNR when a new variable is introduced in the model indicates less signal as compared to noise that the new variable carries.

Strategy 2: Model selection using tSNR

By using the k splits, k different models which differ in terms of samples included as well as SNPs selected are produced. These models are considered non-nested, i.e., two models are non-nested, either partially or strictly, if one model cannot be reduced to the other model by imposing a set of restrictions on the coefficients, β_i

[106]. Therefore, a model selection criterion is needed to select the best model. As mentioned earlier, one of the advantages of tSNR is that it can be used to compare the non-nested models. Hence, by implementing the idea, the fitted PLR models (using the training sets) are ranked based on the highest tSNR to the lowest tSNR values. The SNPs that correspond to the highest ranked model are selected for the next step.

Binary classification

In this thesis, we are focusing on the binary classification using logistic regression as the main classifier. In the classification problem, logistic regression measures the relationship between the outcome (cases coded as '1' or the outcome of interest) and the one or more independent variables, by estimating probabilities using its underlying logistic function. These probabilities (between 0 and 1) are then transformed into either 0 or 1 (binary outcomes) according to the probability threshold specified (usually 0.5).

Commonly, the results for binary classification can be summarised in the two by two table as shown in Table 3.3.

Table 3.3: Binary classification table.

		Classification outcome	
		Controls coded as '0'	**Cases coded as '1'**
True status	**Controls coded as '0'**	True Negative (TN)	False Positive (FP)
	Cases coded as '1'	False Negative (FN)	True Positive (TP)

There are six performance measures that are discussed in this thesis, namely, the Area Under the receiver operating characteristic (ROC) Curve (AUC), the probability of correct classification (PCC), sensitivity, specificity, positive

predictive value (PPV) and negative predictive value (NPV). The ROC curve is a technique that has been widely used to assess classification accuracy. For a given cutoff value of a variable or a subset of variables, the sensitivity and the specificity are employed to quantitatively evaluate the classification performance [107]. By varying the cutoff values, the resulting plot of sensitivity against 1-specificity is a ROC curve. Then the Area Under the ROC Curve (AUC) is calculated as a measure to summarise the overall classification accuracy of a ROC curve. Although the AUC was originally developed for comparing distinct diagnostic tests or biomarkers, it has increasingly been adopted for use in evaluating the incremental effect of an additional biomarker in predicting a binary event via a regression model [108].

The percentage of patients who have the medical condition and correctly classified as cases (true positive) is called sensitivity, whereas the percentage of patients correctly classified as controls (true negative) whom not having medical condition is called specificity [109]. Hence, the sensitivity and specificity are given by,

$$Sensitivity = \frac{TP}{TP+FN} \qquad Specificity = \frac{TN}{TN+FP} \tag{3.13}$$

The accuracy or PCC can be expressed as a sum of true positive and true negative over the total samples (or patients),

$$PCC = \frac{TP + TN}{TP + FN + TN + FP} \tag{3.14}$$

Meanwhile, PPV can be defined as the probability that a patient who correctly classified as case (true positive) really has the medical condition and the NPV is

the probability of a patient who is correctly classified as control (true negative) does not has the medical condition.

$$PPV = \frac{TP}{TP+FP} \qquad NPV = \frac{TN}{TN+FN} \tag{3.15}$$

3.2.3 Statistical concepts related to tSNR

In this section, the relationship between tSNR and other statistical concepts is discussed. This will help to understand the properties of tSNR and its potential use in future studies and research. The discussion includes its relationship to the AIC, BIC and generalised R^2.

The selection of the most informative SNPs is equivalent to the selection of the best model which later is assumed to produce good classification performance [110]. Hence, certain model selection criterion is needed. In this thesis, there are two situations of model selection that need to be considered; nested models and non-nested models. For nested models, the model selection is needed when forward selection method is applied in Strategy 1. Meanwhile, non-nested models are generated when PLR is applied in which k different models are produced through cross-validation (Strategy 2). Therefore, model selection criterion that can deal with both nested and non-nested models will be discussed.

The well-known model selection criteria such as AIC, BIC and generalised R^2 are discussed. The discussion includes brief introduction of each method and its relationship with tSNR. AIC provides a method for assessing the quality of the proposed model through comparison of related models (nested). In GLM, AIC can be written as

$$AIC = Dev\left(y, X_i, \hat{\beta}_i\right) + 2d_i \tag{3.16}$$

where d_i is the bias-correction term representing the number of coefficients in the model. In theory, to compare between two nested models, the model with the smaller value of AIC should be selected. AIC can be used to compare between two nested models, however, it is not interpretable on its own. Given the same models as a comparison, it can easily be shown that AIC has a negative relationship with tSNR. Equation (3.7) can also be written as follows:

$$\widehat{tSNR} = \frac{Dev\left(y, \hat{\beta}_0\right)}{Dev\left(y, X_i, \hat{\beta}_i\right)} - 1 \tag{3.17}$$

Then, by rewriting the definition of AIC in equation (3.16), the relationship between tSNR and AIC is given by,

$$\widehat{tSNR} = \frac{Dev\left(y, \hat{\beta}_0\right)}{AIC - 2d_i} - 1 \tag{3.18}$$

which means that a reduction in the value of AIC will be linked to an increase in tSNR. Therefore, in the situation of nested models comparison, the decision of model selection using either of these two methods are expected to be similar.

The second most used model selection method is the BIC. The difference between AIC and BIC is that BIC generally places a heavier penalty on models with many variables [40]. By referring to equation (3.19), BIC is more stringent in a way that the penalty increases when the sample size (i.e. the number of independent observations), n is getting larger. In this case, BIC tends to select more parsimonious models than AIC [110].

$$BIC = Dev\left(y, X_i, \hat{\beta}_i\right) + d\ log(n) \tag{3.19}$$

Similar to AIC, tSNR can be shown to have a negative relationship with BIC. By inserting the definition of BIC in equation (3.19) into equation (3.17), the tSNR estimator is given by,

$$\widehat{tSNR} = \frac{Dev\left(y, \hat{\beta}_0\right)}{BIC - d\ log(n)} - 1 \tag{3.20}$$

Generalised R^2 is another popular approach for selecting among a set of models that contain a different number of variables. Unlike AIC and BIC, the higher value of R^2 indicates a better model. The generalised R^2 in logistic regression is different from the R^2 in ordinary least square approach. The common generalised R^2 that is used in logistic regression is McFadden's which is written as follows [111]:

$$R^2 = 1 - \frac{Dev\left(y, X_i, \hat{\beta}_i\right)}{Dev\left(y, \hat{\beta}_0\right)} \tag{3.21}$$

From equation (3.21) it can be seen that the numerator is similar to the numerator of tSNR which follows approximately the chi-squared distribution with degrees of freedom equal to the difference between the intercept and the number of coefficients estimated. However, R^2 will always increase with the number of variables added, hence might not be useful from a clinical perspective.

3.3 Combining longitudinal clinical and SNP data for classification

GWAS leveraging cross-sectional phenotypic data has been a useful approach to identifying SNPs that influence the quantitative risk factors relates to the outcome of interest, however the use of cross sectional data does not provide insight into how such risk factors develop over time [112]. Longitudinal studies help in identification of risk profiles of susceptible individuals before prognosis onset. For example, the study about children with focal epilepsy [113] has shown the disorder has a genetic basis and is also age-dependent which suggests the need of having the genetic (cross-section) and longitudinal data.

In biomedical research, the question of the added predictive value of considering molecular data (e.g. SNPs, microarray, RNA) jointly with clinical variables is an open area of research [21]. In certain disease studies (e.g. asthma, type-2 diabetes), clinical variables have been extensively investigated and validated in previous studies. However, the SNPs, which are usually high-dimensional and categorical are often not as well-established. Therefore, it is useful to see how SNP data can improve the classification performance of a conventional model with clinical variables.

The approach to jointly model the clinical and SNP data can be advantageous since the clinical variables are often available and have well-validated predictive ability [114]. This situation allows the focus to be only on the added predictive value of SNP data. Statnikov et al. (2007) [115] show the classification performance of a model with only SNP data, a model which consists of only clinical data and a model with the combination of SNP and clinical data. The data consists of 50 esophageal squamous cell carcinoma patients and 50 controls with 11,542

SNPs. In addition, five clinical variables are recorded. In their study, the classification performance improves slightly using the model with the combination of both types of data.

Frequent clinical interest is in being able to classify patients into various groups corresponding to their prognosis, based on the evolution of variables observed over time. In general, Boulesteix and Sauerbrei (2011) [21] and De Bin et al. (2014)[114] proposed the strategy to combine the clinical and omics data. By adapting similar strategy, we propose to jointly model the longitudinal clinical and SNP data for classification.

Figure 3.2: Stages proposed involved in combining the longitudinal clinical and SNP data for classification.

Figure 3.2 shows the proposed strategy to combine the longitudinal clinical and SNP data. The strategy includes two stages and in each stage the classification performance is evaluated. By evaluating the classification performance, the added predictive value of SNP data to the longitudinal clinical data can be examined.

Firstly, the filter metric tSNR which has been discussed in Section 3.2.1 is applied to select the most informative SNPs. Apart from the SNPs selection, the classification performance is also evaluated at this stage. The approach at this stage allows only a few selected SNPs to be utilised in combination with

longitudinal clinical data. Often, in GWAS only few SNPs are selected which helps in clinical implementation and interpretation.

In the second stage, the approach in which to jointly model the longitudinal clinical and SNP data is discussed. Here, the SNPs are referring to the SNPs that are chosen previously in Stage 1. Due to the longitudinal and cross-sectional nature of the data, the multivariate mixed-effects model needs to be considered. Hence, the analysis of the data using a joint model will follow the work conducted by Hughes et al. [25, 116]. Their research presents a flexible and dynamic (time-dependent) discriminant analysis approach in which multiple variables of various types are jointly modelled for classification purposes by the multivariate generalised linear mixed model. Here the types of variables included i) are of multivariate and longitudinal nature, ii) have a complex correlation structure, (iii) are of varying types of data (e.g. continuous, counts or binary) and, (iv) have different time points at which clinical variables are measured.

By considering the nature of the data, the clinical variables are jointly modelled for classification purposes by the multivariate generalised linear mixed model (MGLMM) [25,116]. These longitudinal models are subsequently used in the dynamic longitudinal discriminant analysis (LoDA) to predict the probability of an individual belonging to a specific group [25, 116].

3.3.1 Overview of Longitudinal Discriminant Analysis (LoDA)

Multivariate generalised linear mixed model (MGLMM)

MGLMM is used to model the longitudinal clinical data, as well as SNP data in Stage 2 (Figure 3.2). The model coefficients generated are subsequently used

within the LoDA procedure [116]. MGLMMs are well-known due to their ability to accommodate the complex nature of datasets.

The following notations were previously defined [25, 116, 117] but there are slight adjustments made, to be consistent throughout this thesis. The aim is to predict, which G phenotype groups a patient is likely to belong to, at a given time point. This is done using information collected at baseline and over time up to that time point. For binary phenotype where $G = 2,\ g = 1$ denotes the cases group (e.g. disease, not achieved remission), meanhwile $g = 0$ denotes the controls group (e.g. healthy, achieved remission), which is only observed at time T. Assume that for each patient, measurements are made on $R \geq 1$ biomarkers at times $t_r = (t_{r,1}, \dots, t_{r,n_r})$, $t_{r,1} < \cdots < t_{r,n_r}$, $r = 1, \dots, R$. Here, R is the number of longitudinal biomarkers in which the type can be binary, continuous or count. For each marker, these longitudinal observations for a particular patient are denoted as $Y_r = (Y_{r,1}, \dots, Y_{r,n_r})$. All information is collected for a patient up until some time $t < T$ to predict the future group, g, to which the patient belongs. The prediction is based on the information gathered about the patient at time t and also all previous data for that patient.

The MGLMMs are fitted to the longitudinal data separately according to the phenotype group (either $g = 0$ or $g = 1$). The expected value (transformed by an appropriate link function) for the j-th longitudinal observation $(j = 1, \dots, n_r)$ for the r-th marker $(r = 1, \dots, R)$ (denoted $Y_{r,j}$) is assumed to follow a distribution from an exponential family (e.g. normal, Poisson, Bernoulli) with a dispersion parameter $\emptyset_r^g$ is given by,

$$\{h_r^{-1}\{E(Y_{r,j}|b, g)\} = x_{r,j}^{gT}\alpha_r^g + z_{r,j}^{gT}b_r, \qquad r = 1, \dots, R, j = 1, \dots, n_r. \qquad (3.22)$$

In (3.22), h_r^{-1} is a known link function used in the GLMM for the r-th marker (e.g. logit for Bernoulli responses, log for Poisson variables), $x_{r,j}^g$ and $z_{r,j}^g$ are covariate vectors used in a model for group g derived from the information on the visit times.

Further, α_r^g are unknown regression coefficients (fixed effects) related to the model for the r-th marker in the group g. It is assumed that a particular subject is characterised by values of a latent random effects vector $b = (b_1, \dots, b_R)$ which accounts for possible correlation between repeated observations of the same marker and also different markers on the same patient. Typically, the random-effects vector is assumed to jointly follow a normal distribution. However, Hughes et al. (2016) [116] proposed to consider different normal mixtures in different groups to allow additional flexibility. That is, they assume

$$b, g \sim \sum_{q=1}^{Q^g} \omega_q^g \, \mathcal{MVN}\left(\mu_q^g, \mathbb{D}_q^g\right), \tag{3.23}$$

where $\mathcal{MVN}(\mu, \mathbb{D})$ stands for a multivariate normal distribution with the mean vector μ and a covariance matrix $\mathbb{D}$. The mixture distributions are weighted by a factor $\omega_q, (q = 1, \dots, Q)$ of which the number of mixture components, Q^g is initially to be known. This multivariate normal distribution has a density denoted as $\varphi(\cdot; \mu, \mathbb{D})$.

To fit this MGLMM, the fixed effects regression coefficients from (3.22) need to be estimated. The estimation is denoted as $\psi^g := \left(\alpha_1^g, \dots, \alpha_R^g, \emptyset_1^g, \dots, \emptyset_R^g\right)$ and additionally mixture related parameters denoted $\theta^g :=$

$\left(\omega^g, \mu_1^g, \ldots, \mu_{Qg}^g, \mathbb{D}_1^g, \ldots, \mathbb{D}_{Qg}^g\right)$. In this thesis, the model with only longitudinal markers in equation (3.22) will be referred as the reference model.

Jointly modelling SNPs with the longitudinal clinical markers

In this thesis, the proposal is to jointly model the SNP data with the longitudinal clinical markers using two different approaches. First, a multivariate model where each longitudinal biomarker (including SNPs), Y_r is modelled using a mixed-effects model. In this first approach, the SNP marker is denoted as $Y_{SNP_i}, (i = 1,2, \ldots p)$ to distinguish it from the longitudinal markers.

The model requires Y_{SNP_i} to be represented by a vector of binary variables ($Y_{SNP_{i1}}$ and $Y_{SNP_{i2}}$) that represent the presence of an allele. Also, it is noted that the genotype of a SNP is constant over time [3], hence the SNP will only be coded for each patient on the first visit. For illustration purpose, assume there is one SNP marker, Y_{SNP_1} and three samples (individuals) with different visits at which the other biomarkers are measured. Table 3.4 shows the transformation of the additive SNP data to binary variables.

Table 3.4: The transformation of additive components of SNP data to binary variables.

Sample	Visits	Original coding (additive)	Coding in joint model	
		Y_{SNP_1}	$Y_{SNP_{11}}$	$Y_{SNP_{12}}$
1	1	0	0	0
1	2	0	NA	NA
1	3	0	NA	NA
1	4	0	NA	NA
2	1	2	0	1
2	2	2	NA	NA
2	3	2	NA	NA

2	4	2	NA	NA
2	5	2	NA	NA
3	1	1	1	0
3	2	1	NA	NA
3	3	1	NA	NA

At this point, each longitudinal marker and the SNP are modelled in each phenotype group using the same set of covariates, $x_{r,j}^{g}$. Unlike the longitudinal clinical markers, the SNP markers were assumed to be independent of both other SNPs and the longitudinal markers. With the additional SNP data as the marker, equation (3.22) can be rewritten as,

$$\begin{cases} h_r^{-1}\{E(Y_{r,j}|b,g)\} = x_{r,j}^{gT}\alpha_r^g + z_{r,j}^{gT}b_r, & r = 1,\dots,R, j = 1,\dots,n_r \\ h_{SNP_i}^{-1}\{E(Y_{SNP_i,j}|g)\} = x_{SNP_i,j}^{gT}\alpha_{SNP_i}^g, & i = 1,\dots,p, j = 1,\dots,n_{SNP_i} \end{cases} \quad (3.24)$$

For the second approach, the SNPs are added as fixed effect covariates. Similar to the reference model, only the longitudinal markers are jointly modelled using a mixed effects model. However, when adding SNP data as the fixed effects to explain the longitudinal evolution of the R longitudinal markers, the covariates information containing the SNP data in equation (3.22) can be written as, $x_{r,j}^{gT} = \left(x_{r,j}^{gT}, X_{SNP_i,j}^{gT}, \dots, X_{SNP_p,j}^{gT}\right)$. By using this approach, the additive component of each SNP needs to be transformed to the binary variables. By using the same example in the previous table (Table 3.4), Table 3.5 shows the transformation of the SNP data when modelled as fixed effects covariates.

Table 3.5: The transformation of additive components of SNP data to binary variables.

Sample	Visits	Original coding (additive)	Coding as fixed effect covariates	
			$X^{g}_{SNP_{11,j}}$	$X^{g}_{SNP_{12,j}}$
1	1	0	0	0
1	2	0	0	0
1	3	0	0	0
1	4	0	0	0
2	1	2	0	1
2	2	2	0	1
2	3	2	0	1
2	4	2	0	1
2	5	2	0	1
3	1	1	1	0
3	2	1	1	0
3	3	1	1	0

Group probabilities for individual patients

Now, the model parameters ψ^{g} and θ^{g} estimated from the MGLMM in each group, can be used to classify a new sample based on their longitudinal history and SNP data. Bayes theorem is applied to calculate the probability of a sample belonging to group g given their longitudinal and covariate data and the model parameters from the MGLMMs fit to samples of known status [117].

$$Pr_{g,new} = \frac{\pi_g \hat{f}_{g,new}}{\sum_{\check{g}=0}^{G-1} \pi_{\check{g}} \hat{f}_{\check{g},new}} \quad g = 0, \dots, G-1, \tag{3.25}$$

where $\hat{f}$ denotes the predictive density of the observed markers given the group and model parameters. The prior probabilities of belonging to each group are denoted by $\pi_g = Pr(g), g = 0, \dots, G-1$. In a Bayesian setting, $f_{g,new}$ is estimated

as the mean of the posterior predictive density estimated from ℵ samples from a Markov Chain Monte Carlo (MCMC) scheme [118].

Prediction methods to specify the predictive density, $f_{g,new}$

There are three different ways to specify the predictive density $f_{g,new}$ considered as mentioned in Hughes et al. (2018) [117]. The three approaches, namely, marginal, conditional and random effects have different focus in predicting the status of the outcome. The *marginal* approach is the most common approach used in LoDA. The new sample (indvidual) is assigned to their specific group to which their longitudinal profiles $Y_{new} = (y_{new,1}, \dots, y_{new,R})$ lie closest. Here, $f_{g,new}$ is taken as the marginal density of Y_{new}. The group membership probabilities are evaluated at each draw of the MCMC procedure and approximate group membership probabilities are calculated as the average across all samples.

For *conditional* approach, $f_{g,new}$ is taken as the conditional density of Y_{new}. Here, the prediction is based on the patient-specific evolution of markers over time, overlooking any error in the variability of the sample's estimated random effects. Similar to the marginal approach, the conditional group membership probabilities are calculated as the average across all samples in the MCMC procedure. Finally, the *random effects* prediction focuses on the sample-specific evolution of the longitudinal markers. Here, $f_{g,new}$ is taken to be the density of the random effects evaluated at the sample and group-specific estimate of the random effect given the marker data. With the MCMC-based Bayesian inference, the estimators of the group probabilities are used for classification.

Classification rules

The sample can now be classified into their specific group by using either the three approaches discussed above. For each approach, the sample is assigned to the group (either $g = 0$ or $g = 1$ for binary phenotype) with the largest probability (i.e. cut off probability of 0.5). Another approach would be to classify a sample into a group only if the probability of belonging to that group is greater than a chosen cut off, $\mathbb{c}$ [25, 116, 117]. The cut off is chosen through the ROC curve of which the best results of classification can be shown (i.e. the closest point to the top left corner).

3.4 Concluding remarks

In this chapter, the methods involved in achieving two main aims were discussed. First, the notations and development of novel filter metric tSNR was explained. The method was proposed not only as univariable variable selection method but also can be used in the multivariable analysis as the model selection criterion. Detailed framework of the application of variable selection and the approach to evaluate the classification performance was shown in Figure 3.1.

Then, the method to jointly model the longitudinal clinical and SNP data was elaborated. Here, detailed description of the modelling approach towards classification using Longitudinal Discriminant Analysis (LoDA) was provided.

In the next chapter, the application of filter metric tSNR will be shown in the simulated datasets, which includes the evaluation of classification performance in both univariable and multivariable settings.

Chapter 4
Simulation Study

4.1 Introduction

This chapter describes the experimental results using simulated datasets. The simulated datasets are useful to evaluate the performance of the proposed method. Normally different settings (e.g. different sample sizes, thresholds, distributions) are determined to carry out specific goal. Having simulated data allows one to evaluate whether a methodology can detect known effects and known associations or group differences [119]. Hence, in this chapter SNP datasets are simulated to evaluate the performance of the novel filter metric, tSNR defined in equation (3.7) (Section 3.2.1).

The datasets are simulated based on the International HapMap Project data as the reference panel [120]. The goal of the project is to map and understand the patterns of common genetic diversity in the human genome in order to accelerate the search for the genetic causes of human disease [121]. Fundamentally, each dataset acquires similar allele frequencies and linkage disequilibrium (LD) structure as the reference panel. Normally, the data is simulated under the null hypothesis with no causal SNPs determined. However, for this study, our datasets are simulated under the alternative hypothesis of which few causal SNPs are specified. The reason is to verify whether tSNR is able to capture the causal SNPs as the informative SNPs during the ranking procedure.

The following section includes the aims of the simulation study (Section 4.2). The data generating mechanism is explained in Section 4.3. Then, in Section 4.4, the methods used for the simulation study are discussed. The results gathered according to the aims are presented in Section 4.5. The chapter is concluded in Section 4.6.

4.2 Aims for simulation study

The aim of the simulation study is threefold: (i) to investigate whether tSNR can capture the causal SNPs as the top SNPs using univariable tSNR ranking, (ii) to explore the tSNR ranking in the multivariable setting, (iii) to evaluate the classification performance in both univariable and multivariable settings.

4.3 Data generating mechanism

The simulation is done using HAPGEN v2.0 [24] software using HapMap3 CEU (Utah residents with Northern and Western European ancestry from the CEPH

collection) as the reference panel. The data is simulated with similar allele frequencies and linkage disequilibrium structure as the reference panel. The simulation generates data for 500 cases (coded as '1') and 500 controls (coded as '0') with 116,415 SNPs on chromosome 1 (ten replicates). The simulation process took 200 to 300 seconds for each replicate on a 3.2 GHz processor PC. However, due to longer process of sample and genotyping QC as well as data pruning, only ten replicates of simulated data are considered. Also, from previous study, Uh et al. (2007) [154], ten replicates of simulated data were able to identify causal SNPs and evaluate classification performance. Two causal SNPs are simulated using a log-additive model as follows:

Table 4.1: The details of the causal SNPs.

SNP	BP	Risk allele	Heterozygote disease risk	Homozygote disease risk	locus index
rs914717	156952983	1	2.20	3.00	66880
rs1130193	200252354	1	5.00	8.30	89466

4.4 Methods

Sample and genotyping QC

Sample and genotyping QC is undertaken as standard data pre-processing procedure. The number of SNPs on chromosome 1 is reduced after applying GWAS QC thresholds based on minor allele frequency (MAF), SNP genotyping rate and test Hardy-Weinberg Equilibrium (HWE) (see Table 4.2). The screening on MAF only includes the SNPs with MAF >0.01. Low MAF SNPs could be more susceptible to genotyping errors and their association signals are less robust [47]. For SNP genotyping rate only SNPs with <10% missing genotypes are included. Further, SNPs that are extremely deviated from HWE (p-value $<10^{-6}$) are removed. At the same time, all 1,000 samples passed the

standard QC procedures (based on rate of missingness, duplication of samples, relatedness and heterozygosity).

Data pruning

The pruning and quality check is done using PLINK software [46]. Similar to Chapter 2 (Literature Review), in this simulation study, the pruning option is undertaken to reduce or eliminate the SNPs that are in approximate LD with each other. The action intuitively can help minimising the computational complexity. Also, it may help focusing on more signals and allow more region of potential interest.

The LD pruning option (150 50 0.9) is applied. For this method, all pairs of SNPs within a window of 150 SNPs, 50 SNPs are compared with each other to measure their pairwise LD. If any pair of SNPs within the window is in LD greater than the R^2 threshold of 0.9, the first SNP in the pair will be inactivated (pruned). Table 4.2 shows the number of SNPs after QC procedure and pruning option are undertaken.

Table 4.2: Number of SNPs after QC and pruning for each replicate (originally there are 116,415 SNPs in each replicate).

Replicate	Number of SNPs after QC	Number of SNPs after pruning
1	96,755	50,047
2	96,887	50,095
3	96,917	50,024
4	96,839	49,989
5	96,632	50,010
6	96,858	50,185
7	96,889	50,133
8	96,829	50,212
9	96,834	50,031
10	96,856	50,037

Univariable ranking using tSNR

The variable selection method, filter metric tSNR (equation 3.7, Section 3.2.1) is used to rank the SNPs in each replicate. As mentioned in Chapter 3 (Methods) the tSNR works within the logistic regression framework. The data for each SNP is fitted in a logistic regression model of which the null and residual deviances are produced. From there, the tSNR is calculated for each SNP. The analysis is implemented using R function '`glm`' [61]. Then, the SNPs are ranked from the highest to the lowest tSNR. Whether tSNR can capture the causal SNPs as the top ranked SNPs is investigated. The workflow of analysis, starting from the univariable tSNR ranking is shown in Figure 4.1 below.

