Detection of Susceptibility to Multiple Sclerosis from Single Nucleotide Polymorphism data

An application of contemporary computational methods and systems for personalised modelling applied in Bioinformatics

Vivienne Breen

A thesis submitted to Auckland University of Technology

in partial fulfilment of the requirements for the degree of

Master of Computer and Information Science (MCIS)

2013

School of Computing and Mathematical Science

Table of Contents

Attestation of authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), no material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Candidate: Vivienne Breen

Acknowledgements

I wish to acknowledge the valued input of both supervisors who commenced this research journey with me: Professor Nikola Kasabov and Dr Raphael Hu, then both of Auckland University of Technology. It is with regret that Dr Hu was not part of the team when the research was completed. Valuable assistance has also been given by Joyce D'Mello, and her encouragement has made the difference to the levels of perseverance required on many occasions.

This study makes use of data generated by the Welcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Welcome Trust under award 076113 and 085475 and is taken from the WTCCC[1] studies. The specific data investigated was that relating to Multiple Sclerosis where the initial research was headed by Alastair Compston, Department of Clinical Neurosciences, University of Cambridge.

Table of figures

Table of tables

Abstract

For many diseases that are genetically based, the date of onset is not predetermined or even predictable. To aid in assisting diagnosis of these diseases it is important to understand the person's susceptibility to developing a particular disease. In this study the susceptibility to Multiple Sclerosis is studied and modelled using as its base SNPs data. A SNP or single nucleotide polymorphism is the name given to a variation is a single base pair in a DNA sequence identified to be at a particular place on a specific chromosome. This data can be obtained using microarray chips which use as their input blood from a sample given by an individual.

To accurately process this data several areas need to be addressed. Firstly, the volume of raw data present, how it is handled, stored and manipulated for later processing. Secondly, how the data can be sensibly reduced to give a more manageable size base from which to model the data, whilst retaining all significant information. Thirdly, the modelling of the data itself and the presentation of these results.. For the purposes of determining susceptibility the modelling draws from the field of classification.

To follow this process of investigation the constructivist research approach is followed allowing for results at an earlier stage to alter testing and inference at later stages of investigation. In the first instance a prioritised list of useful methods of data handling, data reduction and modelling can be produced. In the second instance work can then proceed to use these methods to construct a cohesive system of information processing, from the raw data to a "prediction" of susceptibility ideally for a single individual at one time.

These are the aims and processes undertaken in this research, where the desire is to understand the data, its interactions and inter-relations between data points (SNPs), the best processes of data reduction and accurate modelling.

The analysis of experimental data obtained from the Welcome Trust in the UK and results obtained in this study confirm the hypothesis that is it possible to accurately predict the

susceptibility of an individual to MS using personalised modelling on SNPs data. His is the first study on this data that results in a high predictive accuracy along with discussing the application of different information methods.

The potential for further work to ensure the methods found here can be implemented into a system usable by clinicians to enhance existing medical procedures is huge. The idea of a personalised model of both the disease and an individual interacting to assist doctors may be closer than previously thought.

Chapter 1 Introduction

There are many diseases that are genetically linked which are being investigated in greater detail thanks to the ability researchers now have to interrogate DNA. It is now relatively simple to test a person from a blood sample and determine susceptibility in comparison to previous capabilities. Where does computational type science play a part? Mostly in the analysis of the data obtained, along with assisting to steer future research. Analysing this type of data can take many forms and with the increasing size of datasets becoming available existing statistical techniques alone are not sufficient to mine or dig out all available insights.

1.1 Aims and research questions of the study

The aim of this study was to address how existing methods and algorithms can be utilised (and improved) to obtain a personalised solution in the risk prediction of a genetic disease. In this case the disease in question is Multiple Sclerosis (MS).

The main hypothesis is that personalised modelling methods, with properly selected parameters, can accurately predict the susceptibility to MS from SNPs data. Another hypothesis is that data mining techniques can complement biological approaches by finding new potential test markers, in this case new SNPs that have not been referenced so far in the literature, contributing to new knowledge discovery.

The following research questions are addressed in this study:

1. Issues of handling a large dataset efficiently and effectively?
2. What is the most appropriate pre-processing of the SNPs data related to a problem?
3. How can different methods for personalised modelling be adequately compared?
4. What role does optimisation play in terms of accuracy of a personalised model?
5. Is the final solution of a personalised model robust enough to cope with the addition of data from other sources that have data points not present in the original dataset?
6. As this study is based on SNP (single nucleotide polymorphism) data, the ability of any model to cope with new data where individual SNPs were not present in the base

dataset is a very real and present issue with the on-going development of microarray chips with a vastly increased capacity than that of the one used to obtain the data used in this study.

7. How risk factors, e.g. SNPs, can be ranked in terms of their importance to the problem?
8. What software tools are suitable for personalised modelling on SNPs data?

1.2 Background to the study

The dataset itself comes from a previous study into 7 genetic conditions undertaken on behalf of the WTCCC (Welcome Trust Case Control Consortium) based in the United Kingdom. They investigated SNPs for association with each disease using existing proven statistical methods. Their results have been used for further investigations into a number of the diseases studied, but from a search of available literature very little has been done in the study of MS. It is known that this type of data is undermined in terms of digging out information that existing methods have not been able to address (Bakir-Gungor & Sezerman, 2011). The application of data mining techniques and methodology enhances this and adds increased abilities to model the data for different purposes (Bull & Kovacs, 2005; Sumathi & Sivanandam, 2006).

Why study Multiple Sclerosis?

MS is a debilitating and degenerative disease that can present at any age and early symptoms can be misdiagnosed as other conditions. Understanding the susceptibility of a person allows that person and the medical professionals concerned to take early symptoms into greater account and thus begin to address the problem earlier. To date there are no efficient information methods to accurately predict susceptibility of a person to MS based on SNPs data.

What makes MS a problem to the sufferer?

Multiple sclerosis is a genetic disease that affects the central nervous system. MS gets its name from the many (multiple) scarring (sclerosis) of the nerve casing or myelin sheath. The myelin sheath forms the “insulation” for the nerve and speeds the path of signals carried by

the nerve. When the myelin is infected it heals within weeks or months and leaves scar tissue on the sheath, limiting its function and protection of the nerve itself (Multiple Sclerosis New Zealand, 2011). Figure 1-1 shows the damage done by infection and the resulting scarring of the myelin. Inflammation is caused by the immune system being tricked into thinking normal brain tissue is foreign matter to be attacked and eradicated, in the same way the body detects viruses and goes about combating them. The result being damage to the brain and/or spinal cord which in turn causes difficulties with a number of normal functions such as walking, thinking, and bladder control (*Genes shed new light on cause of MS*). There are cases of MS sufferers having little to no outward affects, while others progress towards total disability very quickly. Most cases fall somewhere in between these two (Multiple Sclerosis New Zealand, 2011). MS is diagnosed more often for people aged between 20 and 45, and in more women than men (Marchiondo, 2010).

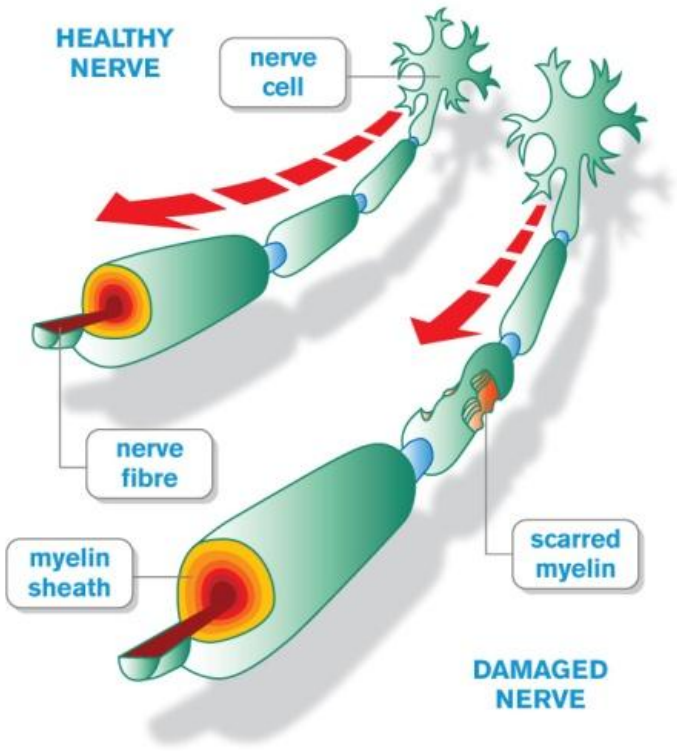


Figure 1 - 2: Damaged nerves due to MS showing direction of "information" flow. (biology111fall2010.wikispaces.com)

How can data mining and computational methods help? Investigating relationships in large volumes of data is more suited to these methods than more traditional statics alone. The ability of incorporate housing the data in databases that are able to "feed" the modelling

components with data in a predetermined format irrespective of the way in which they are stored in the database is of great advantage. Utilising these types of method allows the researcher to implement different techniques/methods at different stages as they are applicable and to combine methods where needed. The ability to dynamically address new data and adjust the previously stored model of a disease or even the profile of a patient is enabled by the application of personalised modelling.

The use of personalised modelling has only in relatively recent times been applied to biological data. It lends itself very much to this area of application, especially in medical decision support systems. To date many medical decision support systems have addressed diagnosis but did not incorporate on-going treatment options, which were implemented in other systems. The desire is to eventually have the ability to construct systems that can track patient information in whatever form it takes, create a patient's risk profile, assist clinicians in diagnosis and monitor treatment; but also to dynamically adjust the underlying models as new data is entered and that will cope with missing information.

A number of studies have developed and optimised various methods for modelling SNP data and compare their results to the WTCCC study, but they do not measure the effectiveness of their method against other methods nor do they address the interaction of methods in a more holistic style of approach. Many of these studies have also taken as their starting point the list of significant SNPs from previous research and not addressed the original dataset as a whole. Doing so limits the new methods of investigation by the limitations of previous research and does not allow for new SNPs to be identified that may contribute to increasing the accuracy of the modelling process. As determining susceptibility, which is the focus of this study, is often difficult the ability to search out new SNPs and combinations of SNPs can increase the accuracy of the result. SNP interaction is an area of investigation that is increasingly being investigated and achieving some surprising results in terms of what SNPs interact and the effects of SNPs that do not lie in chromosomal areas defined as genes.

1.3 Outline of the investigative process undertaken

In very simple terms this study seeks to investigate the *what, when, how, and effectiveness* of various method that address the pre-processing and selection of candidate SNPs, and modelling methods when using SNPs data to determine susceptibility to MS.

Taking as it's starting point the raw data available from WTCCC as it would have appeared before and quality control procedures were applied, a series of pre-processing techniques are needed to both reduce the dataset to a much smaller size determined by significance in explaining the variation in the data, and to prepare the data in a format that is appropriate as input to the various modelling system investigated. From this the modelling methods are assessed and a limited amount of variation allowed to assist in the determination of best method and to indicate where further optimisation would be beneficial.

Utilising this type of investigative process will allow the detection of possible new SNPs and is not limited by previous research and any "blinkering" effect from assuming biological knowledge to be completely correct. Apart from obtaining a list of good methods and options for improvement the style of investigation allows the results form one section of the study to modify others and slightly alter the direction of the details of the research. The overall goal remains in tack but allowing movement within the details permits a more informed conclusion/decision to be made in regard to the understanding of the data, modelling results, and implication for future development.

1.4 Organisation of the study

This thesis is made up of six individual chapters.

Chapter 1 (Introduction) presents a brief outline of the setting for this work and its research aims.

Chapter 2 (Literature review) presents a look into previous research taken from four perspectives; science, medical, software and computational methods. Each of these perspectives has an impact on the research undertaken here at different stages of data processing and modelling. It also highlights areas that affect decisions of method selection, optimisation and eventual implementation.

Chapter 3 (Methodology) presents the setting of the research in its corresponding research approach, outlines how the research is to be undertaken and the identification of limitations.

Chapter 4 (Results) presents the findings from the various stages of investigation and highlights a number of key results, including some that were very unexpected.

Chapter 5 (Discussion of results) gives an interpretation of the key findings and their implications for future work.

Chapter 6 (Conclusion) summarises the study findings and outlines how these affect further research.

1.5 Major contributions of this study

This is the first comprehensive study of using personalised modelling on MS SNPs data that resulted in several contributions to the areas of information science and bioinformatics.

Contributions are made about using different machine learning methods for personalised modelling on SNPs data. Among the tested methods, the WW*k*NN method (Kasabov, 2006) outperformed the rest. The applicability of several software environments were also studied; including NeuCom, WEKA, MATLAB, that offered complementary functions and models, with a slight preference for NeuCom versus WEKA.

This study has found a number of interesting results. One is that using data mining methods for attribute reduction has discovered the same SNPs as a number of previous studies, thus verifying the validity of this approach and that it is possible to apply data mining and knowledge engineering techniques to biological data and achieve at least the same results as existing methods. The difference is that in this study several SNPs were found significant that did not appear in any previous work. Another is that simple methods can be just as effective as their more complex counterparts. A specific contribution to MS SNPs data is the discovery of 5 new SNPs (not listed so far in the literature as important ones for the prediction of MS) through constructive feature selection. These SNPs can be further utilised in new DNA tests by clinicians and experts in MS.

The main results of this study emphasising the new discoveries are being prepared to be published as a journal paper to be submitted within a month after submission of this thesis.

Chapter 2 Literature Review

Research into modelling of biological data has been in existence for many years. In more recent times the application of the computational "grunt" as it were of the computer has made it possible to address issues previously infeasible. The continued growth in computer power and capability along with our understanding of the data allows today's researchers to dig deeper into the data, and this does not seem to be abating in any way. Increased understanding of data modelling and data mining has also enhanced the way researchers interact with data and increased their ability to model it in meaningful ways.

What is meaningful in one context may not be in another and care is needed in determining the relative merits of data management and handling along with the applied modelling systems. The challenge still remains with biological data in particular of incorporating new information as it comes to light in many areas of research. The more that is known about the interaction and interconnection of the various parts of DNA and how variations or SNPs (Single Nucleotide Polymorphisms), which represent an alteration in a single base pair of the DNA coding, affect the functional behaviour of the DNA increases the medical profession's ability to diagnose and monitor various conditions and especially those that can be genetically determined.

Most of the research reviewed to date can be placed into four categories as to their authors perspective; Science, Medical, Software, Computational methods. Much development has been made in the area of computational methods in recent years with the enhanced ability and speed of computers to cope with the complexities required. There are still many gaps in terms of the fluidity of development and usage of methods within the medical profession in particular as application in this domain is very "messy" in terms of initial data where there is much to be done in pre-processing before any method of analysis can be applied. As new computational methods are often developed with artificial or ideal datasets which are well defined and often not large (in comparison to biological research data sizes) it is not a simple matter to translate or scale up such techniques successfully.

2.1 Science & Medical Perspectives

As with many fields of study medicine is both old and new. The ability to understand and treat various conditions and diseases has progressed with both the development of understanding and technology. For example the microscope allowed scientists to view a world not available to the naked eye. The discovery of DNA, its structure and later some of its functions has improved the understanding of many diseases, particularly those with a hereditary basis or component (M. Ban et al., 2009). With the on-going advances in technology and computational understanding and power a return to the synthesis of biological and mathematical fields of study is happening. With the advances in computing and software design a newer discipline is joining in the search of answers.

The advancement of technology has also given rise to an increase in the amount of genomic data (and other medical data) available there is increased crossover between the biological, medical, and computational sciences (Kim, 2002). This interaction is predicted to transform the nature and structure of biomedical knowledge bases, plus the way the data is processed, analysed and interpreted (Li & Schwartz, 2011). Biochips can give rise to tens of thousands of measurements for a single sample and thus lend themselves to the use of computational strategies to cope with the volume of data, and benefits from the application of machine learning methods of analysis (Chen, 2007; Piatetsky-Shapiro, 2003). When reducing the amount of data to be used in the final model it should be kept in mind the final application of any decision or recommendation of the system, and to reduce the number of false negatives it likely that more attributes will be included than would be so if purely computational analysis were used for the final decision (Chen, 2007). The availability and accuracy of data influences much of the early processing from both a computational and biological viewpoint and can lead to the reduction in false positives or other erroneous results which cloud the read results obtained (M. Ban et al., 2009). With only a fraction known from previous studies, mostly Genome Wide Association Studies (GWAS), much work is underway to look into the missing components in diagnosis, treatment and influence of hereditary components on disease (The Wellcome Trust Case Control Consortium, 2010).

Researchers are interested in patterns, associations and models of systems that are different or enhanced in some way in comparison to those already in existence. This is not an easy task

and one researcher has described it as extremely challenging when the volume of data is taken into consideration. There are 10-15 million common genetic variations and billions of rare variations both of which are contributors to the susceptibility and progression of genetic disorders (Sawcer, 2010). They seek to know what we don't already know. To this aim a measure of value of a discovered model of interest is "interestingness", which incorporates a measure of overall pattern value including novelty, validity, usefulness and simplicity in its evaluation (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The most common of uses or objectives for studies involving gene data have been class comparison, class prediction and class discovery, often from microarray data (Chen, 2007). There is also an increase in interest in the detection of more complex traits and gene interactions which take into account loci on different genes and chromosomes as opposed to focusing on the presence of a single disease gene alone (Hoh & Ott, 2003). In the area of disease detection, against a control sample, a marker strongly related to risk does not always guarantee effective discrimination between the two so other loci and markers need to be incorporated into the model (H.-J. Ban, Heo, Oh, & Park, 2010). The on-going analysis of existing data from precious studies, mostly GWAS as they are expensive to undertake and contain much more information than has been extracted to date, can be used to direct further research using newer systems and techniques like next generation sequencing (Wang, Prins, Sober, Laan, & Snieder, 2011), which is also less expensive (comparatively) than precious larger scale investigations. In additions the causal effects of association which are often not known can be investigated for which there is much desire amongst the research community (Todd et al., 2007).

Many sources of data are not only from biological experiments, but come from existing clinical databases. These databases have been setup with a single patient in mind and the need of medical professionals to track their history, medication, conditions etc.; rather than the discovery of prediction models for disease diagnosis or treatment options (Kalatzis et al., 2010). To this a large amount of pre-processing will be needed where data is extracted from existing sources, dealing effectively with data type mismatches, data representation mismatches and inconsistencies, missing data, range inconsistencies and time stamp information (Lin & Haug, 2006; Zhao & Leong, 2000). The data in such data bases is notoriously of poor quality with entry error adding to the difficulty in standardising data before any data mining activities can be carried out (Mitha et al., 2011). Additional reasons for data missing from existing records include their not being collected, omitted from

previous versions of the system, and inapplicability in a specific clinical situation (Lin & Haug, 2006). The incorporation of multiple streams or sources of information into a comprehensive model steers the research towards personalised modelling, where the focus is not only on the discovery of information related to a disease but how that knowledge affects a single person (Kalatzis et al., 2010). Recent rapid technological advances have aided in this endeavour. However these have given rise to some other issues, namely that of the quantity of data now available.

The importance of patient information, genetic data and the biological behaviour or pathway information of the disease (or proteins, amino acids etc.) cannot be ignored nor each treated in isolation (Sawcer, 2010). There still exists much debate on these issues especially in regard to the limited knowledge existing on some (or all) of the cellular processes under consideration, and the relative weights and values of integrating all the data into a single model (Bakir-Gungor & Sezerman, 2011). Existing models focus on strong interactions and associations, leaving much to be found and where it is possible this can be obtained from existing GWAS (Bakir-Gungor & Sezerman, 2011) through the application of data mining and knowledge engineering techniques.

When focusing on multiple sclerosis (MS) in particular, it has been found that after the identification of the DR15 Haplotype on chromosome 6p21 which has been known for over 30 years, it has to date proven difficult to determine other genes adding to susceptibility factors. Researchers are in favour of identifying what common variants of SNPs and their interaction increase risk whilst acknowledging that these variants are likely to have only a modest detectable effect (M. Ban et al., 2009). Several different investigations have been conducted utilising the data collected by the Welcome Trust Case-Control Consortium (WTCCC), which have identified a number of genes that affect the susceptibility and diagnosis of and note that some difficulty in determining the affects sufficiently accurately of rare SNPs and that there is a likelihood of interactions factors being significant (M. Ban et al., 2009; The Welcome Trust Case Control Consortium & The Australo-Anglo-American Spondytitis Consortium, 2007). As the effects of each SNP tested is low, relatively speaking, it makes the application of many of the established statistical methods for modelling at the least inefficient if not outright ineffective. Adding to this the complication that the risk of many genetic diseases, and here MS is included, is very low in proportion to the entire

population, and as existing studies have shown the gender and ethnicity of the individuals influences the likelihood of and level of risk (M. Ban et al., 2009; Sawcer, 2010; The Welcome Trust Case Control Consortium & The Australo-Anglo-American Spondytitis Consortium, 2007).

2.2 Software Perspective

Software has been devised specifically with medical information processing in mind and these often include aspects of machine learning and data analysis found in other systems targeted at different applications. One of the challenges in medicine is dynamic decision making. To assist with this the DynaMOL language was developed. It has been used to develop a data driven dynamic decision analysis system (Cao, Leong, Leong, & Seow, 1998), where the authors highlight the importance of domain expert input into the devising of analysis protocols, others have used existing development environments and still others do not indicate why the specific type of environment or systems have been used it they are noted at all.

Where software packages used have been identified they are listed in the table below. A ✓ in the right most column indicates that the software was developed by the authors. The code listed references table 2-3 at the end of the chapter where a full list of source references can be found. It is followed by a brief snapshot of each of the software systems developed by authors listed in table 2-1.

Software system or package	Code	Developed
MediFAQ	7	
DynaMOL	7	
WEKA	10	
MATLAB	12, 40	
Haploview 3.2	12	
Biomind ArrayGenius	17	✓
Biomind GeneGenius	17	✓
BiomindDB	17	✓
PLINK	18, 19, 38, 39	
R	18, 22, 64, 67	
SNPAnalyser2.0	19	✓
BEAGLE	19	
DynaMOL	20	✓
PAM	3	
Java	34	
DWAS-GMDR	34	✓
SNPy	36	✓
PostgreSQL	36	
Python	36	
GBOOST	39	✓
RS-SNP	40	✓
Stata	67	

Table 2 - 4: Software systems used and developed

DynaMOL is a dynamic decision support system writing language. The authors describe a methodology that supports multiple perspective reasoning and incremental language extension. When comparing this to other decision support systems it is noted that in dynamic decision problems there is an explicit reference to time. Their use of incremental language extensions provides a framework that allows descriptions to be passed to different aspects of the system via translators, allowing the problems addressed to be gradually expanded and all information already existing to be available to each aspect of the framework at one time. Graphical representations are also provided for the different perspectives of the framework.

SNPAnalysier2.0 is web based and focuses primarily on association and linkage disequilibrium analysis. There are four models available; additive model for use with allelic or haplotype association, and genotypic or diplotype association can be analysed using co-dominant model, dominant model or recessive model. Data can be exported in a tab delimited format. This system is free to researchers along with existing datasets (including all data sets

represented in the article). Dataset size is limited only by the RAM available on the executing computer. The authors were able to complete analysis using genotype data with over 100,000 SNPs and 2, 000 samples.

Biomind products (**ArrayGenius, GeneGenius**) are offered by the company with a web interface. Data can be stored on the user's computer or located on the company's servers. The program runs on the company servers and is accessed via the web interface. Each user is required to obtain an account, which is free but services are not. The company reports successful incorporation many consulting contracts to go along with the use of the software. The systems are designed specifically with biologists in mind so that they may take advantage of analysis processes more commonly found in computer science and artificial intelligence fields without having to learn them themselves. The Genius systems combine advanced machine learning algorithms with volumes of background information including biological ontology's, of which BiomindDB is one, to deliver powerful data analysis functionality. ArrayGenius uses microarray data, but GeneGenius is not limited to microarray data and has been extended to handle SNP data.

DWAS-GMDR the authors note as the first parallel computing software developed to perform generalised multifactor dimensionality reduction when determining gene-gene interactions as part of a generalised linear model of covariates to determine classification. It was written in Java and to address the issues of scale involved in GWAS analysis they implemented a more effective memory handling algorithm and increased the efficiency of the GMDR calculations. There is mechanism for reporting multiple candidate SNP combinations with similar gene-gene interactions utilising a weighted version of cross-validation consistency based selection criteria. Four different performance measures to evaluate classifiers have been included giving the user some degree of choice, along with three methods for handling missing genotype information.

SNPy is a relational database developed in PostgreSQL and uses a Python coded interface. It was developed to address issues of large dataset size, data validation, and the ability to rapidly update necessary information. The system manages both SNP data and patient

information in a consolidated manner, drawing on an architectured database approach to manage scalability and increase robustness of data handling. The authors draw on the abilities of a database to perform low level data validation and to detect corruption in the data, especially as they are utilising patient data which is notoriously error prone and mutable. SNP input comes from both the Affymetrix and Illumina platforms and can both be accommodated simultaneously within the system. SNPy is designed to manage data in terms of storage and validation as well as export it in a format useful for analysis by other third party tools, for example PLINK, and as such does not do any direct analysis itself. The inclusion of some analysis directly in the database, which should be faster than in a procedural language environment alone, is indicated by the authors as a potential future development noting that to do this would make the calculations quicker to execute and less computationally intensive yet SNPy itself needs a server to work from so the computational expense is not so great. The SNPy code files are available for download under a GNU General Public License.

GBOOST is the implementation of the BOOST system using the increased processing capacity available with a graphical processing unit (GPU). BOOST was originally implemented on CPU, but utilising a GPUs capacity, speed and increased memory availability the calculation times are vastly reduced. The authors describe a reduction from days to hours for the calculation of pairwise interaction amongst SNPs. Apart from offering an increased speed of calculation, which is always welcome when dealing with very large data sizes, this implementation does not offer much more than the original BOOST software.