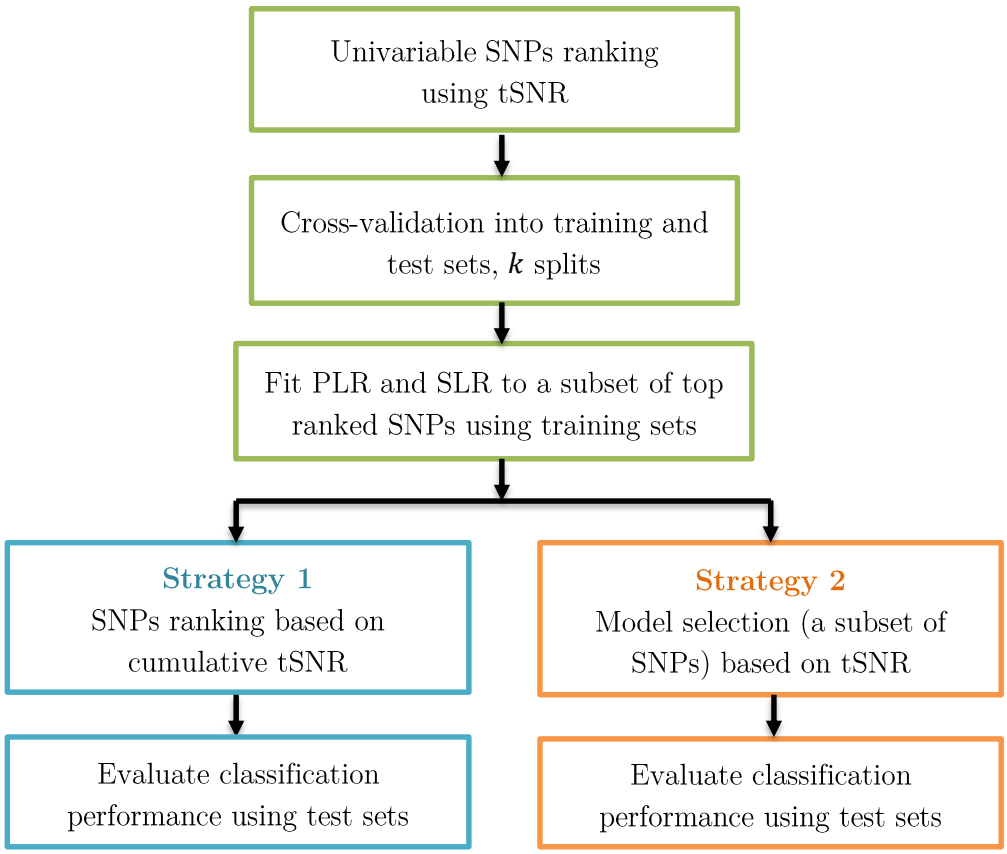


Figure 4.1: Diagram of model building pipeline including (i) univariable tSNR as preselection process; (ii) splits of sample into training and test sets (internal cross-validation); (iii) model building using PLR or SLR; (iv) strategies to select a small subset of SNPs; and (v) model evaluation using test sets.

After the univariable tSNR ranking, a subset of top ranked SNPs (e.g. 100, 200, 300) are selected for the subsequent analysis. Here, the data is divided into training and test sets for 100 times. Particularly in this simulation study, the multivariable selection techniques, penalised logistic regression (PLR) and stepwise logistic regression (SLR) are applied on the training sets. The performance of the two methods are compared. From there, two strategies are proposed to select a subset of SNPs for classification. Strategy 1 mainly ranks the SNPs based on cumulative tSNR (equation 3.12, Section 3.2.2.2). Meanwhile, with Strategy 2, the best model among the non-nested models is selected based on the highest tSNR. Lastly, the classification performance is evaluated on the test sets.

Multivariable selection

The main objective of this thesis is to analyse the SNPs in the multivariable setting. It is deem important to consider the SNPs collectively rather than individually since each SNP might show a different effect when analysed together with other SNPs. In this simulation study, two multivariable selection approaches are compared for model building, namely, penalised logistic regression (PLR) and stepwise logistic regression (SLR).

Penalised Logistic Regression (PLR)

In this thesis, penalised logistic regression (PLR) with lasso penalty is applied for the multivariable selection. PLR is chosen due to its ability to deal with large number of SNPs and works well with categorical variables [60, 65, 103]. PLR applies the embedded approach of variable selection. The penalisation (e.g. Lasso or Ridge) acts as the variable selection method which helps in reducing the

number of SNPs used during the modelling and later classification process. The PLR with lasso penalty is implemented using R function '`glmnet`' [104].

Stepwise Logistic Regression (SLR)

In addition to PLR, SLR is applied to compare the performance between the two methods. The stepwise selection applies both forward selection and backward elimination of the variables. The addition and deletion of each variable is considered using either AIC or BIC. The stepwise selection is attractive since it keeps evaluating the model each time variable is added or eliminated. Park and Hastie (2007) [60] applied the stepwise selection following variable selection using PLR with ridge penalty. They used stepwise selection to further reduce the number of variables in the model. The SLR with AIC as the model selection criterion is implemented using R function '`step`' [61].

Cross-validation

Cross-validation is an important strategy to evaluate the performance of a certain method (e.g. variable selection method). The main idea behind cross-validation is to split data, once or several times of which part of the data (the training set) is used for training the method and the remaining part (the test set) is used for estimating the performance (e.g. error rate, Area under the ROC Curve (AUC)) [122]. Therefore, the cross-validation is an important step that will be considered throughout this thesis. In this simulation study, the data is split into 80% training set (800 samples) and 20% test set (200 samples). The process is repeated for 100 times which produces 100 training and test sets accordingly. The multivariable selection is done using the training sets. Meanwhile, the classification performance is evaluated on the test sets.

Classification performance

As discussed in Section 3.2.2.2, this thesis is focusing on the binary classification using logistic regression as the main classifier. In the classification problem, logistic regression measures the relationship between the outcome (cases coded as '1' or the outcome of interest) and the one or more independent variables, by estimating probabilities using its underlying logistic function. These probabilities (between 0 and 1) are then transformed into either 0 or 1 (binary outcomes) according to the probability threshold specified (usually 0.5). In this chapter, six performance measures that are presented, namely, the Area Under the receiver operating characteristic (ROC) Curve (AUC), the probability of correct classification (PCC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

4.5 Results

In this section, the results from the simulation study are presented. Each analysis is carried out with specific aim that relates to the variable selection method, filter metric tSNR and the workflow presented earlier (Figure 4.1).

4.5.1 Univariable ranking using filter metric tSNR

Here, the SNPs ranking is presented based on the univariable tSNR ranking. The tSNR is calculated for each SNP in each replication using the pruned datasets. The SNPs are ranked from the highest to the lowest tSNR. Table 4.3 shows the ranking of top 10 SNPs for each replicate. The causal SNPs are highlighted in bold. It can be shown that tSNR is able to rank the causal SNPs as the top ranking SNPs using the simulated datasets.

Table 4.3: The univariable tSNR ranking (top 10 SNPs) for the 10 replicates.
The causal SNPs rs1130193 and rs914717 are highlighted in bold in each replicate.

Rank	Replicate 1		Replicate 2		Replicate 3		Replicate 4		Replicate 5	
	SNPs	**tSNR**	**SNPs**	**tSNR**	**SNPs**	**tSNR**	**SNPs**	**tSNR**	**SNPs**	**tSNR**
1	**rs1130193**	0.1071	**rs1130193**	0.1058	**rs1130193**	0.1025	**rs1130193**	0.0976	**rs1130193**	0.0864
2	rs4950760	0.0612	rs10920304	0.0531	rs12756809	0.0756	rs10920304	0.0455	rs10920304	0.0455
3	rs10920304	0.0597	rs4630172	0.0428	rs10920304	0.0562	rs3820439	0.0430	**rs914717**	0.0415
4	rs2819362	0.0504	rs2819362	0.0403	rs2511200	0.0478	rs2819362	0.0386	rs2511200	0.0391
5	rs12731187	0.0404	rs4950760	0.0365	**rs914717**	0.0456	rs432335	0.0317	rs12731187	0.0349
6	**rs914717**	0.0337	**rs914717**	0.0307	rs4630172	0.0423	**rs914717**	0.0290	rs4950760	0.0333
7	rs7516412	0.0319	rs2511200	0.0288	rs2819362	0.0347	rs10489843	0.0273	rs2819365	0.0306
8	rs2511200	0.0304	rs6658647	0.0287	rs6658647	0.0331	rs2511200	0.0265	rs2819362	0.0297
9	rs2735784	0.0255	rs9427715	0.0280	rs10489842	0.0317	rs16849483	0.0246	rs9427715	0.0278
10	rs12023371	0.0237	rs16849483	0.0271	rs10489843	0.0268	rs4950802	0.0226	rs12747653	0.0250

Rank	Replicate 6		Replicate 7		Replicate 8		Replicate 9		Replicate 10	
	SNPs	**tSNR**	**SNPs**	**tSNR**	**SNPs**	**tSNR**	**SNPs**	**tSNR**	**SNPs**	**tSNR**
1	**rs1130193**	0.0790	**rs1130193**	0.1016	**rs1130193**	0.1006	**rs1130193**	0.1078	**rs1130193**	0.1074
2	rs4950760	0.0443	rs10920304	0.0704	rs10920304	0.0564	rs10920304	0.0607	rs10920304	0.0557
3	rs2819362	0.0411	rs2819362	0.0540	**rs914717**	0.0527	rs2819362	0.0520	rs4950760	0.0510
4	rs10920304	0.0399	rs9427715	0.0313	rs2511200	0.0496	rs2511200	0.0320	rs2819362	0.0461
5	rs7516412	0.0376	rs4630172	0.0308	rs4950760	0.0470	rs7516412	0.0317	rs2511200	0.0423
6	**rs914717**	0.0246	**rs914717**	0.0276	rs7516412	0.0467	**rs914717**	0.0305	**rs914717**	0.0365
7	rs857705	0.0241	rs2511200	0.0268	rs2819362	0.0464	rs10489843	0.0282	rs6658647	0.0354
8	rs6658647	0.0236	rs2735784	0.0259	rs6658647	0.0381	rs2735784	0.0279	rs4630172	0.0354
9	rs2735784	0.0231	rs2819360	0.0232	rs2735784	0.0375	rs6658647	0.0263	rs12731187	0.0290
10	rs2819360	0.0221	rs4950760	0.0218	rs12747653	0.0375	rs16849483	0.0253	rs10489843	0.0265

In order to see the relationship between the univariable ranking and the classification performance, the data from Replicate 1 is used for the analysis. The data is divided into training (80%) and test sets (20%). The process is repeated for 100 times. The SNP is added one by one into the logistic regression model based on the univariable tSNR ranking using the training sets. In each addition of SNP, the classification performance is measured on the test sets. Figure 4.2 shows the classification performance (mean) of top 200 SNPs based on the univariable tSNR ranking.

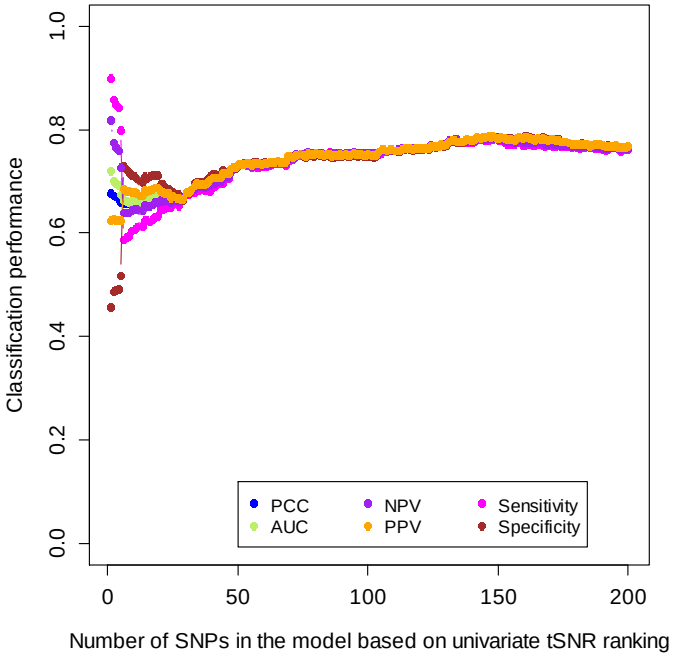


Figure 4.2: The classification performance based on the univariable tSNR ranking for Replicate 1.

As we know, the first SNP is the causal SNP, rs1130193 which was specified as highly informative. Overall, from Figure 4.2 the classification performance ranges between 60% and 80% and remains constant towards the end. Although the simulated causal SNPs are already in the model after six SNPs, the classification performance keeps increasing even the additional SNPs are assumed not relevant. This is because, when simulating the data from a whole chromosome, there are

some SNPs that may be informative by chance. The effect of each simulated SNP is influenced by several factors, namely, sample size, effect size, allele frequency and LD to the causal SNP [24]. Hence, it explains the increasing trend of the classification performance shown in Figure 4.2.

4.5.2 Comparing classification performance between penalised logistic regression (PLR) and stepwise logistic regression (SLR)

Based on the workflow proposed in Figure 4.1, PLR is chosen as the classification method following the univariable ranking. The variable selection within PLR is done using Lasso penalty in order to shrink the coefficients of uninformative variables to zero. On the other hand, SLR is a well-known modelling approach in many fields. The stepwise procedure applies the model selection criterion such as AIC to include or eliminate the variables.

Given the different variable selection approaches in both methods, it will be useful to compare the performance between the two methods. Here, the classification performance using PLR and SLR is compared using the simulated dataset (Replicate 1). The analysis is done in the cross-validation setting by dividing the data into training (80%) and test (20%) sets. The splitting is repeated 100 times. It is important to note, for this comparison, the data is trained only on one training set in order to produce only one model. However, the classification performance is evaluated using all 100 test sets.

The analysis of PLR and SLR are done using the '`glmnet`' and '`step`' functions in R accordingly. The results for both methods are presented in Table 4.4. The initial number of SNPs indicates the number of top SNPs based on the

univariable tSNR ranking in Replicate 1. Then, after applying the variable selection method, these SNPs are reduced to a certain number (refer to column 'Number of SNPs selected'). The classification performance is presented as the average values of PCC, AUC, sensitivity, specificity, PPV and NPV over the 100 test sets.

Table 4.4: The classification performance for PLR and SLR using the SNP data from Replicate 1.

Penalised Logistic Regression (PLR)								
Initial number of SNPs	**Time (sec)**	**Number of SNPs selected**	**PCC**	**AUC**	**Sens**	**Spec**	**PPV**	**NPV**
100	1.83	67	80.97	81.13	83.16	78.78	79.77	82.48
200	3.01	136	89.49	89.55	90.58	88.39	88.69	90.40
500	4.76	252	95.80	95.84	96.74	94.86	94.99	96.68
1000	4.77	357	95.08	95.13	96.38	93.77	93.98	96.29
2000	6.77	370	93.02	93.12	95.05	90.98	91.41	94.83
3000	10.06	358	92.26	92.33	93.82	90.69	91.05	93.62
4000	11.79	237	87.60	87.66	88.57	86.62	86.94	88.38
5000	15.13	22	69.53	69.67	66.46	72.59	70.87	68.47

Note: sec = seconds, Sens = Sensitivity, Spec = Specificity

Stepwise Logistic Regression (SLR)								
Initial number of SNPs	**Time (sec)**	**Number of SNPs selected**	**PCC**	**AUC**	**Sens**	**Spec**	**PPV**	**NPV**
100	42.10	46	80.31	80.33	79.10	81.55	81.00	79.67
200	729.79	102	89.86	89.85	90.27	89.48	89.48	90.20

Note: sec = seconds, Sens = Sensitivity, Spec = Specificity

From the table, it can be shown that computationally, PLR is less complex as compared to SLR based on the time (in seconds) recorded. In terms of the number of SNPs selected, SLR tends to select much lower number of SNPs as compared to PLR (for the first 200 SNPs). However, better classification performance is shown when using PLR in most of the scenarios presented above. Additionally, it is important to note that, the stepwise procedure flagged a

warning message of failed convergence after 500 SNPs were supplied into the model. In general, generalised linear models (GLMs) are fit by maximising the log-likelihood function, where the resultant maximum is referred to as the maximum likelihood estimate (MLE) [123]. Failed convergence occurs whenever the maximising process fails to find the MLE. Thus, based on the results using the simulated dataset presented above, PLR is more reliable when dealing with high-dimensional data as compared to SLR.

4.5.3 Strategy 1: Multivariable ranking using tSNR (cumulative tSNR ranking)

One of the important aspect considered in this thesis is cross-validation. Cross-validation is important to avoid an overoptimistic result when an algorithm is trained and evaluated on the same data [122]. With cross-validation, the data is divided into training set for model building and test set for classification. The process is then repeated multiple times (e.g. 5, 10, 100 or 1000). Normally, a specific model with predetermined number of variables is used. However, variable selection needs to be carried out here before a model is finalised.

As a result, with multiple number of training sets, different number of models (subsets of SNPs) are produced. In this thesis, two strategies are proposed. Strategy 1 mainly ranks the SNPs based on the cumulative tSNR for each SNP. Firstly, the tSNR is calculated for each model based on the 100 split. Then, the tSNR value of each model is multiplied to each of the SNPs presents within the model. The accumulated tSNR of each SNP will then determine the tSNR weight it carries. The SNPs are then ranked based on their cumulative tSNR across 100 models. Usai et al. [124] applied a somewhat similar strategy but by using frequencies of each SNP appeared in multiple training sets.

The ranking of top 10 SNPs using the cumulative tSNR ranking for the ten replicates is shown in Table 4.5. It can be shown that the ranking within the top 10 SNPs changes (compared to Table 4.3). In Replicates 1, 2, 5, 6, 7 and 8, the simulated causal SNPs are ranked higher than the univariable ranking. However, in the other four replicates the SNP rs914717 is not captured as the top ten ranked SNPs. This is expected due to the average odds ratio simulated for the SNP of which other SNPs might hold similar odds ratio as well. In addition, with simulated data, the SNPs may be informative when put in the model jointly, but may be less informative in some cases.

Table 4.5: The cumulative tSNR ranking (top 10 SNPs) for the 10 replicates.
The causal SNPs rs1130193 and rs914717 are highlighted in bold in each replicate.

Rank	Replicate 1		Replicate 2		Replicate 3		Replicate 4		Replicate 5	
	SNPs	Cumulative tSNR	SNPs	Cumulative tSNR	SNPs	Cumulative tSNR	SNPs	Cumulative tSNR	SNPs	Cumulative tSNR
1	**rs1130193**	34.21	**rs1130193**	26.57	**rs1130193**	31.65	**rs1130193**	19.83	**rs1130193**	10.92
2	**rs914717**	34.21	rs2151158	26.23	rs2511200	31.65	rs432335	19.83	**rs914717**	10.92
3	rs7530949	32.99	rs10913887	25.42	rs12410250	27.09	rs10489843	15.24	rs12744678	3.22
4	rs1128400	27.59	rs1339411	23.37	rs12409786	25.49	rs11579514	12.47	rs1324659	2.84
5	rs11803397	26.48	**rs914717**	23.25	rs4845803	24.28	rs2279127	12.14	rs6671643	2.26
6	rs10493765	25.68	rs2686226	20.78	rs644690	23.88	rs973742	11.50	rs12046563	2.15
7	rs41451049	23.73	rs11582767	20.06	rs12045948	23.76	rs12068503	10.46	rs623229	2.08
8	rs7539051	23.50	rs6658647	18.41	rs17112247	22.67	rs482542	9.82	rs6685614	2.06
9	rs2800804	23.04	rs2796077	16.74	rs12026094	22.23	rs4846778	9.69	rs6704463	1.89
10	rs2494606	22.89	rs17443748	15.84	rs16849342	20.53	rs6541199	9.66	rs2352039	1.89

Rank	Replicate 6		Replicate 7		Replicate 8		Replicate 9		Replicate 10	
	SNPs	Cumulative tSNR	SNPs	Cumulative tSNR	SNPs	Cumulative tSNR	SNPs	Cumulative tSNR	SNPs	Cumulative tSNR
1	**rs1130193**	22.75	**rs1130193**	34.98	**rs1130193**	29.89	**rs1130193**	17.57	**rs1130193**	30.55
2	rs17047242	19.56	rs1570089	33.54	**rs914717**	29.89	rs2511200	15.66	rs2511200	30.55
3	**rs914717**	19.52	**rs914717**	32.84	rs1570565	23.17	rs1201157	12.99	rs12085891	26.60
4	rs4650146	18.28	rs2820477	30.54	rs10888744	22.31	rs12747653	12.12	rs16836808	26.44
5	rs4950760	16.51	rs12037907	29.35	rs670318	21.75	rs4847196	10.32	rs2119192	24.24
6	rs1341467	16.08	rs10919407	28.58	rs650755	20.54	rs6704394	9.45	rs4970934	23.05
7	rs6660884	14.77	rs11161686	28.48	rs10798056	20.17	rs17124643	9.38	rs2483688	21.74
8	rs6686126	14.43	rs41431546	26.18	rs6661325	19.42	rs7514221	8.63	rs6658647	21.45
9	rs16857312	14.43	rs2982464	26.05	rs2025582	17.66	rs10494795	8.38	rs7547731	20.73
10	rs6658647	14.28	rs284113	25.31	rs12091463	17.65	rs1406859	8.29	rs6677274	20.37

Similar to the univariable ranking earlier, the relationship between the cumulative tSNR ranking and the classification performance is investigated. The data from Replicate 1 is used for the analysis. The data is divided into training (80%) and test (20%) sets. The SNP is added one by one into the logistic regression model based on the cumulative tSNR ranking using the training sets. In each addition of SNP, the classification performance is measured on the test sets. The process is repeated for 100 times. Figure 4.3 shows the classification performance of top 200 SNPs based on the cumulative tSNR ranking.

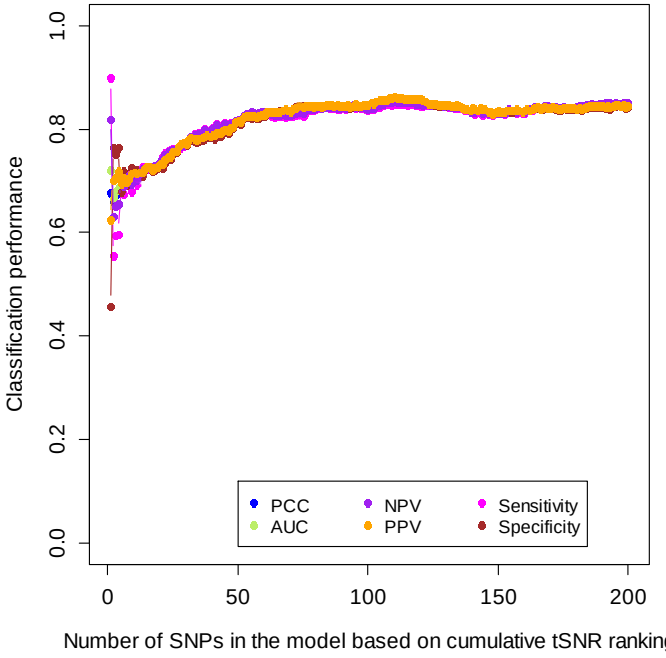


Figure 4.3: The classification performance based on the cumulative tSNR ranking for Replicate 1.

From Figure 4.3 it can be seen that there is an improvement in classification accuracies when using the cumulative tSNR ranking as compared to the univariable ranking (Figure 4.2). For example, the PCC reaches 70% at 8-th SNP and 80% at 45-th SNP when using the cumulative tSNR ranking. However, with the univariable tSNR ranking the SNP reaches 70% at 40-th SNP and still below 80% at 200-th SNP.

Stopping criterion

By using the forward selection procedure in Strategy 1, the final number of SNPs to be in the model needs to be determined. In this thesis, the proposal is to use the adjusted tSNR in equation 3.8 (Section 3.2.1) to determine the number of SNPs to be included in the model. Hence, as an extension to Figure 4.3, the relationship between the classification performance and adjusted tSNR is investigated. Figure 4.4 shows the adjusted tSNR of top 200 SNPs based on the cumulative tSNR ranking .

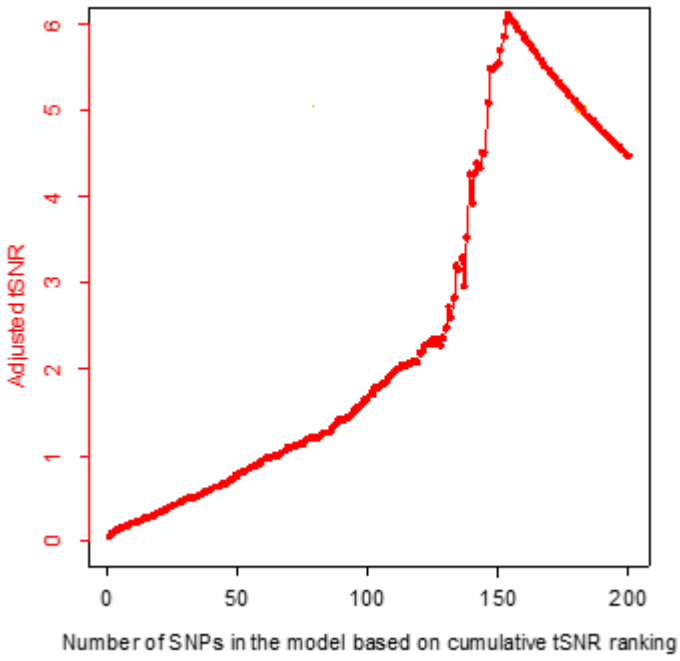


Figure 4.4: The adjusted tSNR against the number of top 200 SNPs (based on cumulative tSNR ranking) for Replicate 1.

It can be seen that the tSNR values keep increasing as the number of SNPs increases. Nevertheless, after 140 SNPs the tSNR values fluctuate indicating the inclusion of the SNP may or may not be informative to the model. Hence, further investigation of the specific SNP can be carried out. Finally, after 154 SNPs the tSNR values show a constant pattern of decreasing which shows the additional SNPs bring less signal as compared to noise. The drop is likely due to the

adjustment related to the number of variables included in the model which is adapted in the adjusted tSNR formula (previously shown in Equation 3.8). Hence, the suggestion is to use the model with 154 SNPs as the final model for classification and to further investigate the SNPs consist in the model.

4.5.4 Strategy 2: Model selection using tSNR

Strategy 2 applies tSNR to select the best model among the 100 models that are produced based on the 100 splits. The tSNR is calculated for each model and from there each model is ranked based on the highest to the lowest tSNR values. Table 4.6 summarises the classification performance for the best model selected in each replicate. On average, the classification performance is satisfactory with more than 80% in each measure.

Table 4.6: Summary of the classification performance (mean and standard deviation) of the model selected by tSNR based on the 100 splits.

	Replicate 1		**Replicate 2**		**Replicate 3**		**Replicate 4**		**Replicate 5**	
Number of SNPs in the selected model	189		179		194		204		170	
Classification performance	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
PCC	85.39	0.03	83.34	0.04	85.85	0.03	86.37	0.03	84.54	0.03
AUC	85.45	0.03	83.39	0.04	85.97	0.03	86.44	0.03	84.66	0.03
Sensitivity	85.60	0.04	83.41	0.04	87.91	0.04	85.73	0.04	86.24	0.04
Specificity	85.18	0.04	83.27	0.05	83.78	0.04	87.01	0.04	82.84	0.04
PPV	85.32	0.04	83.37	0.04	84.50	0.04	86.88	0.04	83.50	0.04
NPV	85.58	0.03	83.42	0.04	87.44	0.04	86.00	0.04	85.82	0.04

	Replicate 6		**Replicate 7**		**Replicate 8**		**Replicate 9**		**Replicate 10**	
Number of SNPs in the selected model	195		184		192		189		187	
Classification performance	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
PCC	84.55	0.04	85.53	0.03	86.42	0.04	86.37	0.03	84.09	0.03
AUC	84.63	0.04	85.60	0.03	86.49	0.04	86.43	0.03	84.23	0.03
Sensitivity	85.77	0.04	84.50	0.04	86.88	0.04	87.13	0.04	86.36	0.04
Specificity	83.33	0.04	86.56	0.04	85.96	0.04	85.61	0.04	81.82	0.04
PPV	83.80	0.04	86.31	0.04	86.17	0.04	85.87	0.03	82.70	0.04
NPV	85.47	0.04	84.90	0.04	86.81	0.04	87.00	0.04	85.76	0.04

4.6 Concluding remarks

In this simulation study, a novel filter metric approach, tSNR is studied. Ten replicates of SNP datasets with binary outcomes were simulated including two causal SNPs. The univariable ranking using tSNR was applied to the datasets of which the ranking of the causal SNPs were observed. The results show that tSNR was able to capture the causal SNPs within the top ten ranked SNPs.

To date, the existing published work on SNPs selection has been focusing on univariable approach. One of the aims of this thesis is to analyse the SNP data in the multivariable setting. Following Figure 4.1 shown above, the PLR was proposed as a multivariable modelling method. The performance between PLR and SLR was compared. The results demonstrated that PLR outperformed the SLR in terms of computing time and its ability to work with high-dimensional data.

Further, following the application of PLR in the cross-validation setting (100 splits), multiple models with different subsets of SNPs were produced. Hence, two strategies were proposed in order to select a subset of SNPs. In Strategy 1, the SNPs are ranked based on the tSNR weighting calculated for each model. The weighting were accumulated across the 100 models. The new ranking is called cumulative tSNR ranking. The multivariable ranking shows different ranks when compared to the univariable tSNR ranking. The cumulative tSNR ranking managed to rank the two causal SNPs higher than the univariable ranking in six of the ten replicates. In addition, the classification performance was improved when including the SNPs based on the cumulative tSNR ranking as compared to the univariable tSNR ranking.

Then in Strategy 2, the best subset of SNPs was selected based on the tSNR value of the model. As mentioned in Chapter 3 (Methods), the tSNR is useful to compare the non-nested models. Hence, by applying the knowledge, the best model was determined based on the highest tSNR among the 100 non-nested models produced by the splitting for cross-validation. The classification performance in the ten replicates is promising with more than 80% for all measures.

From the results shown above, we are confident that the filter metric tSNR and the model building pipeline shown in Figure 4.1 can be applied to real datasets. In the next chapter, the methods will be applied to the datasets from Epilepsy Pharmacogenomics (EpiPGX) and Standard and New Antiepileptic Drugs (SANAD) studies.

Chapter 5
Clinical Applications: tSNR as Variable Selection Method for Classification

5.1 Introduction

Over the years, the interest to study GWAS has grown, and the number of SNPs genotyped has subsequently increased significantly. The objective of a GWAS is twofold; i) to identify the subset of SNPs that best explains the heredity component of the outcome of interest (e.g. disease status or response to treatment), and ii) to generate a rule for classifying patients into their phenotype groups (e.g. healthy or disease, positive or negative response to treatment for binary outcome instance), given their genetic profile and possibly other clinical variables [67, 125]. However, the tasks of variable selection and classification when dealing with a large number of SNPs is challenging.

Normally, GWAS data analyse one SNP at a time by linear or logistic regression [18, 126]. The resulting p-values are then used to rank the SNPs and to select those with a p-value smaller than a pre-specified significance level. However, there are at least two strong reasons for considering all the SNPs or at least a large subset of them simultaneously. First, the marginal effects of SNPs (i.e. the effect of each SNP on the outcome when it is considered alone) may be quite different when the joint effects of multiple SNPs, and therefore their correlation, is taken into account. Second, the predictive power of a single SNP alone can be low.

However, fitting a large number of variables altogether in any regression model may cause instability and overfitting. These problems can be mitigated using penalisation methods, which shrink the coefficients of the regression model towards zero; the extent of shrinkage is controlled by a penalisation parameter (e.g. Lasso, Ridge)[127].

As discussed in Chapter 3 (Methods), a novel variable selection method, tSNR is proposed to select the most informative SNPs. The univariable filter metric tSNR is adopted in a logistic regression framework to measure the signal-to-noise ratio of a variable when added to a model which involves just a constant. The tSNR is used to rank the SNPs which enable only a subset of informative SNPs to be analysed. This step is important to reduce the complexity of the dataset as well as the computational burden.

The objectives of this chapter are:

(i) To explore the tSNR method (equation 3.7, Section 3.2.1) on the real datasets. The methodology is illustrated using two GWAS datasets of patients with epilepsy, where the aim is to identify patients who will not achieve remission of seizures on their first well-tolerated antiepileptic drug (AED).

(ii) To propose a specific workflow to analyse the SNPs by considering multivariable approach. The building of the workflow considers the following; i) univariable or multivariable selection, ii) cross-validation, and iii) evaluation of classification performance.

The datasets that are used here consist of both clinical and genetic information of patients from the Epilepsy Pharmacogenomics (EpiPGX) consortium [128], which includes datasets contributed from several different epilepsy cohorts. The first dataset, or development set, consists of 1,655 patients from the RCSI/Irish, UCL/London, Brussels, EKUT/Tubingen, Glasgow, DoH/ReJuMEC and Melbourne cohorts within EpiPGX. Meanwhile, the second dataset or validation set belongs to 818 patients from the Standard and New Antiepileptic Drugs (SANAD) study cohort [129, 130] also within EpiPGX.

In what follows, a description of each dataset and methods of variable selection are described in Section 5.2. Then, the results from the analysis are presented in Section 5.3. The concluding remarks can be found in Section 5.4.

5.2 Methods

In this section, a detailed description of the datasets and the quality control (QC) process involved are described. The methods used in this chapter are briefly discussed (see Chapter 3 (Methods), for comprehensive description of the methods).

5.2.1 The EpiPGX dataset

It is well known that genetic factors play a role in AED response [131]. The purpose of the EpiPGX consortium study is to identify genome-based predictive biomarkers for use in routine clinical practice to personalise treatment of epilepsy with existing AEDs [128]. The work within EpiPGX study is broken down into work packages (WPs). Specifically, the datasets that are utilised in this chapter are those used for work package 2 (WP02). WP02 involves the study of several pharmacogenomic phenotypes of interest, including failure of the first AED. The patients in this study had been followed-up prospectively from initial diagnosis and treatment initiation until specific reaction associated with their first well-tolerated AED is shown [128].

Only patients with complete phenotype status are considered for analysis, 1,515 patients from the development set and 639 patients from the validation set. The first dataset will be used for variable selection and to develop the model whilst the second dataset will be utilised to externally validate the model generated from the first dataset. The reason for this is that most reports evaluating prediction models focus on the issue of internal validity, leaving the important issue of external validity behind [132]. The external validation is applied with

the aims to address the accuracy of a model in patients from a different but plausibly related population.

Phenotype definition

In this study, the phenotype (i.e. dependent variable) is defined as the remission status (seizure-free) of patients after receiving first well-tolerated AED within a 5-year follow-up period. The phenotype for each patient is coded as '1' if they did not achieve remission after receiving first well-tolerated AED, whilst the phenotype is coded as '0' if they were observed to achieve remission. Figure 5.1 summarises the two datasets within the EpiPGX consortium.

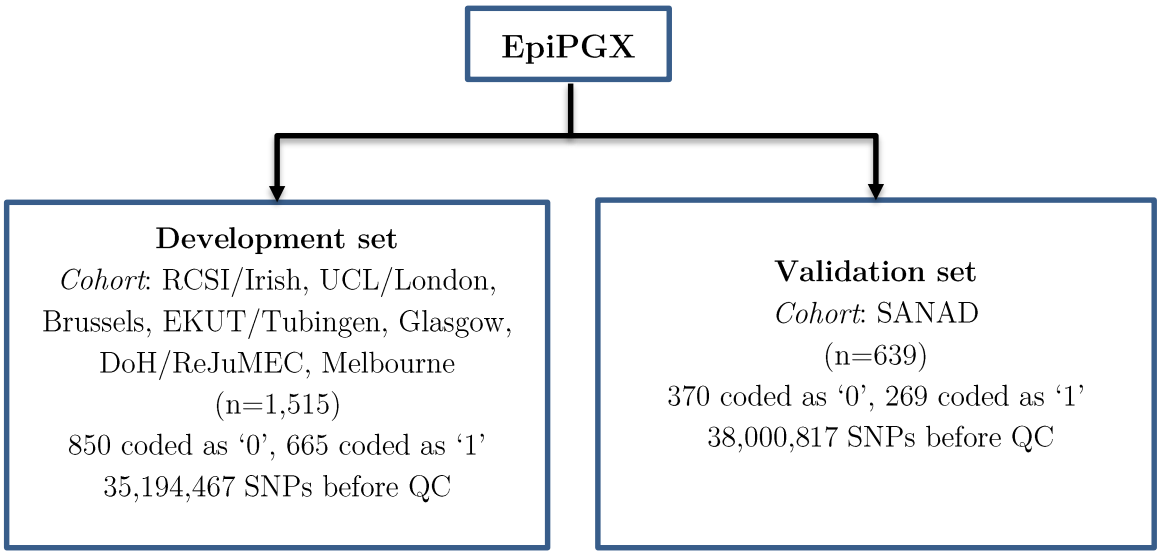


Figure 5.1: Summary of the two datasets; 1) *Development set* for variable selection and model development, and 2) *Validation set* for external validation.

5.2.2 Sample and genotyping QC

Development set

The preparation of genotype data follows a standard quality assurance procedure. Initially, the genotype data consists of 35,194,467 SNPs across 22 chromosomes. Applying standard QC procedures to each SNP, this number is

first reduced to 6,175,331 after applying GWAS thresholds based on minor allele frequency (MAF), SNP genotyping rate and Hardy-Weinberg Equilibrium (HWE). The screening on MAF only includes the SNPs with MAF >0.01. Low MAF SNPs could be more susceptible to genotyping errors and their association signals are less robust [47]. For SNP genotyping rate only SNPs with <10% missing genotypes are included. SNPs that are extremely deviated from HWE are usually removed. Hence, setting the threshold of p-value $<10^{-6}$ implies only one SNP per million be removed when HWE holds. At the same time, all samples passed the standard QC procedure (e.g. inspection for missingness, duplicates or related samples and heterozygosity) of which 1,515 samples were included for further analysis.

Validation set

The external validation is important to make sure the findings can be replicated on 'unseen' data (a different set of patients). The validation set also undergoes similar genotype QC as the development set. Initially, the genotype data consists of 38,000,817 SNPs across 22 chromosomes. This number is reduced to 7,459,851 after applying SNP QC thresholds based on minor allele frequency (including SNPs with MAF > 0.01), SNP genotyping rate (including SNPs with missing genotype rate <10% missing genotypes for the SNP) and for a test for Hardy-Weinberg Equilibrium (including only those with HWE p-value $> 10^{-6}$). Similar to the development set, all samples passed the standard sample QC procedures. Hence, 639 samples were included for further analysis.

5.2.3 Data pruning

Linkage Disequilibrium (LD) pruning is an important quality assurance step for GWAS analysis to reduce the complexity of data. In this thesis, data pruning is

undertaken to reduce the number of SNPs by eliminating the SNPs that are in high LD with each other. It has been shown that some analyses for association obtain better results if the markers used are not in LD with each other [48]. LD refers to the non-random association of alleles at two or more loci in a general population [133]. Therefore, the pruning option is undertaken to reduce or eliminate the SNPs that is in approximate LD with each other thus intuitively reducing the complexity of data.

For both development and validation sets, the pruning option (150 50 0.90) is implemented using PLINK [46] software. For this method, consider a window of 150 SNPs, 50 SNPs are compared with each other to measure their pairwise LD. If any pair of SNPs within the window is in LD greater than the R^2 threshold of 0.9, the first SNP in the pair will be inactivated (pruned). Shift the window 150 SNPs forward and repeat the procedure. After applying the LD pruning, the SNPs in the development and validation sets are reduced to 1,437,725 SNPs and 1,814,949 SNPs accordingly.

Then, in order to make sure that both datasets have similar characteristics, only overlapping SNPs between development and validation sets are considered for further analysis. This action assures that the SNPs selected using the development set are definitely present for validation purposes. The selection of overlapping SNPs brings to the final number of 1,084,548 SNPs in each dataset.

5.2.4 Univariable SNPs ranking using tSNR

Univariable ranking of the SNPs is first performed to reduce the number of SNPs that is analysed during the multivariable approach. This step is important to make sure that only the potentially informative SNPs are considered for further

analysis. The selection of a subset of SNPs will help to reduce the computational burden when dealing with multiple variables in the subsequent step.

By using the pruned dataset, the tSNR value is calculated based upon the residual deviance of a SNP presence in the logistic regression model, for each SNP, $X_i, i = 1,2,\dots,p$ (taking values of 0, 1 or 2 according to the number of minor alleles present). The SNP is coded as such throughout this thesis (unless stated otherwise) assuming additive mode of inheritance.

To recall, the $\widehat{tSNR_i}$ estimate is given by,

$$\widehat{tSNR_i} = \frac{Dev\left(y, \hat{\beta}_0\right) - Dev\left(y, X_i, \hat{\beta}_i\right)}{Dev\left(y, X_i, \hat{\beta}_i\right)} \tag{5.1}$$

where $Dev\left(y, \hat{\beta}_0\right)$ is the null deviance of the null model showing how well the response variable is predicted by a model that includes only the intercept, $\hat{\beta}_0$. Meanwhile, $Dev\left(y, X_i, \hat{\beta}_i\right)$ refers to the deviance of the fitted model which includes the intercept, $\hat{\beta}_0$ and the estimated coefficient, $\hat{\beta}_i$ associated with the i-th SNP, X_i $(i = 1,2,\dots,p)$.

5.2.5 Multivariable approach of SNPs ranking and model selection for classification

Following the univariable tSNR ranking, a subset of top 5,000 SNPs are chosen for the multivariable analysis. Zhou and Wang (2007) [14, 15] has taken similar approach by choosing a maximum of 1,000 SNPs, following a univariable ranking measure across four million SNPs. Figure 5.2 shows the diagram of model building pipeline as discussed in Chapter 3 (Methods).

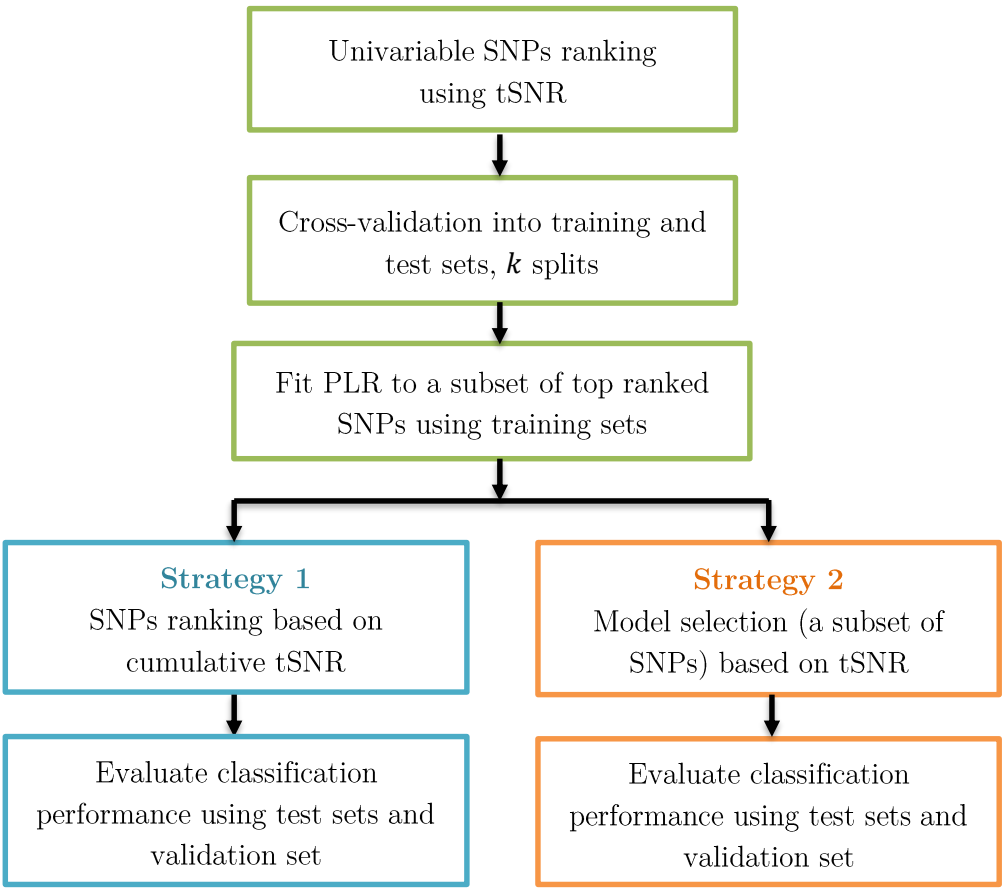


Figure 5.2: Diagram of model building pipeline including (i) univariable tSNR as preselection process; (ii) splits of sample into training and test sets (internal cross-validation); (iii) model building using penalised logistic regression (PLR); (iv) strategies to select a small subset of SNPs; and (v) model evaluation using test sets and validation set (external validation).

Briefly, after the univariable SNPs selection, the subset of SNPs are used for PLR model. Due to the few splits during cross-validation step, two strategies are proposed to select the final predictive model. Finally, the classification performance using the chosen model in each strategy is evaluated internally and externally.

Cross-validation

A cross-validation is also the part of the pipeline shown in Figure 5.2 and it is undertaken to assess the results of the analysis on independent sets. By training and testing the model containing selected variables on different subsets of the data, the strength of prediction for classification can be evaluated [105]. The development set is randomly stratified into training (80% of patients $\mathsf{n = 1{,}212}$) and test (20% of patients $\mathsf{n = 303}$) datasets. The dataset is split k times (i.e. 100 splits) which will then produce 100 different models [134]. In each split, the model (PLR) is fitted to the training set. The classification performance is then evaluated on the test sets.

Penalised Logistic Regression (PLR)

Using the training set, PLR is fitted to a subset of 5,000 SNPs using the R CRAN packages '`glmnet`' [135]. PLR is useful to consider a subset of SNPs simultaneously. The penalisation method, least absolute shrinkage and selection operator (lasso) is applied to reduce the number of SNPs. Lasso works as the variable selection method imposed in the regression models by shrinking the least informative variables' coefficient to zero. The lasso is computationally feasible in dealing with a large number of SNPs simultaneously.

Since the PLR is performed within a cross-validation framework, a final subset of SNPs need to be determined. For that reason, two strategies are analysed and proposed to decide the final subset of SNPs. The strategies apply the tSNR metric as the variable or model selection criterion as detailed below.

Strategy 1: Multivariable ranking using tSNR (cumulative tSNR ranking)

Strategy 1 mainly ranks the SNPs based on the cumulative tSNR for each SNP. Firstly, the tSNR is calculated for each model based on the 100 split. Then, the tSNR value of each model is multiplied to each of the SNPs presents within the model. The weight for each SNP, ω_i is the summation of the tSNR across the k models,

$$\omega_i = \sum_{k=1}^{100} \widehat{tSNR} \tag{5.2}$$

The following Table 5.1 illustrates the calculation of cumulative tSNR :

Table 5.1: Calculation of cumulative tSNR for three SNPs (X_1,X_2 and X_3).

Splits, k	SNPs selected by PLR within each split	tSNR value of the model	Assign tSNR of the model to each SNP		
			X_1	X_2	X_3
1	$X_1 + X_2 + X_3$	1.223	1.223	1.223	1.223
2	X_1	1.012	1.012		
3	$X_1 + X_2$	1.530	1.530	1.530	
:	:	:	:	:	:
100	$X_2 + X_3$	1.683		1.683	1.683
Cumulative tSNR for each SNP, ω_i			ω_1	ω_2	ω_3

The SNPs are then ranked from the highest cumulative tSNR to the lowest. The SNPs are then added to the logistic model one by one of which the classification performance is evaluated. For prediction accuracy and interpretability, only a small subset of SNPs is preferred [60].

In Strategy 1, the final model is determined when there is no increment in the adjusted tSNR value which is given by,

$$\widehat{tSNR}\ adjusted = \frac{Dev\left(y, \hat{\beta}_0\right) - Dev\left(y, X_i, \hat{\beta}_i\right) + d_0 - d_i}{Dev\left(y, X_i, \hat{\beta}_i\right) + d_i} \qquad (5.3)$$

where d_0 is the number of coefficients in the null model, and therefore is equal to one, meanwhile d_i is the number of coefficients in the fitted model.

Strategy 2: Model selection using tSNR

A second strategy for multivariable SNPs selection is based on comparison of PLR models containing multiple SNPs (Figure 5.2). The comparison is formally done via the tSNR values. Briefly, first by using the k splits, k different models which differ in terms of samples included as well as SNPs selected are produced. These models are considered non-nested, i.e., two models are non-nested, either partially or strictly, if one model cannot be reduced to the other model by imposing a set of restrictions on the coefficients, β_i [106]. Therefore, a model selection criterion is needed to select the best model. As mentioned in Chapter 3 (Methods), one of the advantages of tSNR is that it can be used to compare the non-nested models. Hence, by implementing the idea, the fitted PLR models are ranked based on the highest tSNR to the lowest tSNR values. The SNPs that correspond to the highest ranked model are selected for the evaluation of classification performance.

Classification performance

Novel aspects of the analysis includes the use of multivariable logistic regression that examines the contribution of SNPs collectively rather than individually [136]. In the classification problem, logistic regression measures the relationship

between the outcome (patients who will not achieve remission after first AED coded as '1') and the SNPs, by estimating probabilities using its underlying logistic function. These probabilities (between 0 and 1) are then transformed into either 0 or 1 according to the probability threshold specified, 0.5. In this chapter, the mean and standard deviation of six performance measures over 100 splits are presented. The performance measures are the Area Under the receiver operating characteristic (ROC) Curve (AUC), the probability of correct classification (PCC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

External validation

As shown in Figure 5.2, other than the internal cross-validation (i.e. splitting into training and test sets from the same dataset), external validation approach is applied to further validate the results gathered from the analysis using the development set. Here, the dataset from the SANAD cohort is used as the validation set. The univariable ranking procedure is also done on the validation set to investigate any overlapping informative SNPs that appear in both datasets. The SNPs ranking is then compared with the ranking gathered using the development set. Further, the model (involving a subset of SNPs) that is fitted by both Strategy 1 and 2 using the development set is then validated on the validation set accordingly.

5.3 Results

In this section, the results using the proposed methods (from Section 5.2) are discussed. The first part of the analysis includes the univariable tSNR ranking of the development and validation sets. Subsequently, a subset of SNPs from the development set undergoes further reduction before model fitting with smaller number of SNPs. The classification performance using the selected model is then evaluated on the test set from the development set (internal cross-validation) as well as the validation set (external validation).

5.3.1 Univariable SNPs ranking using tSNR

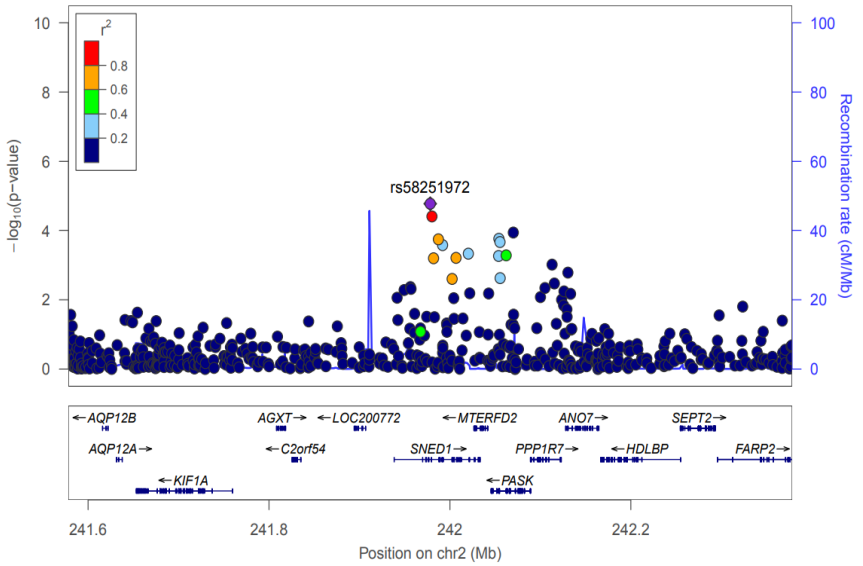
The tSNR is calculated for each SNP in both datasets (1,084,548 SNPs in each dataset). The SNPs are ranked from the highest to the lowest tSNR values. For example, Table 5.2 shows the top 20 SNPs for development and validation sets.

Table 5.2: The top 20 SNPs based on univariable tSNR ranking for development set and validation set.

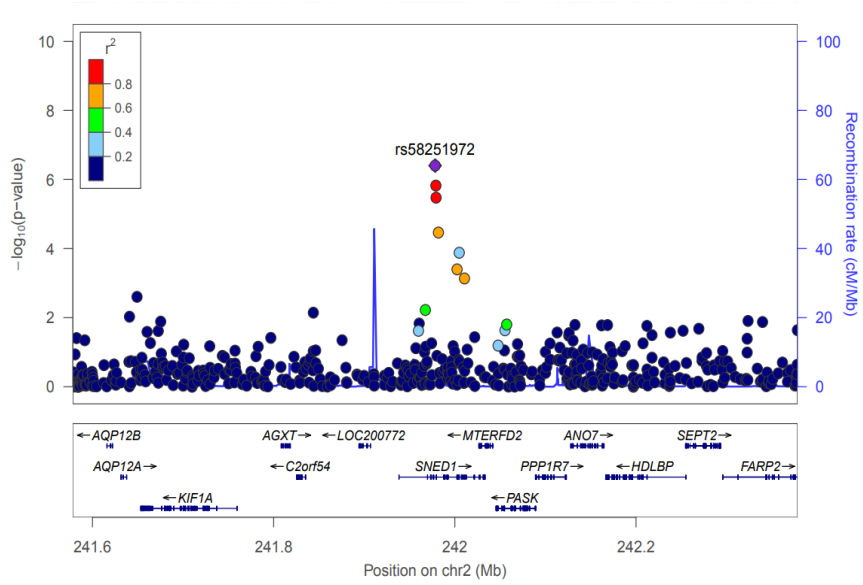
Rank	Development set			Validation set		
	Chr	**SNP**	**tSNR**	**Chr**	**SNP**	**tSNR**
1	19	rs148854414	0.01446	2	rs113868955	0.03052
2	2	**rs58251972**	0.01262	1	rs72708783	0.02971
3	9	rs117807448	0.01104	19	chr19.38803597.D	0.02965
4	9	rs149080038	0.01090	7	rs111299697	0.02680
5	2	rs62191927	0.01080	7	chr7.12536115.I	0.02619
6	7	rs35288464	0.01066	3	rs73024364	0.02586
7	5	rs115058790	0.01059	1	rs79346791	0.02582
8	12	chr12.32523962.I	0.01043	11	rs1774520	0.02484
9	9	rs872374	0.01025	9	rs10405527	0.02451
10	13	rs112771492	0.01023	3	rs73119100	0.02392
11	14	rs12887826	0.01023	16	rs9934340	0.02321
12	20	rs2025046	0.01013	9	chr9.10535674.D	0.02317
13	9	rs869382	0.01009	8	chr8.42378487.D	0.02301
14	2	rs16851377	0.01005	9	rs10738164	0.02270
15	21	rs77386304	0.00990	12	rs117383211	0.02245

Rank	Development set			Validation set		
	Chr	**SNP**	**tSNR**	**Chr**	**SNP**	**tSNR**
16	11	rs770584	0.00980	2	**rs58251972**	0.02241
17	18	rs17515325	0.00978	3	rs76979971	0.02223
18	13	rs140166961	0.00977	7	rs56322358	0.02152
19	18	rs59784013	0.00977	18	rs117380254	0.02142
20	17	rs6502867	0.00976	14	rs74621727	0.02142

Within the top 20 SNPs, only one SNP rs58251972 on chromosome 2 appears in both datasets. Figure 5.3 (a) and (b) show the zoomed in plots of this SNP (purple diamond) and those surrounding it in the development and validation sets accordingly. The y-axis represents the $-\log_{10}$ transformed p-value of each SNP and the x-axis represents its genomic position. As proven in Chapter 3 (Methods), the tSNR and p-value show identical ranking hence the usage of p-values in both plots. SNP rs58251972 resides in gene SNED1. From the database of Single Nucleotide Polymorphisms (dbSNP) [137], SNED1 (Sushi, Nidogen And EGF Like Domains 1) is a Protein Coding gene and the diseases associated with SNED1 include Cortical Thymoma.