RS-SNP implements a random set method for the analysis of GWAS. This method was adopted to address some of the limitations of previously used analysis methods in GWAS where no consideration was possible of the situation where susceptibility to a disease is conferred by a large number of loci each with a small individual effect but with a larger accumulative effect. This is in contrast to the previous GWAS which focused on detecting a small group of SNPs each with large individual effect, which by its nature requires large sample sizes with strong associations to detect an effect. Each SNP is placed in a set respective of its genetic location in relation to the nearest gene, if two are overlapping it is placed in both and if it is over a certain distance from the nearest gene it is discarded. Individual association tests for each SNP are used to computer an enriched score for each set

under five different genetic models. The enrichment score is then analysed to determine final significance. The system was developed in MATLAB and is freely available.

2.3 Computational Methods Perspective

Focusing on the computation and processing involved in attribute selection, model creation and testing has identified a "largish" list of methods and applications. These approaches are listed in the table below in no particular order. The right most column shows a key relating to the method or approach that was used if this was identifiable, and the left most column a key to the sources used (numbers match those of the articles listed in Table 2-3 towards the end of this chapter which lists the source references).

For purpose:

C = classification, A= attribute selection, F = used within measures of fitness or accuracy

Purpose			Method	Code
	C		Selection of new sample based on only a starting set of desired samples	4
	C		Inductive learning by logic minimisation	6
	C		Decision tree (most often C4.5)	6, 10, 12, 20, 21
	C		*k*-NN	6, 10, 9
	C		CN2 rule learning algorithm	6
		F	Relative information score	6
		F	Compression measures	6
			Semi-Markov and/or Markov decision processes	7, 20, 1, 38
		F	Interestingness	8
	C		Fisher linear discriminant analysis	10, 9
	C		Multilayer perceptron	10, 12
	C		Support vector machine	10, 12, 17, 9, 15, 21
	C		Boosting	10
	C		Self-organising maps	10
	C		Hierarchical clustering, or clustering in general	10, 3, 21
	C		Graph theoretic approaches	10
A	C	F	Combining: genetic algorithms, correlation-based heuristics, data mining algorithms (Decision trees, support vector machines, bagging, stacking) and data partitioning	10, 21
A			Automatic relevance determination	12
A			Backwards elimination	12
A			Logistic regression, including regularised logistic regression	12, 9, 1, 67, 21, 32
A	C		Genetic Algorithms, Evolutionary Algorithms (applied to classification,	13, 10

			clustering, data pre-processing)	
		F	Iterative fitness function with wrappers & filters	13
			Pareto optimal solutions	14
	C		Neighbourhood knowledge-based evolutionary algorithm	14
A	C		Artificial neural network	16, 21
	C		Genetic programming	17
A	C		Statistical methods	17, 18, 19, 12, 3, 15, 64, 66
	C		Random forests	18, 38
	C		Gradient boosting machine	18
A			Hardy-Weinberg Equilibrium test	19, 12, 64, 65, 66, 67
A			Bonferroni correction	19, 64, 66
	C		Influence diagrams	20
	C	F	Dynamic decision analysis	20
	C	F	Decision theoretic planning	20
	C	F	Planning in AI	20
A			Principle Component Analysis	3
	C		Relevance-networks	3
			Meta-heuristic algorithms	3
A			Wrappers and filters	9
			Layer ranking algorithms	9
	C		Fuzzy set theory	21
	C		Bayesian Network	21
	C		Sequential covering approach	25
	C		Learning classifier systems	26, 27
A			FCBF filter	32
	C		Grid search	32
A			Multifactor dimensionality reduction	34, 38
	C		Bayesian epitasis association mapping	38
	C		Weighted sum statistics – Sibpair-based and odds ratio	37

Table 2 - 5: Computational methods used to analyse data

Prior knowledge of the application domain is not always available, and if available it is not always in a usable format; yet it can be invaluable to the evaluation of models and pattern identification (Fayyad et al., 1996; Goertzel, Pennachin, Coelho, Shikida, & Queiroz, 2007). Yet knowledge of the data type and meaning drives the determination of what methods are applicable, what has been done and what is missing in the way solutions are derived (Wu et al., 2003). To date statistical methods and latterly data mining has been applied to model development. Other methods developed in a more computationally centred environment are now being applied to biological data analysis, for instance the use of evolutionary algorithms for attribute selection, classification and clustering tasks allows a different perspective to be

taken (Freitas, 2003). It has been traditional for biologists to be trained in statistical analysis but not in machine learning or data mining methods. This gives rise to the production of lower functioning models than can be produced using machine learning, whether or not the data is quantitative or qualitative in nature (Goertzel et al., 2007). This fact has given rise to the development of software designed specifically with biologists in mind, for example the Bioperl extension modules to the Perl programming language. There is also strong development in terms of methods of determining or imputing missing values in genetic data (Marchini, Howie, Myers, McVean, & Donnelly, 2007).

It is not sufficient to have advanced algorithms alone, but they must support whole data mining and discovery process, which is iterative in nature and growingly interactive (Wu et al., 2003). Models must reflect an appropriate level of abstraction for the desired purpose of analysis, and remain meaningful to a human user who will ultimately be interpreting and implementing applicable actions based on model output. The structured and iterative nature of model analysis and decision making involves "a large number of decision factors, relations and numerical parameters that change over time" (Cao et al., 1998), resulting in the need for good supporting data and documentation so that informed alterations can be made when needed, for example patient data which is notorious for being out of date and/or missing but has the potential to add to the understanding of test results and implications for analysis and decision support. How each type and form of data available (either directly or imputed), its usage depends greatly on the implementation of the end result, that is the purpose for which the system of modelling it to be used (Marchini et al., 2007) – decision support, disease progression, quality of life monitoring etc. For the "blind application of methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity leading to the discovery of meaningless patters" (Fayyad et al., 1996) and have an opposing effect to the one desired and thus mislead further research. The application of mapping of the problem space is often an invaluable tool in guiding development of modelling, data handing, data validation, and the approach taken (Bull & Kovacs, 2005).

Methods are emerging to address large datasets (Wu et al., 2003), but the definition of "large" is very much application dependent. Some analysis has been done on as few as 6 attributes (in this case SNPs) (Tomida et al., 2002), but the raw data comes in tens of thousands of

attributes when using biochips, especially microarray data. As datasets get larger the algorithms used to analyse them need to be able to scale up appropriately and remain computationally efficient (Fayyad et al., 1996). Larger datasets especially in the medical and biological domains, impact on current methods of analysis and reflect "our ability to analyse and understand massive datasets lags far behind our ability to gather and store the data" (Fayyad et al., 1996). The more data that is present the higher likelihood of noise or erroneous values appearing. Correct handling of noise, be it errors or outliers need to be appropriately applied as required for the application domain (Gamberger, Lavrac, & Dzeroski, 2000), along with sensible in intelligible handling of missing data (Fayyad et al., 1996; Marchini et al., 2007). It can be noted that "medical datasets represent real-world data usually containing substantial amounts of noise" (Gamberger et al., 2000). Coping appropriately with the high dimensionality of much of the data involved in biological testing, especially genetic information, leads to improvements in handling the reduction of the data in such a way as to reduce the dimensionality yet retain as much valid information as possible (Kwon et al., 2011). Added to this is the challenge of making all this processing fast enough to be done in as short a time as possible along with a reduction in the need for server size processors and computer memory. The use of the processing power of graphics chips to enhance the capability of CPU processing has added much to this arena (Kwon et al., 2011). The issues that high dimensionality creates for analysis is not limited to the sheer size of the dataset. Unfaithfulness can occur where correlation measurements can be distorted, if not exactly erroneous, due to the size of the dataset. Random forests is one technique that has been applied to address specifically the issues of unfaithfulness in datasets with high dimensionality (Yang et al., 2011).

"The problem of knowledge extraction from large databases involves many steps, ranging from data manipulation and retrieval to fundamental mathematical and statistical inference, search, and reasoning" (Fayyad et al., 1996). Each of these steps needs to be balanced against each other to enable the overall goal of the analysis or modelling to be realised. The ability to select the most informative components of dataset, to reduce noise, confusion and complexity, and increase prediction/classification accuracy (Shah & Kusack, 2007) also makes the execution of model faster and taken globally is known as pre-processing. Which techniques to use when pre-processing data is very much dependent on the type of data present, the goal to be achieved (data cleaning, attribute reduction, imputation of missing data

etc.) and the modelling techniques to be used with the resulting dataset (Kharbat, Odeh, & Bull, 2008). When using an iterative approach to attribute selection, it allows for the use of simpler selection methods especially when the attributes are split up into smaller groups randomly each time (Shah & Kusack, 2007). Keeping a larger overall set of attributes than would be done with other methods reduces the chances of erroneously removing significant attributes and attribute interactions. The interaction between attributes when investigating SNP data is becoming increasingly possible with improved techniques for reduction and modelling and is revealing previously hidden relationships between parts of the genome (Bull & Kovacs, 2005; Kwon et al., 2011). Methods of accomplishing this reduction in the overall number of attributes are varied, with and accompanying variability in outcome. One method, Genetic Algorithms, can cope with larger numbers of attributes and perform global searches producing a more advantageous outcome (Freitas, 2003). Other models also employ the use of biological testing, for example linkage disequilibrium, as part of their attribute selection processes (Arshadi, Chang, & Kustra, 2009). Moving to multiple objective problems adds its' own complexity and associated challenges. Again evolutionary algorithms have been successfully used, including the development of one that utilises neighbourhood knowledge between candidate solutions to improve solution quality (Yu, Wong, Wang, & Wei, 2010).

The judging of models and classification algorithms should be done using appropriate measures of accuracy and meaning. One such measure is interestingness – "an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity" (Fayyad et al., 1996), another and more commonly used is overall classification accuracy. This reflects the accuracy of prediction/classification using a simple measure of counting the number of times the model obtains the correct result when tested with data of a known classification. Although accuracy is easier to calculate interestingness can give a more "holistic" or "well rounded" view of the value of the results of the classification methods particularly where prediction is involved in a medical domain (Sumathi & Sivanandam, 2006; Wagholikar, Sundararajan, & Deshpande, 2011). Of note when judging the merits of particular methods is their ability to scale up when dataset sizes increase or do they fail due to over simplicity; is there a lack of proof of their correctness or mathematical rigor; what computational costs are involved especially where large datasets are involved; and is there sufficient explanation of how a conclusion is reached (Wagholikar et al., 2011). The use of evolutionary algorithms,

for example, allows for the application of different quality criteria simultaneously to evaluate a candidate solution, balancing model requirements (Fayyad et al., 1996).

As previously noted reducing the datasets to a meaningful and manageable size along with the corresponding methods of classification or modelling applied often benefits from human interaction within the process. This may not always be the case, as one investigation found when addressing rule induction methods that, "the automatically-discovered rule induction algorithms can avoid some of the human preconceptions and biases embedded in manually-designed rule induction algorithms, possibly leading to more effective algorithms in challenging application domains" (Pappa & Freitas, 2008). Balancing the need for greater or lesser amounts of training data to produce an adequately performing classification model is not an easy task. To this end Papa & Freitas (2008) have extended traditional genetic programming to a grammar-based genetic programme using a grammar to create a population of candidate solutions. One of the main advantages of this system is that prior knowledge can be incorporated into the grammar and used to guide the genetic programme search patterns.

Mixing quantitative and qualitative data together in the same classification model has its own peculiarities and pitfalls. Predictive classification, in a medical domain this would account for systems involved in diagnosis, often has more quantitative data and therefore lends itself to predictive models that cope well with numeric data. Whilst staying in the medical domain there is much more qualitative data involved in the on-going care and monitoring of a patient's condition, adding extra dimensionality and complexity to the models that lie behind any integrated medical decision support system. One such field is the integration of quality of life measures into the treatment of patients with degenerative diseases, for example Multiple Sclerosis. To date three statistical methods are widely accepted and used when monitoring the quality of life measures taken from standardised tests indicating patient perception of their own condition. In the study performed by (Schwartz et al., 2011) one of these methods, Structural Equation Modelling, was shown to be the more successful yielding clear findings. Other data mining methods involving the quantification of non-numeric data have proven equally as successful, yet there is still much debate on the merits of the various quantification methods. The inclusion of other biological data also forms part of this challenge. To include pathway analysis, of the genes, proteins and other coding regions, has been found to

contribute to the improved mapping of particular diseases. Bakir-Gungor and Sezerman (2011) in their investigations into the combination of genetic and pathway analysis note that at the time of their study only one attempt hand been successful in the use of these combined inputs and that was in the study of Multiple Sclerosis. They do caution that this study was limited to known gene sites only, and that as a whole the use of pathway analysis is limited by existing knowledge of cellular processes. This challenges further work into how existing data and knowledge can be fluidly combined with new results for form an on-going picture of a disease, or general type of condition, for example all autoimmune diseases.

In a number of studies machine learning techniques to extract previously unknown knowledge from existing data have proven very successful, the study by (Bull, 2008) is but one of these. The importance of this is increased when the cost of the original studies are taken into account. Taking GWAS as an example, the biochips used costing roughly US$250 per sample (person) (Cortes & Brown, 2010) and with studies often consisting of 2500 plus people the cost of obtaining new data can be rather prohibitive. Whereas being able to investigate existing data in new ways allows for the possibility of uncovering previously undiscovered associations as well as interactions (Quevedo, Bahamonde, Perez-Enciso, & Luaces, 2012). It is possible that further understanding of SNPs from existing data main aid in determining some of the missing heritability of many genetic diseases. From their study of existing SNP data (Quevedo et al., 2012) found that taking into account more SNPs than the original GWAS enhanced the accuracy of the classification process and resulted in significant SNPs in all chromosomes. They note that traditional GWAS is searching for goodness of fit for association of SNPs individually without taking into account any level of interaction between the SNPs. These findings are also borne out by (Wang et al., 2011) in there assertion that existing techniques leaves GWAS data undermined.

A reoccurring theme in a number of studies mentioned already is that of there being a greater collective power in groups of SNPs than in addressing them individually. Feng, Elston & Zhu (2011) extend this in their implementation of weighted sum methods allowing the techniques to be used without specifying a threshold for defining the limits between rare and common variants that would otherwise been previously needed. The ability to use existing techniques, with enhancements, to enable the lowering of prior knowledge of what is in effect the

determination of a classification barrier, allows the system to determine which value is best to use in any one specific situation increasing the degree of flexibility and discovery of previously unknown effects.

As an example of how various requirements of model accuracy, and different techniques can be combined one study (Calcagno et al., 2010) investigating the treatment of MS patients of Caucasian population of southern Italy employed the use of a haplotype, a set of statistically associated SNPs. "It is thought that these associations, and the identification of a few alleles of a haplotype block, can unambiguously identify all other polymorphic sites in its region" (Calcagno et al., 2010). The authors employ a multilayer perceptron of two layers of adaptive weights with full connectivity. They use a logistic regression model fitted using interactive reweighted least squares algorithm to set weights and intercepts. Attribute selection is done by automatic relevance determination and backwards elimination. For model fitting cross-entropy error measures have been used. The model and associated processing was implemented in MATLAB suing the Netlab Toolbox. This is one study that warrants reflection on which genes and SNPs are found to be significant in relation to the comparison of disease identification and treatment uptake, especially as they identified a previously unknown association between two genes.

2.4 Conclusion

Adaptation is the key to success and the incorporation of a number of techniques in a single implementation is a challenge faced by researchers. The volume and dimensionality of datasets being addressed is adding its own complexity to the overall process. An attempt to incorporate different forms of data, especially that which is highly prone to errors either of entry or omission, adds a layer of processing into the pre-processing of the dataset before it can be analysed that is not always present in data mining exercises. The use of databases and data warehouses has the power to alleviate at least some of the complexity form the modelling side of the process. Whatever measures are used for data pre-processing, storage, dimensionality reduction, and modelling the value of all this excellent work can be undone by a lack of attention to the way in which each stage and the system as a whole is assessed for completeness, accuracy, robustness and transparency of decision making. In addition to the development of well-rounded modelling and data handling is the need to address the

computational expenses in terms of data shortage, retrieval time, memory requirements and processor speeds. Irrespective of the perspective from which any solution is viewed it needs to be applicable and executable in a "real world" setting to be of benefit to the maximum audience.

Code	Source	Code	Source
1	(Zhao & Leong, 2000)	**32**	(Quevedo et al., 2012)
3	(Kim, 2002)	**33**	(Wang et al., 2011)
4	(Wu et al., 2003)	**34**	(Kwon et al., 2011)
6	(Gamberger et al., 2000)	**36**	(Mitha et al., 2011)
7	(Cao et al., 1998)	**37**	(Feng et al., 2011)
8	(Fayyad et al., 1996)	**38**	(Yang et al., 2011)
9	(Chen, 2007)	**39**	(Yung, Yang, Wan, & Yu, 2011)
10	(Shah & Kusack, 2007)	**40**	(D'Addabbo et al., 2011)
12	(Calcagno et al., 2010)	**64**	(The Welcome Trust Case Control Consortium & The Australo-Anglo-American Spondytitis Consortium, 2007)
13	(Freitas, 2003)		
14	(Yu et al., 2010)		
15	(H.-J. Ban et al., 2010)		
16	(Tomida et al., 2002)	**65**	(The Wellcome Trust Case Control Consortium, 2010)
17	(Goertzel et al., 2007)		
18	(Arshadi et al., 2009)	**66**	(M. Ban et al., 2009)
19	(Yoo, Lee, Kim, Rha, & Kim, 2008)	**67**	(Todd et al., 2007)
20	(Leong, 1998)		
21	(Wagholikar et al., 2011)		
22	(Li & Schwartz, 2011)		
25	(Pappa & Freitas, 2008)		
26	(Bull & Kovacs, 2005)		
27	(Kharbat et al., 2008)		

Table 2 - 6: Source references for methods appearing in table 2-1 and table 2-2

Chapter 3 Methodology

There are a number of methods that can be used to investigate SNP data as has been outlined in Chapter 2. Not every method is applicable in every situation and where parameter variation is possible method optimisation becomes a critical and non-trivial step. For example, a model developed for one population may not be suited to another as ethnicity is a known factor in the susceptibility to genetic diseases. With the on-going development in microarray chip technology the number of SNPs measureable in a single test is increasing dramatically which in turn puts increased pressure on computational and data manipulation tools. The challenge is to utilise efficient data reduction methods in conjunction with analysis and modelling methods within the limitations of computational cost.

3.1 Research Approach Taken

As outlined in Chapter 2 many different methods of investigation have been applied to microarray and gene expression data. No in-depth study has been made of the comparative usefulness of each method, and no study to date has approached the data from non-biological stand point. The value of addressing biological data from a non-biological standpoint is brought out by Shah & Kusack (2007), where they note that it is advantages not the be "blinded" by what is already known but to let the data explain itself. This leaves two shortcomings in the overall research into the analysis of SNP data which form the on-going focus of this study. Firstly, which methods work well with what data and how is this assessment made; secondly, especially with the increase in number of SNPs measured, how can data mining techniques assist, alongside the biological research, in determining which ones play a part in determining susceptibility to which genetic diseases.

In this study several options for each area of interest have been applied in a mainly explorative manner. This is based around a constructivist approach to the research and enables the investigation of a number of concepts simultaneously with the results giving indicative direction for further in-depth algorithmic and methodological development.

3.2 Research Design

In the constructivist approach to research an artefact is designed, tested, modified and retested repeatedly until a final solution is achieved. In many avenues of research this artefact has been a prototype of a mechanism (for example mechanical or electrical), some software system or other such object. In this study the artefact is a list. A list of possible ways in which SNP data can be moulded and modelled to give a method by which new data can be analysed in a medical application which often involves processing the data for a single person or sample at a time. To judge the value of different methods it is necessary to follow and evaluate each stage in the process of manipulating the original dataset into a format and size that is appropriate for the modelling methods, and then to evaluate the modelling itself.

Taking as an initial standpoint of no prior knowledge is advantages, but does not allow for some of the existing insights into SNP analysis to be applied. So this study has been based around a starting point of little prior knowledge and does not base its decisions on a strong biological understanding. The data is investigated initially to gain some level of understanding of the "what" and "how" of the data fields present and their overall perspective/positions in the original dataset. From there reduction methods are applied and later compared against existing lists of SNPs found to be significant from previous biologically based testing. This reference point is used only after the various reduction methods are applied. A range of modelling techniques are then applied to the reduced datasets and evaluated using the overall and individual class accuracies. The end result is a prioritised list of reduction methods and modelling systems that can be combined and optimised into a complete system.

The desire is to form an optimised system that could take relatively raw inputs, mould the data into what is needed and then apply a classification model to a single sample input representing the information for a single individual. This goal is acknowledge as being ambitious and has the highest probability of not being completed within the time limitations of this study. Having noted this all attempts are made to achieve it.

This study utilises the data from a previous investigation into a number of diseases by the WTCCC, and uses only the data relating to Multiple Sclerosis. Corresponding forms relating to the appropriate use of this data and ethical uses have been signed by all parties involved.

3.2.1 Data exploration

With such a large original dataset, commonly used statistical and graphical methods for summarising data are not completely applicable. To investigate the dataset several forms of summary were used to determine an overall impression of the population. The data was split by region, gender, genotype and SNP score value. A ranking and grouping method was employed to determine whether the SNP score value had a scalable property, i.e. did a high score mean more than a low score for any one SNP.

Other relationships were investigated to answer the following questions. Is there any relationship between the region the sufferer lives in and the number of sufferers? What chromosomes are tested and how much? Does this have a bearing on what has been learnt from existing research and will later analysis conform this? Is there anything else that can be learnt about the data from visual inspection?

3.2.2 Modelling method comparison

In order to improve on existing modelling methods it is necessary to understand how each method functions and how any limitations can be minimised. Modelling methods can be grouped into three categories: global, local, and personal (Kasabov, 2006). It has been suggested that the personalised modelling approach leads to a higher and more robust solution. It investigate this a number of models are tested against each other with some variation applied as deemed appropriate to the purposes of comparison. A full comparison of all possible variations is beyond the scope of this study.

As each implementation has different accuracy statistics available, to enable the comparison of all methods the percentage accuracy of each method overall, and for each class has been used.

Where later studies into the WTCCC dataset have been carried out, the researchers have taken the association results from the original study as their starting point for modelling. It has been argued (D'Addabbo et al., 2011) that this creates possible shortcomings in the results as SNPs previously regarded as unimportant many have significant effects within other methods. To account for this no attribute selection was predetermined from previous studies.

3.2.2.1 Attribute reduction

To reduce the original dataset to a manageable size for modelling and to remove any SNPs that do not significantly contribute to the modelling process attribute selection is necessary. It is known that it is most likely that a large number of SNPs each having a small significant effect will be the outcome of attribute selection rather than a small group each having a large significant effect (D'Addabbo et al., 2011). Many attribute selection methods are centred on the determination of small groups of significant attributes with large effects on outcome, in this case classification of sample, and so a very much lower threshold for acceptance is applied.

A total of five reduced datasets were produced using various methods. Samples were selected from those who came from within the second and third most populated regions for the case class (diseased subjects) and a random selection of 200 from the control class. This gave a total of 424 samples. The most populous region was not used as it has a disproportionate representation of genotype frequency in relation to the other regions which was identified in the data exploration phase. When testing the top ranking methods a random selection of 50 samples each from case and control were used with no regard to region or gender.

The number of SNPs was reduced by several different methods. After initial quality control procedures were applied (refer to section 3.3 Data collection later in this chapter) the total number of SNPs per sample was reduced to 12,374 from the original 15,436. As this is still too large to be accommodated by existing analysis tools further reductions were made in a modular fashion following two main approaches.

The first was to rank the SNPs on their overall average score for the diseased population and to divide this ranked list into groups. Each group comprised 60 SNPs, with the exception of one group that contained the “remainder”, a grouping of 14 as the total was not exactly divisible be 60. From each group 5 SNPs were taken, the SNP with the average score closest

to the calculated group mean, and 4 others randomly selected from the remainder of the group. The smaller group is represented by 3 rather than 5 SNPs chosen in the same manner. This resulted in a list of 1032 SNPs. This list was separated into two streams each of which underwent further selection approaches. On one stream, which yielded what was to be named DS-1, only a correlation calculation was made with the SNPs having the highest correlation coefficients and p-values less than or equal to 0.005 being selected. The second stream underwent selection using a voting system amongst a group of 5 different attribute selection methods implemented in WEKA; correlation, best first search, generic search, linear forward selection and greedy stepwise selection. This resulted in 541 SNPs being selected becoming what is then referred to as DS-2.

The second approach was to address the whole of the data for the two regions from the original dataset. Two basic approaches were used, firstly to rank attributes solely on correlation coefficients & p-values, and secondly to rank attributes on the differential between overall SNP average score for case and control samples. As these two techniques ranked all SNPs within the original dataset, a cut-off point of 200 was used to limit what was to become DS-3 and DS-4 respectively.

All 4 reductions were then further reduced using principle component analysis (PCA), where attributes are ranked according to their "ability" to account for variation in the selected output variable. WEKA was used to perform this analysis as it was able to transform the resulting ranked Eigen values into a ranked list of attributes in their original space. Two options for termination of PCA are implemented in WEKA, one enabling the designation of a number of variables (e.g. the top 20), the other specifying what percentage of variation is to be explained before the test is terminated. The default setting of a 95% of variation explained was used allowing for the different number of SNPs in each of the four reductions.

All four reductions were further reduced to the levels indicated by the PCA results. A combination of all four sets was formed and also reduced by PCA, again using the same 95% variability explained setting to create DS-5.

3.2.2.2 Model selection

Models are grouped into global, local and personal combinations. To compare several different types of method in each group three implementation environments were used. Firstly, WEKA which is an internationally accepted data mining/knowledge engineering software package; secondly, NeuCom which was developed by researchers at AUT working within KEDRI again for data mining/knowledge engineering tasks; and lastly MATLAB, which is a programming environment popular with engineers and others who need to perform detailed numerical calculations and is used here as the algorithms tested are those versions implemented by researchers at KEDRI.

In total 11 different methods have been investigated, including variants where possible resulting in a total of 544 individual tests performed on each of the five reduced datasets.