(a)



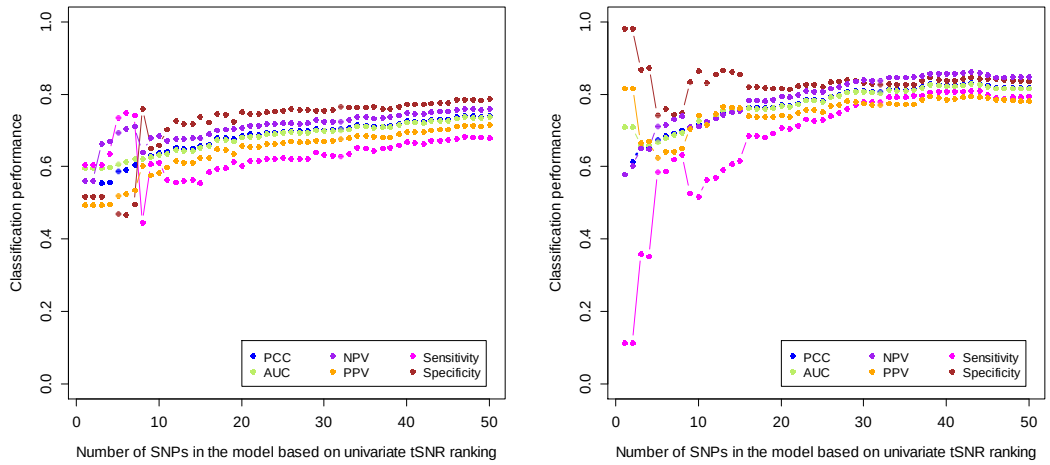
(b)

Figure 5.3: Regional plots for SNP rs58251972 that appears in (a) Development set and (b) Validation set based on the univariable tSNR ranking.

Note: The plots are created with LocusZoom (version 1.1) with linkage disequilibrium (LD) data taken from the 1000 Genomes Project, HG19, March, 2012.

At this point, it will be useful to see the classification performance in both datasets using the univariable tSNR ranking. Figure 5.4 (a) and (b) show the classification performance using logistic regression for development and

validation sets accordingly. The plots are drawn by adding one SNP at a time based on their univariable ranking (top 50 SNPs). The classification performance is comparatively good in both datasets which at some points could reach more than 80% accuracy. However, it can be observed that the values will keep increasing every time a new SNP is introduced in the model. The situation can be due to the overfitting issue.



(a) (b)

Figure 5.4: Classification performance (mean) of the top SNPs selected by univariable tSNR ranking for (a) Development set and (b) Validation set.

Hence, the classification procedure will be impractical due to large number of SNPs. Further reduction of the SNPs is required to resolve the high-dimensionality and overfitting issues. Though, the univariable ranking is still useful to lessen the computational complexity by selecting a subset of top-ranked SNPs to be utilised in the multivariable analysis.

5.3.2 Multivariable approach of SNPs ranking and model selection for classification

The analysis starts with a subset of top 5,000 SNPs from the development set which are ranked in section 5.3.1. The subset of SNPs is then modelled using PLR and cross-validation based on 100 splits (training sets). There are two important settings that need to be determined when using R CRAN packages '`glmnet`'; i) the number of maximum variables ever to be nonzero in the model, *dfmax*, and ii) the complexity parameter, λ. To minimise the number of variables, *dfmax* is set to 200 and for each split, the minimum λ is chosen through an internal cross-validation (default 10-folds).

The 100 splits (which generate 100 training and test sets) produces 100 different models of which the tSNR value is calculated for each model. Here, Strategy 1 and Strategy 2 is applied to select the final subset of SNPs for classification.

Strategy 1: Multivariable ranking using tSNR (cumulative tSNR ranking)

Following Strategy 1, each of the SNPs will be assigned a weighting based on the tSNR value calculated for each model it belongs to. The accumulated tSNR of each SNP will then determine the tSNR weight it carries. The SNPs are then ranked based on their cumulative tSNR. For illustration, Table 5.3 shows the top 20 SNPs based on the cumulative tSNR ranking.

Table 5.3: The top 20 SNPs based on cumulative tSNR ranking from the development set.

Rank	Chr	Base-pair position	SNP	Cumulative tSNR
1	19	10527666	rs148854414	24.1875
2	2	241978231	rs58251972	23.9526
3	9	134642577	rs869382	21.9703
4	21	46487635	rs77386304	20.8868
5	14	101117587	rs12887826	20.3452
6	7	11477000	rs35288464	19.3331
7	12	32523962	chr12.32523962.I	19.2856
8	20	19856187	rs2025046	19.0655
9	18	11695551	rs17515325	18.7487
10	18	46265448	rs59784013	18.7470
11	13	48778115	rs140166961	18.3232
12	17	5420328	rs6502867	17.0284
13	12	116773835	rs12423990	16.8446
14	16	64616524	rs112557806	16.7252
15	2	214889035	rs62191927	16.6525
16	2	48454553	rs80160664	16.4822
17	9	90136937	rs872374	16.3989
18	18	3725553	rs7235163	15.9781
19	19	58830709	rs56725489	15.7411
20	9	135672707	rs117807448	15.4306

From the ranking, the top two SNPs rs148854414 and rs58251972 are similar with the SNPs recorded by the univariable tSNR ranking for the development set. In particular, rs148854414 resides in gene PDE4A. Though there is no specific evidence from past research that relates the SNP to epilepsy, the gene is believed to be associated with the neurological disease, Parkinson [138]. The study involved 12 patients with Parkinson disease and 12 healthy individuals with no history of neurological or psychiatric disorders. The findings suggested that the loss of PDE4 expression in the stratio-thalamo-cortical circuit, associated with deficits of spatial working memory in patients with Parkinson disease.

Another SNP to highlight within the top 20 SNPs is rs17515325 which resides in gene GNAL. The protein encoded within the gene is widely expressed in the central nervous system [139]. Mutations in this gene have been associated with dystonia 25 and this gene is located in a susceptibility region for bipolar disorder and schizophrenia. Dystonia is defined as hyperkinetic movement disorders, characterised by involuntary sustained muscle contractions affecting one or more sites of the body, which lead to twisting and repetitive movements or abnormal postures of the affected body part [140]. There are past research that reported the association between dystonia and seizure or epilepsy [140, 141].

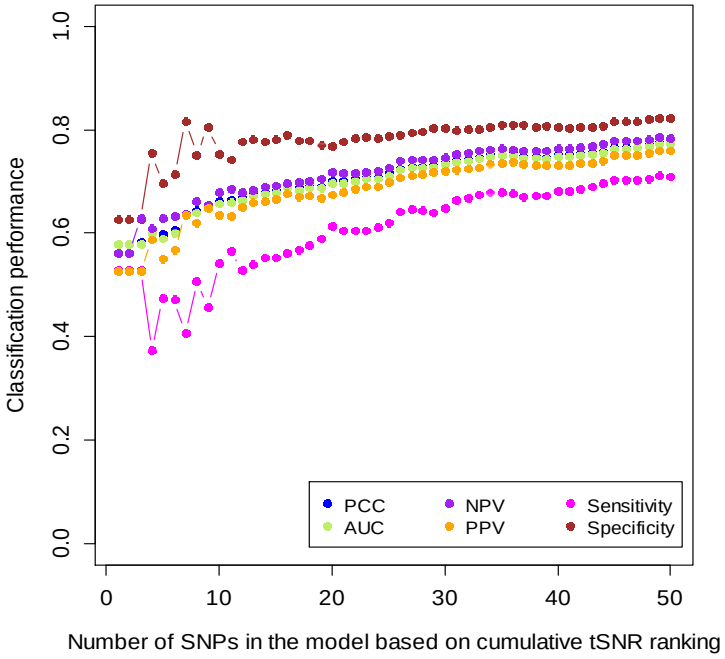
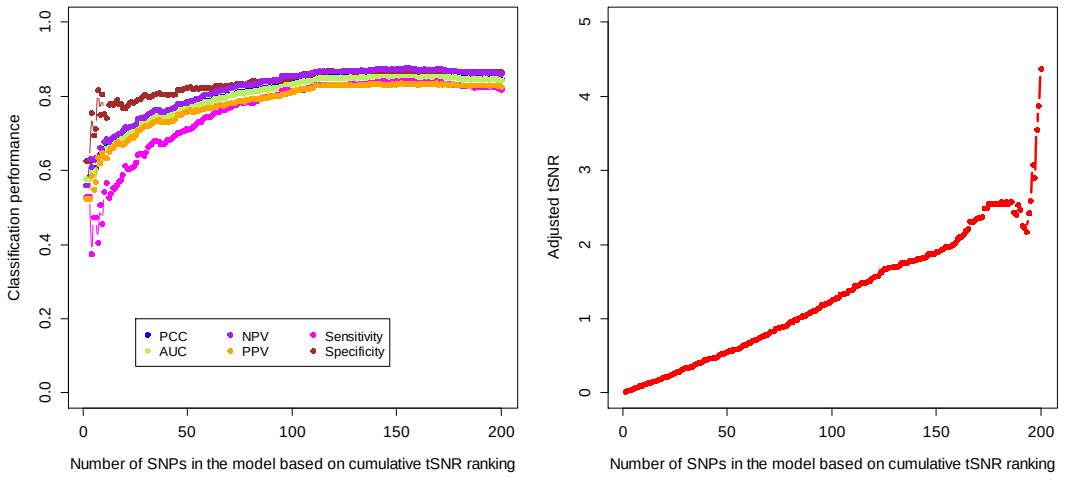


Figure 5.5: Classification performance (mean) of the top SNPs selected by cumulative tSNR ranking from the development set.

As a comparison for classification performance based on univariable ranking (Figure 5.4 (a)), Figure 5.5 shows the classification performance of the top-ranked SNPs that were selected based on the cumulative tSNR ranking using the development set. The classification performance is slightly better than the top-ranked SNPs that were selected based on univariable tSNR ranking. By

using the cumulative tSNR ranking, the PCC and AUC achieved 70% at 22 SNPs. Meanwhile, the PCC and AUC achieved 70% at 29 SNPs when using univariable tSNR ranking (Figure 5.4).

From the cumulative tSNR ranking, it is important to determine the number of SNPs to be selected for classification. The proposal is to evaluate the number of top SNPs selected by adjusted tSNR (from equation (5.3)). Similar to other model selection criteria (e.g. AIC, BIC) the value of adjusted tSNR can determine when the inclusion of variable into the model can be stopped. In this case, the decreasing value of adjusted tSNR tells that the extra variable to be included in the model has less signal as compared to noise. Figure 5.6 (a) shows the classification performance against the number of top SNPs based on the cumulative tSNR ranking. Separately, Figure 5.6 (b) illustrates the pattern of adjusted tSNR following the inclusion of each top SNPs.



(a) (b)

Figure 5.6: Top 200 SNPs based on cumulative tSNR ranking from the development set against (a) Classification performance (mean) and (b) Adjusted tSNR (mean).

As shown in Figure 5.6 (b) there is a slight drop of adjusted tSNR after 187 SNPs. While there is an increment in adjusted tSNR after the descent, the classification accuracies remain constant afterwards. Hence, from the pattern shown in Figure 5.6, the final model containing 187 top SNPs based on the cumulative tSNR ranking is selected. The summary of classification performance of the top SNPs is shown in Table 5.4. The classification performance is satisfactory with average more than 80% in each measure.

Table 5.4: Summary of the classification performance (mean and standard deviation) of top 187 SNPs using the development set.

Classification performance	**Mean**	**Standard deviation**
PCC	84.86	0.02
AUC	84.71	0.02
Sensitivity	82.46	0.03
Specificity	86.75	0.03
PPV	83.04	0.03
NPV	86.39	0.02

Strategy 2: Model selection using tSNR

Strategy 2 applies tSNR to select the best model among the 100 models that are produced based on the 100 splits. The tSNR is calculated for each model and from there each model is ranked based on the highest to the lowest tSNR values. Figure 5.7 illustrates the distribution of tSNR across the 100 models.

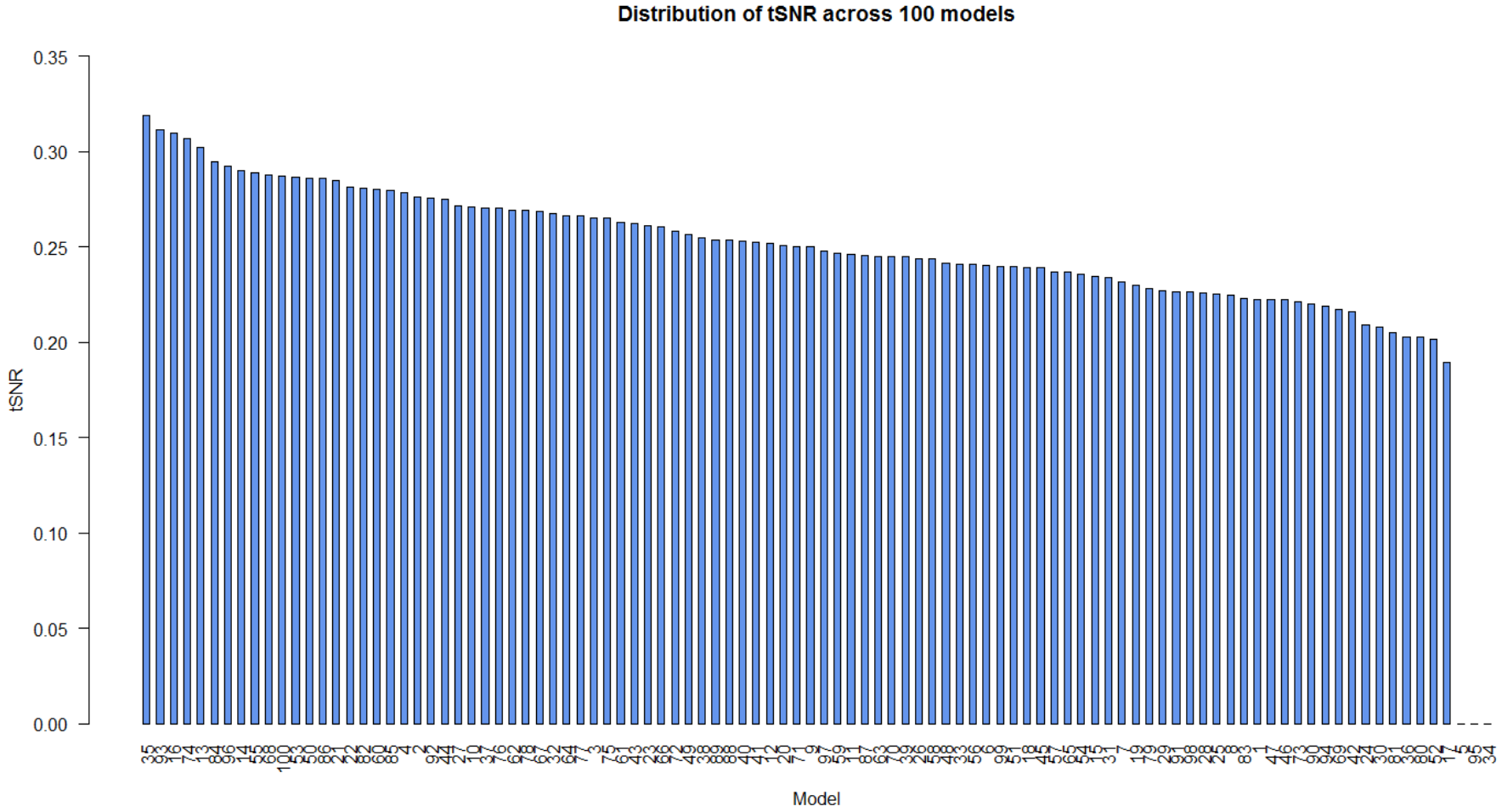


Figure 5.7: Distribution of tSNR across 100 models in descending order.

To highlight a few, Table 5.5 shows the information gathered for each model. Based on the ranking, the model with the highest tSNR consists of 189 SNPs. The lowest ranked model is among the 12 models that consist less than two SNPs in each model, hence the low tSNR.

Table 5.5: The summary of models (100 models based on the 100 splits) ranked from the highest to lowest tSNR.

Rank	Model no.	SNPs	No of SNPs	tSNR
1	35	rs72640613, rs11120890, rs35622037, chr1.17702979.D, rs144278461, … , rs76081526	189	0.3190
2	93	rs72640613, chr1.17702979.D, rs35730900, rs11263857, rs4246520, … , rs78851025	189	0.3112
3	16	rs79346791, rs2236055, rs399628, rs11211072, rs35619352, … , rs2284093	203	0.3097
:	:	:	:	:
100	34	rs4773561	1	0.0000

By using the selected model (ranked first) from the PLR, the classification performance is calculated on the test sets. The classification performance is presented as mean and standard deviation of PCC, AUC, sensitivity, specificity, NPV and PPV based on the 100 splits.

Table 5.6: Summary of the classification performance using PLR on the development set (the first ranked model).

Classification performance	Mean	Standard deviation
PCC	81.29	0.03
AUC	83.31	0.03
Sensitivity	65.48	0.05
Specificity	93.66	0.02
PPV	88.94	0.04
NPV	77.67	0.03

From Table 5.6 it can be observed that classification accuracies are adequate with average of 80% accuracies. However, it is noted that the sensitivity value of 65.48%. is quite low. This may be due to the imbalance sample size between the two phenotype groups.

External validation

The two strategies applied earlier provide the alternative on how to select the best subset of SNPs for classification. The models are validated using an internal cross-validation approach by splitting the dataset into training and test sets. However, to avoid optimistic assessments of the models, further validation approach (i.e. external validation) is needed [142]. To assess whether the approach for classification that is presented in Figure 5.2 is acceptable, the validation set (SANAD cohort) is used to validate the models selected by both Strategy 1 and 2.

The final model used by Strategy 1 and 2 are fitted on the validation set. The classification performance is measured by PCC, AUC, sensitivity, specificity, NPV and PPV. The classification performance using the validation set is summarised in Table 5.7. The classification performance increased significantly using the model selected using Strategy 1 (multiple logistic regression). Meanwhile, the classification performance was slightly increased when using the model selected using Strategy 2 (PLR). It is important to note that the sample size in the validation set is much lower $(n = 639)$ than the development set $(n = 1{,}515)$. Hence, with 187 and 189 SNPs (SNPs selected by Strategy 1 and Strategy 2 accordingly) used in the model, the smaller number of samples might influence the results presented below.

Table 5.7: Summary of the classification performance on the validation set (external validation) using models selected by Strategy 1 (187 SNPs) and Strategy 2 (189 SNPs).

Classification performance	**Strategy 1**	**Strategy 2**
PCC	94.59	82.78
AUC	94.52	84.14
Sensitivity	93.12	68.40
Specificity	95.67	93.24
PPV	94.00	88.04
NPV	95.04	80.23

5.4 Concluding remarks

As shown in this chapter, the tSNR variable selection can be advantageous to reduce the dimensionality of the data and to help reduce the computational complexity. It also leads to improving the classification accuracy rate of a classifier. Particularly in this study, the variable selection method tSNR is applied on the EpiPGX and SANAD datasets. The roles of tSNR in this study are twofold; i) as a filter metric to select the most informative SNPs by using univariable ranking, and ii) as a model selection criterion for non-nested models.

The filter approach is known due to its ability to select the variables which are most important for classification and so reduce the number of dimensions necessary for classification [7]. The desired effect of this is to speed up algorithms and to make the subsequent analysis more effective by only focusing on the most relevant variables in the dataset. Hence, by using the filter metric tSNR as the variable selection method, the classification performance shown in this chapter was adequate (more than 70%) when either internally or externally validated.

In the multivariable setting, two strategies were proposed to select the best model. Strategy 1 applies the similar strategy as the univariable ranking by using the cumulative tSNR as the ranking measure. The cumulative tSNR ranking is superior as compared to the univariable ranking as the SNPs are considered together during the modelling process by PLR. This method has shown better classification performance in which the PCC achieved 70% at 22 SNPs as compared to 29 SNPs in the univariable setting.

Meanwhile, Strategy 2 was applied to compare the non-nested models that were produced by the cross-validation approach. The models were ranked from the highest to the lowest tSNR. The higher the value of tSNR indicates the model shows higher signal than the noise. From the result, the classification performance shown was adequate with average of 80% accuracies. This analysis confirms the advantage of PLR since it not only able to reduce the number of variables but also give good classification accuracy.

Although the accuracy of prediction can be improved substantially using the multivariable approach, the number of SNPs selected is still large and may be difficult to interpret. However, from the pharmacogenomic point of view, dealing with hundreds of SNPs compared to the initial more than a million SNPs allows researchers to focus on fewer SNPs that can be informative for personalising treatment. Also it is still important to include a large number of SNPs in a predictive model to capture the most genetic variance [143].

An important strength of this study is the fact that it is validated in a different cohort of patients (i.e. external validation). The method proposed is internally and externally validated which helps in confirming the results. However, the

sample size used for external validation is half of the development set. This scenario may lead to differences in model performance. Therefore, a further research with larger samples in validation dataset is needed. In addition, it will be useful to see how our methods work in other disease scenarios.

Chapter 6

Clinical Applications: Combining Longitudinal Clinical and SNP Data for Classification

6.1 Introduction

Genetic studies have shown that Single Nucleotide Polymorphisms (SNPs) have become essential variables to predict individual's belonging to a particular class of complex diseases and different reactions to medications and treatments [4, 20]. The outcome usually is not only influenced by the genetics information, but may be from the interaction with clinical and environmental variables. Statnikov et al. (2007) [115] show the classification performance of a model with only genetic data, a model which consists of only clinical data and a model with the combination of genetic and clinical data. The data consists of 50 esophageal squamous cell carcinoma patients and 50 controls all genotyped at 11,542 SNPs. In addition, five clinical variables are recorded. In their study, the area under

the ROC curve (AUC) obtained when using only genetic data in the model is 0.51 and 0.60 for only clinical data. The classification performance improves slightly to 0.62 using the model with the combination of both types of data.

Predicting the risk of individuals to develop a disease or have a particular response to treatment given their genetic sequence (e.g. SNPs) is a desirable goal, yet the current ability to make such predictions is relatively poor on its own [23]. One explanation is in the complexity of the data structure; spatial, high-dimensional and categorical. Secondly, the challenge lies on how to best combine the clinical and SNP data.

In certain disease studies (e.g. epilepsy, asthma, type-2 diabetes), the discriminatory (predictive or diagnostic) strength of clinical variables such as sex, age, time to treatment are often extensively investigated as potential predictors and well-validated [21]. However, the question on added predictive value of genetic data given the availability of classical clinical variables, has long been under-considered in the bioinformatics literature [21].

Hence, in this chapter, the analysis revolves around combining longitudinal clinical and SNP data for classification. The objective of this chapter is twofold:

(i) To select the most informative SNPs using tSNR.

(ii) To jointly model the longitudinal clinical and SNP data for the purpose of classification. The data will be jointly modelled via a longitudinal discriminant analysis approach with multivariate generalised linear mixed models and the classification performance will be evaluated.

6.1.1 Challenges of SANAD dataset and motivation

SANAD is a dataset generated from an unblinded randomised controlled trial in hospital-based outpatient clinics in the UK [129, 130]. The study aimed to assess the efficacy of antiepileptic drugs on patients with different types of seizures with regards to longer-term outcome, quality of life, and health economic outcomes. In the study, two treatment arms were used (Table 6.1).

Table 6.1: Summary of SANAD study in arm A and arm B.

Arm	**Antiepileptic Drugs (AED)**	**No of Patients**	**Outcomes**
Arm A	Carbamezapine, Gabapentine, Lamotrigine, Oxcarbazepine, Topiramate	1,721 (standard treatment: Carbamezapine)	Time to treatment failure and time to 1-year remission
Arm B	Valproate, Lamotrigine, Topiramate	716 (standard treatment: Valproate)	Time to treatment failure and time to 1-year remission

Epilepsy is a neurological disorder which causes repeated seizures. Seizures are physical findings or changes in behaviour that occur after an episode of abnormal electrical activity in the brain. In the UK, epilepsy is estimated to affect more than 500,000 people [144]. The causes underlying the development of epilepsy are not well understood. In some cases, genetic factors are clearly evident as the cause of epilepsy [145]. However, in other cases the cause is not easily identified.

The efficacy of the treatment in SANAD has been studied via survival models, with most previous work concentrating on clinical variables as the predictive variables [129, 130, 146, 147]. Using the SANAD dataset, Speed et al. [148] reported the first GWAS which represents a comprehensive analysis of genetic effects on the prognosis of newly treated epilepsy. Meanwhile, Hughes et al. [25, 116] developed a discriminant analysis approach to predict whether a new

patient would either achieve remission or would not achieve remission within five years of commencing treatment. A patient is defined as being in remission if they had continuous 12-month period without any seizures within five years from diagnosis.

In this chapter, the group of patients not achieving remission is referred as the refractory group. The prediction is made based on their baseline characteristics as well as longitudinally gathered biomarkers information (e.g. whether a patient had seizures or not since their last visit, total number of seizures since their last visit and the number of adverse events experienced since the last visit). To the best of our knowledge, the existing approaches use either SNP data or longitudinal clinical data, but not both. Therefore, we propose a multivariate approach that incorporate both type of data as an extension of the work by Hughes et al. [25, 116].

This chapter is organised as follows. The multivariate approach is discussed in Section 6.2. Then, in Section 6.3 detailed results and discussions from the analysis are presented. In this chapter, the classification performance is measured by calculating the probability of correct classification (PCC), area under the ROC curve (AUC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Also, the prediction time at which patients are correctly classified as refractory is also presented. Then, the chapter is concluded in Section 6.4.

6.2 Methods

In this section, a detailed description of the dataset will be given and summary of the methods that will be used to analyse the dataset will be discussed.

6.2.1 The SANAD dataset

Particularly in this study, the SANAD dataset can be divided into two categories of data. The first dataset consists of longitudinal clinical data of 1,752 patients. The dataset was used in previous studies by Hughes et al. [25, 116] which undertook the selection procedure as shown in Figure 6.1.