The tests covered a range of approaches and methods. A full list of tests and the corresponding variations applied is listed below in table 3-1. Where no variation is set the system default values for each implementation are used.

Classification method		Variation		Environment
Rules	OneR			WEKA
	Decision Table	Best first		WEKA
Trees	J48	Confidence Factor = 0.5 Confidence Factor = 0.25 Confidence Factor = 0.1		WEKA
Bayesian	Naive Bayes			WEKA
	Bayesian Logistic regression			WEKA
Nearest Neighbour	*k*NN	No distance weighting	K=1 K=3 K=5 K=7 K=9 K=11 K=15 K=20 K=30 K=50	WEKA

		Weight by 1/distance	K=1 K=3 K=5 K=7 K=9 K=11 K=15 K=20 K=30 K=50		WEKA
		Weight by 1-distance	K=1 K=3 K=5 K=7 K=9 K=11 K=15 K=20 K=30 K=50		WEKA
	W*k*NN	Threshold=0.01 Threshold=0.05 Threshold=0.1 Threshold=0.25 Threshold=0.35 Threshold=0.40 Threshold=0.45 Threshold=0.50 Threshold=0.55 Threshold=0.60 Threshold=0.65 Threshold=0.75 Threshold=0.90 Threshold=0.95 Threshold=0.99	K=1 K=3 K=5 K=7 K=9 K=11 K=15 K=20 K=30 K=50		MATLAB
	WW*k*NN	Features=100% Features=95% Features=90% Features=75% Features=50%	Threshold=0.35 Threshold=0.40 Threshold=0.45 Threshold=0.50 Threshold=0.55 Threshold=0.60 Threshold=0.65	K=1 K=3 K=5 K=7 K=9 K=11 K=15 K=20 K=30 K=50	MATLAB
Support Vector Machine	Polynomial				WEKA NeuCom
	Linear				WEKA NeuCom

	Radial Base Function	WEKA NeuCom
Multilayer Perceptron		WEKA NeuCom
Multiple Linear Regression		NeuCom
Evolving Classification Function		NeuCom
Evolving Clustering Method for Classification		NeuCom

Table 3 - 3: Classification methods and variations tested

3.2.2.3 Description of models

Rules and trees are similar in that they both test for a set of conditions and result in a conclusion, often a classification. The distinction is that rules are evaluated for a particular set of conditions and when new conditions arise new rules must be formed; where trees evaluate their conditions in a progressive manner allowing branching dependent on some subset of conditions allowing for more flexibility in accommodating new conditions. Rules and trees can be robust in execution, but also tend to replicate tests which make execution times longer and more computationally expensive.

The basic approach in determining rules is to divide the instances that can be classified under a single condition, remove these from further consideration and repeat the process until no instances remain. Two rule based classification systems were tested; One R and a Decision Table.

One R generates rules based on a single attribute at a time, and then moves on to the next attribute until all have been processed. It is based on the theory that a single attribute can determine class quite accurately in many real world datasets. A very simple method that is quick to execute.

Decision Table rule discovery applies a best first search and is thus less likely to get stuck in local maximum than other methods. It produces a table of attribute values and classifications that can be "looked up" to determine new instance.

This table represents the rules behind determining the classification without displaying the rules themselves.

A single tree method was tested, J48 which is the implementation in WEKA of the C4.5 algorithm (revision 8). This is the last public release of this algorithm before the commercial version C5 appeared.

J48 applies a top down divide and conquer approach in a recursive fashion. The root node of the tree is selected based on the greatest possible split between instances that then appear on each side of the condition based on attribute value. The instances that appear in each side or branch of the root node are then split in a similar fashion until a classification is determined for all instances. A confidence factor is applied to each branching condition to give a statistical measure of validity for the inclusion of the instance within which branch. This has been varied to assist in the determination of robustness of this method.

Bayes theorem is a statistical tool based on probability theory to determine the probability of a future event given some, or many, known probabilities for events that have already occurred. Bayes theorem states that the probability of event H (or class classification) given evidence E

$$Pr[H|E] = \frac{Pr[E|H]Pr[H]}{Pr[E]}$$

This approach uses all training instances to build a model and only needs to pass through the data once, making it a useful tool when datasets are large. Two methods were tested that employ this Bayesian approach: Naive Bayes and Bayesian Logistic Regression.

Naive Bayes constructs a model based on a normal distribution and assumes that all the variables (attributes) are independent. These two assumptions may not hold well or at all but with datasets large enough they allow a reasonable approximation to be made which yield relatively good results dealing well with unknown or missing values.

Bayesian Logistic Regression employs a Bayesian approach to a learning binomial logistic regression function, which is suitable for learning logistic regression models for high-dimensional problems.

The **Nearest Neighbour** method is what is known as an instance based learner in that it computes the classification of each new instance at a time. To do this it compares the new instance attribute values to those of existing instances and determines the classification of the new instance based on the shortest distance to its nearest neighbour. To determine which neighbour is the closest two parameters are used; *k* the number of neighbours to compare against and how the distance between them is to be measured, known as the distance metric. The distance metric commonly used, especially with numeric attributes, is the Euclidian measure of distance (Abdleazeem & El-Sherif, 2008; E Alpaydin & C Kaynak, 1998; E. Alpaydin & C. Kaynak, 1998; Duin, van Breukelen, Tax, & den Hartog, 1998). The Euclidean distance is computed by taking the difference of the values for each attribute by subtracting them, giving rise to a measure of the distance between each instance.

*k*NN is the implementation of the basic nearest neighbour method described above. Variations applied to the testing were to alter the weighting given to the distance metric, and the value of *k*.

Weighted *k*NN uses the weighted average of k nearest neighbours to mitigate noise in the dataset and smooth the impact of any one instance. Each instance requires the formation of a local model that describes that instance that can be done without reference to any existing instances. The weighting is termed threshold in the implementation of this method tested.

Weighted-Weighted *k*NN

Neural networks have been successfully applied to many classification problems. They can process both numeric and nominal data and are not hindered by dependencies among attributes.

Multi-Layer Perceptron is a neural network that utilises back propagation to incorporate findings from previous instances to adjust the model employed to classify the next instance. This method can be very accurate, but execution time increases greatly with increases in dataset size.

Linear models are very popular in statistics as they are relatively simple to calculate, but real world data does not always fall into neat linearly separable classifications. A blend of linear modelling with instance based learning is employed by a classification algorithm known as Support Vector Machine to translate data from the original problem space where classification boundaries are non-linear into a different space where a linear boundary can be applied. To do this SVMs take a selection of critical boundary instances, known as support vectors, from each class and build a linear discriminant function that separates the classes as widely as possible. How this done is dependent upon the kernel applied within this method.

SVM kernels tested were linear, polynomial and the radial based function.

Multiple Linear Regression is a statistical method that applies a least squares fit to multivariate data for each class in the dataset.

Clustering is a method of grouping data such a way that the variation within each group is minimised and the distance between each group is maximised. Two evolving methods that employ clustering have been tested.

Evolving Classification Function

Evolving Clustering Method for Classification

3.2.2.4 Model classification

Each of the models tested can be grouped according to Kasabov's (2006) definitions of global, local and personalised. The SVM method although designated a global method in other research here it is classified according to the kernel employed.

The distribution of methods reflects the development work that has previously gone into model development and a growing shift away from a dependence upon global models.

Type	Method
Global	
	OneR Decision Table
	J48
	Naive Bayes Bayesian Logistic Regression
	SVM – Linear SVM – Polynomial
	Multiple Linear Regression
	Multi-Layer Perceptron
Local	
	Evolving Classification Function
	Evolving Clustering Method for Classification
	SVM – Radial Based Function
Personalised	
	k-Nearest Neighbour
	Weighted *k*-Nearest Neighbour
	Weighted-Weighted *k*-Nearest Neighbour

Table 3 - 4: Classification methods by type

3.2.2.5 Cross validation applied

Cross validation was applied to each test in one of two forms dependent upon which was available within the implementation environment. Where available a *k*-fold cross validation was applied to each method, using a *k* value of 10. Where this was not possible leave-one-out cross validation was applied. Both WEKA and NeuCom have implemented the k-fold cross validation approach, but this was not available for the models tested using their MATLAB codes. With the use of additional code the leave-one-out cross validation method could be successfully utilised instead.

In the *k*-fold cross validation approach the training data is "sliced" or folded into *k* units in either a random or sequential fashion. Random was selected here to minimise any effects of the order of the classes in the dataset. A model representing the data is built using whichever method is being tested using *k*-1 of the folds and tested against the remaining instances in the last fold. This is repeated so that all *k* folds have been used as the test fold, and the results from each evaluated and combined to form the reported model and accuracy results.

Leave-one-out is similar in that it slices the data into training and testing divisions, with the exception that here the testing division contains only a single instance. This method was not used in this study as it was desired to test the model with against a larger number of samples.

3.2.3 Implementation environment

It was necessary to utilise three implementation environments to test all of the desired methods.

WEKA http://www.cs.waikato.ac.nz/ml/index.html

The Waikato Environment for Knowledge Analysis, or WEKA, was developed by the machine learning group at Waikato University in New Zealand and enjoys worldwide success. Weka can be described as "a collection of machine learning algorithms for data mining tasks" (WEKA). It is implemented in the Java language and is distributed under the GNU General Public License for open source software. Released in this form a number of other developers worldwide have contributed their own algorithms for inclusion and development of WEKA, which is in direct reference to the desire to make WEKA a good environment for developing further machine learning schemes.

NeuCom www.theneucom.com

Developed at the Knowledge Engineering and Discovery Research Institute at Auckland University of Technology, NeuCom is a software environment for neuro-computing, data mining and intelligent system design. It is based on the theory of Evolving Connectionist Systems (ECOS) as published in the book titled "Evolving connectionist systems: methods and applications in bioinformatics, brain study and intelligent machines"(Kasabov, 2007). "NeuCom learns from data, thus evolving new connectionist modules. The modules can adapt to new incoming data in an on-line incremental, life-long learning mode, and can extract meaningful rules that would help people discover new knowledge in their respective fields" (NeuCom). NeuCom is freely available to education and researchers and is currently being used in 20 universities around the world and in research lavatories.

MATLAB http://www.mathworks.com/products/matlab/

Developed by MathWorks, MATLAB® is a high-level programming language and development environment. It is frequently used in the areas of numerical computation, visualisation and engineering programming, by those both in industry and academia. The range of applications for which MATLAB has been used include signal processing and communications, image and video processing, control systems, test and measurement, computational finance, and computational biology (MATLAB).

3.3 Data Collection

The data used in this investigation was collected by the Welcome Trust Case Control Consortium (WTCCC) in an investigation into a number of genetic diseases with their findings published in (The Welcome Trust Case Control Consortium & The Australo-Anglo-American Spondytitis Consortium, 2007; The Wellcome Trust Case Control Consortium, 2010).

3.3.1 Data acquisition

The tested population is white European in origin residing in the United Kingdom. This population pool is useful in relation to the higher proportion of instances of Multiple Sclerosis in those with this ethnic association as compared to other ethnic groupings. Each participating person contributed a blood sample which was tested using one of two microarray chips. The diseases being investigated were divided into two sections, each being tested on a different chip, one testing just over 15,000 SNPs the other just over 500,000 SNPs. Control samples were tested on both chips.

The Multiple Sclerosis group was tested on the chip yielding 15,436 SNPs, which is a labelling number as the chip is designated 15k, but in actually yields 15,463 per sample. Geographically the samples are spread over 12 regions, see figure 3-1.

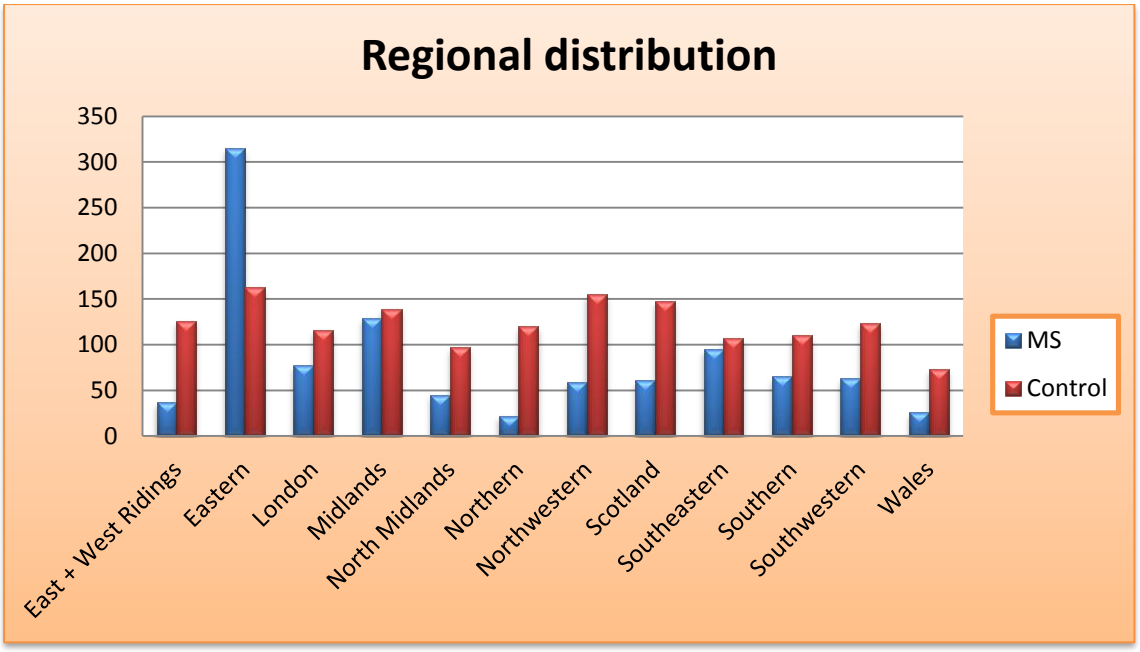


Figure 3 - 3: Regional distribution of original samples

3.3.2 Quality control

Two phases of analysis in the original study yielded two sets of exclusions from the final association study. Phase 1 focused on individuals and exclusions were made for individuals who were found to have been putatively related, those with questionable ancestry (outside the UK-European origins), and those with more than 10% missing genotypes. Of these the missing genotypes contributed to the most exclusions, see figure 3-2.

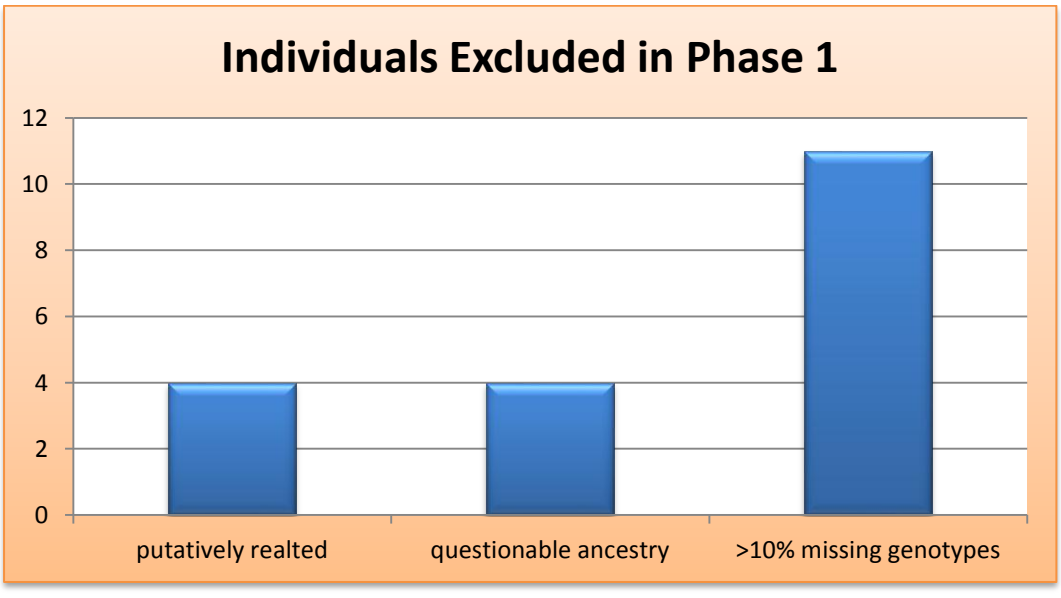


Figure 3 - 4: Reasons for exclusion of samples

In phase 2 the focus was on the SNPs and their accuracy and reliability of measurement. The software used to analyse the genotype data, GenCall, assigns a quality score to each locus and a confidence score for each individual genotype. This is the "Score" listed in the raw data. Initially samples found to have more than 50% of loci score below 0.7 were removed along with those with a quality score below 0.2. Two additional filtering criteria were applied where individual genotypes with a confidence score of less than 0.15 and any further SNP having more than 20% of its samples with confidence scores below 0.15. These criteria were used with a view to optimising the genotype accuracy while still minimising any uncalled genotypes(The Welcome Trust Case Control Consortium & The Australo-Anglo-American Spondytitis Consortium, 2007).

There are 2470 samples (persons) passing the first two phases of quality control procedures that form the case (MS diseased) and control (non-disease) instances in what is here referred to as the original dataset. There is a spread of 12 regions with a total of 1017 males and 1453 females. The total SNPs per sample was reduced to 12,374 from the original 15,436 recorded by the chip.

.

3.3.3 Manipulation of raw data

The raw data was downloaded from the WTCCC in compressed and encrypted format totalling 31.1GB. Once processed, the raw data consisted of several files of plain text representing the original investigation data. To pre-process this into a format compatible with analysis procedures and tools the data was imported into a database for the sample SNP data, and a spreadsheet for the chip definition, sample specifications, and regional locations data.

From the database, data was exported in several formats as each analysis tool used needed different formats and alignments of data and to cope with the restrictions of file size within each tool. For some aspects of data extraction up to four separate data processing stages were needed, some of which were required to be performed manually which was very time consuming. This is a factor in much of the research in bioinformatics field and is by no means limited to this study (Mitha et al., 2011; Witten, Frank, & Hall, 2011).

The database itself is very large and was needed to be spit up into smaller units and dynamically linked to enable processing to occur without overloading the available computer resources. It was not possible to store the SNP data in a transposed form suitable for input onto analysis tools directly in the database as this would exceed the system capabilities of the database application used.

3.3.4 Ethics

To use this data a release and ethical approval and data access agreement is signed by all researchers before the data is released by the Welcome Trust Case-Control Consortium.

3.4 Data Analysis

As the original dataset is very large it is necessary to find a way of reducing the number of attributes so that existing systems of analysis can be used. In order to do this ways were sought to reduce the data without relying on existing biological knowledge in an endeavour not to eliminate any data points that may not have already been associated with the disease but play a part in determining the degree of susceptibility.

Part of the difficulties of handling data of the size of the original dataset is that it is difficult to store the data in a format accessible to the software used to analyse it. The full dataset was stored in a database in the same field alignment that was originally supplied, which is not compatible with software like WEKA. To transform the data into a form and format that other software, even statistical analysis packages, can utilise is a lengthy process. To date much of this has been done semi-manually with the aid of intermediate formats.

3.4.1 Attribute selection

As noted in section 3.2.2.1 Attribute reduction, no one method of selecting the “best” set of attributes has been applied. This to enable a comparison of the various methods of attribute selection and their feasibility in dealing with the sheer volume of attributes being produced by modern microarray chips. To date the simpler measurements have proven very

advantageous, reliable and accurate in lowering the number of attributes to use as an initial input to more robust methods of determining attribute selection.

Issues present when dealing with SNPs data give rise to results that do not follow the behaviour of other types of data. This is especially so in that it is not uncommon to find that many SNPs having low individual relative significance gives the best results in comparison to a smaller number of more highly significant SNPs. The "rule of thumb" for many real world datasets is that a small group of attributes can adequately describe the whole population (Witten et al., 2011). This simply does not hold true for SNPs data. The effects SNP interaction have not been addressed in this study in terms of attribute selection or model comparison as this is outside its scope.

3.4.2 Model comparison

Based around the definition global, local, and personalised methods of modelling the data given by Kasabov (2006) several different modelling methods were selected for comparison. This focused around the desire to determine which types of methods were best for further investigation into disease susceptibility detection.

There are numerous ways of determining which test is the "best" method, but these are not always implemented in the same way in different packages. Some, like WEKA, are more defined in terms of what analysis statistics they provide and how they are implemented. Others, for example the methods developed at KEDRI, have very little limited only to measures of class and overall accuracy. Some of these difficulties results directly from the stage of public release that each method and implementation has currently reached. As WEKA is a freely available program and has been used in a number of published studies (Kharbat et al., 2008; Shah & Kusack, 2007) has much more comparative statistics available to the researcher. As WEKA is an open-source development environment many developers from around the world have now contributed to its development and the implementation of many of its analysis systems.

Noting this difficulty, it was determined that overall model accuracy along with the accuracy of each class, case and control, be used for model comparison as these measurements were available in all implementation environments.

Chapter 4 Results on personalised modelling for MS susceptibility prediction based on SNPs data

These are several steps in the analysis of any data. What the data is and what it relates to, how it is represented and how voluminous it is, all contribute to how the data is to be managed. This is all before any analysis can take place. In the case of SNP data the sheer volume of it causes its own issues along with decisions on how to reduce it to both a manageable size and one that contains relevant data to the question of interest. A different question would require different reductions of the raw data. In this case we address the issues of determination of susceptibility, and to do so we attempt to build a model that best captures insights from samples where the diagnosis is known. This is defined as a classification problem.

There are several steps required in determining a classification model. They are to reduce the number of attributes to those of significance, to select a classification method that is appropriate and to optimise this method for the best accuracy possible. Robustness will also need to be included in the decisions of optimisation and classification method. Full optimisation was not possible under the current study as the attribute reduction and classification method selection processes were of greater importance at this stage in the analysis of the data taking into account time limitations. Without a good attribute list and fairly robust classification method (or methods) to start from, no real optimisation is possible.

The results presented in this chapter follows the same outline as standard data mining processes; firstly the raw data is addressed and pre-processing performed before attribute reduction is done. Secondly, several different classification methods are applied to the reduced datasets and their accuracy results presented.

4.1 The data

As stated in chapter 3 the original data for this investigation has been obtained from the Welcome Trust Case-Control Consortium (WTCCC). It represents data obtained from a study in to 12 genetic diseases, and corresponds here to only one specific disease – Multiple Sclerosis.

4.1.1 The original data

The original data downloaded form WTCCC is 31.1 GB in size and is both compressed and encrypted. Encryption codes are released upon request to preapproved persons who have previously obtained permission to use the data and have signed the appropriate confidentiality ethical agreement forms.

The data is separated into case (diseased) and control (non-diseased) sections. Each section has been analysed by the same chip, the Infinium 15K genotyping chip manufactured by Illumina[1], and as such the file descriptions for the detailed content of the original data set are the same for both case and control.

For each section, case and control, the data is comprised of a collection of 7 files. These contain data on SNPs, the samples, intensity measurements from the determination of genotype, and information on the chip used. In addition there were two files describing each phase of the quality control procedures applied and which samples or SNPs are to be removed as the original downloaded data was supplied at the point before these procedures were applied in the original study.

In this "raw" form the SNP data was contained in a plain text file with tab delimited rows of data. Each row pertains to a single SNP for a particular sample. There are 15, 436 SNP entries, lines, per sample. An example of the data is as follows.

SNP	SAMPLE	GENOTYPE	SCORE
rs1234567	WTCCC14145	GC	0.9243
rs3445568	WTCCC14565	CC	0.8657
rs5678569	WTCCC24545	CA	0.7823

Each SNP has an identifying code which is a shortened form of any name that may have been assigned either by the scientific community or the chip manufacturer. Each sample is coded for identification purposes so that anonymity of the participants can be maintained. The genotype is represented by the genetic coding for the two corresponding alleles. The score value reflects a combination of measures of intensity gained in determining the genotype along with other factors and has been normalised to represent values between zero and one.

[1] Illumina is a leading developer, manufacturer, and marketer of life science tools and integrated systems for large-scale analysis of genetic variation and function. International headquarters in San Diego, California. www.illimina.com

There were 13 genotypes listed with NN denoting that the testing could not determine the genotype of the SNP in a specific sample. Both case and control samples have NN genotypes, but these are noticeably small in relation to the overall distribution of genotypes. This can be seen in figure 4-1. After the two quality control phases (see later in this chapter) are applied the number of NN genotypes found is vastly reduced.

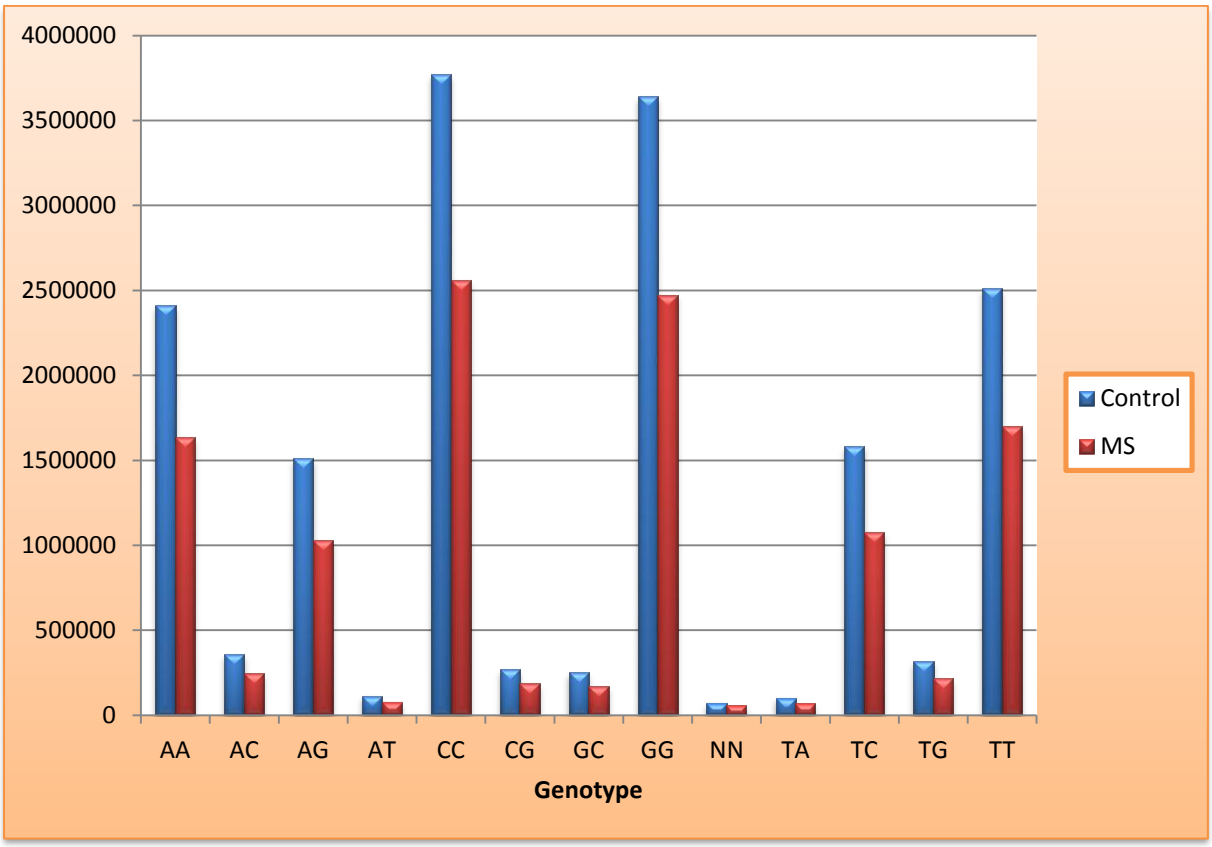


Figure 4 - 13: Genotype within original WTCCC dataset

The sample file contained information on each sample, or person, giving their gender as either 1 for males or 2 for females, the cohort to which they belonged as MS for case and 58C for control, the supplier of the source sample (often the code for a hospital), the geographic region from which the sample originates, in addition to the plate and well number of the sample. Age of onset for the diseases studied by the WTCCC is included in the sample files where this information is available; unfortunately it was not available for the MS data group.

An example of sample information data is as follows.

Sample	Gender	Cohort	Supplier	plate_well	Region
WTCCC92727	1	MS	ADD_NU	12835a1	East + West Ridings
WTCCC92937	1	MS	ADD_NU	12835a10	Eastern
WTCCC92956	1	MS	ADD_NU	12835a11	Midlands
WTCCC92974	1	MS	ADD_NU	12835a12	Midlands
WTCCC92739	2	MS	ADD_NU	12835a2	Eastern
WTCCC92760	2	MS	ADD_NU	12835a3	Midlands
WTCCC92788	2	MS	ADD_NU	12835a4	Northwestern
WTCCC92806	1	MS	ADD_NU	12835a5	Southeastern
WTCCC92838	2	MS	ADD_NU	12835a6	Eastern
WTCCC92862	1	MS	ADD_NU	12835a7	Southern
WTCCC66261	1	58C	1958OC	11025a2	East + West Ridings
WTCCC66262	1	58C	1958OC	11025a3	Northern
WTCCC66061	1	58C	1958OC	11025a5	Scotland
WTCCC66062	1	58C	1958OC	11025a6	Midlands
WTCCC66273	1	58C	1958OC	11025b2	North Midlands
WTCCC66274	1	58C	1958OC	11025b3	Scotland
WTCCC66073	1	58C	1958OC	11025b5	Scotland
WTCCC66074	1	58C	1958OC	11025b6	North Midlands
WTCCC66285	1	58C	1958OC	11025c2	Eastern
WTCCC66286	1	58C	1958OC	11025c3	Midlands
WTCCC66085	1	58C	1958OC	11025c5	Northern

It is likely to be clear to a resident of the UK where regional boundaries lie, but with no regional definition given it is impossible to determine exactly. Figure 4-2 shows the distribution of samples amongst the regions, and has had no smoothing applied to limit any distortion caused by the differences in original numbers of samples taken in each case and control groups. The distribution of both case and control samples along with gender identification is listed in table 4-1.

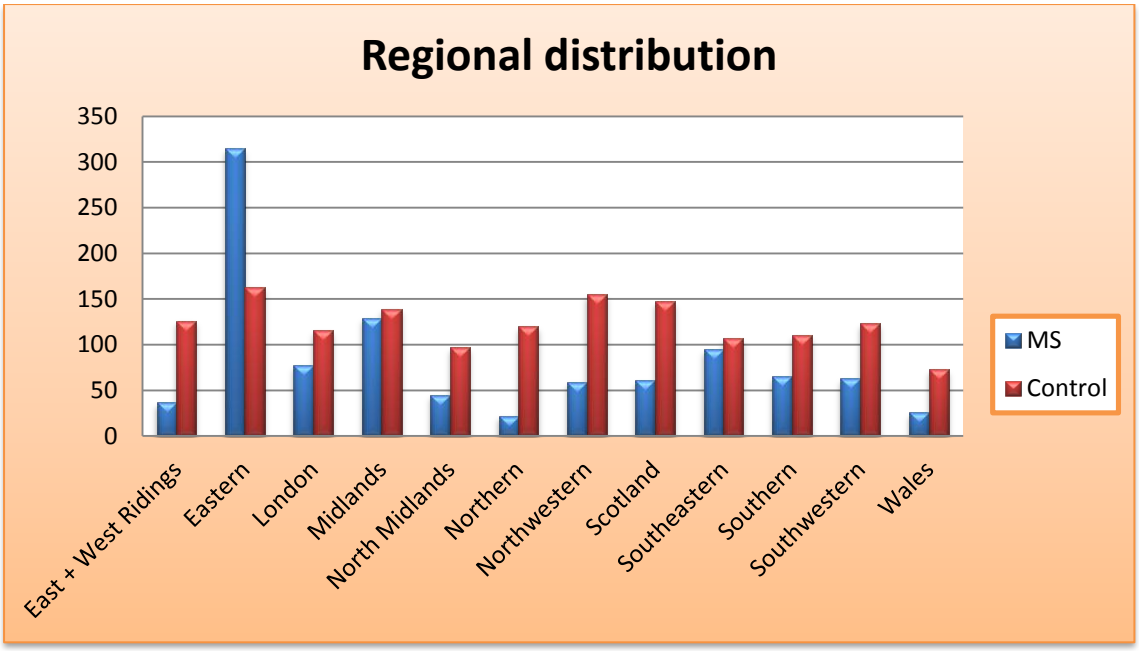


Figure 4 - 14: Regional breakdown of WTCCC dataset

Region	Case (MS)			Control		
	Total	Male	Female	Total	Male	Female
East + West Ridings	37	11	26	126	59	67
Eastern	315	96	219	163	82	81
London	77	16	61	116	56	60
Midlands	129	34	95	139	71	68
North Midlands	45	15	30	97	55	42
Northern	22	8	14	120	62	58
Northwestern	59	17	42	155	77	78
Scotland	61	20	41	147	75	72
Southeastern	95	24	71	107	55	52
Southern	65	18	47	110	54	56
Southwestern	63	12	51	123	60	63
Wales	26	7	19	73	33	40
Totals	994	278	716	1476	739	737

Table 4 - 19: Regional breakdown on samples in original WTCCC study

4.1.2 Processing the data

Data was imported into a database from the SNP files and into spreadsheets from the files containing smaller amounts of information. From a combination of data in the exclusions files both samples (phase 1) and SNP (phase 2) removals were made in accordance with the appropriate quality control procedures which were indicated as necessary by the WTCCC before further analysis can take place.

Phase 1 focused on individuals and exclusions were made for individuals who were found to have been putatively related, those with questionable ancestry (outside the UK-European origins) and those with more than 10% missing genotypes. Of these the missing genotypes contributed to the most exclusions, see figure 4-3.

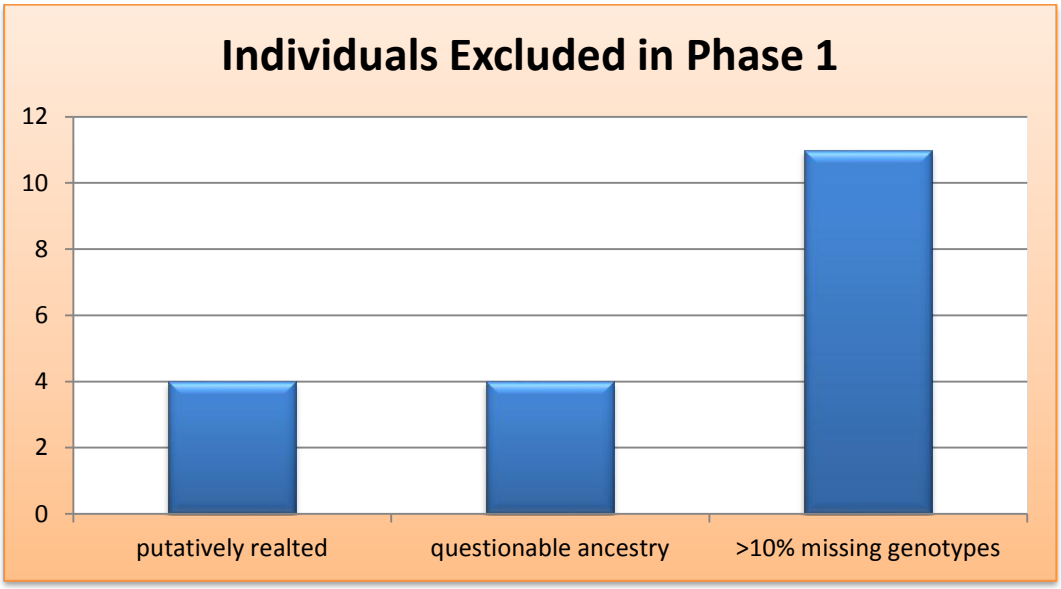


Figure 4 - 15: Samples excluded after phase 1 analysis

In phase 2 the focus was on the SNPs and their accuracy and reliability of measurement. The software used to analyse the genotype data, GenCall, assigns a quality score to each locus and a confidence score for each individual genotype. Initially samples found to have more than 50% of loci score below 0.7 were removed along with those with a quality score below 0.2. Two additional filtering criteria were applied where individual genotypes with a confidence score of less than 0.15 and any further SNP having more than 20% of its samples with confidence scores below 0.15. These criteria were used with a view to optimising the genotype accuracy while still minimising any uncalled genotypes(The Welcome Trust Case Control Consortium & The Australo-Anglo-American Spondytitis Consortium, 2007).

4.1.3 Production of reduced datasets

After the initial application of the quality control exclusions there remained a total of 12,374 SNPs per sample, 975 case samples and 1476 control samples.

As this represents a dataset that is still very large it was necessary to find a way of reducing the number of attributes (SNPs) so that existing systems of analysis can be used. In order to

do this ways were sought to reduce the data without relying on existing biological knowledge in an endeavour not to eliminate any data points that may not have already been associated with the disease but play a part in determining degree of susceptibility.

Part of the difficulties of handling data of the size of the original dataset is that it is difficult to store the data in a format accessible to all other software needed. The full dataset was placed in a database in the same format that was originally supplied, which is not compatible with software like WEKA. To transform the data into a form and format that other software, even statistical analysis packages, can utilise is a lengthy process. To date much of this has been done semi-manually with the aid of intermediate formats, which is very costly in terms of time. For example, in the production of DS-2 after the data was extracted from the database it took approximately 6 hours to transform the data into the format required for further analysis.

4.1.3.1 Methods of reduction

Two starting points were employed when addressing the data in its original size.

The first was to rank the SNPs on their overall average score for the diseased population and to divide this ranked list into groups. Each group comprised 60 SNPs, with the exception of one group that contained the "remainder" a grouping of 14 as the total was not exactly divisible be 60. From each group 5 SNPs were taken as representing that group, the SNP with the average score closest to the calculated group mean, and 4 others randomly selected from the remainder of the group. The smaller group is represented by 3 rather than 5 SNPs chosen in the same manner. This resulted in a list of 1032 SNPs. This list was separated into two streams each of which underwent further selection approaches. On one stream, only a correlation calculation was made with the SNPs having the highest correlation coefficients and p-values less than or equal to 0.005 being selected which became DS-1. The second stream underwent selection using a voting system amongst a group of 5 different attribute selection methods implemented in WEKA. This resulted in 541 SNPs being selected which became DS-2. This approach was used to determine whether the score value has any scalable factors.

Does a larger score mean more than a lower one? To assist in answering this question the SNPs were ranked on their score value using the mean of the SNPs for all samples. This ranking was used to group the data for both case and control classes and is represented in figure 4-4. Although means differed amongst case and control classes the variance displayed within each class is very similar.

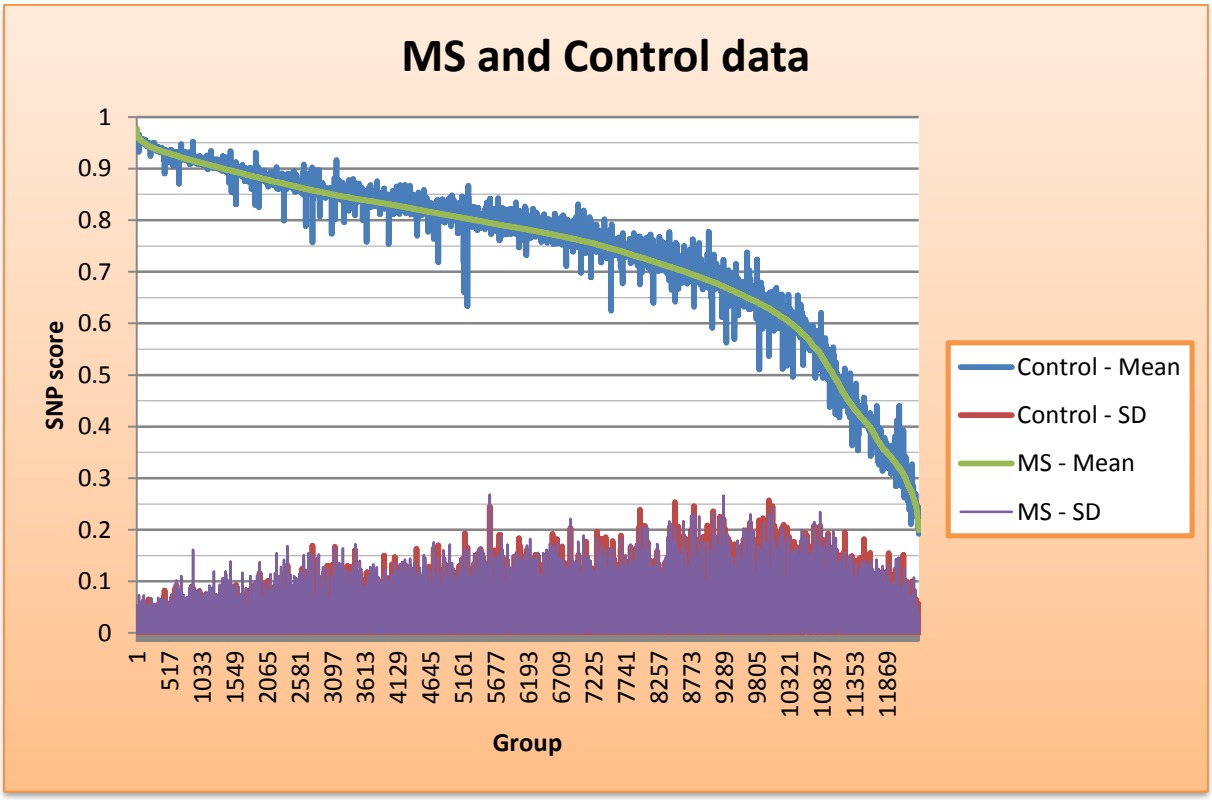


Figure 4 - 16: Comparison of SNP score mean and standard deviation between case and control samples

The second approach was to address the original dataset as a whole. Two basic approaches were used, firstly to rank attributes solely on correlation coefficients & p-values, and secondly to rank attributes on the differential between overall SNP average score for diseased and control samples. As these two techniques ranked all SNPs within the original dataset, a cut-off point of 200 was used to limit what was to become DS-3 and DS-4 respectively.

All 4 reductions were then further reduced using principle component analysis (PCA), where attributes are ranked according to their "ability" to account for variation in the selected output variable. WEKA was used to perform this analysis as it was able to transform the resulting

ranked Eigen values into a ranked list of attributes in their original space. Two options for termination of PCA are implemented in WEKA, one enabling the designation of a number of variables (e.g. the top 20), the other specifying what percentage of variation is to be explained before the test is terminated. The default setting of a 95% of variation explained was used allowing for the different number of SNPs in each of the four reductions.

Attributes	Before PCA applied	After PCA applied	Reduced by
DS-1	141	76	46.10%
DS-2	541	150	72.27%
DS-3	510	93	18.24%
DS-4	300	105	65.00%
DS-5	380	125	67.11%