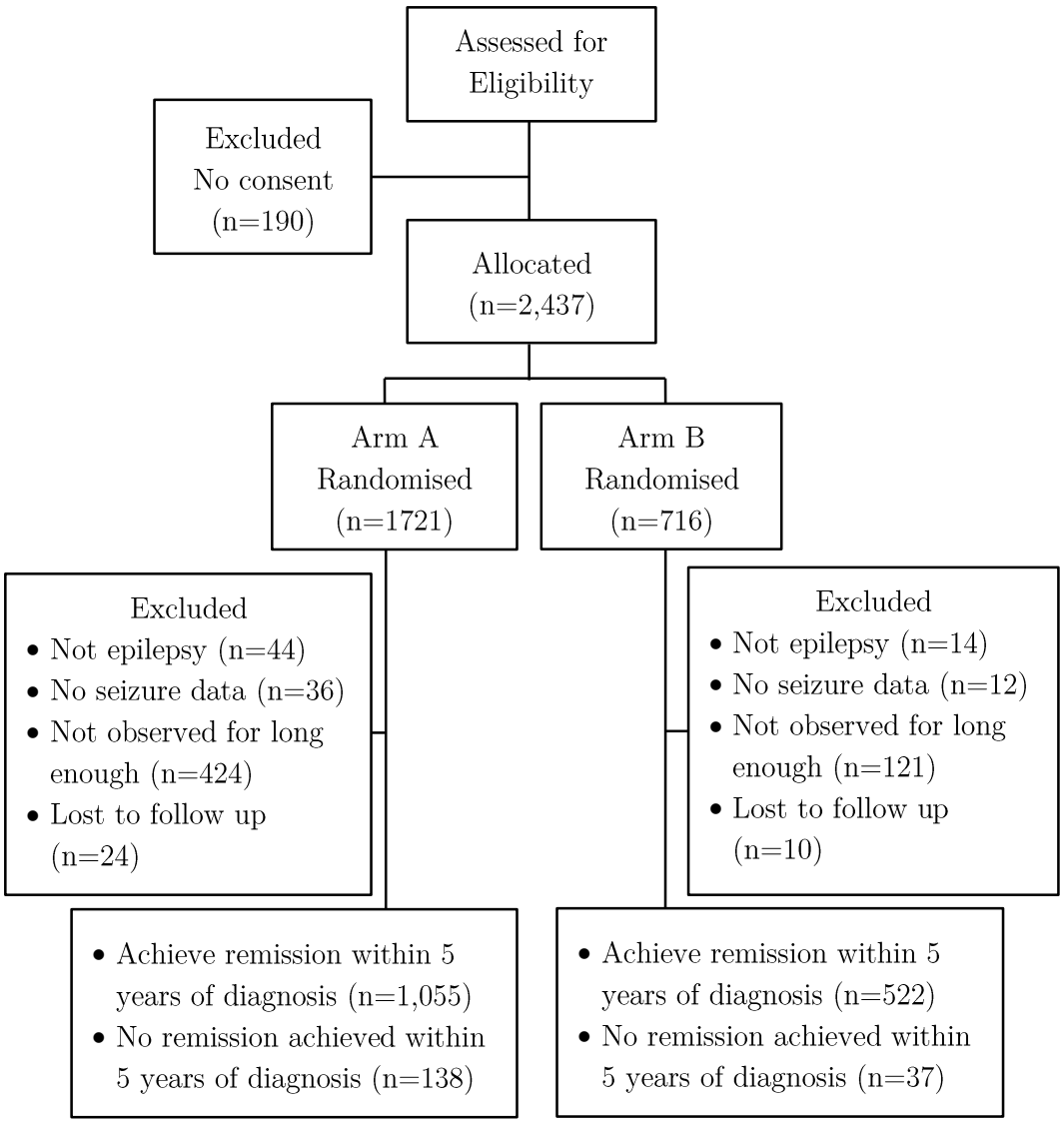


Figure 6.1: Patients selection criteria for SANAD dataset with longitudinal information [149].

Meanwhile, the second dataset consists of genotype data (SNPs) of 818 patients within the SANAD dataset. For this study, only patients with complete information of longitudinal clinical information, SNP data and phenotype status are considered. In order to achieve that, both datasets are merged which resulted in 573 patients are chosen in the final dataset. Figure 6.2 shows the summary of the merged dataset that contains longitudinal clinical information, SNP data and phenotype status.

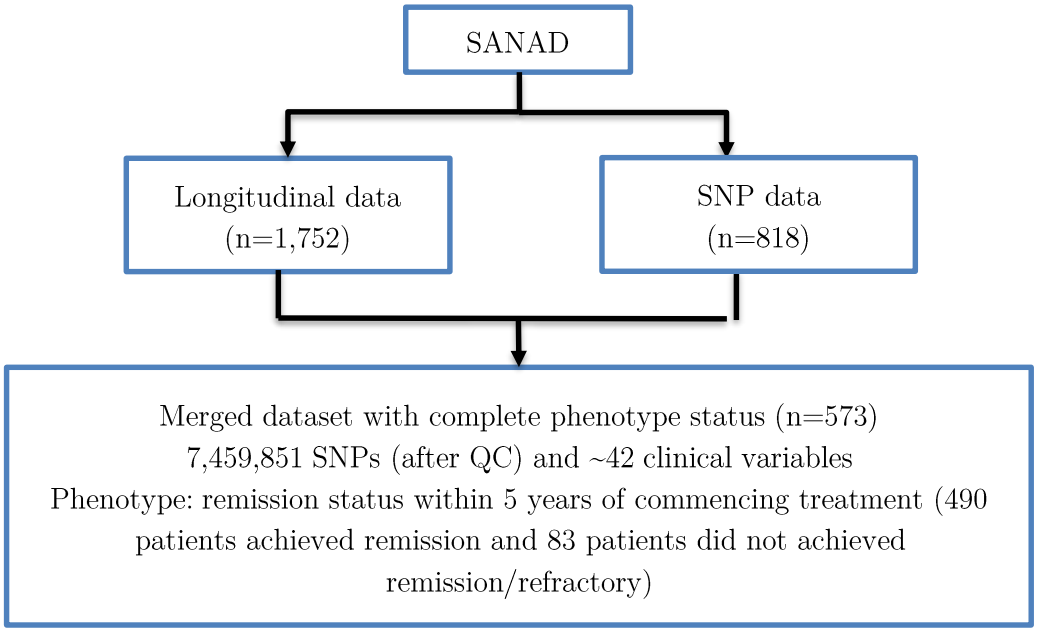


Figure 6.2: Merged dataset that contains 573 patients with longitudinal clinical information, SNP data and phenotype status.

6.2.2 Phenotype definition

In this study, the interest is to identify patient who will not achieve remission within five years of commencing treatment diagnosis. The phenotype for each patient is coded as '1' if they did not achieve remission (refractory/disease group), whilst the phenotype is coded as '0' if they were observed to achieve remission within five years of diagnosis (healthy group).

6.2.3 Overview of the SNPs selection process and classification

The analysis of the SANAD dataset will follow the workflow presented in Figure 6.3 below which has been discussed in Section 3.3. Frequent clinical interest is in being able to classify patients into various groups defined based on their future clinical status, based on the evolution of variables observed over time. This chapter is focusing on the approach to jointly model the longitudinal clinical and SNP data from the SANAD dataset in order to identify patients who will achieve a 12-months remission within five years of diagnosis. In general, Boulesteix and Sauerbrei (2011) [21] and De Bin et al. (2014)[114] proposed the strategy to combine the clinical and omics data.

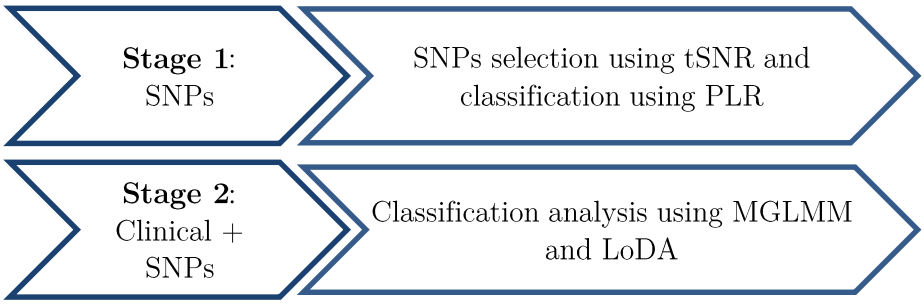


Figure 6.3: Stages proposed involved in combining the longitudinal clinical and SNP data for classification.

Here, the strategy is modified to fit the longitudinal clinical and SNP data accordingly. Figure 6.3 shows the proposed strategy to combine the longitudinal clinical and SNP data. The strategy includes two stages in which the classification performance is evaluated in each stage.

In Stage 1, the variable selection process will go through the proposed approach as shown in Figure 6.4. The approach follows the workflow presented in Section

3.2.2. However, for the strategies proposed in multivariable analysis, only Strategy 1 (cumulative tSNR ranking) is included. This is because, the main concern in this chapter is to select only the top ranked SNPs which will be utilised in Stage 2. Firstly, the univariable filter metric tSNR is applied to rank the SNPs and a subset of 5,000 SNPs is selected.

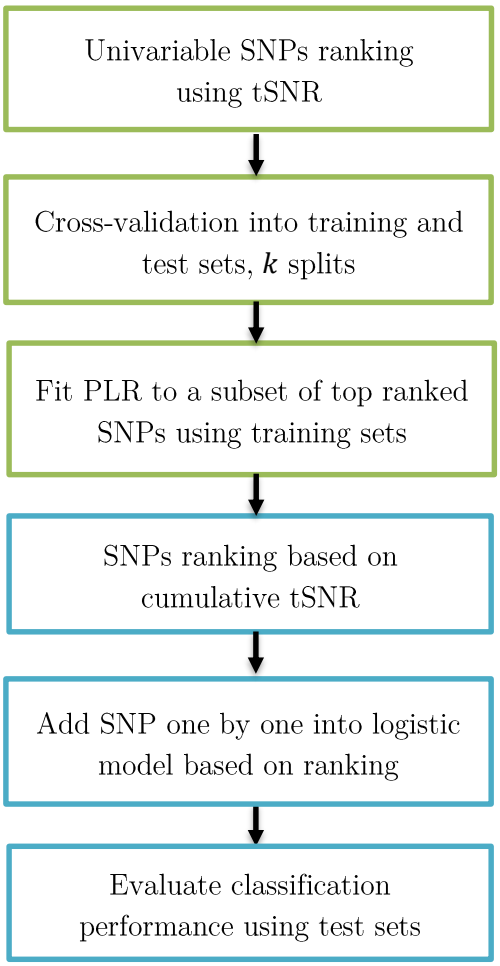


Figure 6.4: Diagram of model building pipeline including (i) univariable tSNR as pre-selection process; (ii) splits of sample into training and test sets; (iii) model building using PLR; (iv) strategy to select a small subset of SNPs; and (v) model evaluation using test sets.

From here, the process is contained within cross-validation structure of which 70% of the data is randomly chosen into the training set and the remaining 30% as test set. The split is repeated 100 times which will produce 100 training and

test sets accordingly. The model building steps are performed in each training set and the classification performance is performed using the test sets. The penalised logistic regression (PLR) is implemented using the '`glmnet`' function in R. However, at this point the number of SNPs in each model is still large. Hence, the cumulative tSNR is proposed to further rank and select a smaller subset of SNPs. Here, the adjusted tSNR is used as a stopping criterion on how many number of SNPs to be included in the final model. The classification performance is measured using PCC, AUC, sensitivity, specificity, PPV and NPV.

In Stage 2, the longitudinal clinical data and SNP data are jointly modelled. At this stage, the process will follow the work conducted by Hughes et al. [25, 116]. The work considered the complexities of the data which include; i) their multivariate and longitudinal nature, ii) their complex correlation structure, iii) the varying types of data (e.g. continuous, counts, binary, categorical) and, iv) different clinical variables are measured at different time points for the same patient (and also across patient). By taking these complexities into account, the longitudinal clinical and SNP data are jointly modelled for classification purposes by the multivariate generalised linear mixed model (MGLMM). These longitudinal models are subsequently used in the dynamic longitudinal discriminant analysis (LoDA) to predict the probability of belonging to a specific group.

From now on, the model with only longitudinal clinical data is referred as the *reference* model. Few additional models (combination of clinical and SNPs) will be evaluated and compared with the reference model.

6.2.4 Classification with LoDA

This section will be discussing the method covered in Section 3.3 when applied to SANAD data. The analysis starts with a *reference* model which consists of multiple outcomes (longitudinal biomarkers) for the SANAD study [116]. These correlated outcomes were longitudinally measured comprised of whether or not the patient experienced seizures since the last clinic visit (logistic model), total numbers of seizures since the last clinic visit under the transformation log(1 + total seizures) (Gaussian model) and the numbers of adverse events experienced since the previous clinic visit (log-Poisson model). Figure 6.5 shows the change over time (days) for a sample of 20 patients for each longitudinal biomarker.

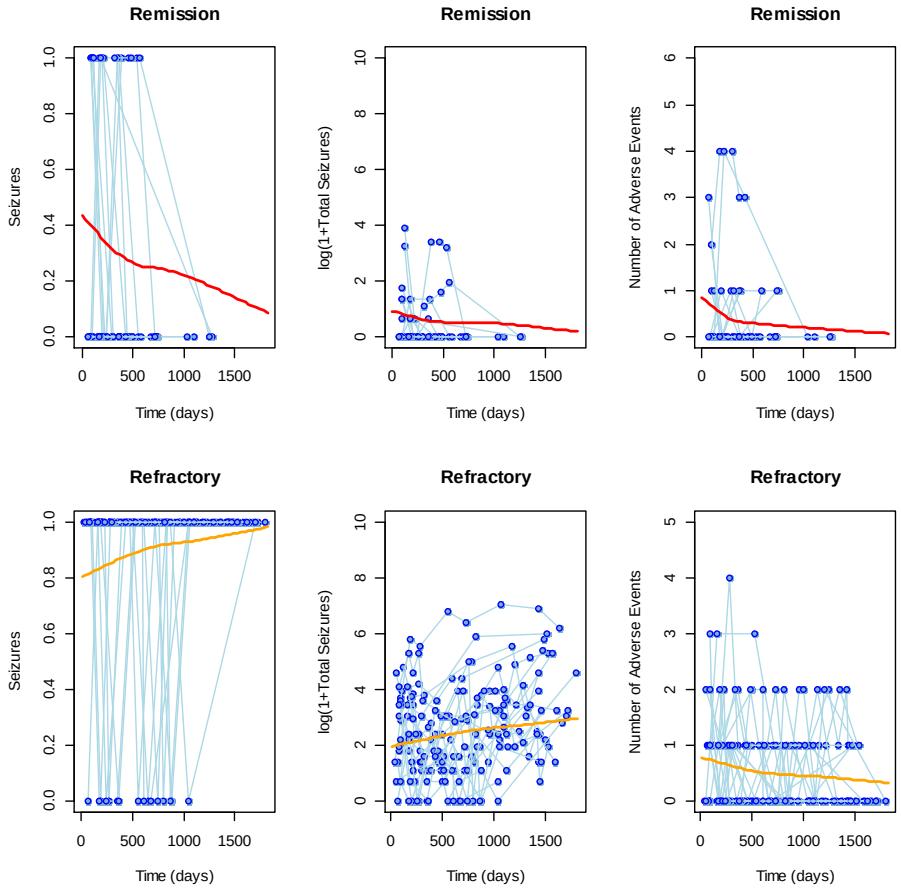


Figure 6.5: Longitudinal profiles whether patient had seizures, log(1+total seizures) and the number of adverse events for patients from Remission group (first row) and the Refractory group (second row). Solid bold lines show LOESS smoothed profiles calculated using data from all patients.

The MGLMMs are fitted to the longitudinal data separately according to the phenotype group. The fitting of MGLMMs is done using the '`mixAK`' package in R [157]. The expected value (transformed by an appropriate link function) for the j-th longitudinal observation $(j = 1, \dots, n_r)$ for the r-th marker $(r = 1, \dots, R)$ (denoted $Y_{r,j}$) is assumed to follow, given g (0 or 1 according to the phenotype group) and given b, a distribution from an exponential family (e.g. normal, Poisson, Bernoulli) with a dispersion parameter $\emptyset_r^g$ is given by,

$$h_r^{-1}\{E(Y_{r,j}|b, U = g)\} = x_{r,j}^{gT}\alpha_r^g + z_{r,j}^{gT}b_r \quad r = 1, \dots, R, j = 1, \dots, n_r \qquad (6.1)$$

Two models with the structure described by equation (6.1) are modelled, one for each prognostic group (remission and refractory). For each of these models, the three longitudinal biomarkers $R = 3$ are considered to predict whether a patient achieved remission after five years of commencing treatment. The following illustrates the expansion of equation 6.1 according to the prognostic group.

Refractory

$$E(Y_{1,j}|b, U = 1) = h_1^{-1}(x_{1,j}^{gT}\alpha_1^g + z_{1,j}^{gT}b_1)$$
$$E(Y_{2,j}|b, U = 1) = h_2^{-1}(x_{2,j}^{gT}\alpha_2^g + z_{2,j}^{gT}b_2)$$
$$E(Y_{3,j}|b, U = 1) = h_3^{-1}(x_{3,j}^{gT}\alpha_3^g + z_{3,j}^{gT}b_3)$$

Remission

$$E(Y_{1,j}|b, U = 0) = h_1^{-1}(x_{1,j}^{gT}\alpha_1^g + z_{1,j}^{gT}b_1)$$
$$E(Y_{2,j}|b, U = 0) = h_2^{-1}(x_{2,j}^{gT}\alpha_2^g + z_{2,j}^{gT}b_2)$$
$$E(Y_{3,j}|b, U = 0) = h_3^{-1}(x_{3,j}^{gT}\alpha_3^g + z_{3,j}^{gT}b_3)$$

Each of these three longitudinal biomarkers is modelled for each prognostic group involving a set of fixed effects, $x_{r,j}^g$; 1) time since last visit, 2) time since

diagnosis, 3) age at diagnosis, 4) epilepsy type, 5) sex, and 6) randomisation period. With respect to the random effects structure, the multivariate model for each prognostic group contains random intercepts. This means $z_{r,j}^{g} = 1$ and a three-dimensional random effects vector $b = (b_1, b_2, b_3)^T$(random intercepts for the three markers) is involved.

Once the two multivariate models are derived, they can be used to allocate a new patient into group its belongs to. Bayes theorem is applied to calculate the probability of a patient belonging to group g given their longitudinal and covariate data and the model parameters from the MGLMMs fit to patients of known status.

$$Pr_{g,new} = \frac{\pi_g \hat{f}_{g,new}}{\sum_{\check{g}=0}^{G-1} \pi_{\check{g}} \hat{f}_{\check{g},new}} \quad g = 0, \dots, G-1, \tag{6.2}$$

where $\hat{f}$ denotes the predictive density of the observed markers given the group and model parameters. The prior probabilities of belonging to each group are denoted by $\pi_g = Pr(g), g = 0, \dots, G-1$. In a Bayesian setting, $f_{g,new}$ is estimated as the mean of the posterior predictive density estimated from ℵ samples from a Markov Chain Monte Carlo (MCMC) scheme [118]. Here, an R package '`coda`' [158] is used to evaluate the convergence of MCMC simulation. For illustration of implemented convergence check, the estimated autocorrelations and traceplots for deviance, fixed effects vector α_r^g and dispersion parameter $\emptyset_r^g$ are appended in section Appendices. The convergence check is done using a subset of samples for each prognostic groups.

As discussed in Section 3.3, the predictive density $f_{g,new}$ can be specified by either marginal, conditional or random effects approach. By using the *marginal*

approach, the new patient is assigned to their specific group to which their longitudinal profiles $Y_{new} = (y_{new,1}, \dots, y_{new,R})$ lie closest [116, 118]. Here, $f_{g,new}$ is taken as the marginal density of Y_{new}. Meanwhile, for *conditional* and *random effects* prediction, it is necessary to represent the new patient also by the values (unobservable) of the random effect vector b_{new} for which assumed joint distribution, given the group allocation.

The allocation of patient into either the refractory or remission group is defined as follows. Consider the first visit for each patient. If the estimated probability of being in the refractory group is greater than a chosen cut off, $\mathbb{c}$, then the patient will be assigned to the refractory group and stop predicting for this patient. If the probability is lower than $\mathbb{c}$, proceed to the next visit, and the patient remains *under observation*, repeating the process until either the patient has been grouped as refractory or all their visits have been used.

The classification performance is then measured using AUC, PCC, sensitivity, specificity, PPV and NPV. The cross-validation method is applied in which 70% of the data is randomly chosen for training set and the remaining 30% to test the classification accuracy. The process is repeated 100 times. For each split of data into training and test sets, the classification performance measures are calculated and the values are then averaged across the 100 splits.

6.2.5 Jointly modelling SNPs with longitudinal clinical markers

This section explains the way in which the SNP data are included in the MGLMM model. Two approaches are followed: (i) to jointly model the SNPs and the longitudinal markers, Y_r, and (ii) to add the SNPs as additional fixed

effects variables, X_r when modelling the longitudinal markers. For each approach, two top SNPs that were chosen by univariable tSNR ranking and cumulative tSNR ranking accordingly, will be jointly modelled with the longitudinal clinical markers. Since, the idea is to investigate the added predictive value of the SNPs, we focus on the top two SNPs of which ranked based on their importance.

With approach (i), the top two SNPs, represented by Y_{SNP_1} and Y_{SNP_2}, are added to the model together with the longitudinal markers. Although, the SNP is not a longitudinal marker, it can be modelled as constant across all visits for each patient using the model below:

$$\begin{cases} h_r^{-1}\{E(Y_{r,j}|b,g)\} = x_{r,j}^{gT}\alpha_r^g + z_{r,j}^{gT}b_r, & r = 1,\dots,R, j = 1,\dots,n_r \\ h_{SNP_i}^{-1}\{E(Y_{SNP_i,j}|g)\} = x_{SNP_i,j}^{gT}\alpha_{SNP_i}^g, & i = 1,\dots,p, j = 1,\dots,n_{SNP_i} \end{cases} \quad (6.3)$$

On the other hand, with approach (ii), the top two SNPs, coded as X_{SNP_1} and X_{SNP_2}, are added to the model as fixed effects. However, when adding SNP data as the fixed effects to explain the longitudinal evolution of the R longitudinal markers, the covariates information containing the SNP data in equation (6.1) can be written as, $x_{r,j}^{gT} = \left(x_{r,j}^{gT}, X_{SNP_i,j}^{gT}, \dots, X_{SNP_p,j}^{gT}\right)$. Table 6.2 illustrates the coding for one SNP when jointly modelled with the longitudinal clinical markers and as the fixed effects.

Table 6.2: The coding for SNP when jointly modelled with longitudinal clinical markers and as fixed effects (using rs680730 as an example for patients id 34, 1266 and 1268).

ID	Visits	Rem status	Original Coding (additive)	Coding as Y_{SNP_i}		Coding as X_{SNP_i}	
			rs680730	$Y_{SNP_{11}}$	$Y_{SNP_{12}}$	$X_{SNP_{11}}$	$X_{SNP_{12}}$
34	1999-04-12	0	0	0	0	0	0
34	1999-07-12	0	0	NA	NA	0	0
34	1999-10-04	0	0	NA	NA	0	0
34	2000-03-13	0	0	NA	NA	0	0
34	2000-05-26	0	0	NA	NA	0	0
34	2000-08-21	0	0	NA	NA	0	0
34	2001-02-02	0	0	NA	NA	0	0
34	2001-05-25	0	0	NA	NA	0	0
34	2002-03-19	0	0	NA	NA	0	0
34	2003-01-30	0	0	NA	NA	0	0
1266	2002-10-16	1	2	0	1	0	1
1266	2003-01-23	1	2	NA	NA	0	1
1266	2003-05-08	1	2	NA	NA	0	1
1266	2003-11-06	1	2	NA	NA	0	1
1266	2004-07-01	1	2	NA	NA	0	1
1268	2002-11-21	1	1	1	0	1	0
1268	2003-02-21	1	1	NA	NA	1	0
1268	2003-05-21	1	1	NA	NA	1	0

6.3 Results

This section presents the results of the methodology discussed in Section 6.2.3 to Section 6.2.5 when applied to the SANAD dataset (Section 6.2.1). As described earlier, the following research tasks are addressed: (i) to select the most informative SNPs using tSNR, and (ii) to jointly model the longitudinal clinical and SNP data. The aim is to develop a predictive model that allows to identify patients that will not achieve remission after five years of commencing treatment (refractory). Within each analysis, the classification performance is evaluated which allows the inspection of the added predictive value of the SNPs.

6.3.1 Selection of the most informative SNPs using tSNR

The first objective is to apply the filter metric tSNR (equation 3.7, Section 3.2.1) to the SANAD dataset and select the most informative SNPs to be jointly modelled with the longitudinal clinical variables in the following section. By using the ranking measure, the most informative SNPs may aid better classification accuracies.

Sample and genotyping QC

Initially, the genotype data consists of 38,000,817 SNPs across 22 chromosomes. Similar to the previous chapters, standard SNP QC procedures are applied to each SNP. This number is first reduced to 7,459,851 after filtering based on minor allele frequency (MAF), SNP genotyping rate and test Hardy-Weinberg Equilibrium (HWE). The screening on MAF only includes the SNPs with MAF >0.01. Low MAF SNPs could be more susceptible to genotyping errors and their association signals are less robust [47]. For SNP genotyping rate only SNPs with <10% missing genotypes are included. Further, SNPs that are extremely deviated from HWE (p-value $<10^{-6}$) are removed. At the same time, all 573 samples passed the standard sample QC procedure (based on rate of missingness, duplication of samples, relatedness and heterozygosity).

Data Pruning

The pruning option (150 50 0.90) is implemented using PLINK [46] software. For this method, consider a window of 150 SNPs, calculate the LD between each pair of SNPs in the window. If any pair of SNPs within the window are in LD greater than R^2 threshold of 0.9, the first SNP in the pair will be inactivated (pruned). Shift the window 150 SNPs forward and repeat the procedure. After

applying the LD pruning to the SANAD dataset, the SNPs in the dataset are reduced to 1,816,214 SNPs.

Univariable ranking using tSNR

Using the pruned dataset, univariable tSNR is calculated for each of the 1,816,214 SNPs. The SNPs are ranked based on the highest to the lowest tSNR. Table 6.3 shows an extract of the ranking (top 20 SNPs) which indicates the name, chromosome and tSNR value of each SNP. The first ranked SNP, rs680730 resides in gene DSCAML1 which is involved in neuronal differentiation. It has been linked to neuronal disorders such as Gilles de la Tourette and Jacobsen syndromes [150].

Table 6.3: The list of top 20 SNPs based on univariable tSNR ranking.

Rank	chr	rs	Base-pair position	tSNR
1	11	rs680730	117475233	0.0541
2	8	chr8.80678532.D	80678532	0.0512
3	13	rs17085098	70026654	0.0511
4	10	rs7908691	97686064	0.0510
5	20	rs75097987	62611904	0.0507
6	8	rs11785119	4244283	0.0507
7	1	rs115685211	156001168	0.0503
8	7	rs13227274	95695805	0.0502
9	8	rs2285266	17080623	0.0497
10	3	rs11714754	16434903	0.0496
11	2	rs62143849	68225099	0.0488
12	18	rs147294858	71489889	0.0486
13	4	rs151039268	65193598	0.0485
14	4	chr4.65098663.D	65098663	0.0481
15	10	rs4406763	97563394	0.0474
16	3	rs1093947	155693990	0.0469
17	1	rs61825964	218108423	0.0463
18	8	rs62497331	17059056	0.0458
19	3	rs62236347	16460617	0.0451
20	2	rs17725471	52892813	0.0447

Multivariable ranking using cumulative tSNR

In order to find the best predictive SNP the ranking based on cumulative tSNR (Figure 6.4) is applied. It ranks the SNPs based on the multivariable model produced by PLR. The analysis starts off with a subset of top 5,000 SNPs which are selected from the univariable tSNR. Then PLR is fitted on the 70% of the dataset in order to select the most informative SNPs. The remaining 30% of the data will be used to evaluate the classification performance. The splitting procedure is repeated for 100 times.

In the workflow (Figure 6.4), the tSNR is calculated for each model produced by PLR (100 models due to the 100 splits). The tSNR value for each model is then serves as the weighting for each SNP within the model. Then, the sum of weighting for each SNP across 100 models is produced. Now, a ranking of the SNPs based on the cumulative tSNR can be produced. The ranking is then used to determine the SNPs that will be jointly modelled with the longitudinal clinical variables.