Table 4 - 20: Size of reduced datasets before and after the application of PCA

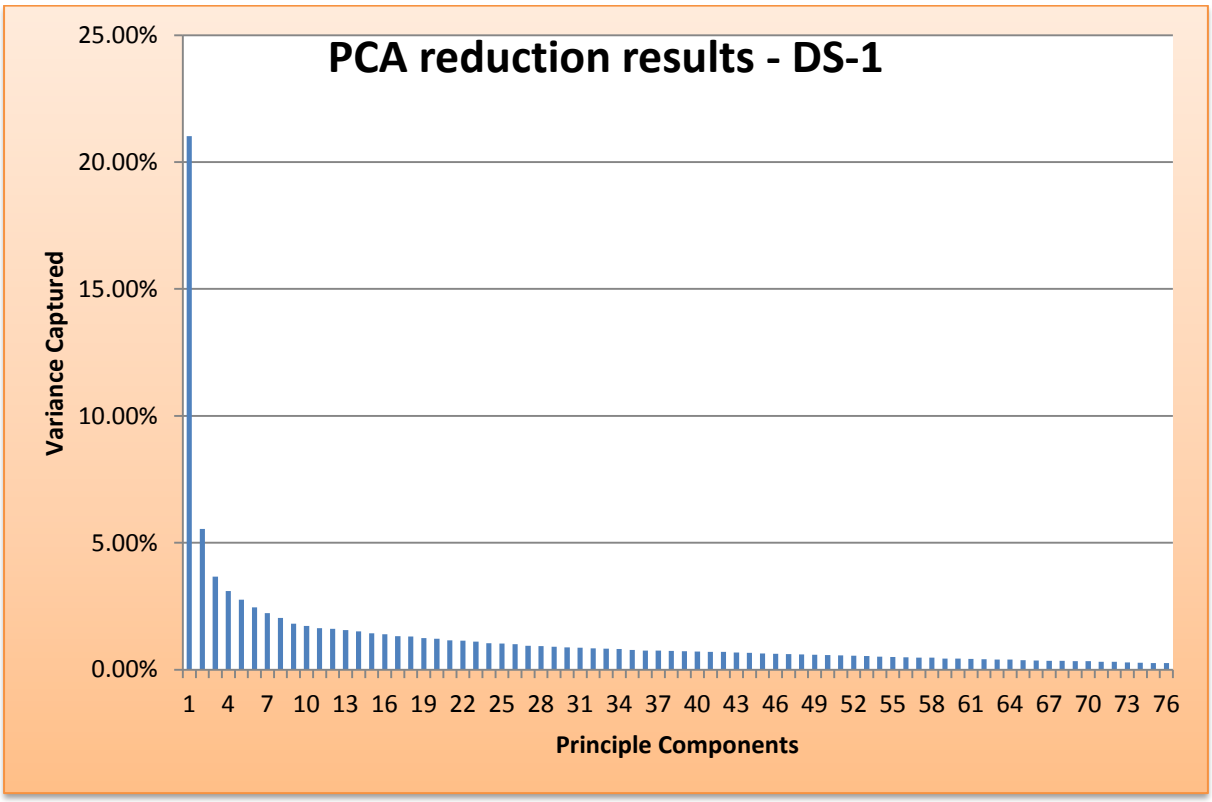


Figure 4 - 17: Principle Component Analysis results for DS-1

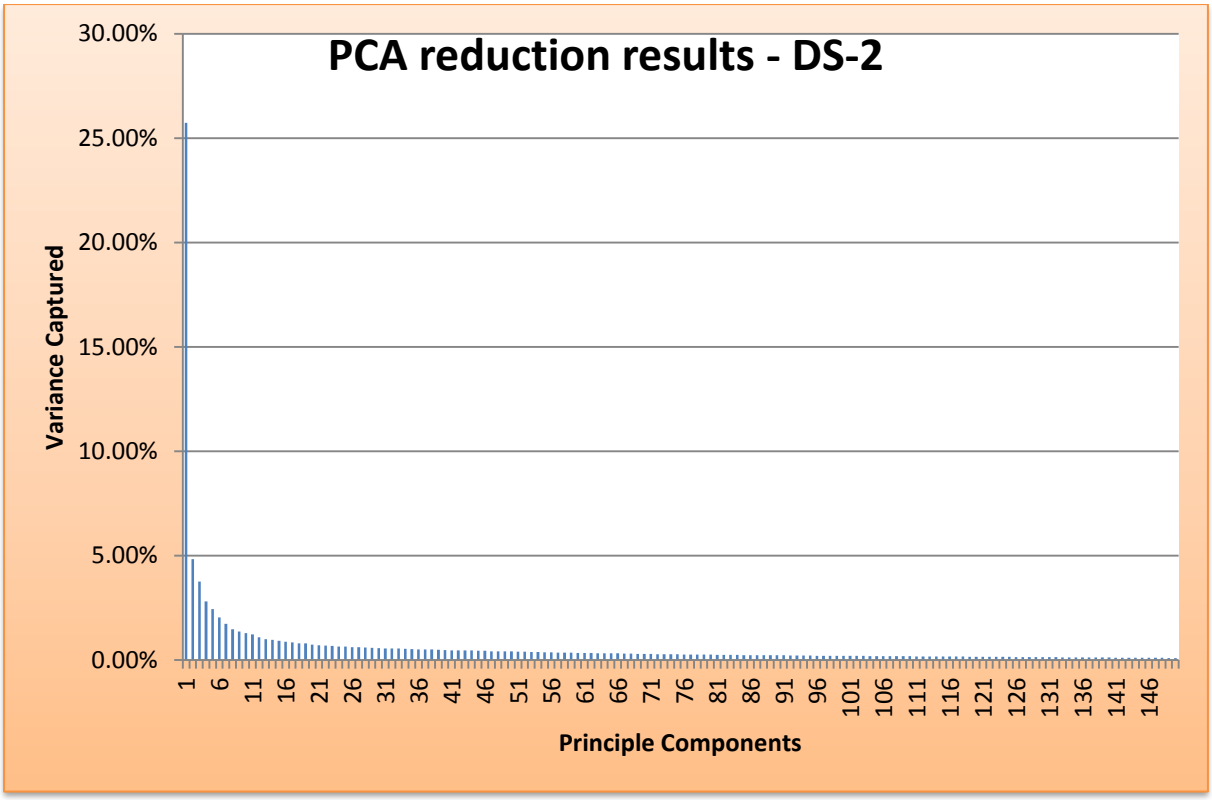


Figure 4 - 18: Principle Component Analysis results for DS-2

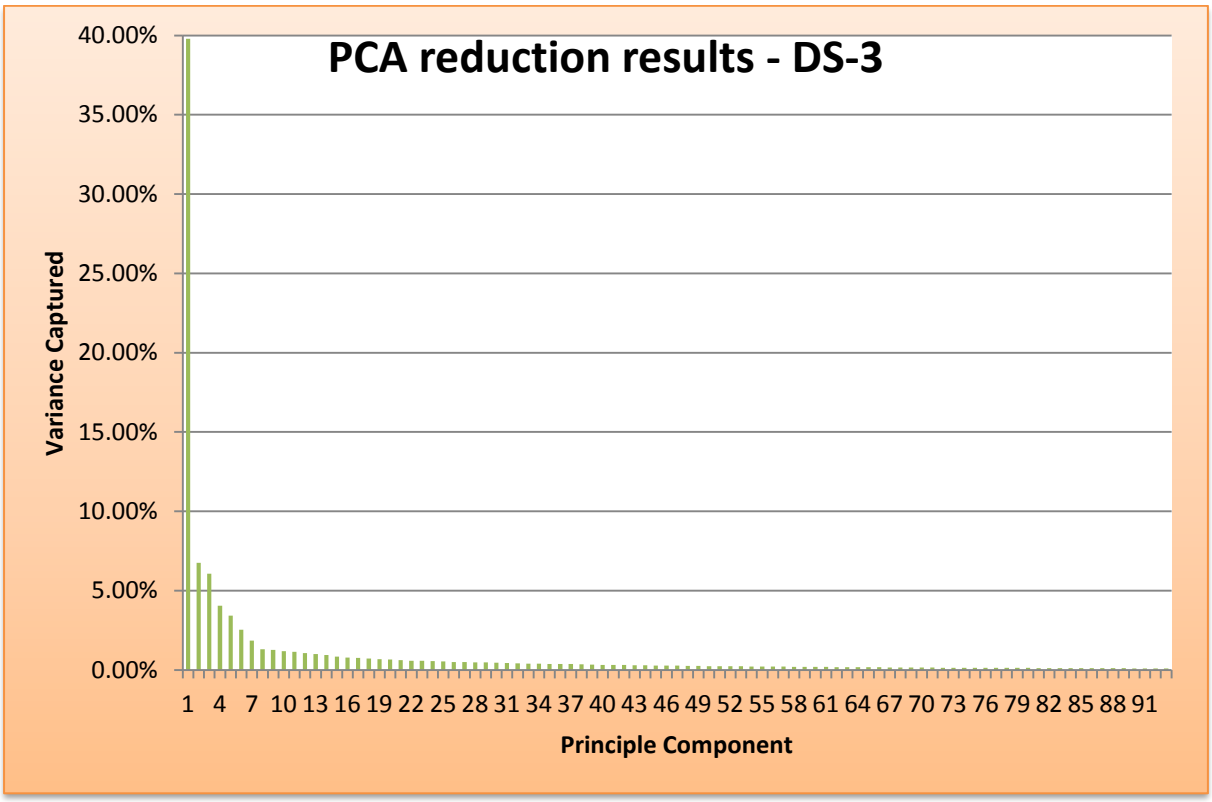


Figure 4 - 19: Principle Component Analysis results for DS-3

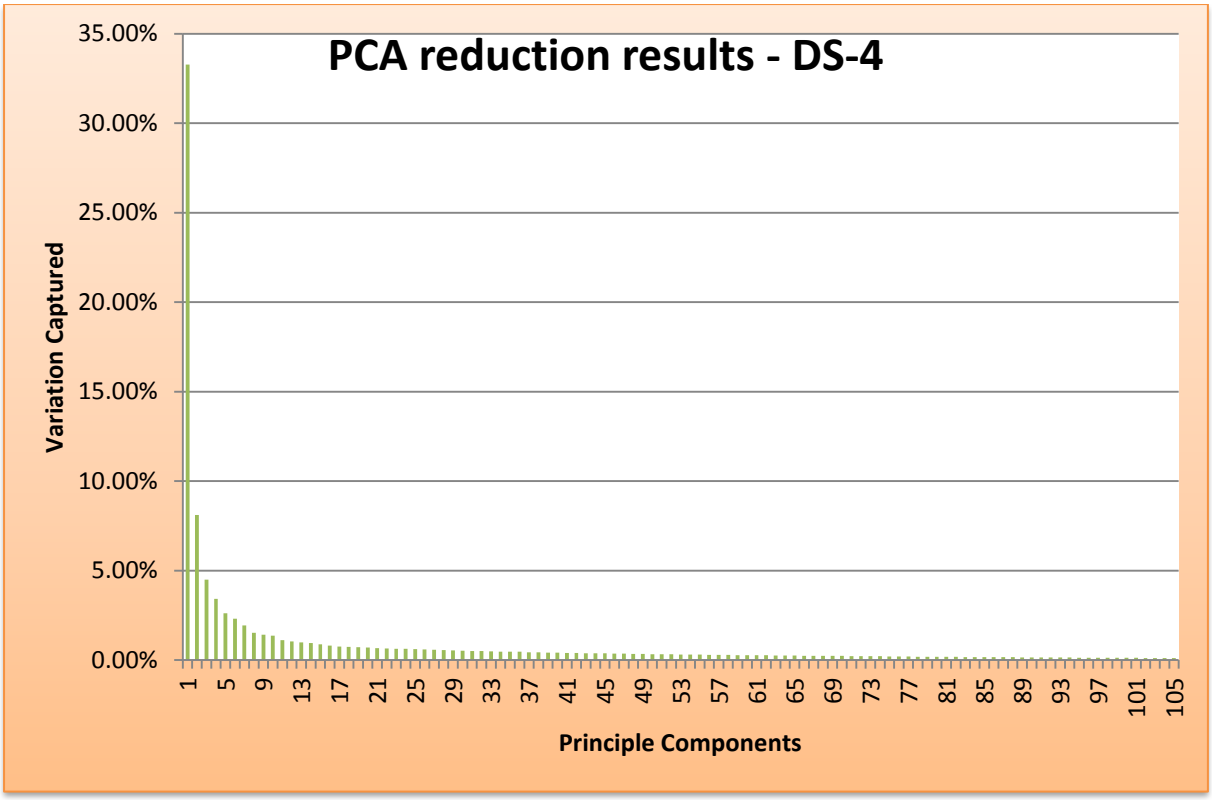


Figure 4 - 20: Principle Component Analysis results for DS-4

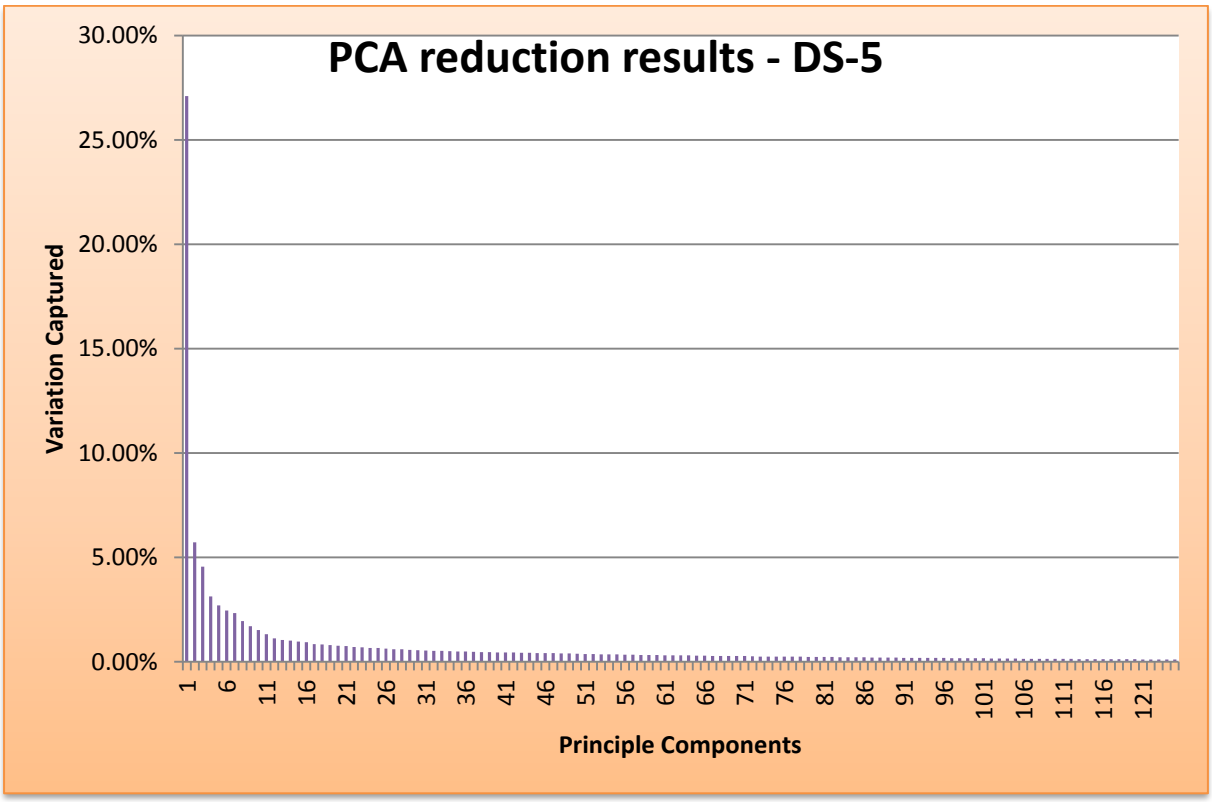


Figure 4 - 21: Principle Component Analysis results for DS-5

All four reductions were further reduced to the levels indicated by the PCA results. A combination of all four sets was reduced by PCA again using the same 95% variability explained setting to create DS-5.

These five reductions, hereafter referred to themselves as datasets, are used in a comparison of methods of classification to determine how each method and attribute selection methods "reacted" together in terms of performance. Of interest was that on viewing the genotype mix present in these datasets there was a large reduction in NN values found, with some datasets even reporting none present.

4.1.3.2 Comparison & overlap

Utilising these attribute selection methods it was possible for a single SNP to be present in more than one dataset. Table 4-3 shows the number of SNPs in each set and where any overlaps exist. From this table it can be seen that 16 SNPs that are in DS-1 also appear in DS2; that these is only a single SNP is DS-3 and none in DS-4 that appear in DS-1; and that 24 SNPs that appear in DS-1 are also found in DS-5.

	DS-1	DS-2	DS-3	DS-4	DS-5
DS-1	76	16	1	0	24
DS-2		150	3	4	77
DS-3			93	19	14
DS-4				105	25

Table 4 - 21: Number of SNPs in crossover between datasets

It is to be expected that DS-1 and DS-2 shared a number of overlaps as they were drawn from the same original sampling. This is also to be expected from DS-3 and DS-4. It is interesting to note where overlaps existed between these two separate streams of dataset reductions.

DS-5 as a compilation set is included in an attempt to limit the effects of any one attribute selection method over another. Table 4-4 shows the makeup of DS-5 after the PCA reduction has been applied in terms of its contributing datasets. The percentages in the table represent the percentage contribution from each of the originating datasets not their percentage makeup of DS-5 itself, for example SNPs that appear in DS-1 and DS-5 (24 of them) give rise to 5.94% of DS-5 and represent 31.58% of DS-1. The total percentage of DS-5 exceeds 100% as there are overlaps in the SNP contributions from the other datasets.

Original set	Number included	Percentage of DS-5	Percentage of original
DS-1	24	5.94%	31.58%
DS-2	77	76.24%	51.33%
DS-3	14	12.17%	15.05%
DS-4	25	17.86%	23.81%

Table 4 - 22: Source and makeup of DS-5 the combined dataset

4.1.4 Summarising the data

The Illumina chip tests SNPs on all chromosomes, but due to limited size it cannot test the entire genome. The number of SNPs sampled summarised by chromosome can be found in figure 4-10. When the individual reduced datasets are compared they reveal a similar but not exactly representative sample of SNPs per chromosome. This is most likely to be due to the development of the chip to be used with a number of genetic diseases and other analysis of SNP data. The distribution of SNPs amongst chromosomes in the reduced datasets can be found in Figure 4-11. There is a notably higher amount of SNPs found on the X chromosome in relation to many others, with nothing being selected from the SNPs tested on the Y chromosome. This may be a contributing factor in relation to the higher proportion of MS sufferers being women; further research will be needed to determine if this is the case.

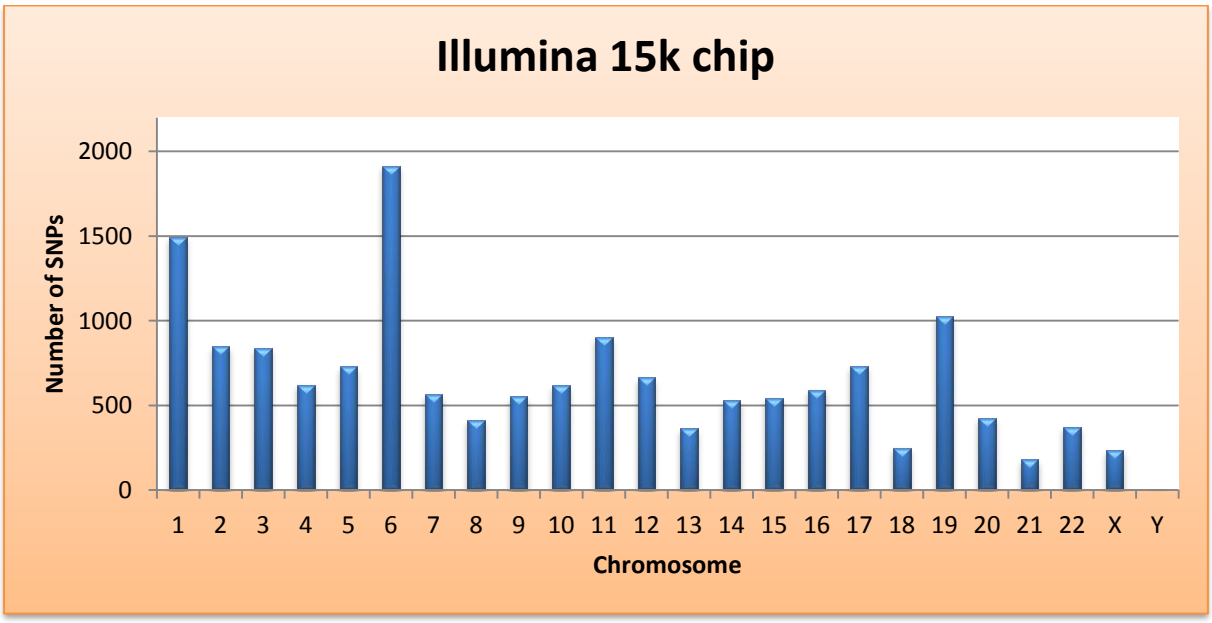


Figure 4 - 22: SNPs by chromosome tested on Infinium 15K chip

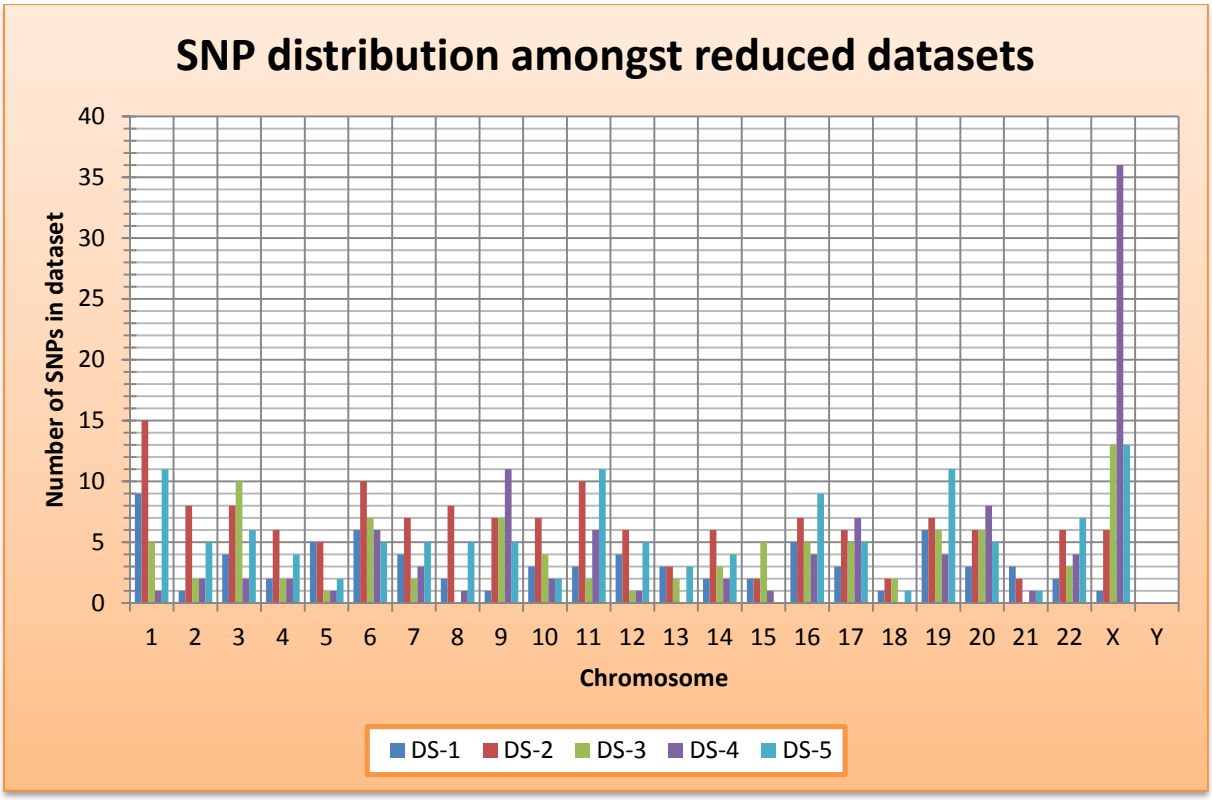


Figure 4 - 23: SNPs by chromosome in reduced datasets

A similar comparison can be made using genotype. A comparison of genotype spread amongst case and control sample is found in Figure 4-12. This reveals a very similar distribution of genotype. The differences in numerical amounts are a reflection of the different sizes of the sample classes, control being larger by approximately one third than the case class.

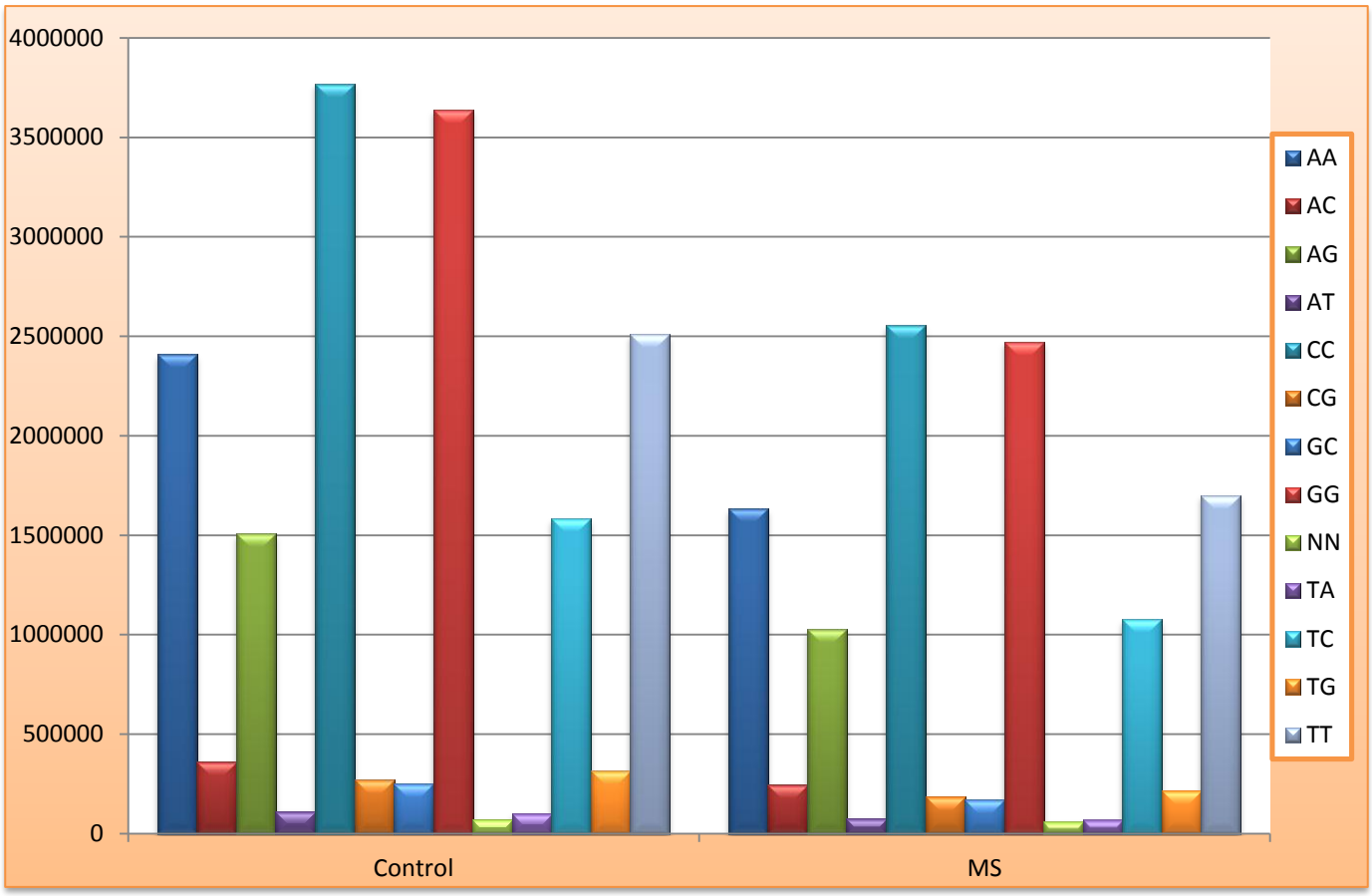


Figure 4 - 24: Genotypes found in case and control samples

4.2 Method Comparison

4.2.1 How to define the "best" result

Although there are many measures of how to determine which classification method, variation and dataset are "best", overall accuracy remains the most well used. Using overall accuracy gives a measure of how well the classification method has performed irrespective of how well it has performed for each respective class. In this application where a determination of classification is to be applied to a health condition it is more critical that the accuracy of determining the case or diseased individual is of greater importance than the accuracy of the control or non-diseased individual. The accuracy measures are very simple to calculate where only a count of correctly classified instances is needed to calculate a percentage figure. Percentages have been used against "raw" numbers as this is how much of the data in previous studies has been reported and as such it would be difficult to compare them otherwise, it also negates any differences in the number of instances in the datasets.

The classification results are not intended as a diagnosis but an aid to determining susceptibility in future patients from the analysis of known individuals. To this end lower accuracy in detecting non-susceptibility is more acceptable. Some results may need to be reassessed on this basis as their overall accuracy is lower, but individual accuracies are more compliant with this desire to have greater individual accuracy for one class over another.

4.2.2 Comparison of results

Tables 4-5, 4-6, 4-7 and 4-8 show the overall, class and control accuracy of reach classification method tested. These tables appear as reduced images of their full-scale versions which can be found in Appendix A. (This has been done for page size considerations and does not interfere with determining performance characteristics.) Colour has been used to assist in the identification of trends and the best performing variation of each test. Shaded cell have been calculated horizontally and cells there the text or border intensities have been changed have been calculated vertically. In the tables a green shaded cell represents the best overall accuracy for a particular variation which assists in the identification of the best performing dataset. A blue shaded cell represents the best classification accuracy for an individual classification, and an apricot shaded cell represents the worst classification accuracy for an individual classification. Cells where the figures are in bold with a red border indicate that they are the best classification accuracy for that dataset, and where the text is also red this represents the best overall classification accuracy for each implementation environment.

4.2.2.1 The datasets

It is easily seen from the green shaded cells in tables 4-5 – 4-8, that DS-4 is by far the best and most consistent performing dataset. This is further reflected in the dominance of blue shaded cells in DS-4. The worst performing datasets are DS-1 and DS-2, and this reflected in the numerous apricot shaded cells within these two datasets. These patterns are consistent amongst each of the implementation environments and classification methods.

A trend in the variation of best performing classification method within each dataset shows a clear correspondence to the origins of each dataset. DS-1 and DS-2 predominantly display similar characteristics and often have the same best classification method and variation within that method. This was also found to be true for DS-3 and DS-4. This trend breaks up when

the more personalised methods are considered as the thresholding can make the distinction between variations. DS-5 follows the trend of either the DS-1/DS-2 grouping or the DS-3/DS-4 grouping dependent upon which method us used. This reflects greatly the origins of DS-5 as a compilation of the other datasets.

From an overall perspective DS-4 is the best performing and most consistently performing of all five datasets tested. The only one to exceed DS-4 in overall accuracy for any test was DS-5 using the J48 tree methods and Bayesian Logistic Regression with a difference of 1.41% and 0.24% respectively. The feature reduction capability available in WW*k*NN was tested and found that it did not yield any significant improvement after more than a 5% reduction.