The ranking of the cumulative tSNR (top 20 SNPs) is shown in Table 6.4. From the ranking, rs115685211 from chromosome 1 is found to be the most informative SNP based on the cumulative tSNR ranking. The SNP resides in gene called BGLAP. BGLAP is a protein coding gene mainly associated with diseases like Osteitis Fibrosa and Glucocorticoid-Induced Osteoporosis [151]. Meanwhile, the second top ranked SNP, rs75097987 exists within gene PRPF6 which has been associated with disease retinitis pigmentosa [152].

Table 6.4: The list of top 20 SNPs based on cumulative tSNR ranking.

Rank	chr	rs	Base-pair position	tSNR
1	1	rs115685211	156001168	4805.517
2	20	rs75097987	62611904	4550.501
3	1	rs61825539	229791727	4472.484
4	2	rs62143849	68225099	4146.340
5	19	rs117795722	22050297	3857.903
6	6	rs186948829	157867510	3677.574
7	2	chr2.80014517.D	80014517	3603.490
8	12	rs11061883	1741282	3339.295
9	6	rs116770746	32598268	3269.336
10	3	rs147906012	151142271	3263.983
11	9	rs148836393	74614961	3210.241
12	12	rs7956369	23853906	3203.066
13	8	rs111776025	126410966	3009.624
14	3	rs77748476	4946359	3005.977
15	4	chr4.65098663.D	65098663	3004.736
16	8	rs11785119	4244283	2955.329
17	4	chr4.25508711.I	25508711	2915.395
18	2	rs150553138	88296647	2874.237
19	1	rs138862060	53168108	2807.607
20	11	rs74766182	86923572	2671.426

The classification performance on the remission status of the patients is evaluated. The SNP is added into the logistic regression model one by one based on the ranking. The values of PCC, AUC, sensitivity, specificity, PPV and NPV are recorded every time a new SNP is entering the model. The inclusion of the SNPs is stopped when there is very small or no added improvement in any of the classification measures. For now, the classification performance of top 150 SNPs is shown in Figure 6.5.

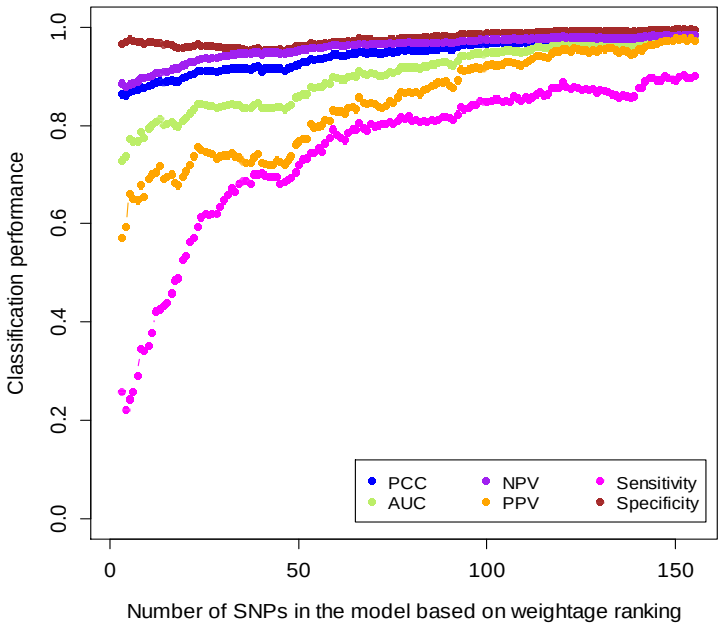


Figure 6.6: Classification performance (mean) of the top SNPs ranked by cumulative tSNR.

The values of PCC and AUC are ranging from 70% to 90%. It is important to note that after 50 SNPs the classification performance tends to increase every time a new SNP is added into the model. Although, there is a slight drop observed for sensitivity and PPV towards the end. Dealing with SNP data which is high-dimensional can be challenging in the context of prediction or classification. The results are often affected by over-fitting in which a model with too many variables begins to describe the random error (e.g. noise) rather than the underlying relationships between variables. An over-fitted model will produce erroneously high, or even perfect classification, which is misleading.

In order to investigate the extent of possible overfitting, adjusted tSNR is calculated (averaged over 100 splits). To visualise the scenario, Figure 6.6 shows the adjusted tSNR (in red) on top of the classification performance of the top 150 SNPs. It can be shown that there is no increment of adjusted tSNR after 50

SNPs which indicates possible overfitting of the data when more SNPs are added to the model afterwards.

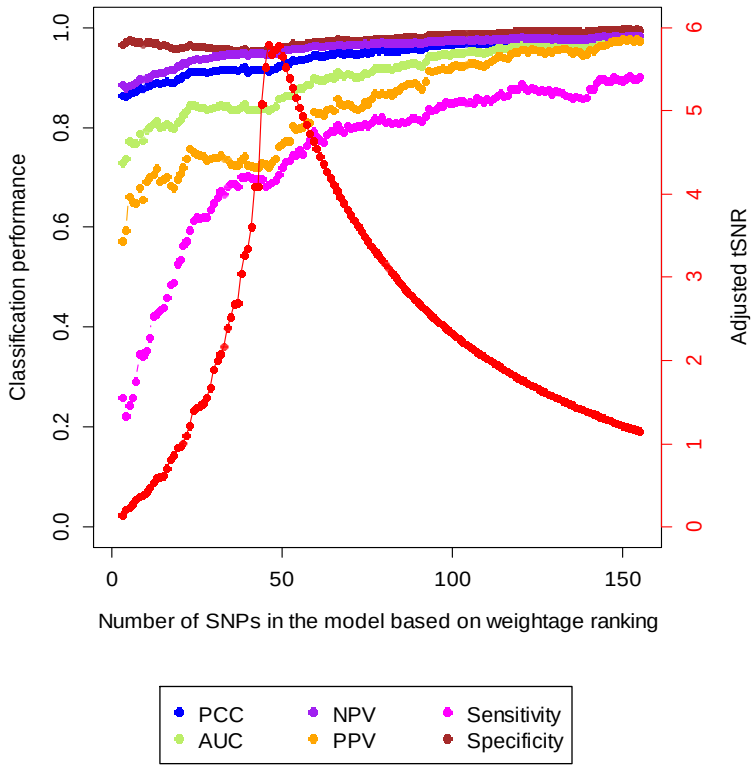


Figure 6.7: Classification performance (mean) of the top SNPs selected by cumulative tSNR ranking against adjusted tSNR.

To summarise, when considering adjusted tSNR as a stopping criterion (i.e. how many SNPs to retain in the model), the classification performance with 50 SNPs is shown in Table 6.5.

Table 6.5: Summary of the classification performance based on the cumulative tSNR by using adjusted tSNR as a stopping criterion (50 SNPs).

Classification performance	Mean	Standard deviation
PCC	0.93	0.02
AUC	0.86	0.05
Sensitivity	0.72	0.10
Specificity	0.96	0.02
PPV	0.77	0.09
NPV	0.95	0.02

6.3.2 Jointly modelling SNPs with longitudinal clinical data

In the previous section, SNPs are ordered by their predictive importance (Section 6.3.1). Here, their inclusion into the longitudinal modelling is considered. This section will show the results obtained when the longitudinal clinical data and SNP data are jointly modelled. The interest lies on how this combination of data can aid the classification performance of when using the longitudinal clinical data alone.

First, only longitudinal clinical data $(R = 3)$ are fitted using MGLMM as the reference model. The classification is then done applying LoDA (Section 6.2.4). Table 6.6 shows the classification performance of the reference model. Here, three prediction approaches namely; marginal, conditional and random effects are considered. The results are consistent with the results presented in Hughes et al. [116] in which the marginal prediction approach shows the best classification performance as compared to the other two approaches.

Table 6.6: Summary of the classification performance of LoDA for each marginal, conditional and random effects approaches for reference model.

	Marginal	**Conditional**	**Random effects**
Cutoff	0.82	0.41	0.17
Sensitivity	0.85	0.90	0.90
Specificity	0.92	0.88	0.88
PCC	0.92	0.88	0.89
AUC	0.97	0.95	0.94
PPV	0.67	0.57	0.58
NPV	0.99	0.98	0.98
Mean lead time (days)	692	782	891
Mean prediction time (days)	845	755	643

Lead time is defined as the time, before clinical classification can be confirmed, of which the LoDA can correctly predict a patient as belonging to the refractory

group. For the marginal prediction approach, the mean lead time is 692 days. This indicates that the patients who will not achieve remission from seizures (refractory) can be identified almost two years before they are clinically confirmed as such. The duration of two years is good which allows the clinicians to consider different forms of treatment, so that the patients do not have to continue with any treatment that is not suitable for them [116].

Meanwhile, prediction time is the average time since diagnosis at which patients are correctly identified as belonging to refractory group [116]. The mean prediction time, is 845 days for the marginal prediction approach, which denotes that the patients are correctly identified as being refractory approximately two years and three months after diagnosis.

The reference model with only longitudinal markers is used for comparison with other models where SNPs are added. In this analysis, the SNPs are selected by univariable tSNR ranking and cumulative tSNR ranking (multivariable). Often in Genome-wide Association Studies (GWAS), only few SNPs are concluded to be associated with the phenotype [148]. For this study only two SNPs (first and second top SNPs) are chosen to be modelled with the longitudinal clinical data in order to explore differences in classification accuracy.

The improvement in classification when two SNPs (selected using univariable selection from Table 6.3) are added is summarised in Table 6.7. It shows the classification performance when the clinical variables are jointly modelled with the SNPs. The SNPs rs680730 and chr8.80678532.D are the top two SNPs from the univariable tSNR ranking. Firstly, the SNPs are included into the model as the additional fixed effects, X_{SNP_1} and X_{SNP_2} accordingly. Secondly, the SNPs are

jointly modelled with the longitudinal markers, coded as Y_{SNP_1} and Y_{SNP_2} accordingly.

Table 6.7: Comparison of the models based on 70%-30% cross-validation with 100 splits. (The SNPs added based on univariable tSNR ranking)

	Reference model	**SNPs as fixed effects**			**SNPs in joint model**		
		X_{SNP_1}	X_{SNP_2}	$X_{SNP_1} + X_{SNP_2}$	Y_{SNP_1}	Y_{SNP_2}	$Y_{SNP_1} + Y_{SNP_2}$
Cutoff	0.82	0.80	0.84	0.86	0.82	0.80	0.81
Sensitivity	0.95	0.95	0.95	0.94	0.96	0.96	0.96
Specificity	0.92	0.92	0.91	0.90	0.92	0.92	0.92
PCC	0.92	0.92	0.91	0.91	0.93	0.92	0.93
AUC	0.97	0.97	0.96	0.96	0.97	0.97	0.97
PPV	0.67	0.67	0.64	0.64	0.68	0.67	0.68
NPV	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Mean lead time (days)	692	687	765	767	700	745	729
Mean prediction time (days)	845	849	773	770	840	794	811

From the table it can be seen that there is no improvement in the classification performance when the SNPs are added as the fixed effects. However, there is a slight improvement (e.g. PCC and PPV) when adding the first-ranked SNP, rs680730 independently (column Y_{SNP_1}) as well as in combination with the second-ranked SNP, chr8.80678532.D (column $Y_{SNP_1} + Y_{SNP_2}$).

There is very little and no improvement in mean lead time when modelling the SNPs as fixed effects or jointly modelled with the longitudinal biomarkers. For example, when adding both SNPs as fixed effects (column $X_{SNP_1} + X_{SNP_2}$) the mean lead time (i.e., the time to correctly predict a patient as belonging to the

refractory group) is only longer by 75 days. Similarly, the mean prediction time at which patients are correctly identified as belonging to the refractory group is 75 days sooner than the reference model.

Table 6.8: Comparison of the models based on 70%-30% cross-validation with 100 splits. (The SNPs added are based on cumulative tSNR ranking)

	Reference model	**SNPs as fixed effects**			**SNPs in joint model**		
		x_{SNP_1}	x_{SNP_2}	$x_{SNP_1} + x_{SNP_2}$	Y_{SNP_1}	Y_{SNP_2}	$Y_{SNP_1} + Y_{SNP_2}$
Cutoff	0.82	0.86	0.94	0.93	0.80	0.81	0.81
Sensitivity	0.95	0.94	0.93	0.93	0.96	0.96	0.95
Specificity	0.92	0.86	0.83	0.81	0.92	0.92	0.92
PCC	0.92	0.87	0.85	0.83	0.92	0.93	0.93
AUC	0.97	0.93	0.91	0.89	0.97	0.97	0.97
PPV	0.67	0.60	0.60	0.54	0.68	0.69	0.69
NPV	0.99	0.99	0.98	0.98	0.99	0.99	0.99
Mean lead time (days)	692	747	747	760	732	728	748
Mean prediction time (days)	845	790	791	777	808	811	791

The improvement in classification of two SNPs which were selected using multivariable analysis (from Table 6.4) is summarised in Table 6.8. It presents the classification performance of the joint model between the longitudinal clinical data and the SNPs selected using weighting tSNR ranking. The SNPs are the top two SNPs rs115685211 (X_{SNP_1} and Y_{SNP_1}) and rs75097987 (X_{SNP_2} and Y_{SNP_2}) selected by the cumulative tSNR ranking. Similar to Table 6.7, there is no improvement when adding the SNPs as fixed effects. However, there is a slight improvement (e.g. sensitivity, PCC and PPV) when jointly modelled the SNPs

with the longitudinal markers either individually (columns Y_{SNP_1} and Y_{SNP_2}) or together (column $Y_{SNP_1} + Y_{SNP_2}$) as compared to the reference model.

Further in Table 6.8, adding both SNPs (Y_{SNP_1} and Y_{SNP_2}) chosen by cumulative tSNR ranking has substantially improved the mean lead time from 692 to 748 days. The mean lead time 748 days indicates that the patients who will not achieve remission from seizures (refractory) can be identified two years before they are clinically observed as such. The time gain allows the clinicians to consider different treatments for the patients. The mean prediction time is positively reduced to 791 days as compared to the reference model. This means that the model with the addition two SNPs as the longitudinal markers can predict the patients belonging to the refractory group approximately two years and two months earlier than waiting for five years to determine their status.

From the results, the SNPs selected by cumulative tSNR contribute positively in evaluating the classification performance.

6.4 Concluding remarks

In this chapter, two objectives were tackled. Firstly, the variable selection method, filter metric tSNR was applied to select the most informative SNPs and the classification performance of the most informative set of SNPs was investigated. Secondly, the added predictive value of the selected SNPs when jointly modelled with longitudinal clinical data was explored.

From the SNPs selection procedure, two highest-ranked SNPs were chosen from each univariable tSNR ranking and cumulative tSNR ranking. Two different SNPs were chosen by each method. Building a predicting model with only two

SNPs may not be practical due to low classification performance. Hence, the proposal was to jointly model the SNPs and the longitudinal clinical data which was often well-validated and known to produce good classification accuracy.

Although the reference model (with only clinical variables) showed good classification accuracy, adding the SNPs as longitudinal markers improved the mean lead time and the mean prediction time of the patients significantly. The results also suggested that the model with SNPs chosen based on cumulative tSNR ranking provided better results as compared to the univariable tSNR ranking.

Chapter 7
Discussion

7.1 Introduction

The classification using single nucleotide polymorphism (SNP) data mainly aims to assign each sample (or individual) correctly to the group it belongs to. For example, in the situation of binary phenotypes, one is either interested to classify the samples into cases (e.g. disease, negative response to treatment) or controls (e.g. healthy, positive response to treatment). However, dealing with the high-dimensional problem of SNP data is challenging as it may cause overfitting and be computationally expensive. Hence, a typical procedure is to use a variable selection approach, often univariable, where the primary aim is to select the most important SNPs associated with an outcome of interest. Although the univariable approach is computationally inexpensive, it assumes complete

independence of SNPs [5]. Therefore, multivariable selection has gained lots of interest due to its ability to consider the correlation between SNPs which may contribute to an increment in the classification accuracy [23, 57].

There were three main objectives set out in Chapter 1 which were to be investigated in this thesis. The first objective was to develop a novel variable selection method for classification by considering the multivariable nature of SNP data. The second objective was to propose a framework for multivariable selection. The final objective was to jointly model the SNP data and longitudinal clinical data for classification.

In Chapter 2, a detailed review of the variable selection methods for classification using SNP data was undertaken. Following this review, a novel variable selection method, tSNR was proposed in Chapter 3. In addition, a multi-step framework that involved univariable and multivariable selection in a cross-validation setting was proposed. How to combine the SNP and longitudinal clinical data with the aim to improve classification performance was also explored in this chapter.

The filter metric tSNR and the proposed framework were studied using simulated datasets in Chapter 4. In Chapter 5 the methods were applied to a dataset from Epilepsy Pharmacogenomics (EpiPGX) study. Then, in Chapter 6, the method involved in combining the SNP and longitudinal clinical data were evaluated using the Standard and New Antiepileptic Drugs (SANAD) dataset.

In this chapter, the main contributions of this thesis presented from Chapter 2 to Chapter 6 are highlighted and general conclusions from the findings are discussed (Section 7.2). Then, the limitations are listed in Section 7.3. In order

to help researchers to benefit from this study, recommendations for practice are given in Section 7.4. Then, further works related to variable selection and classification are discussed in Section 7.5. Finally, the chapter is concluded in Section 7.6.

7.2 Discussion of thesis results

In this section, the results and key points from each chapter are summarised.

7.2.1 Implications of the literature review

The literature review was undertaken by reviewing papers related to variable selection for classification with specific application to categorical SNP data and binary phenotypes. In the literature review, a simple example using a simulated dataset was used to help explain the various variable selection and classification methods. Based on the review five main statistical challenges were identified (Section 2.3 in Chapter 2): high-dimensional data, high correlation between SNPs, categorical type of data, reproducibility issue, and determining the thresholds of SNPs selection. These challenges were then addressed in the subsequent chapters by developing the novel variable selection and model building framework.

Variable selection methods based on a univariable filter metric have been widely used during the preselection process to reduce the high-dimensionality of SNP data. Due to the multivariable nature of the data, SNPs selection (i.e. ranking) multivariable selection processes are also considered in the literature [9]. Hence, from the literature review, an ideal variable selection method would be able to perform not only in a univariable and multivariable setting. It is also ideal for the method to address some of the said statistical challenges mentioned above.

In terms of a specific method reviewed, logistic regression was not only performed well as a variable selection method (i.e. ability to select the simulated five causal SNPs), and also performed as well as other classifiers (e.g. SVM). In addition, with the knowledge that the underlying algorithm of logistic regression is well-known and widely accepted by researchers, the method was applied as the framework in developing a novel variable selection method in this thesis.

7.2.2 Implications of the proposed methodology

Following the literature review, a novel filter metric based on signal-to-noise ratio (SNR), tSNR was proposed. The SNR estimator proposed by Czanner et al. [17] was extended so that it can be applied within the framework of SNP data analysis. SNR is a measure that compares the level of desired signal to the level of background noise [11]. SNR is a well-known measure of fidelity in physical systems, for examples, audio and image processing [17]. Although, the idea of using SNR for variable selection in Genome-wide Association Study (GWAS) has been presented in the literature (e.g. t-test, Hotelling T^2), these previous methods require the data in continuous form (e.g. gene expression data) [98].

The tSNR filter metric does not only work in the univariable setting. This thesis explored tSNR in the multivariable setting, (for the purpose of model selection). It was shown that the tSNR has a negative relationship with AIC and BIC, while a positive relationship with generalised R^2. These three model selection criteria are long established. Their general rule is to compare models involving the same set of patients and fully nested models. However, their ability to compare between non-nested models is still debatable. On the other hand, tSNR can be used for both nested and non-nested models, and across different sets of patients

which is one of the main advantages. Particularly in this thesis, tSNR is applied to compare several models (non-nested) which are produced with the implementation of cross-validation in penalised logistic regression (PLR) usage.

The workflow introduced in this thesis is useful when the researcher is interested in investigating the effect of specific variables and not just the mere classification or prediction problem. The tSNR algorithm itself can be interpreted as the process of finding the set of SNPs that carries the greatest signal compared to noise in relation to the considered phenotype.

In addition, in order to improve the classification accuracy the approach to jointly model the SNP data and longitudinal clinical data was proposed. The variables were jointly modelled for classification purposes using the multivariate generalised linear mixed model (MGLMM). These longitudinal models were subsequently used in the dynamic longitudinal discriminant analysis (LoDA) to predict the probability of an individual belonging to a specific group.

7.2.3 Implications of the simulation study

In this simulation study analysis, the statistical properties of the novel filter metric approach, tSNR were studied. The performance of tSNR was measured by its ability to capture the causal SNPs which were predetermined in the simulated datasets, and tSNR was able to capture the causal SNPs in the top ten ranked SNPs using the simulated datasets.

Further, we tested the ability of multivariable modelling approaches such as PLR and stepwise logistic regression (SLR) to further reduce the number of SNPs (after the univariable ranking). It was found that PLR is more stable and

faster when applied to a large number of SNPs compared to SLR. Since, we required a large number of SNPs to be considered after the univariable ranking, PLR was chosen as the method in the multivariable setting.

The classification performance between univariable tSNR and cumulative tSNR ranking were compared in the simulated scenario. The results suggested that the classification performance using SNPs selected by cumulative tSNR (multivariable) was better than the classification performance using SNPs based on univariable tSNR ranking. Specifically, in our multivariable approach (cumulative ranking), PLR incorporated the classifier while selecting the SNPs (i.e. embedded method). Hence, the multivariable approach can offer better classification performance compared to when analysing the SNPs one at a time (univariable).

7.2.4 Implications of the clinical findings: tSNR as variable selection for classification

The methodology proposed in this thesis was illustrated using two SNP datasets of patients with epilepsy, where the aim was to identify patients who will not achieve remission of seizures on first well tolerated antiepileptic drugs (AEDs). The datasets consists of clinical and genetic information of patients in the EpiPGX study.

As shown in Chapter 5, the tSNR variable selection can be advantageous to reduce the dimensionality of the data and to help reduce the computational complexity. It leads to improving the classification accuracy of the classifier. The roles of tSNR in this study were twofold; i) as a filter metric to select the most

informative SNPs by using the univariable ranking, and ii) as selection criterion for the non-nested models to derive the model (combination of SNPs).

The filter approach is known as the pre-processing step to rank and select the most informative SNPs which eventually reduce the number of dimensions necessary for classification. The desired effect of this is to speed up algorithms and to make the subsequent analysis more effective by only focusing on the most relevant variables in the dataset. Hence, by using the filter metric tSNR as the variable selection method, fewer SNPs were analysed in the subsequent analysis with the knowledge that only informative SNPs (with assumption most of the noisy SNPs were eliminated) were selected for subsequent analysis.

In the multivariable setting, two strategies were proposed to select the best model. Strategy 1 applies a similar strategy as the univariable ranking by using the cumulative tSNR as the ranking measure. The cumulative tSNR ranking is superior compared to the univariable ranking as the SNPs are considered together during the modelling process by PLR. The PCC achieved 70% accuracy with a subset of 22 SNPs when using the cumulative tSNR ranking compared to 29 SNPs when univariable tSNR ranking was used.

Meanwhile, Strategy 2 was applied to compare the non-nested models that were produced by the cross-validation approach. The models were ranked from the highest to the lowest tSNR. From the result, the classification performance shown was adequate with an average of 80% accuracy. This analysis confirms the advantage of PLR since it not only enables to reduce the number of variables but leads to better classification accuracy.

Another important factor considered in this chapter was external validation. The models selected by Strategy 1 and Strategy 2 were externally validated. The classification performance increased significantly using the model chosen by Strategy 1. Meanwhile, a slight increment was observed for the model chosen by Strategy 2. However, the sample sizes in the external validation dataset were much lower than the development set which might influence high classification performance using model chosen by Strategy 1.

7.2.5 Implications of the clinical findings: combining SNP and longitudinal clinical data for classification

In Chapter 6, two objectives were considered. Firstly, the variable selection method, filter metric tSNR was applied to select the most informative SNPs and the classification performance of the most informative set of SNPs was investigated. Secondly, the added predictive value of the selected SNPs when jointly modelled with longitudinal clinical data was explored. The methods were applied to the SANAD dataset. In this study, the interest is to identify patient who will not achieve remission from seizures within five years of commencing treatment diagnosis. Patients who achieve a continuous 12-month period free from seizures within 5 years of diagnosis are regarded as being in "remission," whereas patients who do not are referred to as "refractory" [25].

The two highest-ranked SNPs were chosen from each univariable tSNR ranking and cumulative tSNR ranking. Two different SNPs were chosen by each method. The aim was to jointly model the SNPs and the longitudinal clinical data, which was known to produce good classification accuracy.

Although the reference model (with only clinical variables) showed good classification accuracy, adding the SNPs as longitudinal markers improved the mean lead time (i.e. the time required to correctly predict a patient as belonging to the refractory group) and, therefore also improved the mean prediction time (i.e. average time since diagnosis at which patients are correctly identified as belonging to refractory group) of the patients significantly. The early prediction allows the clinicians to consider different type of treatments since the initial treatment may not be suitable for the patients. Consequently, the patients do not have to endure any side effects due to unsuitable treatment. The results also suggested that the model with SNPs chosen based on cumulative tSNR ranking provided better results compared to the univariable tSNR ranking. As mentioned in Section 7.2.3 above, when developing the cumulative tSNR ranking, a large amount of SNPs are considered simultaneously using PLR which incorporates correlations among SNPs. Hence, this consideration which resulted in a different ranking of SNPs compared to the univariable tSNR ranking may have affected the classification performance positively.

7.3 Limitations

There are a few areas for improvement and optimisation in our approach. Firstly, in this thesis due to the large number of SNPs only the top ranked SNPs are chosen for the next analysis (forward selection). It is known that stepwise and backward selection are much preferred as they consider all variables from the start. The approaches only work if the number of variables is not too large. However, if the number of variables is large (ten thousand or more), it is not practical to start a backward or stepwise selection [153]. The forward selection procedure is feasible because in many classification problems a small number of

variables will be enough for plausible classification. Hence, with forward selection, it is not required to wait until all variables are added into the model.

Secondly, computer memory limitation was experienced with the statistical programming language R when trying to loop the variable selection using PLR with 5,000 SNPs. The number of loops or splits was set to 100 due to memory limitation. Although more splits are needed, the number of splits is still acceptable in reporting the average classification performance [134]. In a future study it may be interesting to consider more SNPs (more than 5,000) and more splits of cross-validation. This may not be possible on a Windows PC but it is possible to increase the memory that R uses by running the program on a Linux PC.

Thirdly, another potential limitation lies when jointly modelling the longitudinal clinical and SNP data. In the dataset used, the reference model with only longitudinal clinical data already demonstrated high classification accuracy, which left only a small room for improvement. Thus, the improvement for classification performance after adding the SNP data was up to one percent for the classification measures. In future studies in which clinical data show lower accuracy there may be an opportunity for the SNPs to improve predictive accuracy to a greater extent.

7.4 Recommendation for practice

Based on the results presented in this thesis, the novel filter metric tSNR may be employed for genetic research. However, a couple of recommendations should be considered when applying the filter metric. First, when applying the penalised logistic regression (PLR) using '`glmnet`' [135], values for λ and *dfmax* should be

considered. In PLR, λ is a constant to adjust the amount of the coefficient shrinkage. It can be manually specified or automatically using the built in cross-validation method within the package. Meanwhile, the *dfmax* determines the number of maximum variables to be non-zero in the model. Depending on the number of variables at the beginning, it is advisable for researchers to try multiple values of *dfmax* (e.g. 50, 100, 200) and study how the classification accuracies differ and then choose the value of *dfmax* that gives the highest accuracy.

The second recommendation for practice includes the stopping criterion involving adjusted tSNR. Depending on the interval (number of top ranked SNPs selected), the local or global maximum of the adjusted tSNR can be observed. In this situation, it is advisable to select the global maximum of adjusted tSNR to determine the final model.

7.5 Further perspective

Based on the results presented in this thesis, some topics are noteworthy for further exploration. First, some studies (particularly in machine learning) have considered bag of variables (i.e. clustering the variables according to their similarities) as the initial step when selecting a subset of variables from the original large dataset [127]. The action allows a bigger subset of variables considered at the pre-selection stage. Hence, it will be useful to consider a bigger subset of SNPs for multivariable analysis (e.g. PLR). However, few things need to be considered when submitting a large number of variables to a model; (i) whether the model will capture the informative SNPs or whether they might be missed due to the large number of variables considered, and (ii) the computational complexity involved. Subsequently, it will be interesting to

compare the performance of the proposed filter method, tSNR and other multivariable filter methods.

Second, normally when calculating the univariable p-value for each SNP the value is produced only once (without resampling). Hence, it is desirable to investigate the univariable tSNR ranking based on a bootstrapping technique. The bootstrapping technique is believed to help in producing a robust prioritisation of the SNPs which then enhance the SNPs ranking for classification [23]. Dealing with a large number of SNPs, the bootstrapping process might take a while. However, with high throughput computing, the univariable analysis is feasible since the system allows applications to run on over multiple computers which eventually reduces the time needed for certain analysis.

Third, stopping criteria in the model selection is another important area of research to be further explored. Although, adjusted tSNR is suggested as the stopping criterion when adding the variables based on ranking, the number of variables in the model is still large. It is worth to investigate further when imposing a more stringent penalisation for variable selection. For example, in AIC the penalisation is done with the coefficients multiplied by two, meanwhile in BIC the penalisation is done using log which is quite stringent. Hence, it will be interesting to impose a higher adjustment within the adjusted tSNR, which may aid in selecting a more parsimonious model.

Fourth, while this thesis is focusing on categorical SNP data with binary outcomes, it will be useful to see the application of tSNR filter metric using SNP data with continuous or time-to-event outcomes. Also, it will be interesting to

evaluate the performance of filter metric tSNR on continuous genetic data (e.g. microarray) which are known to be more flexible compared to categorical data.

7.6 Concluding remarks

Large scale simultaneous SNP selection is a statistically and computationally challenging task [5]. To this end, a novel filter metric based on signal-to-noise ratio, tSNR is introduced. From the results presented in this thesis, it can be concluded that, our proposed variable selection method, tSNR statistic and the multivariable modelling framework towards classification are all promising tools for applications to SNP data. The main advantage of the implementations using filter metric tSNR within the framework lies in its simplicity. The analysis is quick and the underlying framework of logistic regression is well-known, implemented in all statistical software and widely accepted.

In addition, the tools can be used to select the top ranked SNPs which can be jointly modelled with longitudinal clinical data. In order to improve the classification performance, we provide the methods to best combine the two type of data. In our clinical application, the combination improved the classification performance (specifically the prediction time).

REFERENCES

1. Elston R, Olson J, Palmer L: **Biostatistical Genetics and Genetic Epidemiology**. West Sussex, England John Wiley & Sons Ltd; 2002.
2. Genoma: Molecular Genetics Laboratories Group: **What are SNPs** [http://www.nutrigenetica.it/che-cosa-sono-gli-snps]; 2014. (Access date September 19, 2018)
3. Schwender H, Ickstadt K, Rahnenfuhrer J: **Classification with high-dimensional genetic data: assigning patients and genetic features to known classes**. *Biometrical Journal* 2008, **50**(6):911-926.
4. Batnyam N, Gantulga A, Oh S: **An Efficient Classification for Single Nucleotide Polymorphism (SNP) Dataset**. *Computer and Information Science* 2013, **493**:171-185.
5. Zuber V, Silva APD, Strimmer K: **A novel algorithm for simultaneous SNP Selection in high-dimensional genome-wide association studies**. *BMC bioinformatics* 2012, **13**(284):1-8.
6. Fridley BL: **Bayesian Variable and Model Selection Methods for Genetic Association Studies** *Genetic epidemiology* 2009, **33**:27-37.
7. Mathhew Shardlow: **An Analysis of Feature Selection Techniques** [https://studentnet.cs.manchester.ac.uk/pgt/COMP61011/goodProjects/Shardlow.pdf], 2016 (Access year 2017)
8. Guyon I, Elisseeff A: **An Introduction to Variable and Feature Selection**. *Journal of Machine Learning Research* 2003, **3**:1157-1182.
9. Saeys Y, Inza I, Larranaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507-2517.
10. Hira ZM, Gillies DF: **A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data**. *Advances in Bioinformatics* 2015, **2015**:1-13.
11. Welvaert M, Rosseel Y: **On the Definition of Signal-To-Noise Ratio and Contrast-To-Noise Ratio for fMRi Data**. *PloS one* 2013, **8**(11):1-10.
12. Alkuhlani A, Nassef M, Farag I: **A comparative study of feature selection and classification techniques for high-throughput DNA methylation data.** In: *International Conference on Advanced Intelligent Systems and Informatics 2016: 2016*: Springer, Cham; 2016: 793-803.
13. Mishra D, Sahu B: **Feature selection for cancer classification: a signal-to-noise ratio approach**. *International Journal of Scientific & Engineering Research* 2011, **2**(4):1-7.

14. Zhou N, Wang L: **A Modified T-test Feature Selection Method and Its Application on the HapMap Genotype Data**. *Genomics, Proteomics & Bioinformatics* 2007, **5**(3-4):242-249.
15. Zhou N, Wang L: **Effective selection of informative SNPs and classification on the HapMap genotype data**. *BMC bioinformatics* 2007, **8**:484.
16. Wang L: **Feature Selection in Bioinformatics**. *SPIE Proceedings, Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering X* 2012, **8401**:1-6.
17. Czanner G, Sarma SV, Ba D, Eden UT, Wu W, Eskandar E, Lim HH, Temereanca S, Suzuki WA, Brown EN: **Measuring the signal-to-noise ratio of a neuron**. *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112**(23):7141-7146.
18. He Q, Lin DY: **A variable selection method for genome-wide association studies**. *Bioinformatics* 2011, **27**(1):1-8.
19. Lin HY, Desmond R, Bridges SL, Jr., Soong SJ: **Variable selection in logistic regression for detecting SNP-SNP interactions: the rheumatoid arthritis example**. *European journal of human genetics : EJHG* 2008, **16**(6):735-741.
20. H.Hennings-Yeomans P, Cooper GF: **Improving the prediction of clinical outcomes from genomic data using multiresolution analysis**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2012, **9**(5):1442-1450.
21. Boulesteix A-L, Sauerbrei W: **Added predictive value of high-throughput molecular data to clinical data and its validation.** *Briefings in Bioinformatics* 2011, **12**(3):215-229.
22. Kohlmann M: **Discriminant Analysis for Longitudinal Data with Application in Medical Diagnostics**. Munich, Germany: Ludwig-Maximilians-University of Munich 2010.
23. Manor O, Segal E: **Predicting Disease Risk Using Bootstrap Ranking and Classification Algorithms**. *PLOS Computational Biology* 2013, **9**(8):1-10.
24. Su Z, Marchini J, Donnelly P: **HAPGEN2: simulation of multiple disease SNPs**. *Bioinformatics* 2011, **27**(16):2304-2305.
25. Hughes DM, Komárek A, Bonnett LJ, Czanner G, García-Fiñana M: **Dynamic classification using credible intervals in longitudinal discriminant analysis**. *Statistics in medicine* 2017:1-17.
26. Visscher PM, Brown MA, McCarthy MI, Yang J: **Five Years of GWAS Discovery**. *American Journal of Human Genetics* 2012, **90**(1):7-24.

27. Balding DJ: **A tutorial on statistical methods for population association studies**. *Nature Reviews Genetics* 2006, **7**(10):781-791.
28. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R *et al*: **Genome-wide association study identifies novel breast cancer susceptibility loci**. *Nature* 2007, **447**(7148):1087-1095.
29. Rabie HS, Saunders IW: **A simulation study to assess a variable selection method for selecting single nucleotide polymorphisms associated with disease**. *Journal of computational biology : a journal of computational molecular cell biology* 2012, **19**(10):1151-1161.
30. Schwender H, Zucknick M, Ickstadt K, Bolt HM, network TG: **A pilot study on the application of statistical classification procedures to molecular epidemiological data**. *Toxicology letters* 2004, **151**(1):291-299.
31. Huang LC, Hsu SY, Lin E: **A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data**. *Journal of translational medicine* 2009, **7**(81):1-8.
32. Lewis CM, Knight J: **Introduction to genetic association studies**. *Cold Spring Harbor protocols* 2012, **2012**(3):297-306.
33. Bush WS, Moore JH: **Chapter 11: Genome-Wide Association Studies** *PLOS Computational Biology* 2012, **8**(12):1-11.
34. Ziegler A, Konig IR, Thompson JR: **Biostatistical aspects of genome-wide association studies**. *Biometrical Journal* 2008, **50**(1):8-28.
35. Foulkes AS: **Applied Statistical Genetics with R**: Springer; 2009.
36. Schwender H, Ickstadt K: **Identification of SNP interactions using logic regression**. *Biostatistics* 2008, **9**(1):187-198.
37. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT: **Basic statistical analysis in genetic case-control studies**. *Nature protocols* 2011, **6**(2):121-133.
38. Schwender H, Rabstein S, Ickstadt K: **Do You Speak Genomish?** *Chance* 2006, **19**(3):3-10.
39. Liang Y, Kelemen A: **Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases**. *Statistics Surveys* 2008, **2**(0):43-60.
40. James G, Witten D, Hastie T, Tibshirani R: **An Introduction to Statistical Learning with applications in R**. New York: Springer; 2013.
41. Buhlmann P, van de Geer S: **Statistics for High-Dimensional Data Methods, Theory and Applications**. Heidelberg: Springer; 2011.

42. Nguyen TT, Huang J, Wu Q, Nguyen T, Li M: **Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests**. *BMC genomics* 2015, **16 Suppl 2**:S5.
43. Pearson TA, Manolio TA: **How to interpret a genome-wide association study**. *Jama* 2008, **299**(11):1335-1344.
44. Fadista J, Manning AK, Florez JC, Groop L: **The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants**. *European Journal of Human Genetics* 2016, **24**(8):1202-1205.
45. McCormack M, Alfirevic A, Bourgeois S, Farrell JJ, Kasperavičiūtė D, Carrington M, Sills GJ, Marson T, Jia X, Bakker PIWd *et al*: **HLA-A★3101 and Carbamazepine-Induced Hypersensitivity Reactions in Europeans**. *The New England Journal of Medicine* 2011, **364**(12):1134-1143.
46. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, Bakker PIWd, Daly MJ *et al*: **PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses**. *The American Journal of Human Genetics* 2007, **81**(3):559-575.
47. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT: **Data quality control in genetic case-control association studies**. *Nat Protoc 2010 Sep; 5(9):* 2010, **5**(9):1564-1573.
48. Darby Kameraad: **Determining the best LD Pruning options** [http://blog.goldenhelix.com/dkammeraad/determining-best-ld-pruning-options/]; 2016. (Access date July 1, 2018)
49. Assawamakin A, Prueksaaroon S, Kulawonganunchai S, Shaw PJ, Varavithya V, Ruangrajitpakorn T, Tongsima S: **Biomarker selection and classification of "-omics" data using a two-step bayes classification framework**. *BioMed research international* 2013, **2013**:1-9.
50. Pirooznia M, Seifuddin F, Judy J, Mahon PB, The Bipolar Genome Study (BiGS) Consortium JBP, Zandi PP: **Data Mining Approaches for Genome-Wide Association of Mood Disorders**. *Psychiatr Genet* 2012, **22**(2):55-61.
51. Cao K-AL, Boitard S, Besse P: **Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems**. *BMC bioinformatics* 2011, **12**(253):1-16.
52. Hemphill E, Lindsay J, Lee C, Mandoiu II, Nelson CE: **Feature selection and classifier performance on diverse bio-logical datasets**. *BMC bioinformatics* 2014, **15**(Suppl 13:S4):1-14.

53. Saurav Kaushik: **Introduction to Feature Selection methods with an example (or how to select the right variables?)** [https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/]; 2016.(Access date May 1, 2017)
54. Janecek AGK, Gansterer WN: **On the relationship between feature selection and classification accuracy**. *JMLR: Workshop and Conference Proceedings* 2008, **4**:90-105.
55. Wang Y-T, Sung P-Y, Lin P-L, Yu Y-W, Chung R-H: **A multi-SNP association test for complex diseases incorporating an optimal P-value threshold algorithm in nuclear families**. *BMC genomics* 2015, **16**(381):1-10.
56. LeBlanc M, Goldman B, Kooperberg C: **Methods for SNP Regression Analysis in Clinical Studies Selection, Shrinkage, and Logic**. In: *Handbook of Statistics in Clinical Oncology.* Third Edition edn. Edited by Crowley J, Hoering A: Chapman and Hall; 2012: 591-604.
57. Petrovski S, Szoeke C, Sheffield L, D'souza W, Huggins R, O'brien T: **Multi-SNP pharmacogenomic classifier is superior to single-SNP models for predicting drug outcome in complex diseases**. *Pharmacogenetics and Genomics* 2009, **19**(2):147-152.
58. Ratner B: **Variable selection methods in regression: Ignorable problem, outing notable solution**. *Journal of Targeting, Measurement and Analysis for Marketing* 2010, **18**(1):66-75.
59. Bursac Z, Gauss CH, Williams DK, Hosmer DW: **Purposeful selection of variables in logistic regression**. *Source code for biology and medicine* 2008, **3**:17.
60. Park MY, Hastie T: **Penalized logistic regression for detecting gene interactions**. *Biostatistics* 2007, **9**(1):30-50.
61. worldwide RCTac: **R statistical functions**. In., 3.6.0 edn; 2018.
62. Ayers KL, Cordell HJ: **SNP selection in genome-wide and candidate gene studies via penalized logistic regression**. *Genetic epidemiology* 2010, **34**(8):879-891.
63. Sun H, Wang S: **Penalized logistic regression for high-dimensional DNA methylation data with case-control studies**. *Bioinformatics* 2012, **28**(10):1368-1375.
64. Tibshirani R: **Regression Shrinkage and Selection via the Lasso**. *Journal of the Royal Statisical Society Series B (Methodological)* 1996, **58**(1):267-288.

65. Wu TT, Chen YF, Hastie T, Sobel E, Lange K: **Genome-wide association analysis by lasso penalized logistic regression**. *Bioinformatics* 2009, **25**(6):714-721.
66. Winkler C, Krumsiek J, Buettner F, Angermüller C, Giannopoulou EZ, Theis FJ, Ziegler A-G, Bonifacio E: **Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes**. *Diabetologia* 2014, **57**(12):2521-2529.
67. Sambo F, Trifoglio E, Camillo BD, Toffolo GM, Cobelli C: **Bag of Naïve Bayes: biomarker selection and classification from genome-wide SNP data**. *BMC bioinformatics* 2012, **13**(S2):1-10.
68. Zhang Z: **Naive Bayes classification in R**. *Annals of Translational Medicine* 2016, **4**(12):241-245.
69. Liu X, Wang Y, Sriram T: **Determination of sample size for a multi-class classifier based on single-nucleotide polymorphisms : a volume under the surface approach** *BMC bioinformatics* 2014, **15**(190):1-8.
70. Hartley SW, Sebastiani P: **PleioGRiP:genetic risk prediction with pleiotropy**. *Bioinformatics* 2013, **29**(8):1086-1088.
71. Setsirichok D, Piroonratana T, Assawamakin A, Usavanarong T, Limwongse C, Wongseree W, Aporntewan C, Chaiyaratana N: **Small Ancestry Informative Marker panels for complete classification between the original four HapMap populations.** *International Journal of Data Mining and Bioinformatics* 2012, **6**(6):651-674.
72. Malovini A, Barbarini N, Bellazzi R, Michelis Fd: **Hierarchical Naïve Bayes for genetic association studies**. *BMC bioinformatics* 2011, **13**(Suppl 14):1-11.
73. Miller PJ, Duraisamy S, Newell JA, Chan PA, Tie MM, Rogers AE, Ankuda CK, Walstrom GMv, Bond JP, Greenblatt MS: **Classifying variants of CDKN2A using computational and laboratory studies**. *Human Mutation* 2011, **32**(8):900-911.
74. Long N, Gianola D, Rosa GJ, Weigel KA, Avendaño S: **Comparison of classification methods for detecting associations between SNPs and chick mortality**. *Genetics Selection Evolution* 2009, **41**(18):1-14.
75. Long N, Gianola D, Rosa GJM, Weigel KA, Avendano S: **Machine learning classification procedure for selecting SNPs in genomic selection : application to early mortality in broilers**. *Journal of Animal Breeding Genetics* 2007, **124**(6):377-389.
76. Malovini A, Nuzzo A, Ferrazzi F, Puca AA, Bellazzi R: **Phenotype forecasting with SNPs data through gene-based Bayesian networks**. *BMC bioinformatics* 2009, **10**(Suppl 2):1-9.

77. Zhang J, Rowe WL, Struewing JP, Buetow KH: **HapScope: a software system for automated and visual analysis of functionally annotated haplotypes**. *Nucleic Acids Research* 2002, **30**(23):5213-5221.
78. Keith JM, Davey CM, Boyd SE: **A Bayesian method for comparing and combining binary classifiers in the absence of a gold standard**. *BMC bioinformatics* 2012, **13**:1-11.
79. Majka M: **High Performance Implementation of the Naive Bayes Algorithm**. In: *R CRAN Package*. 0.9.2 edn; 2018.
80. Shazadi K, Petrovski S, Roten A, Miller H, Huggins RM, Brodie MJ, Pirmohamed M, Johnson MR, Marson AG, O'Brien TJ *et al*: **Validation of a multigenic model to predict seizure control in newly treated epilepsy**. *Epilepsy Research* 2014, **108**(10):1797-1805.
81. Savan Patel: **Chapter 4: K Nearest Neighbors Classifier** [https://medium.com/machine-learning-101]; 2017. (Access date March 3, 2017)
82. Ripley B, Venables W: **Functions for Classification**. In: *R CRAN Package*. 7.3-14 edn; 2015.
83. OpenCV: **Introduction to Support Vector Machines** [http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html]; 2011-2014. (Access date August 1, 2015)
84. Kumar A, Rajendran V, Sethumadhavan R, Purohit R: **Identifying novel oncogenes : A machine learning approach**. *Interdisciplinary Sciences : Computational Life Sciences* 2013, **5**(4):241-246.
85. Upstill-Goddard R, Eccles D, Ennis S, Rafiq S, Tapper W, Fliege J, Collins A: **Support vector machine classifier for estrogen receptor positive and negative early-onset breast cancer**. *PloS one* 2013, **8**(7):1-8.
86. Yoon Y, Song J, Hong SH, Kim JQ: **Analysis of multiple single nucleotide polymorphisms of candidate genes related to coronary heart disease susceptibility by using support vector machines.** *Clinical Chemistry Laboratory Medicine* 2003, **41**(4):529-534.
87. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C-C, Lin C-C: **Support Vector Machines**. In: *R package*. 1.7-0 edn; 2018.
88. Zhang H, Wang M, Chen X: **Willows : a memory efficient tree and forest construction package** *BMC bioinformatics* 2009, **10**(130):1-6.
89. Garciıa-Magarinos M, Lopez-de-Ullibarri I, Cao R, Salas A: **Evaluating the Ability of Tree-Based Methods and Logistic Regression for the Detection of SNP-SNP Interaction**. *Annals of Human Genetics* 2009, **73**(Pt.3):360-369.

90. JS B-S, X G, C Z-J, NJ M, TR R: **Decision tree-based modeling of androgen pathway genes and prostate cancer risk.** *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2011, **20**(6):1146-1155.
91. Lunetta KL, Hayward LB, Segal J, Eerdewegh PV: **Screening large-scale association study data : exploiting interactions using random forests**. *BMC genetics* 2004, **5**(32):1-13.
92. Jiang Y, Brennan JS, Calixte R, He Y, Nyirabahizi E, Zhang H: **Novel tree-based method to generate markers from rare variant data**. *BMC proceedings* 2011, **5**(Suppl 9):1-6.
93. Therneau T, Atkinson B, Ripley B: **Recursive Partitioning and Regression Trees**. In: *R Package.* 4.1-13 edn; 2018.
94. Tsalenko A, Ben-Dor A, Cox N, Yakhini Z: **Methods for analysis and visualization of SNP genotype data for complex diseases**. *Pacific Symposium on Biocomputing* 2003, **8**:548-561.
95. Zhang X, Xue F, Liu H, Zhu D, Peng B, Wiemels JL, Yang X: **Integrative Bayesian variable selection with gene-based informative priors for genome-wide association studies** *BMC genetics* 2014, **15**(130):1-11.
96. Gary Altunian: **What is signal-to-noise ratio?** [https://www.lifewire.com/signal-to-noise-ratio-3134701]; 2016. (Access date 28 September 2017)
97. Smith SW: **The Scientist and Engineer's Guide to Digital Signal Processing** Second edn. San Diego, CA: California Technical Publishing; 1999.
98. Lu Y, Liu PY, Xiao P, Deng HW: **Hotelling's T2 multivariate profiling for detecting differential expression in microarrays**. *Bioinformatics* 2005, **21**(14):3105-3113.
99. Xiong M, Zhao J, Boerwinkle E: **Generalized T^2 test for genome association studies**. *The American Society of Human Genetics* 2002, **70**:1257-1268.
100. Dobson AJ: **An introduction to generalized linear models**, 2 edn: Chapman & Hall/CRC; 2002.
101. Wakefield J: **Bayes factors for genome-wide association studies: comparison with p-values** *Genetic epidemiology* 2009, **33**(1):79-86.
102. Bouhamed H, Lecroq T, Rebai A: **New filter method for categorical variables' selection.** *IAENG International Journal of Computer Science* 2012, **9**:10-19.

103. Croiseau P, Cordell HJ: **Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach**. *BMC proceedings* 2009, **3**(Suppl 7):S61.
104. Friedman J, Hastie T, Tibshirani R: **Regularization paths for generalized linear models via coordinate descent**. *Journal of Statistical Software* 2010, **33**(1):1-22.
105. Tibshirani RJ, Tibshirani R: **A bias correction for the minimum error rate in cross-validation**. *The Annals of Applied Statistics* 2009, **3**(2):822-829.
106. Clarke KA: **Testing nonnested models of international realtions: reevaluating realism.** . *American Journal of Political Science* 2001, **45**(3):724-744.
107. Wang B: **Variable selection in ROC regression**. *Computational and mathematical methods in medicine* 2013, **2013**:436493.
108. Seshan VE, Gönen M, Begg CB: **Comparing ROC curves derived from regression models** *Statistics in medicine* 2013, **32**(9):1483-1493.
109. Kohl M: **Performance measures in binary classification** *International Journal of Statistics in Medical Research* 2012, **1**:79-81.
110. Frommlet F, Ruhaltinger F, Twaróg P, Bogdan M: **Modified versions of Bayesian Information Criterion for genome-wide association studies**. *Computational Statistics & Data Analysis* 2012, **56**(5):1038-1051.
111. **Introduction to SAS. UCLA: Statistical Consulting Group.** [https://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/] (Access date November 15, 2017)
112. EN S, W C, M K, J K, T L, L P, OT R, RM S, NJ S, M S *et al*: **Longitudinal genome-wide association of cardiovascular disease risk factors in the Bogalusa heart study.** *PLoS genetics* 2010, **6**(9):1-11.
113. Bray PF, Wiser WC: **Hereditary characteristics of familial temporal-central focal epilepsy**. *Pediatrics* 1965, **36**(2):207-211.
114. Bin RD, Sauerbrei W, Boulesteix A-L: **Investigating the prediction ability of survival models based on both clinical and omics data: two case studies**. *Statistics in medicine* 2014, **33**:5310-5329.
115. Statnikov A, Li C, Aliferis CF: **Effects of environment, genetics and data analysis pitfalls in an Esophageal cancer genome-wide association study.** . *PloS one* 2007, **2**(9):1-5.
116. Hughes DM, Komárek A, Czanner G, Garcia-Fiñana M: **Dynamic longitudinal discriminant analysis using multiple longitudinal**

markers of different types. *Statistical Methods in Medical Research* 2016, **0**(0):1-21.

117. DM H, R ES, M G-F: **A comparison of group prediction approaches in longitudinal discriminant analysis** *Biometrical Journal* 2018, **60**(2):307-322.
118. Komárek A, Komárková L: **Clustering for multivariate continuous and discrete longitudinal data**. *The Annals of Applied Statistics* 2013, **7**(1):177-200.
119. Dudek SM, Motsinger AA, Velez DR, Williams SM, Ritchie MD: **Data simulation software for whole-genome association and other studies in human genetics.** *Pacific Symposium on Biocomputing* 2006, **11**:499-510.
120. International HapMap Project: **HapMap** [https://hapmap.ncbi.nlm.nih.gov/index.html.en]; 2005. (Access date January 15, 2015)
121. G. T, A. S, L. K, Stein L: **A Users Guide to the International HapMap Project Web Site**. In.; 2005.
122. Arlot S: **A survey of cross-validation procedures for model selection**. *Statistics Surveys* 2010, **4**:40-79.
123. Williamson T, Eliasziw M, Fick GH: **Log-binomial models: exploring failed convergence**. *Emerging Themes in Epidemiology* 2013, **10**(14).
124. Usai MG, Carta A, Casu S: **Alternative strategies for selecting subsets of predicting SNPs by LASSO-LARS procedure**. *BMC proceedings* 2012, **6 Suppl 2**:S9.
125. Ku CS, Loy EY, Pawitan Y, Chia KS: **The pursuit of genome-wide association studies: where are we now?** *Journal of human genetics* 2010, **55**(4):195-206.
126. Szymczak S, Holzinger E, Dasgupta A, Malley JD, Molloy AM, Mills JL, Brody LC, Stambolian D, Bailey-Wilson JE: **r2VIM: A new variable selection method for random forests in genome-wide association studies**. *BioData Mining* 2016, **9**(7):1-15.
127. Tharmaratnam K, Sperrin M, Jaki T, Reppe S, Frigessi A: **Tilting the lasso by knowledge-based post-processing**. *BMC bioinformatics* 2016, **17**(344):1-9.
128. Sanjay Sisodiya: **Epilepsy Pharmacogenomics: delivering biomarkers for clinical use** [http://cordis.europa.eu/docs/results/279/279062/final1-epipgx_final_report_final.pdf]; 2016. (Access date 19 December 2017)
129. Anthony G Marson, Asya M Al-Kharusi, Muna Alwaidh, Richard Appleton, Gus A Baker, David W Chadwick, Celia Cramp, Oliver C Cockerell, Paul N Cooper, Julie Doughty *et al*: **The SANAD study of**

effectiveness of carbamazepine,gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. *Lancet* 2007, **369**:1000-1015.