	DS-1			DS-2			DS-3			DS-4			DS-5		
Method	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control
Rules															
OneR	85.14%	93.75%	75.50%	85.14%	93.75%	75.50%	85.14%	93.75%	75.50%	91.27%	95.09%	87.00%	88.92%	94.64%	82.50%
Decision Table – Best First	91.51%	84.54%	75.50%	91.51%	86.08%	93.50%	91.51%	93.30%	89.50%	95.28%	99.11%	91.00%	89.39%	90.18%	88.50%
Trees															
J48 – confidence factor = 0.25	93.40%	95.88%	91.00%	93.40%	95.88%	91.00%	95.75%	96.88%	94.50%	96.23%	95.98%	96.50%	97.64%	97.32%	98.00%
J48 – Confidence factor = 0.5	93.15%	94.85%	91.50%	93.15%	94.85%	91.50%	95.99%	96.88%	95.00%	96.70%	95.98%	97.50%	97.41%	96.88%	98.00%
J48 – Confidence factor = 0.1	93.65%	95.88%	91.50%	93.65%	95.88%	91.50%	95.75%	96.88%	94.50%	96.23%	95.98%	96.50%	97.64%	97.32%	98.00%
Bayesian															
Naive Bayes	61.17%	29.90%	91.50%	61.17%	29.90%	91.50%	65.80%	40.63%	94.00%	74.06%	54.46%	96.00%	68.63%	46.88%	93.00%
Bayesian Logistic Regression	70.05%	70.10%	70.00%	70.05%	70.10%	70.00%	79.72%	80.80%	78.50%	88.44%	95.98%	80.00%	88.68%	97.32%	79.00%
Nearest Neighbour															
No distance weighting															
K = 1	77.16%	85.05%	69.50%	77.16%	85.05%	69.50%	87.26%	87.05%	87.50%	94.81%	94.64%	95.00%	87.26%	89.29%	85.00%
K = 3	75.38%	87.63%	63.50%	75.38%	87.63%	63.50%	87.26%	88.39%	86.00%	93.63%	94.64%	92.50%	88.92%	93.75%	83.50%
K = 5	75.89%	88.14%	64.00%	75.89%	88.14%	64.00%	88.68%	92.86%	84.00%	93.63%	95.54%	91.50%	87.74%	94.20%	80.50%
K = 7	73.10%	89.69%	57.00%	73.10%	89.69%	57.00%	88.68%	94.64%	82.00%	93.16%	96.43%	89.50%	87.03%	95.09%	78.00%
K = 9	70.81%	89.69%	52.50%	70.81%	89.69%	52.50%	89.39%	95.09%	83.00%	92.69%	96.88%	88.00%	87.97%	97.77%	77.00%
K = 11	70.30%	91.24%	50.00%	70.30%	91.24%	50.00%	88.21%	95.54%	80.00%	92.22%	96.88%	87.00%	86.79%	98.21%	74.00%
K = 15	68.02%	92.27%	44.50%	68.02%	92.27%	44.50%	87.97%	96.43%	78.50%	90.80%	96.88%	84.00%	84.43%	98.21%	69.00%
K = 20	69.04%	91.24%	47.50%	69.04%	91.24%	47.50%	88.68%	96.43%	80.00%	90.57%	95.98%	84.50%	84.67%	97.77%	70.00%
K = 30	69.29%	91.24%	48.00%	69.29%	100.00%	48.00%	83.96%	95.09%	71.50%	88.92%	97.77%	79.00%	80.42%	98.21%	60.50%
K = 50	65.48%	96.39%	35.50%	65.48%	96.39%	35.50%	83.73%	93.30%	73.00%	84.20%	100.00%	66.50%	78.07%	98.66%	55.00%
Weight by 1/distance															
K = 1	77.16%	85.05%	69.50%	77.16%	85.05%	69.50%	87.26%	87.05%	87.50%	94.81%	94.64%	95.00%	87.26%	89.29%	85.00%
K = 3	75.63%	87.63%	64.00%	75.63%	87.63%	64.00%	87.97%	88.84%	87.00%	94.58%	95.54%	93.50%	89.15%	93.75%	84.00%
K = 5	76.14%	87.63%	65.00%	76.14%	87.63%	65.00%	90.33%	93.30%	87.00%	94.34%	95.54%	93.00%	88.92%	95.54%	81.50%
K = 7	73.86%	89.69%	58.50%	73.86%	89.69%	58.50%	89.62%	94.20%	84.50%	94.58%	96.88%	92.00%	87.50%	95.98%	78.00%
K = 9	71.32%	89.69%	53.50%	71.32%	89.69%	53.50%	90.09%	94.20%	85.50%	93.87%	97.32%	90.00%	88.44%	97.77%	78.00%
K = 11	71.07%	91.75%	51.00%	71.07%	91.75%	51.00%	88.68%	94.64%	82.00%	93.40%	97.32%	89.00%	88.44%	97.77%	78.00%
K = 15	70.30%	92.78%	48.50%	70.30%	92.78%	48.50%	90.57%	96.43%	84.00%	91.75%	96.43%	86.50%	86.32%	98.21%	73.00%
K = 20	70.56%	93.30%	48.50%	70.56%	93.30%	48.50%	90.33%	96.43%	83.50%	91.51%	96.88%	85.50%	86.32%	98.21%	73.00%
K = 30	70.56%	95.88%	46.00%	70.56%	95.88%	46.00%	89.15%	97.32%	80.00%	90.57%	98.21%	82.00%	82.55%	98.66%	64.50%
K = 50	67.51%	95.88%	40.00%	67.51%	95.88%	40.00%	87.50%	95.09%	79.00%	87.03%	100.00%	72.50%	81.37%	98.66%	62.00%
Weight by 1-distance															
K = 1	77.16%	85.05%	69.50%	77.16%	85.05%	69.50%	87.26%	87.05%	87.50%	94.81%	94.64%	95.00%	87.26%	89.29%	85.00%
K = 3	75.38%	87.63%	63.50%	75.38%	87.63%	63.50%	87.26%	88.39%	86.00%	93.63%	94.64%	92.50%	88.92%	93.75%	83.50%
K = 5	75.89%	88.14%	64.00%	75.89%	88.14%	64.00%	88.68%	92.86%	84.00%	93.63%	95.54%	91.50%	87.74%	94.20%	80.50%
K = 7	73.10%	89.69%	57.00%	73.10%	89.69%	57.00%	88.68%	94.64%	82.00%	93.16%	96.43%	89.50%	87.03%	95.09%	78.00%
K = 9	70.81%	89.69%	52.50%	70.81%	89.69%	52.50%	89.39%	95.09%	83.00%	92.69%	96.88%	88.00%	87.97%	97.77%	77.00%
K = 11	70.30%	91.24%	50.00%	70.30%	91.24%	50.00%	88.21%	95.54%	80.00%	92.22%	96.88%	87.00%	86.79%	98.21%	74.00%
K = 15	68.02%	92.27%	44.50%	68.02%	92.27%	44.50%	87.97%	96.43%	78.50%	90.80%	96.88%	84.00%	84.43%	98.21%	69.00%
K = 20	69.29%	93.30%	46.00%	69.29%	93.30%	46.00%	88.44%	96.88%	79.00%	91.04%	96.88%	84.50%	84.43%	98.21%	69.00%
K = 30	68.78%	94.33%	44.00%	68.78%	94.33%	44.00%	85.14%	97.32%	71.50%	89.15%	98.21%	79.00%	80.42%	98.66%	60.00%
K = 50	65.23%	96.39%	35.00%	65.23%	96.39%	35.00%	84.43%	95.09%	72.50%	84.20%	100.00%	66.50%	78.07%	98.66%	55.00%
Support Vector Machine															
Polynomial	50.76%	0.00%	100.00 %	50.76%	87.63%	63.00%	52.83%	100.00%	0.00%	52.83%	100.00%	0.00%	52.83%	100.00 %	0.00%
Linear	64.97%	75.77%	54.50%	75.13%	0.00%	100.00%	86.56%	88.84%	84.00%	86.56%	95.54%	89.00%	91.51%	97.77%	54.50%
Radial Based Function	76.14%	37.63%	75.00%	76.14%	85.57%	67.00%	89.62%	100.00%	0.00%	96.23%	100.00%	59.00%	89.39%	100.00 %	0.00%
Multilayer Perceptron	76.14%	85.57%	67.00%	76.14%	85.57%	67.00%	89.62%	91.52%	87.50%	96.23%	96.43%	96.00%	89.39%	92.41%	86.00%

Table 4 - 23: Classification testing methods implemented in WEKA

Method	DS-1 Overall	DS-1 Case	DS-1 Control	DS-2 Overall	DS-2 Case	DS-2 Control	DS-3 Overall	DS-3 Case	DS-3 Control	DS-4 Overall	DS-4 Case	DS-4 Control	DS-5 Overall	DS-5 Case	DS-5 Control
Support Vector Machine															
Polynomial	55.55%	73.20%	75.50%	52.02%	84.02%	71.00%	59.30%	87.95%	84.00%	70.77%	95.54%	94.50%	54.06%	95.54%	88.00%
Linear	75.11%	74.74%	75.50%	77.38%	82.99%	72.00%	89.16%	90.63%	87.50%	95.04%	97.32%	92.50%	92.69%	98.21%	86.50%
Radial Based Function	74.87%	73.20%	76.50%	76.69%	78.35%	75.00%	86.31%	89.73%	82.50%	93.17%	96.43%	89.50%	91.50%	95.54%	87.00%
Multilayer Perceptron	74.85%	76.17%	73.50%	76.09%	81.96%	70.50%	89.87%	91.52%	88.00%	96.21%	95.98%	96.50%	91.51%	93.75%	89.00%
Evolving clustering method for classification	59.21%	60.32%	57.00%	69.58%	77.84%	61.50%	77.81%	75.00%	81.00%	94.09%	95.54%	92.50%	85.64%	90.18%	80.50%
Multiple Linear Regression	70.33%	69.59%	71.00%	69.29%	70.90%	67.00%	82.33%	82.14%	82.50%	92.44%	92.41%	92.50%	86.09%	89.29%	82.50%
Evolving Classification Function	55.55%	76.80%	87.00%	52.02%	4.64%	98.00%	59.30%	37.05%	84.00%	70.77%	52.68%	91.00%	54.06%	16.52%	96.00%

Table 4 - 25: Classification testing methods implemented in NeuCom

Method		DS-1 Overall	DS-1 Case	DS-1 Control	DS-2 Overall	DS-2 Case	DS-2 Control	DS-3 Overall	DS-3 Case	DS-3 Control	DS-4 Overall	DS-4 Case	DS-4 Control	DS-5 Overall	DS-5 Case	DS-5 Control
WkNN																
Threshold = 0.45	*K=1*	64.72%	59.28%	70.00%	79.44%	86.60%	72.50%	89.15%	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=3	69.29%	63.92%	74.50%	78.43%	87.11%	70.00%	87.26%	88.39%	86.00%	93.63%	95.09%	92.00%	88.44%	93.57%	82.50%
	K=5	70.30%	67.01%	73.50%	78.68%	89.18%	68.50%	88.21%	91.07%	85.00%	93.87%	96.43%	91.00%	87.74%	94.64%	80.00%
	K=7	68.27%	64.43%	72.00%	77.66%	89.69%	66.00%	87.26%	91.52%	82.50%	92.92%	97.32%	88.00%	87.50%	95.09%	79.00%
	K=9	68.53%	64.43%	72.50%	76.14%	90.72%	62.00%	86.56%	90.18%	82.50%	91.98%	96.43%	87.00%	85.85%	94.64%	76.00%
	K=11	70.81%	77.32%	64.50%	71.07%	97.94%	45.00%	87.50%	94.64%	79.50%	90.09%	98.21%	81.00%	82.55%	96.88%	66.50%
	K=15	70.56%	73.71%	67.50%	70.05%	97.94%	43.00%	87.74%	94.64%	80.00%	89.39%	98.66%	79.00%	82.55%	98.66%	64.50%
	K=20	72.84%	71.65%	74.00%	72.84%	99.48%	47.00%	72.84%	97.32%	81.00%	91.04%	98.66%	82.50%	82.08%	98.66%	63.50%
	K=30	72.59%	72.16%	73.00%	68.02%	99.48%	37.50%	86.08%	95.98%	75.00%	87.03%	99.11%	73.50%	78.30%	99.11%	55.00%
	K=50	71.07%	71.13%	71.00%	61.68%	98.45%	26.00%	84.43%	92.86%	75.00%	82.78%	100.00%	63.50%	74.06%	99.11%	46.00%
Threshold = 0.50	*K=1*	64.72%	59.28%	70.00%	79.44%	86.60%	72.50%	89.15%	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=3	69.29%	63.92%	74.50%	78.43%	87.11%	70.00%	87.26%	88.39%	86.00%	93.63%	95.09%	92.00%	88.44%	93.57%	82.50%
	K=5	70.30%	67.01%	73.50%	78.68%	89.18%	68.50%	88.21%	91.07%	85.00%	93.87%	96.43%	91.00%	87.74%	94.63%	80.00%
	K=7	68.27%	64.43%	72.00%	77.66%	89.69%	66.00%	87.26%	91.52%	82.50%	92.92%	97.32%	88.00%	87.50%	95.09%	79.00%
	K=9	68.53%	64.43%	72.50%	76.14%	90.72%	62.00%	86.56%	90.18%	82.50%	91.98%	96.43%	87.00%	85.85%	94.64%	76.00%
	K=11	70.05%	65.46%	74.50%	74.37%	91.24%	58.00%	87.26%	91.52%	82.50%	91.98%	96.43%	87.00%	85.61%	95.98%	74.00%
	K=15	72.08%	67.01%	77.00%	74.37%	94.85%	54.50%	87.97%	93.30%	82.00%	91.27%	96.88%	85.00%	84.20%	97.32%	69.50%
	K=20	72.59%	67.53%	77.50%	74.87%	97.42%	53.00%	88.92%	93.75%	83.50%	91.98%	96.43%	87.00%	84.43%	98.21%	69.00%
	K=30	74.62%	67.01%	82.00%	71.83%	98.45%	46.00%	83.96%	88.84%	78.50%	89.39%	98.21%	79.50%	80.19%	97.77%	60.50%
	K=50	70.30%	66.49%	74.00%	66.24%	94.85%	38.50%	83.96%	87.50%	80.00%	86.56%	99.11%	72.50%	79.01%	98.21%	57.50%
Threshold = 0.55	*K=1*	64.72%	59.28%	70.00%	79.44%	86.60%	72.50%	89.15%	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=3	69.29%	63.92%	74.50%	78.43%	87.11%	70.00%	87.26%	88.39%	86.00%	93.63%	95.09%	92.00%	88.44%	93.57%	82.50%
	K=5	70.30%	67.01%	73.50%	78.68%	89.18%	68.50%	88.21%	91.07%	85.00%	93.87%	96.43%	91.00%	87.74%	94.63%	80.00%
	K=7	68.27%	64.43%	72.00%	77.66%	89.69%	66.00%	87.26%	91.52%	82.50%	92.92%	97.32%	88.00%	87.50%	95.09%	79.00%
	K=9	68.27%	63.92%	72.50%	76.14%	90.72%	62.00%	86.56%	90.18%	82.50%	92.69%	96.43%	88.50%	85.85%	94.64%	76.00%
	K=11	70.05%	56.70%	83.00%	73.10%	78.35%	68.00%	84.43%	83.93%	85.00%	92.22%	93.75%	90.50%	87.26%	92.86%	81.00%
	K=15	70.56%	60.31%	80.50%	72.08%	83.51%	61.00%	87.26%	88.39%	86.00%	91.89%	94.64%	89.00%	85.85%	94.20%	76.50%
	K=20	72.34%	63.92%	80.50%	76.65%	91.24%	62.50%	86.32%	87.50%	85.00%	91.57%	98.21%	84.50%	85.61%	97.32%	72.50%
	K=30	73.60%	63.92%	83.00%	73.86%	92.78%	55.50%	82.55%	83.04%	82.00%	90.57%	96.88%	83.50%	82.08%	96.88%	65.50%
	K=50	73.35%	63.40%	83.00%	70.05%	91.75%	49.00%	81.60%	77.23%	86.50%	88.44%	98.66%	77.00%	82.55%	96.43%	67.00%
Threshold = 0.60	*K=1*	64.72%	59.28%	70.00%	79.44%	86.60%	72.50%	89.15%	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=3	69.29%	63.92%	74.50%	78.43%	87.11%	70.00%	87.26%	88.39%	86.00%	93.63%	95.09%	92.00%	88.44%	93.57%	82.50%
	K=5	67.51%	54.64%	80.00%	79.19%	82.99%	75.50%	86.56%	83.93%	89.50%	94.58%	94.64%	94.50%	89.86%	91.52%	88.00%
	K=7	69.80%	64.43%	87.00%	76.90%	89.69%	76.00%	82.31%	91.52%	87.50%	93.40%	97.32%	94.50%	89.62%	95.09%	88.00%
	K=9	71.57%	55.67%	87.00%	77.16%	82.47%	72.00%	84.43%	83.93%	85.00%	92.92%	92.86%	93.00%	90.33%	92.41%	88.00%
	K=11	70.05%	56.70%	83.00%	73.10%	78.35%	68.00%	84.43%	83.93%	85.00%	92.22%	93.75%	90.50%	87.26%	92.86%	81.00%
	K=15	70.81%	57.73%	83.50%	74.62%	79.90%	69.50%	85.61%	84.38%	87.00%	93.16%	93.75%	92.50%	87.03%	92.96%	80.50%
	K=20	70.05%	57.22%	82.50%	76.14%	85.05%	67.50%	86.56%	85.71%	87.50%	91.98%	92.41%	91.50%	85.14%	92.86%	76.50%
	K=30	72.59%	60.82%	84.00%	77.66%	88.66%	67.00%	84.20%	79.46%	89.50%	91.04%	94.20%	87.50%	83.96%	92.86%	74.00%
	K=50	73.86%	62.37%	85.00%	72.34%	86.08%	59.00%	80.42%	72.32%	89.50%	90.80%	95.54%	85.50%	83.02%	94.20%	70.50%
Threshold = 0.65	*K=1*	64.72%	59.28%	70.00%	79.44%	86.60%	72.50%	89.15%	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=3	69.29%	63.92%	74.50%	78.43%	87.11%	70.00%	87.26%	88.39%	86.00%	94.10%	95.09%	93.00%	88.44%	93.57%	82.50%
	K=5	62.18%	37.63%	86.00%	78.43%	75.26%	81.50%	82.31%	73.66%	92.00%	93.16%	91.07%	95.50%	89.39%	87.05%	92.00%
	K=7	69.80%	52.06%	87.00%	76.90%	77.84%	76.00%	82.31%	77.68%	87.50%	93.40%	92.41%	94.50%	89.62%	91.07%	88.00%
	K=9	71.57%	55.67%	87.00%	77.16%	82.47%	72.00%	84.43%	83.93%	85.00%	92.92%	92.86%	93.00%	90.33%	92.41%	88.00%
	K=11	69.29%	48.97%	89.00%	71.07%	61.34%	80.50%	79.25%	71.88%	87.50%	91.04%	87.50%	95.00%	88.44%	88.39%	88.50%
	K=15	70.30%	55.15%	85.00%	73.60%	72.68%	74.50%	82.55%	78.13%	87.50%	92.22%	91.07%	93.50%	86.56%	91.07%	81.50%
	K=20	68.02%	51.03%	84.50%	74.11%	73.71%	74.50%	82.31%	75.89%	89.50%	91.95%	88.84%	94.50%	86.08%	88.84%	83.00%
	K=30	70.81%	55.15%	86.00%	73.60%	72.68%	74.50%	79.25%	69.20%	90.50%	88.92%	86.16%	92.00%	83.73%	88.39%	78.50%
	K=50	73.10%	60.31%	85.50%	72.59%	77.84%	67.50%	79.01%	68.30%	91.00%	88.44%	87.50%	89.50%	80.90%	83.93%	77.50%

Table 4 - 24: Classification testing methods implemented in MATLAB - WkNN method

Method		DS-1 Overall	DS-1 Case	DS-1 Control	DS-2 Overall	DS-2 Case	DS-2 Control	DS-3 Overall	DS-3 Case	DS-3 Control	DS-4 Overall	DS-4 Case	DS-4 Control	DS-5 Overall	DS-5 Case	DS-5 Control
WWkNN – 100% Features																
Threshold = 0.35	K=3	66.50%	66.49%	66.50%	79.70%	90.21%	69.50%	89.15%	91.07%	87.00%	94.34%	95.54%	93.00%	88.21%	92.86%	83.00%
	K=5	68.27%	73.20%	63.50%	80.46%	73.20%	71.00%	91.98%	95.09%	88.50%	96.46%	97.32%	95.50%	89.62%	94.64%	84.00%
	K=7	69.54%	75.26%	64.00%	78.17%	75.26%	66.00%	90.33%	92.86%	87.50%	95.99%	96.43%	95.50%	90.33%	95.54%	84.50%
	K=9	71.57%	76.29%	67.00%	80.71%	76.29%	67.00%	92.45%	95.98%	88.50%	95.97%	97.77%	93.50%	89.39%	95.54%	82.50%
	K=11	70.05%	75.77%	64.50%	80.46%	95.88%	65.50%	91.75%	95.98%	87.00%	95.99%	97.77%	94.00%	92.57%	97.77%	82.50%
	K=15	68.02%	76.80%	59.50%	77.41%	76.80%	59.00%	91.27%	96.43%	85.50%	95.99%	98.66%	93.00%	90.09%	98.21%	81.00%
	K=20	70.56%	79.90%	61.50%	75.38%	79.90%	53.50%	91.51%	97.77%	84.50%	94.58%	97.21%	90.50%	88.44%	98.21%	77.50%
	K=30	71.32%	81.44%	61.50%	72.08%	98.45%	46.50%	91.27%	98.66%	83.00%	92.45%	99.11%	85.00%	86.08%	97.77%	73.00%
	K=50	72.08%	84.02%	60.50%	68.78%	98.97%	39.58%	88.21%	98.66%	76.50%	90.09%	100.00%	79.00%	83.02%	98.21%	66.00%
Threshold = 0.40	K=3	65.74%	63.92%	67.50%	79.70%	88.66%	71.00%	89.39%	90.63%	88.00%	94.34%	95.54%	93.00%	88.68%	92.86%	84.00%
	K=5	69.29%	72.16%	66.50%	79.95%	89.18%	71.00%	91.75%	94.20%	89.00%	96.70%	97.32%	96.00%	90.09%	94.20%	85.50%
	K=7	71.07%	74.74%	67.50%	79.19%	89.69%	69.00%	90.80%	92.41%	89.00%	96.23%	96.43%	96.00%	91.04%	94.20%	87.50%
	K=9	71.32%	74.74%	68.00%	81.73%	92.78%	71.00%	92.45%	94.20%	90.50%	96.46%	97.77%	95.00%	90.57%	95.54%	85.00%
	K=11	71.05%	74.23%	68.00%	80.71%	93.81%	68.00%	91.75%	94.64%	88.50%	96.70%	97.32%	96.00%	90.80%	96.43%	84.50%
	K=15	70.05%	73.71%	66.50%	78.17%	95.36%	61.50%	91.51%	95.98%	86.50%	96.93%	98.66%	95.00%	90.57%	97.32%	83.00%
	K=20	70.81%	75.26%	66.50%	76.90%	96.91%	57.50%	91.51%	96.88%	85.50%	95.28%	97.77%	92.50%	88.92%	97.77%	79.00%
	K=30	71.83%	76.29%	67.50%	76.65%	97.94%	56.00%	90.57%	97.32%	83.00%	92.92%	98.66%	86.50%	88.44%	97.77%	78.00%
	K=50	71.83%	78.57%	68.00%	70.56%	98.97%	43.00%	88.92%	98.21%	78.50%	91.51%	99.11%	83.00%	85.85%	97.77%	72.50%
Threshold = 0.45	K=3	65.23%	91.34%	69.00%	79.44%	87.11%	72.00%	89.62%	89.73%	89.50%	94.32%	95.09%	93.50%	89.15%	92.41%	85.50%
	K=5	68.78%	67.53%	70.00%	79.70%	87.63%	72.00%	91.27%	93.30%	89.00%	96.46%	96.88%	96.00%	90.33%	93.75%	86.50%
	K=7	71.07%	73.20%	69.00%	79.19%	87.63%	71.00%	90.57%	91.52%	89.50%	96.46%	96.43%	96.50%	91.27%	93.75%	88.50%
	K=9	72.34%	74.23%	70.50%	80.96%	90.21%	72.00%	91.98%	92.86%	91.00%	96.70%	97.77%	95.50%	91.04%	95.09%	86.50%
	K=11	72.34%	72.68%	72.00%	80.20%	91.24%	69.50%	91.56%	92.86%	90.00%	96.93%	97.32%	96.50%	91.04%	95.09%	86.50%
	K=15	72.59%	70.10%	75.00%	78.68%	93.30%	64.50%	91.98%	94.64%	89.00%	97.41%	98.66%	96.00%	90.80%	95.54%	85.50%
	K=20	73.10%	70.62%	75.50%	79.44%	95.36%	64.00%	91.98%	95.54%	88.00%	95.28%	96.88%	93.50%	90.09%	96.43%	83.00%
	K=30	71.57%	70.10%	73.00%	78.17%	97.94%	59.00%	91.04%	95.98%	85.50%	94.10%	97.77%	90.00%	89.39%	96.88%	81.00%
	K=50	73.35%	68.04%	78.50%	75.81%	98.45%	52.50%	91.04%	98.21%	83.00%	92.45%	98.66%	85.50%	86.32%	96.88%	74.50%
Threshold = 0.50	K=3	61.97%	59.79%	70.00%	79.44%	86.60%	75.50%	89.15%	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=5	69.04%	64.95%	73.00%	80.71%	87.63%	74.00%	90.80%	91.52%	90.00%	96.46%	96.88%	96.00%	90.80%	93.75%	87.50%
	K=7	70.56%	68.56%	72.50%	78.68%	84.54%	73.00%	89.86%	89.73%	90.00%	96.46%	96.43%	96.50%	91.98%	93.75%	90.00%
	K=9	72.34%	71.13%	73.50%	80.96%	86.08%	76.00%	91.98%	91.96%	92.00%	96.70%	97.77%	95.50%	91.98%	94.20%	89.50%
	K=11	71.07%	67.53%	74.50%	80.46%	88.66%	72.50%	91.51%	91.96%	91.00%	96.93%	97.32%	96.50%	91.98%	94.64%	89.00%
	K=15	71.07%	65.46%	76.00%	80.20%	91.24%	69.50%	90.80%	92.41%	89.00%	96.93%	97.77%	96.00%	91.51%	95.09%	87.50%
	K=20	71.57%	65.46%	77.50%	80.71%	93.30%	68.50%	92.22%	94.64%	89.50%	96.23%	96.88%	95.50%	91.51%	95.98%	86.50%
	K=30	73.60%	64.95%	82.00%	78.68%	94.85%	63.00%	91.98%	95.54%	88.00%	94.58%	97.32%	91.50%	90.80%	96.43%	84.50%
	K=50	74.11%	65.98%	80.22%	79.44%	97.94%	61.50%	91.51%	95.09%	87.50%	92.92%	98.66%	86.50%	88.44%	96.88%	79.00%
Threshold = 0.55	K=3	64.72%	56.70%	72.50%	78.68%	85.05%	72.50%	89.39%	87.95%	91.00%	94.58%	94.64%	94.50%	88.68%	91.07%	86.00%
	K=5	67.77%	61.34%	74.00%	80.96%	86.60%	75.50%	90.33%	90.18%	90.50%	96.46%	96.88%	96.00%	90.80%	92.41%	89.00%
	K=7	69.80%	64.43%	75.00%	78.68%	82.47%	75.00%	90.57%	89.73%	91.50%	96.70%	96.43%	97.00%	91.75%	92.41%	91.00%
	K=9	72.08%	66.49%	77.50%	80.96%	84.54%	77.50%	91.98%	91.52%	92.50%	96.46%	96.88%	96.00%	91.98%	93.75%	90.00%
	K=11	71.32%	65.98%	76.50%	80.71%	86.08%	75.50%	91.04%	90.18%	92.00%	96.93%	96.88%	97.00%	92.69%	93.75%	91.50%
	K=15	70.81%	62.37%	79.00%	81.73%	90.21%	73.50%	90.33%	90.63%	90.00%	97.17%	96.88%	97.50%	91.98%	94.64%	89.00%
	K=20	71.07%	61.86%	80.00%	80.46%	90.72%	70.50%	91.51%	91.96%	91.00%	96.23%	95.98%	96.50%	92.45%	95.54%	89.00%
	K=30	72.59%	61.34%	83.50%	82.23%	93.30%	71.50%	91.98%	94.20%	89.50%	94.81%	95.54%	94.00%	91.98%	95.98%	87.50%
	K=50	73.60%	64.43%	82.50%	80.96%	94.33%	68.00%	89.62%	89.73%	89.50%	93.63%	98.21%	88.50%	91.51%	96.88%	85.50%
Threshold = 0.60	K=3	64.21%	54.12%	74.00%	78.68%	84.54%	73.00%	89.39%	87.05%	92.00%	94.34%	93.75%	95.00%	88.21%	90.18%	86.00%
	K=5	67.01%	57.73%	76.00%	80.71%	84.02%	77.50%	90.09%	88.84%	91.50%	96.26%	96.43%	96.00%	90.57%	91.52%	89.50%
	K=7	69.54%	60.82%	78.00%	79.19%	80.93%	77.50%	89.39%	87.05%	92.00%	96.46%	95.98%	97.00%	91.98%	91.96%	92.00%
	K=9	71.57%	61.86%	81.00%	81.22%	84.02%	78.50%	90.33%	87.95%	93.00%	96.70%	96.43%	97.00%	91.27%	91.96%	90.50%
	K=11	70.30%	60.82%	79.50%	80.96%	83.52%	78.50%	89.62%	87.05%	92.50%	97.17%	96.43%	98.00%	92.69%	93.30%	92.00%
	K=15	71.32%	59.28%	83.00%	81.22%	85.57%	77.00%	90.09%	88.39%	92.00%	96.70%	95.98%	97.50%	92.92%	92.41%	83.50%
	K=20	70.05%	57.73%	82.00%	80.46%	86.60%	74.50%	90.33%	89.29%	91.50%	95.99%	95.09%	97.00%	92.45%	93.75%	91.00%
	K=30	71.57%	56.70%	86.00%	81.47%	88.66%	74.50%	90.09%	89.29%	91.00%	95.75%	94.64%	97.00%	92.22%	94.20%	90.00%
	K=50	72.59%	61.34%	83.50%	80.96%	90.21%	72.00%	86.32%	82.59%	90.50%	93.63%	96.43%	90.50%	90.80%	94.20%	87.00%
Threshold = 0.65	K=3	64.72%	53.61%	75.50%	78.68%	82.99%	74.50%	89.15%	86.16%	92.50%	94.58%	93.75%	95.50%	88.68%	89.73%	87.05%
	K=5	66.50%	53.61%	79.00%	79.95%	81.96%	78.00%	90.33%	88.39%	92.50%	95.75%	95.54%	96.00%	90.33%	90.63%	90.00%
	K=7	65.58%	51.03%	79.50%	78.93%	78.35%	79.50%	88.92%	84.82%	93.50%	96.46%	95.98%	97.00%	91.75%	90.63%	93.00%
	K=9	68.02%	52.58%	83.00%	81.71%	81.44%	80.00%	88.68%	84.38%	93.50%	96.93%	96.43%	97.50%	90.57%	90.18%	91.00%
	K=11	69.04%	54.64%	83.00%	82.49%	82.99%	82.00%	88.44%	84.38%	93.00%	96.93%	95.95%	98.00%	91.98%	91.52%	92.50%
	K=15	69.04%	53.09%	84.50%	81.98%	81.96%	82.00%	89.62%	86.16%	93.50%	96.46%	95.54%	97.50%	93.63%	92.41%	95.00%
	K=20	69.04%	52.58%	85.00%	78.43%	78.35%	78.50%	89.86%	87.05%	93.00%	95.99%	94.64%	97.50%	92.69%	91.96%	93.50%
	K=30	69.80%	52.06%	87.00%	80.20%	81.96%	78.50%	89.15%	87.05%	91.50%	94.34%	91.07%	98.00%	92.17%	91.52%	91.00%
	K=50	70.30%	55.67%	84.50%	81.47%	84.54%	78.50%	83.49%	76.79%	91.00%	93.40%	91.52%	95.50%	89.15%	84.82%	94.00%

Table 4 - 26: Classification testing methods implemented in MATLAB - WWkNN with no feature reduction

4.2.2.2 The classification methods

When considering the classification methods, these in the most part performed exceedingly well. It was not expected that accuracies above 95% would be likely at this stage in the investigation especially as very little work was done to fully optimise each method. Of the best 76 variations tested 51 were found to have overall accuracy levels greater than 95%, and 74 over 90%. The highest individual overall accuracy in each of the three categories is listed in table 4-9.

Type	Method	Overall accuracy
Global	J48 – Tree	97.64%
Local	Support Vector Machine (Radial Based Function)	96.23%
Personalised	WW*k*NN (Threshold = 0.45, K = 15)	97.41%

Table 4 - 27: Highest overall accuracy per type of classification method

By this summary it could be interpreted that the global methods were better, although not by much, than the personalised. But this is an anomaly as the personalised methods ranked consistently higher than those of any other class. For a full listing of the best results from each method see table 4-11.

Where the overall accuracy for any one method variation was low, hovering between 50% and 55% particularly, it was found that the method correctly classified all of one class and none of the other. In this situation the method classified all instances in the same class. This was true of all but one of the data sets for the Support Vector Machine with a Polynomial kernel implemented in WEKA. The WEKA implementation of the SVM over the three variations and five datasets produced on 7 occasions this classification situation. NeuCom in comparison did not do this on any occasion.

Overall classification alone is not necessarily the best measure of performance. As stated earlier the accuracy of each class is also important in medical situations and in this study when the class classifications are added to the picture an overall accuracy result can change from a very good to a medium level performance. Adding class accuracy to the data found in table 4-9, results in a change of perspective as shown in table 4-10. The performance of the global method is fairly even over both classes with the control class slightly higher. The personalised model shows a similar pattern but with the case class having a higher accuracy and by a larger margin. The local model has a great result for case accuracy but a terrible one for control accuracy, leading to a downgrading of the validity of accepting the overall accuracy. This can be a failing of localised methods that can get trapped in a localised solution and not find the best overall solution that is desired.

Type	Method	Overall accuracy	Case class	Control class
Global	J48	97.64%	97.32%	98.00%
Local	SVM	96.23%	100.00%	59.00%
Personalised	WW*k*NN	97.41%	98.66%	96.00%

Table 4 - 28: Accuracy levels for highest performing methods

Method	Variation	Implementation, & Dataset		Overall	Case	Control
Rules	Decision Table – Best First	WEKA	DS-4	95.28%	99.11%	91.00%
Trees – J48	confidence factor = 0.1	WEKA	DS-5	97.64%	97.32%	98.00%
Trees – J48	confidence factor = 0.25	WEKA	DS-5	97.64%	97.32%	98.00%
Bayesian Logistic Regression		WEKA	DS-5	88.68%	97.32%	79.00%
Nearest Neighbour	No distance weighting, K=1	WEKA	DS-4	94.81%	94.64%	95.00%
Nearest Neighbour	Weight by 1/distance, K=1	WEKA	DS-4	94.81%	94.64%	95.00%
Nearest Neighbour	Weight by 1-distance, K=1	WEKA	DS-4	94.81%	94.64%	95.00%
Support Vector Machine	Radial Based Function	WEKA	DS-4	96.23%	100.00%	59.00%
Support Vector Machine	Linear	NeuCom	DS-4	95.04%	97.32%	92.50%
Multilayer Perceptron		WEKA	DS-4	96.23%	96.43%	96.00%
Multilayer Perceptron		NeuCom	DS-4	96.21%	95.98%	96.50%
Evolving clustering method for classification		NeuCom	DS-4	94.09%	95.54%	92.50%
Multiple Linear Regression		NeuCom	DS-4	92.44%	92.41%	92.50%
Evolving Classification Function		NeuCom	DS-4	70.77%	52.68%	91.00%
W*k*NN	Threshold = 0.45, K=1	MATLAB	DS-4	94.34%	95.09%	93.50%
W*k*NN	Threshold = 0.5, K=1	MATLAB	DS-4	94.34%	95.09%	93.50%
W*k*NN	Threshold = 0.55, K=1	MATLAB	DS-4	94.34%	95.09%	93.50%
W*k*NN	Threshold = 0.6, K=5	MATLAB	DS-4	94.58%	94.64%	94.50%
W*k*NN	Threshold = 0.65, K=1	MATLAB	DS-4	94.34%	95.09%	93.50%
WW*k*NN, Features 100%	Threshold = 0.35, K=5	MATLAB	DS-4	96.46%	97.32%	95.50%
WW*k*NN, Features 100%	Threshold = 0.40, K=15	MATLAB	DS-4	96.93%	98.66%	95.00%
WW*k*NN, Features 100%	Threshold = 0.45, K=15	MATLAB	DS-4	97.41%	98.66%	96.00%
WW*k*NN, Features 100%	Threshold = 0.50, K=11	MATLAB	DS-4	96.93%	97.32%	96.50%
WW*k*NN, Features 100%	Threshold = 0.50, K=15	MATLAB	DS-4	96.93%	97.77%	96.00%
WW*k*NN, Features 100%	Threshold = 0.55, K=15	MATLAB	DS-4	97.17%	96.88%	97.50%
WW*k*NN, Features 100%	Threshold = 0.60, K=11	MATLAB	DS-4	97.17%	96.43%	98.00%
WW*k*NN, Features 100%	Threshold = 0.65, K=9	MATLAB	DS-4	96.93%	96.43%	97.50%
WW*k*NN, Features 100%	Threshold = 0.65, K=11	MATLAB	DS-4	96.93%	95.95%	98.00%
WW*k*NN, Features 95%	Threshold = 0.35, K=5	MATLAB	DS-4	96.93%	96.88%	97.00%
WW*k*NN, Features 95%	Threshold = 0.40, K=5	MATLAB	DS-4	96.93%	96.88%	97.00%
WW*k*NN, Features 95%	Threshold = 0.40, K=7	MATLAB	DS-4	96.93%	96.43%	97.50%
WW*k*NN, Features 95%	Threshold = 0.40, K=9	MATLAB	DS-4	96.93%	97.77%	95.00%
WW*k*NN, Features 95%	Threshold = 0.40, K=15	MATLAB	DS-4	96.23%	98.66%	93.50%
WW*k*NN, Features 95%	Threshold = 0.45, K=9	MATLAB	DS-4	97.17%	97.77%	95.50%
WW*k*NN, Features 95%	Threshold = 0.50, K=7	MATLAB	DS-4	96.93%	96.43%	97.50%
WW*k*NN, Features 95%	Threshold = 0.50, K=9	MATLAB	DS-4	96.93%	97.32%	95.50%
WW*k*NN, Features 95%	Threshold = 0.50, K=11	MATLAB	DS-4	96.46%	96.80%	96.00%
WW*k*NN, Features 95%	Threshold = 0.50, K=15	MATLAB	DS-4	96.46%	97.77%	95.00%
WW*k*NN, Features 95%	Threshold = 0.55, K=7	MATLAB	DS-4	96.93%	96.43%	97.50%
WW*k*NN, Features 95%	Threshold = 0.55, K=11	MATLAB	DS-4	96.46%	96.43%	96.50%
WW*k*NN, Features 95%	Threshold = 0.55, K=15	MATLAB	DS-4	96.93%	97.77%	96.00%
WW*k*NN, Features 95%	Threshold = 0.60, K=7	MATLAB	DS-4	96.70%	95.98%	97.50%
WW*k*NN, Features 95%	Threshold = 0.60, K=9	MATLAB	DS-4	96.70%	95.98%	97.50%
WW*k*NN, Features 95%	Threshold = 0.65, K=7	MATLAB	DS-4	96.46%	95.54%	97.50%
WW*k*NN, Features 95%	Threshold = 0.65, K=9	MATLAB	DS-4	96.46%	95.54%	97.50%

WW*k*NN, Features 90%	Threshold = 0.35, K=5	MATLAB	DS-4	95.28%	98.09%	95.50%
WW*k*NN, Features 90%	Threshold = 0.35, K=11	MATLAB	DS-4	95.28%	97.32%	93.00%
WW*k*NN, Features 90%	Threshold = 0.40, K=11	MATLAB	DS-4	95.75%	97.32%	94.00%
WW*k*NN, Features 90%	Threshold = 0.45, K=11	MATLAB	DS-4	95.99%	96.88%	95.00%
WW*k*NN, Features 90%	Threshold = 0.50, K=11	MATLAB	DS-4	95.99%	96.43%	95.50%
WW*k*NN, Features 90%	Threshold = 0.50, K=20	MATLAB	DS-4	95.99%	96.88%	95.00%
WW*k*NN, Features 90%	Threshold = 0.55, K=20	MATLAB	DS-4	95.99%	96.43%	95.50%
WW*k*NN, Features 90%	Threshold = 0.60, K=9	MATLAB	DS-4	95.28%	95.54%	95.00%
WW*k*NN, Features 90%	Threshold = 0.60, K=15	MATLAB	DS-4	95.28%	94.64%	96.00%
WW*k*NN, Features 90%	Threshold = 0.65, K=9	MATLAB	DS-4	95.05%	95.09%	95.00%
WW*k*NN, Features 90%	Threshold = 0.65, K=15	MATLAB	DS-4	95.05%	93.30%	97.00%
WW*k*NN, Features 90%	Threshold = 0.65, K=20	MATLAB	DS-4	95.05%	92.86%	97.50%
WW*k*NN, Features 75%	Threshold = 0.35, K=5	MATLAB	DS-4	93.40%	95.98%	90.50%
WW*k*NN, Features 75%	Threshold = 0.35, K=7	MATLAB	DS-4	93.40%	96.88%	89.50%
WW*k*NN, Features 75%	Threshold = 0.35, K=9	MATLAB	DS-4	93.40%	98.21%	88.00%
WW*k*NN, Features 75%	Threshold = 0.40, K=7	MATLAB	DS-4	93.63%	96.88%	90.00%
WW*k*NN, Features 75%	Threshold = 0.45, K=7	MATLAB	DS-4	94.10%	96.88%	91.00%
WW*k*NN, Features 75%	Threshold = 0.50, K=15	MATLAB	DS-4	95.05%	97.77%	92.00%
WW*k*NN, Features 75%	Threshold = 0.55, K=11	MATLAB	DS-4	95.52%	97.77%	93.00%
WW*k*NN, Features 75%	Threshold = 0.60, K=9	MATLAB	DS-4	95.05%	96.43%	93.50%
WW*k*NN, Features 75%	Threshold = 0.60, K=11	MATLAB	DS-4	95.05%	95.98%	94.00%
WW*k*NN, Features 75%	Threshold = 0.65, K=9	MATLAB	DS-4	95.05%	95.54%	94.50%
WW*k*NN, Features 75%	Threshold = 0.65, K=15	MATLAB	DS-4	95.05%	93.75%	96.50%
WW*k*NN, Features 50%	Threshold = 0.35, K=9	MATLAB	DS-4	92.69%	97.32%	87.50%
WW*k*NN, Features 50%	Threshold = 0.40, K=9	MATLAB	DS-4	93.16%	97.32%	88.50%
WW*k*NN, Features 50%	Threshold = 0.45, K=9	MATLAB	DS-4	93.40%	96.88%	89.50%
WW*k*NN, Features 50%	Threshold = 0.50, K=7	MATLAB	DS-4	93.63%	94.64%	92.50%
WW*k*NN, Features 50%	Threshold = 0.50, K=9	MATLAB	DS-4	93.63%	95.98%	91.00%
WW*k*NN, Features 50%	Threshold = 0.55, K=11	MATLAB	DS-4	94.10%	95.09%	93.00%
WW*k*NN, Features 50%	Threshold = 0.60, K=11	MATLAB	DS-4	94.34%	94.64%	94.00%
WW*k*NN, Features 50%	Threshold = 0.65, K=9	MATLAB	DS-4	93.63%	91.52%	96.00%

Table 4 - 29: Highest accuracy results for all methods tested

When all three levels of accuracy are considered, the methods that are classed as personalised models regularly outperform all others. For the WW*k*NN method this is true irrespective of the dataset used. This is very promising as the determination of the best SNPs is not very accurate in this study, and this method still has great potential for future development.

To allow for direct comparison amongst the results the following tables only take into account the results using 100% of the datasets, i.e. there is no further attribute reduction with the WW*k*NN method. This takes the 76 best results down to 28. Tables 4-12, 4-13 and 4-14

record the best performances of the classification methods arranged by type where G represents global methods, L local methods and P personalised methods.

The reduction of attributes with WW*k*NN showed a decrease in accuracy when increasing numbers of attributes that were removed. The accuracy stayed high for 100% and 95% levels, which reflect the relative accuracy of the attribute selection methods applied. It also reflects the expectation that a large number of SNPs will have a small individual effect, but in combination have a large effect, which is also seen in the PCA reductions Figures 4-4 to 4-9 .

Method	Variation	Implementation		Overall	Case	Control
Trees – J48	confidence factor = 0.1	WEKA	G	97.64%	97.32%	98.00%
Trees – J48	confidence factor = 0.25	WEKA	G	97.64%	97.32%	98.00%
Multilayer Perceptron		WEKA	G	96.23%	96.43%	96.00%
Multilayer Perceptron		NeuCom	G	96.21%	95.98%	96.50%
Rules	Decision Table – Best First	WEKA	G	95.28%	99.11%	91.00%
Support Vector Machine	Linear	NeuCom	G	95.04%	97.32%	92.50%
Multiple Linear Regression		NeuCom	G	92.44%	92.41%	92.50%
Bayesian	Bayesian Logistic Regression	WEKA	G	88.68%	97.32%	79.00%

Table 4 - 30: Accuracies of best performing global methods

Method	Variation	Implementation		Overall	Case	Control
Support Vector Machine	Radial Based Function	WEKA	L	96.23%	100.00%	59.00%
Evolving clustering method for classification		NeuCom	L	94.09%	95.54%	92.50%
Evolving Classification Function		NeuCom	L	70.77%	52.68%	91.00%

Table 4 - 31: Accuracies of best performing local methods

Only two methods resulted in an overall accuracy less than 90%, which as stated earlier is rather surprising. If these two methods are removed from comparison only one method, SVM with a Radial Based Function kernel, results in either the case or control accuracy levels falling below 90%. After removing this method, only one of the linear methods remains in contention, the evolving clustering method for classification.

Method	Variation	Implementation		Overall	Case	Control
WW*k*NN	Threshold = 0.45, K=15	MATLAB	P	97.41%	98.66%	96.00%
WW*k*NN	Threshold = 0.55, K=15	MATLAB	P	97.17%	96.88%	97.50%
WW*k*NN	Threshold = 0.60, K=11	MATLAB	P	97.17%	96.43%	98.00%
WW*k*NN	Threshold = 0.40, K=15	MATLAB	P	96.93%	98.66%	95.00%
WW*k*NN	Threshold = 0.50, K=11	MATLAB	P	96.93%	97.32%	96.50%
WW*k*NN	Threshold = 0.50, K=15	MATLAB	P	96.93%	97.77%	96.00%
WW*k*NN	Threshold = 0.65, K=9	MATLAB	P	96.93%	96.43%	97.50%
WW*k*NN	Threshold = 0.65, K=11	MATLAB	P	96.93%	95.95%	98.00%
WW*k*NN	Threshold = 0.35, K=5	MATLAB	P	96.46%	97.32%	95.50%
Nearest Neighbour	No distance weighting, K=1	WEKA	P	94.81%	94.64%	95.00%
Nearest Neighbour	Weight by 1/distance, K=1	WEKA	P	94.81%	94.64%	95.00%
Nearest Neighbour	Weight by 1-distance, K=1	WEKA	P	94.81%	94.64%	95.00%
W*k*NN	Threshold = 0.6, K=5	MATLAB	P	94.58%	94.64%	94.50%
W*k*NN	Threshold = 0.45, K=1	MATLAB	P	94.34%	95.09%	93.50%
W*k*NN	Threshold = 0.5, K=1	MATLAB	P	94.34%	95.09%	93.50%
W*k*NN	Threshold = 0.55, K=1	MATLAB	P	94.34%	95.09%	93.50%
W*k*NN	Threshold = 0.65, K=1	MATLAB	P	94.34%	95.09%	93.50%

Table 4 - 32: Accuracies of best performing personalised methods

The overall accuracy of the global methods ranges from 97.64% to 92.44% and averages 94.90%. In comparison to the personalised methods which range from 97.41% to 94.34% with an average of 95.85%, the global methods are less consistent and have a wider range. Comparing the accuracies for case and control classes between all three types of classification method the averages are similar but the personalised methods have a reduced range. Tables 4-15 and 4-16 show these comparisons with methods removed as stated above. This reduction in accuracy range indicates that the accuracy is more consistent amongst the various methods tested.

Type	Maximum	Minimum	Average	Range
Global	99.11%	92.41%	96.56%	4.70%
Local	95.54%			
Personalised	98.66%	94.64%	96.14%	4.02%

Table 4 - 33: Case class accuracies by type

Type	Maximum	Minimum	Average	Range
Global	98.00%	91.00%	94.93%	7.00%
Local	92.50%			
Personalised	98.00%	93.50%	95.50%	4.50%

Table 4 - 34: Control class accuracies by type

An additional advantage of the personalised methods tested here is that they were found to be computationally faster to execute. This may, or may not, have to do with executing them as native code in MATLAB as against the other environments; although considerable work has gone into the optimisation and speed of execution in the two publicly available packages. This is another area that would benefit from further development.

4.3 Random test sample

The top 76 method variations were tested against a random selection of 50 case samples and 50 control samples. These were selected at random so that the region of origin would not be a factor and to give a more accurate indication of the relative accuracies previously obtained using samples from only two regions. Two test datasets were constructed based on the SNPs in DS-4 and DS-5 as these datasets contributed to the best accuracies recorded previously. Tables 4-17 and 4-18 show the accuracies for each method tested including only the results for WW*k*NN that did not add any further attribute reduction. It can be seen from these results that the DS-4 based test dataset out performed the one based on DS-5 on every test.

The relative accuracies of each test were lower than their counterparts in the previous section, but remained consistent with very little change in the ranking of any one method. There was a difference in the relative merits of different threshold levels for WW*k*NN and this should be investigated in further research.

The accuracy levels obtained with this test sample are more consistent with expectation but are still relatively high with 8 variations achieving an overall accuracy of over 90%. In the dataset based on DS-4 there was one method, WW*k*NN with a threshold of 0.55 and a *k* value of 15 that achieved a case accuracy of 100%. This is offset by a control sample accuracy for the same method of 82%. This does focus the accuracy on the case (diseased) samples where more accuracy is needed in medical applications, but a decision on the level of acceptable penalty paid in terms of control accuracy needs to be set and this can only be done by the medical profession. In a computational sense the higher the class accuracy the better, but the relative importance of each class is dependent on the specific application of the model and the incorporation of other relevant data to the overall decision making process.

Method	Variation	Implementation		Overall	Case	Control
Trees – J48	confidence factor = 0.1	WEKA	G	92.00%	88.00%	96.00%
Trees – J48	confidence factor = 0.25	WEKA	G	92.00%	88.00%	96.00%
Multilayer Perceptron		WEKA	G	87.00%	82.00%	92.00%
Multilayer Perceptron		NeuCom	G	87.00%	84.00%	90.00%
Rules	Decision Table – Best First	WEKA	G	90.00%	94.00%	94.00%
Support Vector Machine	Linear	NeuCom	G	86.00%	90.00%	82.00%
Multiple Linear Regression		NeuCom	G	62.00%	68.00%	56.00%
Bayesian Logistic Regression		WEKA	G	84.00%	90.00%	78.00%
Support Vector Machine	Radial Based Function	WEKA	L	75.00%	56.00%	94.00%
Evolving clustering method for classification		NeuCom	L	87.00%	84.00%	90.00%
Evolving Classification Function		NeuCom	L	70.00%	82.00%	58.00%
WW*k*NN	Threshold = 0.45, K=15	MATLAB	P	82.00%	78.00%	86.00%
WW*k*NN	Threshold = 0.55, K=15	MATLAB	P	91.00%	100.00%	82.00%
WW*k*NN	Threshold = 0.60, K=11	MATLAB	P	89.00%	98.00%	80.00%
WW*k*NN	Threshold = 0.40, K=15	MATLAB	P	82.00%	76.00%	88.00%
WW*k*NN	Threshold = 0.50, K=11	MATLAB	P	84.00%	84.00%	84.00%
WW*k*NN	Threshold = 0.50, K=15	MATLAB	P	87.00%	88.00%	86.00%
WW*k*NN	Threshold = 0.65, K=9	MATLAB	P	88.00%	98.00%	78.00%
WW*k*NN	Threshold = 0.65, K=11	MATLAB	P	88.00%	98.00%	78.00%
WW*k*NN	Threshold = 0.35, K=5	MATLAB	P	91.00%	92.00%	90.00%
Nearest Neighbour	No distance weighting, K=1	WEKA	P	86.00%	86.00%	86.00%
Nearest Neighbour	Weight by 1/distance, K=1	WEKA	P	86.00%	86.00%	86.00%
Nearest Neighbour	Weight by 1-distance, K=1	WEKA	P	86.00%	86.00%	86.00%
W*k*NN	Threshold = 0.6, K=5	MATLAB	P	88.00%	90.00%	86.00%
W*k*NN	Threshold = 0.45, K=1	MATLAB	P	88.00%	90.00%	86.00%
W*k*NN	Threshold = 0.5, K=1	MATLAB	P	88.00%	90.00%	86.00%
W*k*NN	Threshold = 0.55, K=1	MATLAB	P	88.00%	90.00%	86.00%
W*k*NN	Threshold = 0.65, K=1	MATLAB	P	88.00%	90.00%	86.00%

Table 4 - 35: Model accuracies of randomly selected test sample based on DS-4 SNPs

Method	Variation	Implementation		Overall	Case	Control
Trees – J48	confidence factor = 0.1	WEKA	G	87.00%	84.00%	90.00%
Trees – J48	confidence factor = 0.25	WEKA	G	89.00%	86.00%	92.00%
Multilayer Perceptron		WEKA	G	71.00%	62.00%	80.00%
Multilayer Perceptron		NeuCom	G	70.00%	76.00%	64.00%
Rules	Decision Table – Best First	WEKA	G	83.00%	84.00%	82.00%
Support Vector Machine	Linear	NeuCom	G	85.00%	86.00%	84.00%
Multiple Linear Regression		NeuCom	G	49.00%	50.00%	48.00%
Bayesian Logistic Regression		WEKA	G	83.00%	88.00%	78.00%
Support Vector Machine	Radial Based Function	WEKA	L	56.00%	14.00%	98.00%
Evolving clustering method for classification		NeuCom	L	63.00%	56.00%	70.00%
Evolving Classification Function		NeuCom	L	50.00%	100.00%	0.00%
WW*k*NN	Threshold = 0.45, K=15	MATLAB	P	77.00%	68.00%	86.00%
WW*k*NN	Threshold = 0.55, K=15	MATLAB	P	84.00%	86.00%	82.00%
WW*k*NN	Threshold = 0.60, K=11	MATLAB	P	82.00%	90.00%	74.00%
WW*k*NN	Threshold = 0.40, K=15	MATLAB	P	73.00%	56.00%	90.00%
WW*k*NN	Threshold = 0.50, K=11	MATLAB	P	75.00%	72.00%	78.00%
WW*k*NN	Threshold = 0.50, K=15	MATLAB	P	80.00%	76.00%	84.00%
WW*k*NN	Threshold = 0.65, K=9	MATLAB	P	80.00%	86.00%	74.00%
WW*k*NN	Threshold = 0.65, K=11	MATLAB	P	80.00%	90.00%	70.00%
WW*k*NN	Threshold = 0.35, K=5	MATLAB	P	73.00%	66.00%	80.00%
Nearest Neighbour	No distance weighting, K=1	WEKA	P	75.00%	72.00%	78.00%
Nearest Neighbour	Weight by 1/distance, K=1	WEKA	P	75.00%	72.00%	78.00%
Nearest Neighbour	Weight by 1-distance, K=1	WEKA	P	75.00%	72.00%	72.00%
W*k*NN	Threshold = 0.6, K=5	MATLAB	P	80.00%	84.00%	76.00%
W*k*NN	Threshold = 0.45, K=1	MATLAB	P	80.00%	84.00%	76.00%
W*k*NN	Threshold = 0.5, K=1	MATLAB	P	80.00%	84.00%	76.00%
W*k*NN	Threshold = 0.55, K=1	MATLAB	P	80.00%	84.00%	76.00%
W*k*NN	Threshold = 0.65, K=1	MATLAB	P	80.00%	84.00%	76.00%

Table 4 - 36: Model accuracies of randomly selected test sample based on DS-5 SNPs

4.4 Conclusion

In any attempt to compare classification methods it is necessary to carefully and accurately prepare the data so that any shortcomings in the data are minimised, for example missing values, and so that the data is in a format that is consistent with the method chosen, i.e. does the classification method work on numeric or nominal data, so the method itself is not hampered unnecessarily by the composition of the dataset. Of particular importance is the reduction of attributes from raw data to a size that is sufficiently small to enable existing tools

to analyse it. To do this for the current investigations it was necessary to prepare five reductions, from which results showed that one particular reduction outperformed all the others on a consistent basis. This dataset was prepared using the simplest of measures, which was to rank the difference in composite mean for each SNP (attribute) between the case (diseased) and control (non-diseased) classes.