130. Anthony G Marson, Asya M Al-Kharusi, Muna Alwaidh, Richard Appleton, Gus A Baker, David W Chadwick, Celia Cramp, Oliver C Cockerell, Paul N Cooper, Julie Doughty *et al*: **The SANAD study of effectiveness of valproate, lamotrigine,or topiramate for generalised and unclassifi able epilepsy: an unblinded randomised controlled trial**. *Lancet* 2007, **369**:1016-1026.
131. Depondt C, Shorvon SD: **Genetic association studies in epilepsy pharmacogenomics: lessons learnt and potential applications** *Pharmacogenomics* 2006, **7**(5):731-745.
132. Bleeker SE, H.AMoll, Steyerberg EW, A.R.TDonders, Derksen-Lubsen G, Grobbee DE, MMoons KG: **External validation is necessary in prediction research: A clinical example**. *Journal of Clinical Epidemiology* 2003, **56**:826-832.
133. Goode EL: **Linkage Disequilibrium**. In: *Encyclopedia of Cancer*. edn. Edited by Schwab M. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011: 2043-2048.
134. Algamal ZY, Lee MH: **Applying penalized binary logistic regression with correlation based elastic net for variables selection**. *Journal of Modern Applied Statistical Methods* 2015, **14**(1):168-179.
135. Jerome Friedman, Trevor Hastie, Simon N, Qian J, Tibshirani R: **glmnet : Lasso and Elastic-Net Regularized Generalized Linear Models**. In., vol. R Package version 2.0-13; 2017.
136. Winkler C, Krumsiek J, Buettner F, Angermuller C, Giannopoulou EZ, Theis FJ, Ziegler AG, Bonifacio E: **Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes**. *Diabetologia* 2014, **57**(12):2521-2529.
137. **Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine.** [http://www.ncbi.nlm.nih.gov/SNP/]
138. Niccolini F, Wilson H, Pagano G, Coello C, Mehta MA, Searle GE, Gunn RN, Rabiner EA, Foltynie T, Politis M: **Loss of phosphodiesterase 4 in Parkinson disease: Relevance to cognitive deficits**. *Neurology* 2017, **89**(6):586-593.
139. Zigman JM, Westermark GT, Mendola JL, Boel E, Steiner DF: **Human G(olf) alpha: complementary deoxyribonucleic acid structure and expression in pancreatic islets and other tissues outside the**

olfactory neuroepithelium and central nervous system. *Endocrinology* 1993, **133**(6):2508-2514.

140. Charlesworth G, Bhatia KP, Wood NW: **The genetics of dystonia: new twists in an old tale** *Brain* 2013, **136**(7):2017-2037.
141. O'Riordan S, Ozelius LJ, Aguiar PdC, Hutchinson M, King M, Lynch T: **Inherited Myoclonus–Dystonia and Epilepsy: Further Evidence of an Association?** *Movement disorders* 2004, **19**(12):1456-1459.
142. König IR: **Validation in genetic association studies**. *Briefings in Bioinformatics* 2010, **12**(3):253-258.
143. Cosgun E, Limdi NA, Duarte CW: **High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans**. *Bioinformatics* 2011, **27**(10):1384-1389.
144. NHS: **Epilepsy** [http://www.nhs.uk/conditions/Epilepsy/Pages/Introduction.aspx]; 2011. (Access date May 17, 2017)
145. Office of Communications and Public Liaison National Institute of Neurological Disorders and Stroke National Institutes of Health Bethesda, MD 20892: **Epilepsy: Hope Through Research** [https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Hope-Through-Research/Epilepsies-and-Seizures-Hope-Through#3109_4]; 2015. (Access date November 17, 2017)
146. Bonnett L, Tudor-Smith C, Smith D, Williamson P, Chadwick D, Marson AG: **Prognostic factors for time to treatment failure and time to 12 months of remission for patients with focal epilepsy: post-hoc, subgroup analyses of data from the SANAD trial** *Lancet Neurology* 2012, **4**:331-340.
147. Bonnett LJ, Smith CT, Smith D, Williamson PR, Chadwick D, Marson AG: **Time to 12-month remission and treatment failure for generalised and unclassified epilepsy**. *Journal of Neurology, Neurosurgery, and Psychiatry* 2014, **85**:603-610.
148. D S, C H, S P, I T, A C, A J, H E, M DI, M T, T D *et al*: **A genome-wide association study and biological pathway analysis of epilepsy prognosis in a prospective cohort of newly treated epilepsy.** *Human Molecular Genetics* 2014, **23**(1):247-258.
149. Hughes DM, Bonnett LJ, Czanner G, Komárek A, Marson AG, García-Fiñana M: **Early identification of patients who will not achieve seizure remission within 5 years on AEDs (Accepted)**. *NEUROLOGY* 2017.
150. Agarwala KL, Ganesh S, Tsutsumi Y, Suzuki T, Amano K, Yamakawa K: **Cloning and Functional Characterization of DSCAML1, a**

Novel DSCAM-like Cell Adhesion Molecule That Mediates Homophilic Intercellular Adhesion. *Biochemical and Biophysical Research Communications* 2001, **285**(3):760-772.

151. **BGLAP gene** [www.genecards.org]. (Access date August 10, 2018)
152. **PRPF6 gene** [www.gencards.org]. (Access date August 10, 2018)
153. Bonev BI: **Feature selection based on information theory**. University of Alicante: University of Alicante; 2010.
154. Uh H-W, Mertens BJ, van der Wijk HJ, Putter H, van Houwelingen HC, Houwing-Doustermaat JJ: **Model selection based on logistic regression in a highly correlated candidate gene region**. *BMC Proceedings* 2007, **I**(Suppl I):S114.
155. Panagiotou OA, Ioannidis JP, Genome-wide Significance Project: **What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations.** *International Journal of Epidemiology* 2012, **41**:273-286
156. Wei-Yin Loh: **Fifty Years of Classification and Regression Trees.** *International Statistical Review* 2014, **82**(3):329-348
157. Arnošt Komárek: **Multivariate Normal Mixture Models and Mixtures of Generalized Linear Mixed Models Including Model Based Clustering**. In: *R package.* Version 5.1; 2018.
158. Plummer M, Best N, Cowles K, Vines K, Sarkar D, Bates D, Almond R, Magnusson A: **Output Analysis and Diagnostics for MCMC**. *R package.* Version 0.19-2; 2018.

APPENDICES

1.0 Data QC and pruning using PLINK

for i in `seq 1 22`; do echo "plink --bfile /ph-users/shima/SANAD_PLINK/chr$i --geno 0.10 --maf 0.01 --hwe 0.000001 --noweb --make-bed --out /ph-users/shima/SANAD_QC/chr$i.snps" > /ph-tmp/shima/script$i.sh; qsub /ph-tmp/shima/script$i.sh; done

for i in `seq 1 22`; do echo "plink --bfile /ph-users/shima/SANAD_QC/chr$i.snps --indep-pairwise 150 50 0.9 --allow-no-sex --noweb --out /ph-users/shima/SANAD_PRUNE0.9/chr$i " > /ph-tmp/shima/script$i.sh; qsub /ph-tmp/shima/script$i.sh; done

2.0 Data simulation using HAPGEN v2.0

hapgen2_submit -h CEU.chr1.hap -l hapmap3.r2.b36.chr1.legend -m genetic_map_chr1_combined_b36.txt -o chr1_rep1 -n 500 500 -dl 156952983 1 2.2 3.0 200252354 1 5.0 8.3

3.0 R Codes

3.1 Univariable tSNR

```
args <- commandArgs(trailingOnly=TRUE)
input_file <- args[1]
output_file <- args[2]

chr <- read.table(input_file, header=TRUE)

chr$PHENOTYPE[chr$PHENOTYPE==1] <- 0
chr$PHENOTYPE[chr$PHENOTYPE==2] <- 1
chr$PHENOTYPE <- as.numeric(as.character(chr$PHENOTYPE))
names(chr) <- sub('_\\d+$', '', names(chr))

SNPs <- names(chr)[7:ncol(chr)]
SNPsMatrix <- chr[,is.element(names(chr),SNPs)]
```

```
xfunction <- function(Geno)
{
glm.snp <- glm(PHENOTYPE~Geno, family=binomial, data=chr)
null <- glm.snp$null.deviance
residual <- glm.snp$deviance
SNR <-((null-residual)/residual)
return(SNR)
}
result <- apply(SNPsMatrix,2,xfunction)
write.table(result,output_file)
```

3.2 PLR with cross-validation

```
library(plyr)
library(glmnet)
library(Matrix)

#Create 100 training and 100 test datasets (80%-20% or 70%-30%)

first_seed <- 3009

x_train <- list()
y_train <- list()
x_test <- list()
y_test <- list()

PCC <- matrix(NA, nrow=100, ncol=100)
AUC <- matrix(NA, nrow=100, ncol=100)
Sensitivity <- matrix(NA, nrow=100, ncol=100)
Specificity <- matrix(NA, nrow=100, ncol=100)
PPV <- matrix(NA, nrow=100, ncol=100)
NPV <- matrix(NA, nrow=100, ncol=100)
tSNR <- matrix(NA, nrow=100, ncol=1)

cv <- list()
glmnet_mod <- list()
```

```
for (i in 1:100)
          {
          cat("count=",i,"\n")
          first_seed <- first_seed+i

          smp_size <- 0.8
          train <- ddply(data, .(PHENOTYPE), function(.,seed) {
          set.seed(seed); .[sample(1:nrow(.), trunc(nrow(.) *
          smp_size)), ] }, seed = first_seed)
          test <- ddply(minusID, .(PHENOTYPE), function(.,seed) {
          set.seed(seed); .[-sample(1:nrow(.), trunc(nrow(.) *
          smp_size)), ] }, seed = first_seed)

          x_train[[i]] <- sparse.model.matrix(PHENOTYPE~.,train)[,-1]
          y_train[[i]] <- as.numeric(train$PHENOTYPE)

          x_test[[i]] <- sparse.model.matrix(PHENOTYPE~.,test)[,-1]
          y_test[[i]] <- as.numeric(test$PHENOTYPE)

          cv[[i]] <- cv.glmnet(x_train[[i]],y_train[[i]], dfmax=200,
          family="binomial")
          glmnet_mod[[i]] <- glmnet(x_train[[i]],y_train[[i]],
          alpha=1, dfmax=200, lambda=cv[[i]]$lambda.min,
          family="binomial")
          null <- glmnet_mod[[i]]$nulldev
          residual <- deviance(glmnet_mod[[i]])
          SNR <-((null-residual)/residual)
          tSNR[i,] <- c(SNR)
          }


          for(i in 1:100)
          {
          for(j in 1:100)
          {
```

```
glm.probs                                                    <-
predict(glmnet_mod[[i]],x_test[[j]],lambda=cv[[i]]$lambda.
min, type="response")
glm.pred <- rep("0",nrow(glm.probs))
glm.pred[glm.probs>.5] <- "1"

rightPred <- glm.pred == y_test[[j]]
t <- table(glm.pred,y_test[[j]])
auc <- auc(as.numeric(glm.pred),y_test[[j]])
pcc <- sum(rightPred)/nrow(glm.probs)

totalpos <- sum(y_test[[j]]==1)
totalneg <- sum(y_test[[j]]==0)
truepos <- sum((y_test[[j]]==1)*(glm.pred==1))
trueneg <- sum((y_test[[j]]==0)*(glm.pred==0))
predpos <- sum(glm.pred==1)
predneg <- sum(glm.pred==0)

sens <- truepos/totalpos
spec <- trueneg/totalneg
ppv <- truepos/predpos
npv <- trueneg/predneg

PCC[i,j] <- c(pcc)
AUC[i,j] <- c(auc)
Sensitivity[i,j] <- c(sens)
Specificity[i,j] <- c(spec)
PPV[i,j] <- c(ppv)
NPV[i,j] <- c(npv)
}
}
```

3.3 Calculating cumulative tSNR

```
mtrx <- data.frame(variable=rep(0,5000))
rownames(mtrx) <- colnames(data[,2:5001])
```

```
variable <- NULL
for (i in 1:100)
{
variable[i] <- as.data.frame(as.matrix(glmnet_mod[[i]]$beta))
mtrx[,i] = variable[i]
}

mtrx[mtrx > 0] <- 1
mtrx[mtrx < 0] <- 1

colsum <- colSums (mtrx, na.rm = FALSE, dims = 1)

names(tSNR)[1] <- paste("variable")
tr_mtrx <- t(mtrx)
tr_tSNR <- t(tSNR)

multiply_mtrx <- sweep(tr_mtrx,MARGIN=1,tr_tSNR,`*`)

sum <- colSums(multiply_mtrx)
list <- head(sort(sum, decreasing=TRUE),5000)
convert <- data.frame(rs = names(list), weightage = list)
rownames(convert) <- NULL
```

3.4 Classification performance measures

```
library(matrixStats)
#repeat for AUC, Sensitivity, Specificity, PPV, NPV
sd_PCC <- transform(PCC, SD=rowSds(PPV, na.rm=TRUE))
mean_PCC <- rowMeans(PCC, na.rm = TRUE)
```

3.5 Graph for classification performance for univariable or cumulative tSNR ranking

```
##plot the first axis for PCC, AUC, Sensitivity, Specificity, PPV,
NPV
par(mar=c(5,5,4,4)+.1)
```

```
plot(Accuracy$SNP, Accuracy$PCC_mean, ylim=c(0,1.0), cex.lab=1.2,
xlab="Number of SNPs in the model based on cumulative tSNR
ranking", cex.axis=1.2, ylab="Classification performance",
type='b', col="blue", lwd=3, pch=16)
lines(Accuracy$SNP,Accuracy$AUC_mean,type='b',pch=16,
col='darkolivegreen2')
lines(Accuracy$SNP,Accuracy$Sens_mean, type='b', pch=16,
col='magenta')
lines(Accuracy$SNP,Accuracy$Spec_mean, type='b', pch=16,
col='brown') lines(Accuracy$SNP,Accuracy$NPV_mean, type='b',
pch=16, col='purple') lines(Accuracy$SNP,Accuracy$PPV_mean,
type='b', pch=16, col='orange')

legend(50,0.12,c("PCC","AUC","NPV",
"PPV","Sensitivity","Specificity"), ncol=3, cex=1.0,
col=c("blue","darkolivegreen2","purple","orange","magenta","brown")
,bty = "o", pch=c(16,16,16,16,16,16))

#To include axis on the right for adjusted tSNR
par(new = T)
with(Accuracy, plot(SNP, tSNR_adj, pch=16, axes=F, xlab=NA,
ylab=NA, ylim=c(0,6), col="red", col.axis="red", type="o"))
axis(side = 4, col="red", col.axis="red", cex.axis=1.2)
mtext(side = 4, line = 3, 'Adjusted tSNR', cex=1.2)
```

3.6 MGLMM

```
library(mixAK)
load("observationsSANAD0.dat"

mod0<-GLMM_MCMC(y=SANAD_0Train[,c("ltotsez","NumAdv","seizures")],
dist=c("gaussian","poisson(log)","binomial(logit)"),
id=SANAD_0Train[,"id"],
x=list(ltotsez=SANAD_0Train[,c("Time_LFU","time","Age.x","type","se
x","randP")],NumAdv=SANAD_0Train[,c("Time_LFU","time","Age.x","type
","sex","randP")],seizures=SANAD_0Train[,c("Time_LFU","time","Age.x
","type","sex","randP")]),
```

```
z=list(ltotsez="empty",NumAdv="empty",seizures="empty"),
random.intercept = c(ltotsez=TRUE,NumAdv=TRUE,seizures = TRUE),
prior.b=list(Kmax=1),
nMCMC=c(burn = 5000,keep=10000,thin=10,info=500),
PED=FALSE)

mod1<-GLMM_MCMC(y =
SANAD_1Train[,c("ltotsez","NumAdv","seizures")],
dist=c("gaussian","poisson(log)","binomial(logit)"),
id=SANAD_1Train[,"id"],
x=list(ltotsez=SANAD_1Train[,c("Time_LFU","time","Age.x","type","se
x","randP")],NumAdv=SANAD_1Train[,c("Time_LFU","time","Age.x","type
","sex","randP")],seizures=SANAD_1Train[,c("Time_LFU","time","Age.x
","type","sex","randP")]),
z=list(ltotsez="empty",NumAdv="empty",seizures="empty"),
random.intercept = c(ltotsez=TRUE,NumAdv=TRUE,seizures = TRUE),
prior.b = list(Kmax = 1),

nMCMC = c(burn = 5000, keep = 10000, thin = 10, info = 500),
PED=FALSE)
```

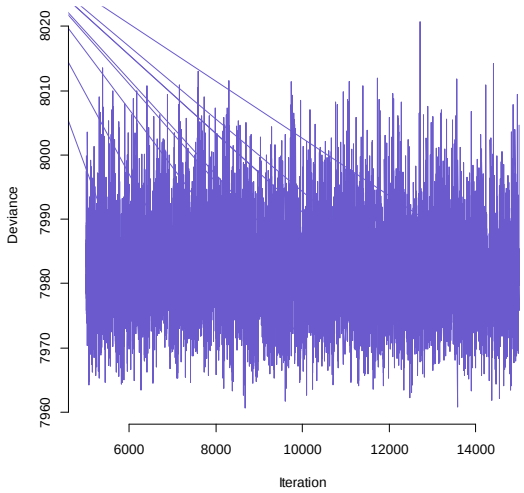
3.7 Convergence diagnostics check

```
##Check autocorrelation
library("coda")
DevChains <- mcmc.list(mcmc(mod0$Deviance))
autocorr(DevChains)
tracePlots(mod0, param = "Deviance")
tracePlots(mod0, param = "alpha")
tracePlots(mod1, param = "sigma_eps")
```

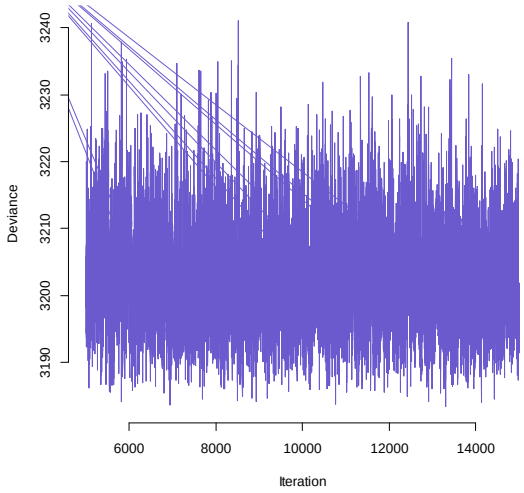
Table on estimated autocorrelations in the MCMC of the model deviances for each of the diagnostic groups (using a subset of samples).

	Achieved remission	Refractory
Lag 0	1.000000000	1.000000000
Lag 1	0.235788416	0.241577584
Lag 5	0.115316631	0.083228046
Lag 10	0.060002877	0.045431298
Lag 50	0.007749035	0.006682352

Traceplots of the model deviance

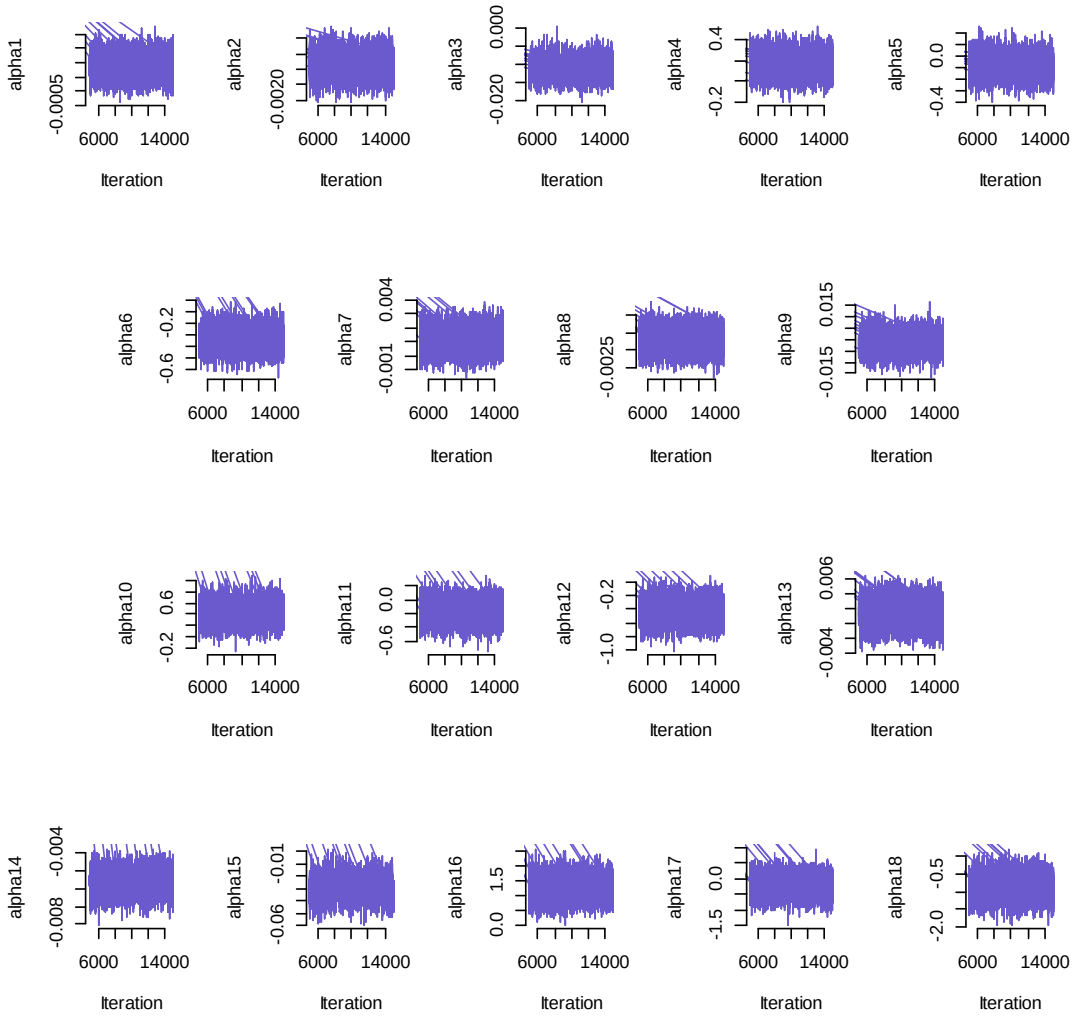


$g = 0$

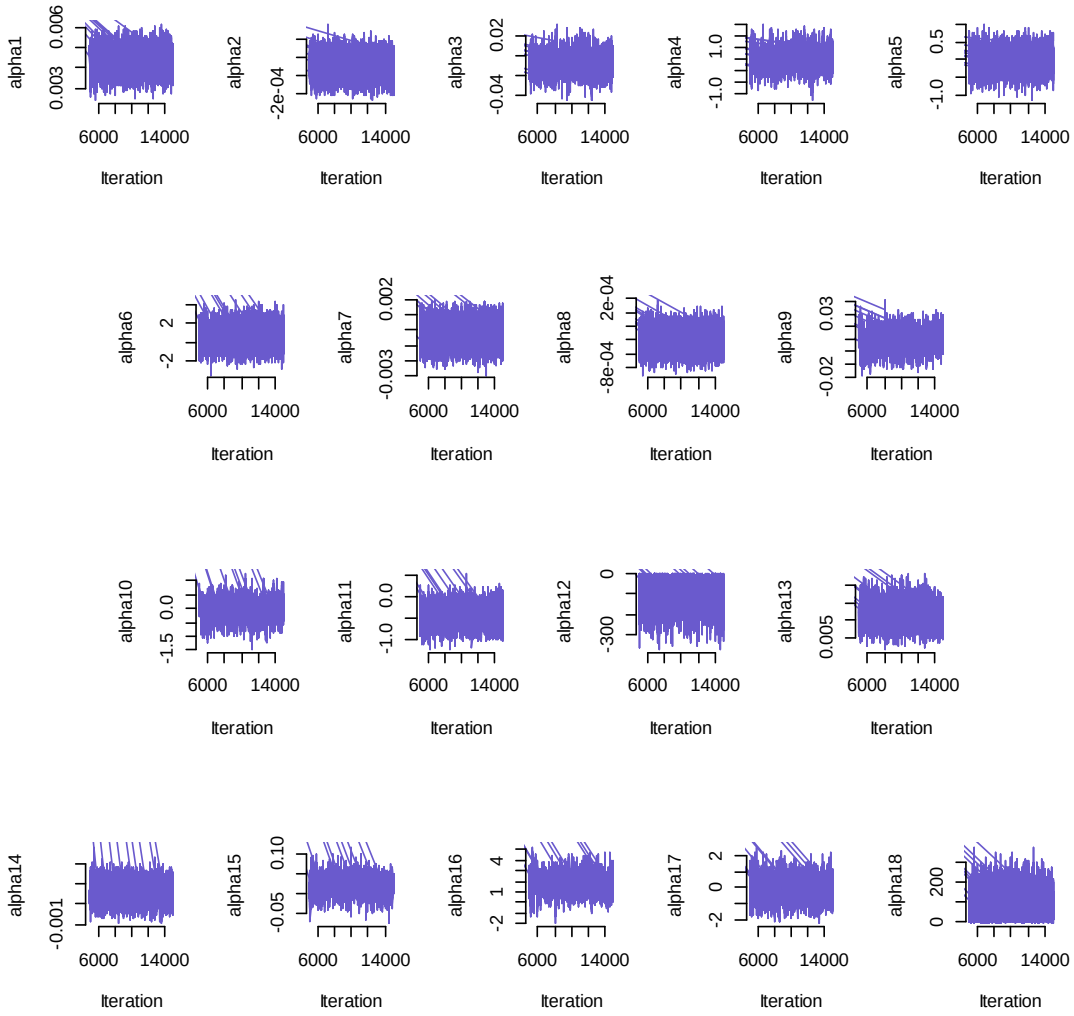


$g = 1$

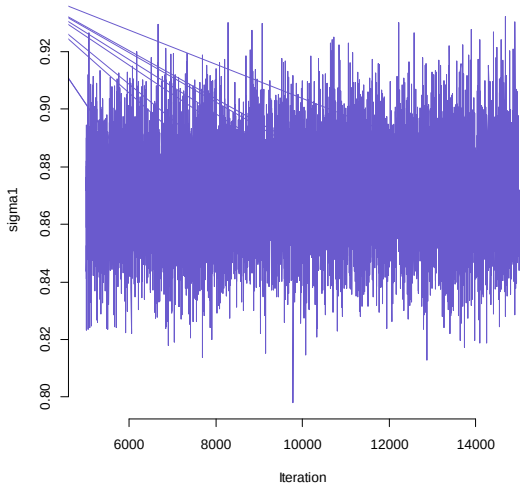
Traceplots of the fixed effects vector α_r^g $(g = 0)$



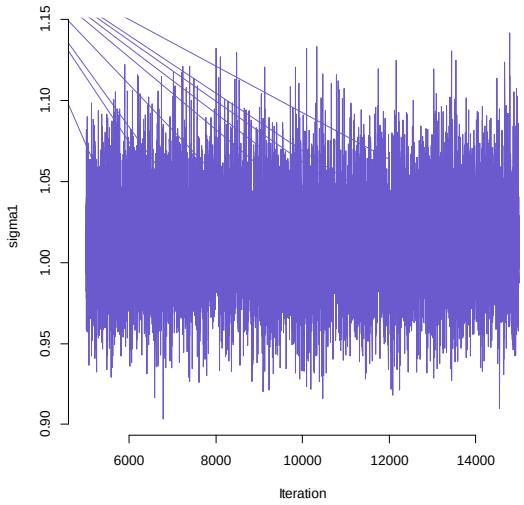
Traceplots of the fixed effects vector α_r^g $(g = 1)$



Traceplots of the dispersion parameters $\emptyset_r^g$



$g = 0$



$g = 1$