The results obtained from the various models tested shows that the way the data is reduced from its original raw form into that used to model susceptibility to the disease has a large effect on the overall and individual class accuracies obtained. Of the models tested those that have been here defined as personalised models outperformed those of each of the other categories, global and local methods, with only a single exception which deserves further investigation. The personalised methods also outperformed all others in terms of execution time. When methods were tested against a random selection of samples the same patterns emerged with the personalised methods performing more consistently than either the local or global methods.

Chapter 5 Discussion of Results and Findings

Focusing on model development has been a large part of many studies utilising SNPs data as described in chapter 2. Unfortunately none of the studies done with the WTCCC study data has undertaken a comparison of methods for modelling the data, nor have they looked into different ways of determining the significance of any one SNP or groups of SNPs. These comparisons and investigations form an integral part of the research undertaken in this study.

A number of questions arise when comparing methods; how can they be compared accurately, what constitutes the "best" result, and is the way the method is implemented a factor are but a few of these. When a focus on knowledge discovery is utilised previously accepted assumptions can be challenged and either refuted or verified. Much of previous work investigating this data has utilised purely statistical methods. Questions arise as to the application of data mining methods and approaches to modelling and analysis. Is it better, worse, or no different from plain statistics?

Decisions on how the data is processed can affect later analysis especially where attribute reduction is being employed. Previous studies that have used the WTCCC data start from the reduced SNPs identified in the original investigation which used statistical procedures alone determining variation within each SNP one at a time and did not look at any SNP interaction.

5.1 The data

The biggest challenge with a large dataset is to manipulate it in such a way as to maintain information value whilst reducing the overall size. To do so a number of decisions are made to determine the worth of each component of the data. In data mining terms this is referred to as attribute reduction. There are numerous methods of achieving this and the selection of reduction method is as important as the reduction itself in many ways. Each method will have its strengths and weaknesses and the goal is to determine which method, or combination of methods, are best suited to the particular type of data such that the effect of weaknesses in the methods are minimised.

5.1.1 Manipulating the data

The first obstacle to handling datasets of the size of the original WTCCC study is that many existing software programs do not have the capacity to store the data in a way that is easily manipulated. The only option for storage and initial manipulation of the SNP data was to place it in a database. Even so it was not possible to do all transformations within the database. Many of the data manipulation functions within the database were very useful, for example the ability to select a group of samples by sample number, or region, or just summative data on all occurrences of a single SNP. These selection tools were used to extract the data pertaining to the samples from the second and third most populous regions to enable further investigation on this group alone, and the extraction of the randomly selected control samples, as well as applying the restrictions placed on the data by the quality control procedures applied by the original WTCCC study.

Once data was extracted from the database the next manipulation challenge was to transpose it from a single SNP measurement per row layout to a single sample per row layout with all SNPs of that sample in the same row. This, unfortunately, was by necessity done manually and even with the assistance of transposition abilities in Microsoft Excel this took a long time to perform. In this format the data could be input into other software packages for further analysis. WEKA takes data in this format and so does Minitab[2] and both packages were used to do the initial reduction of SNPs. When using Minitab, there is a limit to the amount of data a single worksheet can take so to accommodate this the data was split over 8 worksheets to enable all original SNPs to the tested for correlation. Even so it took three days to get all results due to the limitations of calculation within the packages.

It is not very surprising that existing software packages do not have the capacity to deal effectively with large datasets, unless very large and expensive server based options are employed. Even so much would still need to be done manually in the initial cases and appropriate code developed to manipulate the data effectively in a more automated manner. This dataset is not considered to be excessively large in comparison to others obtained by similar chips. Seven of the diseases on the original WTCCC study were investigated using a

[2] Minitab is a statistical analysis package produced my Minitab Inc.

different chip which has a 500k capacity, in comparison to the chip used in this section of the study having only a 15k capacity. There is likely in the future to only be more data analysed per chip (Cortes & Brown, 2010) with these increasingly large raw datasets needing to be incorporated with data from previous research done with chips that have a much lower capacity. There comes a challenge in terms of integrating the separate sets of data and dealing with what is now missing data points where different SNPs were not included on previous chips. As the chips are not inexpensive to manufacture and can only be used once, it is inevitable that from a research as well as a manufacturing point of view being able to test as much on a single chip as possible is more cost effective. Like the different diseases investigated in this WTCCC study many SNPs are important to only a small groups of diseases or even to only one disease and separating out the ones that pertain to a specific disease is a non-trivial exercise and is likely to get more challenging as the volume of data increases.

5.1.2 Dataset reduction

Not only were the methods of modelling the data investigated in this study, but the comparative effects of different attribute reduction methods. The first reduction made was to limit the sample size. To do this the case data from the second and third most populous areas, Midlands and Southeastern (*sic*) regions, were chosen as they represented nearly 200 samples making it simpler to balance with 200 control samples. The Eastern region which contained the highest number of case samples was not used as this was deemed to be rather disproportionate in terms of geography and the total number of sufferers represented. Although interesting, the investigation of the effect of geographical region on the distribution of disease sufferers was beyond the scope of this study.

The second type of reduction involved reducing the number of SNPs to be tested. As stated above only a small number of the original SNPs are pertinent to any one disease and these need to be found before accurate modelling can take place. It was not the aim of this study to form an ideal set of SNPs but to get a rough set that performed well enough and was robust enough to cope with the natural variations that exist amongst samples. It was known from previous research (D'Addabbo et al., 2011) that the most likely situation was that a large

number of SNPs each having a small individual effect, but greater collective effect would be significant. With this in mind the cut off points for both the correlations and differential comparisons used when forming the datasets were taken to be higher than would otherwise been deemed suitable. Ideally the *p*-value used in the case of measured correlation would have been lower demonstrating an increased significance of the SNP, but as this investigation was assuming little to no prior knowledge of the data and its behaviour in such tests a higher value was deemed appropriate.

The different methods employed to form the reduced datasets have all been used successfully in many other situations and are common to this type of exercise, refer to table 2-2 in chapter 2. The value of ranking the data as done for DS-1 and DS-2 has shown that the score value of the SNP is not to be treated as if it were a linear measurement as would the measurement of height for example where a larger number represented a taller person, the higher the SNP score did not indicate a more significant SNP. The SNPs found in the midrange were more prevalent in the final datasets than were those in the groups at either end. From Fig 4-4 in chapter 4 it can be seen that this region also contains the largest variations between mean score values for each SNP between case and control samples.

Once the four datasets were finalised a comparison of their content revealed a number of overlaps as stated in chapter 4. This was expected for the datasets taken from the same source selections, that being DS-1 with DS-2 and DS-3 with DS-4. Of interest was to note the overlaps between these two groups. The highest level of overlap was between DS-2 and the DS-3/DS-4 combination with a total of 7 SNPs in common. DS-2 also had the highest contribution to the final makeup of DS-5, the dataset formed by combining the other 4 datasets and performing a further principle component analysis. A number of the SNPs that appeared in more than one dataset also appeared in DS-5. This could easily give indication to their level of importance, but no level of certainty can be attached at this stage.

5.1.3 Comparison with published data

There are a number of sources of published data listing SNPs that have been found to be significant to Multiple Sclerosis. A total of 175 from 15 studies have been compiled and table 5-1 shows the breakdown of each dataset in relation to these previously published results. The studies used in this comparison were published on the GWAS central website coded here as HGVR and a number (GWAS Central), the OMMIM website coded here as OMMIM (Online Mendelian Inheritance in Man), a precious study into the effects of genetics in predicting MS coded here as TH-029 (Sawcer, 2010) and WTCCC coded here as TH-064 (The Welcome Trust Case Control Consortium & The Australo-Anglo-American Spondytitis Consortium, 2007). SNPs not found in these published studies are listed in table 5-2.

SNP	Source	Datasets present in
rs10201872	HGVRS1735	DS-4
rs10411936	HGVRS1871	DS-3
rs10466829	HGVRS1735	DS-4
rs10492972	HGVRS179	DS-1
rs1062158	HGVRS1735	DS-4
rs1077667	HGVRS1735	DS-2
rs10866713	HGVRS1871	DS-2
rs10936599	HGVRS1735	DS-4
rs1109670	HGVRS173	DS-2
rs11129295	HGVRS1735	DS-2
rs11154801	HGVRS1735	DS-4
rs1132200	TH-064	DS-4
rs11574422	TH-064	DS-1
rs11581062	HGVRS1735	DS-1
rs11865121	HGVRS643	DS-1
rs12025416	HGVRS1396	DS-3
rs12048904	HGVRS1735	DS-3
rs12143502	OMIM	DS-3
rs12212193	HGVRS1735	DS-4
rs12466022	HGVRS1735	DS-3
rs1250540	HGVRS643	DS-3, DS-4
rs1250542	HGVRS1871	DS-3, DS-4
rs12708716	TH-029	DS-4
rs12722489	HGVRS1871, OMMIM	DS-4
rs132630295	OMIM	DS-4
rs1335532	HGVRS650	DS-4

rs1356122	HGVRS102	DS-4
rs140522	HGVRS1735	DS-1
rs1458175	HGVRS173	DS-1
rs151719	HGVRS102	DS-1
rs1529316	HGVRS173	DS-1
rs17009792	TH-064	DS-1
rs17066096	HGVRS1735	DS-2
rs170934	HGVRS1871	DS-1, DS-2
rs17174870	HGVRS1735	DS-2
rs17368528	OMIM	DS-3
rs1738074	HGVRS1735, HGVRS1871	DS-2
rs17445836	HGVRS643, TH-029	DS-3, DS-4
rs1755289	HGVRS173	DS-2
rs17824933	HGVRS643, TH-029	DS-2
rs1790100	HGVRS643	DS-2
rs1800437	TH-064	DS-3
rs1800693	HGVRS643, TH-029	DS-3
rs180515	HGVRS1735	DS-4
rs1841770	HGVRS173	DS-2
rs2019960	HGVRS1735	DS-4
rs2040406	HGVRS1036	DS-2
rs2104286	HGVRS643, HGVRS650, OMMIM,TH-029	DS-1, DS-2
rs2119704	HGVRS1735	DS-1, DS-2
rs2150702	HGVRS1871	DS-2
rs2157082	HGVRS102	DS-2
rs2248359	HGVRS1735	DS-2
rs2281868	TH-064	DS-1, DS-2
rs2283792	HGVRS1735	DS-1, DS-2
rs228614	HGVRS1735	DS-4
rs2293152	HGVRS1871	DS-2
rs2300603	HGVRS1735	DS-2
rs2300747	HGVRS1871, HGVRS643, TH-029	DS-2
rs2303137	OMIM	DS-2
rs2303759	HGVRS1735	DS-2
rs233100	HGVRS1735	DS-3
rs241427	HGVRS102	DS-3
rs2503875	HGVRS1063	DS-3
rs2523393	HGVRS643	DS-3
rs2523485	HGVRS102	DS-3
rs2546890	HGVRS1735, HGVRS1871	DS-3
rs2596437	HGVRS102	DS-3

rs2596517	HGVRS102	DS-3
rs2596571	HGVRS102	DS-3
rs2621383	HGVRS102	DS-3
rs2647046	HGVRS102	DS-3
rs2681424	HGVRS1871	DS-3
rs2736177	HGVRS102	DS-3
rs2744148	HGVRS1735	DS-3
rs281380	HGVRS1735	DS-3
rs2857154	HGVRS102	DS-3
rs2857161	HGVRS102	DS-3
rs2894249	HGVRS102	DS-3
rs2905747	HGVRS102	DS-3
rs290986	HGVRS1735	DS-3
rs307896	HGVRS1735	DS-4
rs3087456	OMIM	DS-4
rs3095238	HGVRS102	DS-4
rs3115537	HGVRS102	DS-4
rs3129768	HGVRS102	DS-4
rs3129889	HGVRS1871	DS-4
rs3129900	HGVRS102	DS-4
rs3129932	HGVRS102	DS-4
rs3129934	HGVRS102, HGVRS202	DS-4
rs3130287	HGVRS102	DS-3, DS-4
rs3130532	HGVRS102	DS-3, DS-4
rs3130952	HGVRS102	DS-3, DS-4
rs3131294	HGVRS102	DS-3, DS-4
rs3131631	HGVRS102	DS-3, DS-4
rs3135338	HGVRS975	DS-3, DS-4
rs3135377	HGVRS102	DS-3, DS-4
rs3135388	HGVRS102	DS-3, DS-4
rs34536443	TH-029	DS-4
rs354033	HGVRS1735	DS-4
rs3745672	HGVRS1063	DS-4
rs3780792	HGVRS1063	DS-4
rs3817511	TH-064	DS-3
rs388706	TH-064	DS-4
rs397020	HGVRS173	DS-4
rs3997982	HGVRS102	DS-4
rs4075958	HGVRS1735	DS-4
rs4149584	HGVRS643	DS-4
rs4285028	HGVRS1735	DS-4

rs4409785	HGVRS1735	DS-4
rs4410871	HGVRS1735	DS-4
rs4680534	HGVRS643	DS-4
rs486416	HGVRS102	DS-4
rs4902647	HGVRS1735	DS-4
rs4939490	HGVRS1396	DS-4
rs6062314	HGVRS1735	DS-3, DS-4
rs6074022	HGVRS1871, HGVRS650	DS-4
rs630923	HGVRS1735	DS-3
rs644045	HGVRS102	DS-4
rs6470147	TH-064	DS-3, DS-4
rs651477	HGVRS173	DS-3, DS-4
rs6604026	HGVRS650	DS-4
rs669607	HGVRS1735	DS-4
rs6718520	HGVRS1871	DS-4
rs6871748	OMIM	DS-4
rs6896969	HGVRS643	DS-4
rs6897932	HGVRS643, OMMIM, TH-029, TH-064	DS-4
rs6936204	HGVRS102	DS-3, DS-4
rs6984045	HGVRS650	DS-4
rs70384	TH-029	DS-4
rs703842	HGVRS650	DS-4
rs7090512	HGVRS1735	DS-4
rs7191700	HGVRS1871	DS-1
rs7194	HGVRS102	DS-1
rs7238078	HGVRS1735	DS-4
rs7255066	HGVRS1735	DS-3
rs7382297	HGVRS102	DS-3
rs744166	HGVRS975	DS-3
rs7453920	HGVRS102	DS-3
rs756699	HGVRS1735	DS-1
rs7592330	HGVRS1871	DS-1
rs7595037	HGVRS1735	DS-1
rs760293	HGVRS102	DS-1
rs7672826	HGVRS173	DS-3
rs771767	HGVRS1735	DS-3
rs78778622	OMIM	DS-4
rs7923837	HGVRS1735	DS-4
rs802734	HGVRS1735	DS-4
rs8049603	HGVRS650	DS-4
rs806321	HGVRS1735	DS-4

rs8070463	HGVRS1871	DS-4
rs874628	HGVRS1735	DS-3
rs882300	HGVRS643	DS-3
rs908821	HGVRS173	DS-3, DS-4
rs910049	HGVRS102	DS-3, DS-4
rs9260489	HGVRS1871	DS-4
rs9267971	HGVRS102	DS-4
rs9268402	HGVRS102	DS-4
rs9268557	HGVRS102	DS-4
rs9268560	HGVRS102	DS-4
rs9268877	HGVRS102	DS-4
rs9270986	HGVRS102	DS-4
rs9271366	HGVRS1063	DS-4
rs9275572	HGVRS102	DS-4
rs9275765	HGVRS102	DS-4
rs9275772	HGVRS102	DS-4
rs9275793	HGVRS102	DS-4
rs9276431	HGVRS102	DS-4
rs9282641	HGVRS1735	DS-4
rs931555	HGVRS1396	DS-4
rs9321490	HGVRS1735	DS-4
rs9368716	HGVRS102	DS-4
rs9469220	HGVRS102	DS-4
rs9523762	HGVRS173	DS-4
rs9596270	HGVRS1871	DS-4
rs9657904	HGVRS1036	DS-4

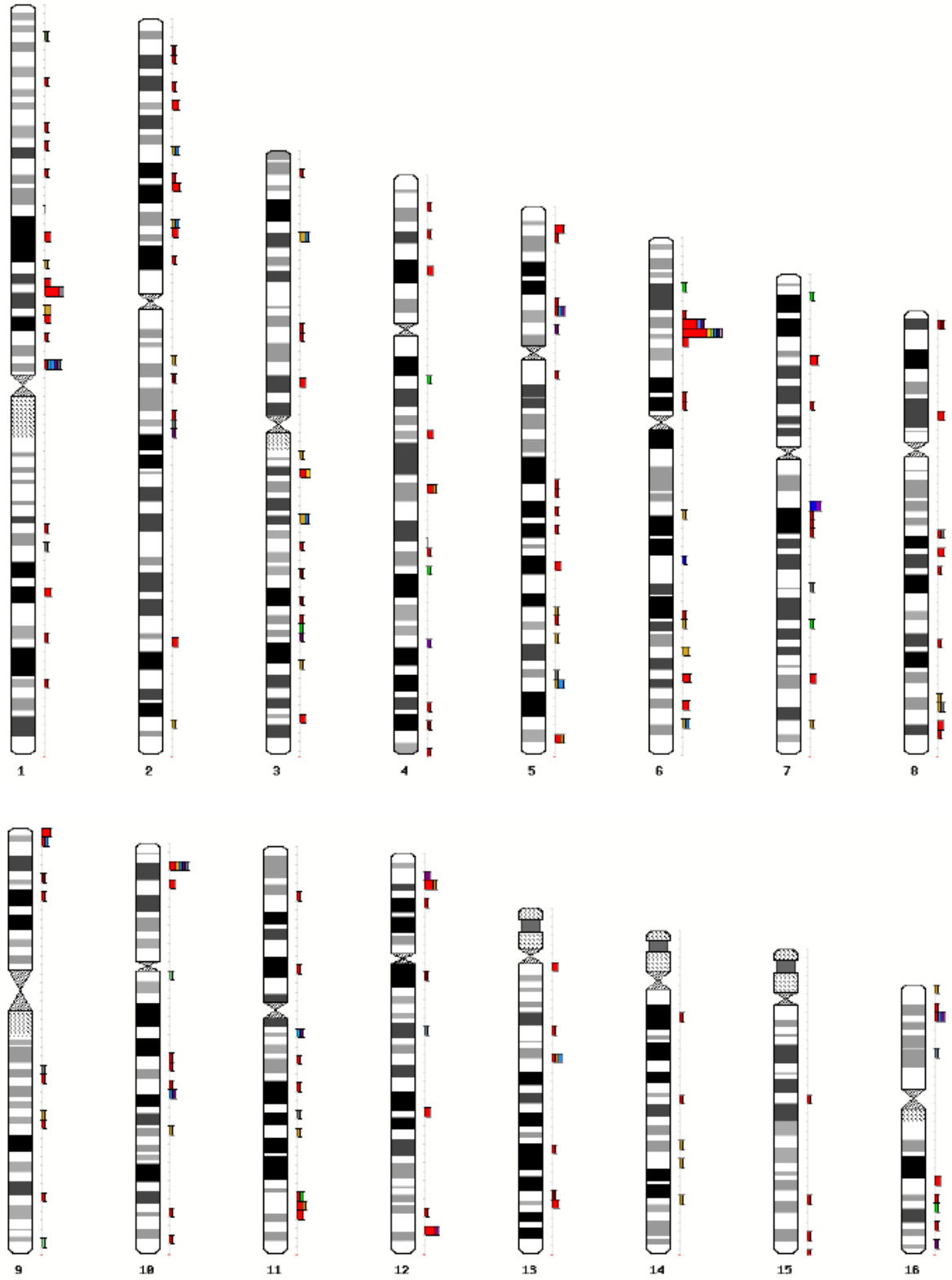
Table 5 - 3: Comparison of SNP sources

Number of new SNPs per dataset		SNPs not listed in previous studies
1	DS-1	Gender
0	DS-2	
1	DS-3	BRCA2-1420
5	DS-4	rs10195632
		NCBI35_X_104984443
		rs10146906
		BRCA1-695
		BRCC36-74

Table 5 - 4: SNPs present in reduced datasets but not in published research

There is a surprisingly small number of SNPs that do not correspond with previous studies into Multiple Sclerosis. Given the approach in this study, it was expected that a much greater separation between this studies results and those preciously undertaken would be present. It was very pleasing, and again somewhat unexpected, to note that all SNPs identified in the previous research are represented amongst the datasets. This gives rise to the conclusion that the data mining approach to attribute selection is a valid one and does not hinder any other biological approach to the data. It would be interesting to investigate how incorporating biological data not represented as SNPs into the selection process would affect the identification of significant attributes to be used in modelling susceptibility, but this comes unfortunately beyond the scope of this investigation.

When comparing the chromosomal location of SNPs selected with that of the published data, they are very similar in makeup but do differ in a number of items. For example the number of SNPs found in this study on the X chromosome is larger than previous studies, and the number found on chromosome 6 is lower. Although not strictly a direct comparison, the best pictorial representation of the chromosomal locations of previously identified SNPs can be found in Figure 5-1, taken from the GWAS central results where each colour represents a different study of origin, and that of this study in Figure 5-2. From Figure 5-2 it can also be seen that on four occasions DS-4 contains a higher number of SNPs on a specific chromosome. On two of these chromosomes, 9 and X, the difference is more pronounced with chromosome X showing a more dramatic increase. The reasons behind the difference on chromosome X deserves further investigation as the reasons may be more biological in nature rather than statistical. None of the reduction processes found any significant SNPs on chromosome Y. As these two chromosomes, X and Y, determine the gender and it is known that more females develop MS than males then this difference may be both biological and statistical.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

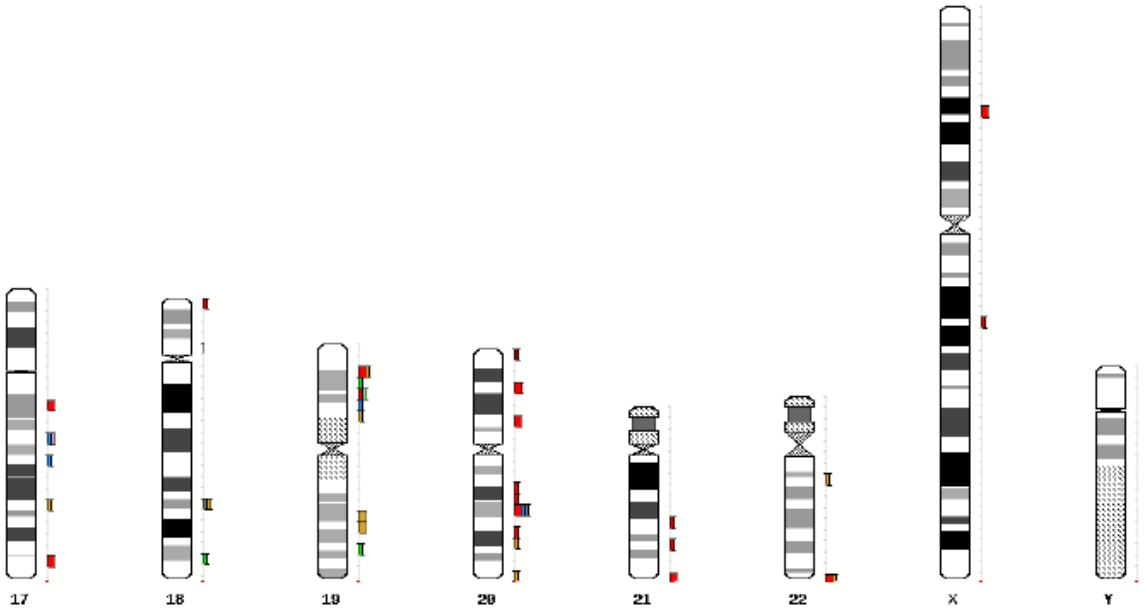


Figure 5 - 3: Chromosomal locations of identified SNPs

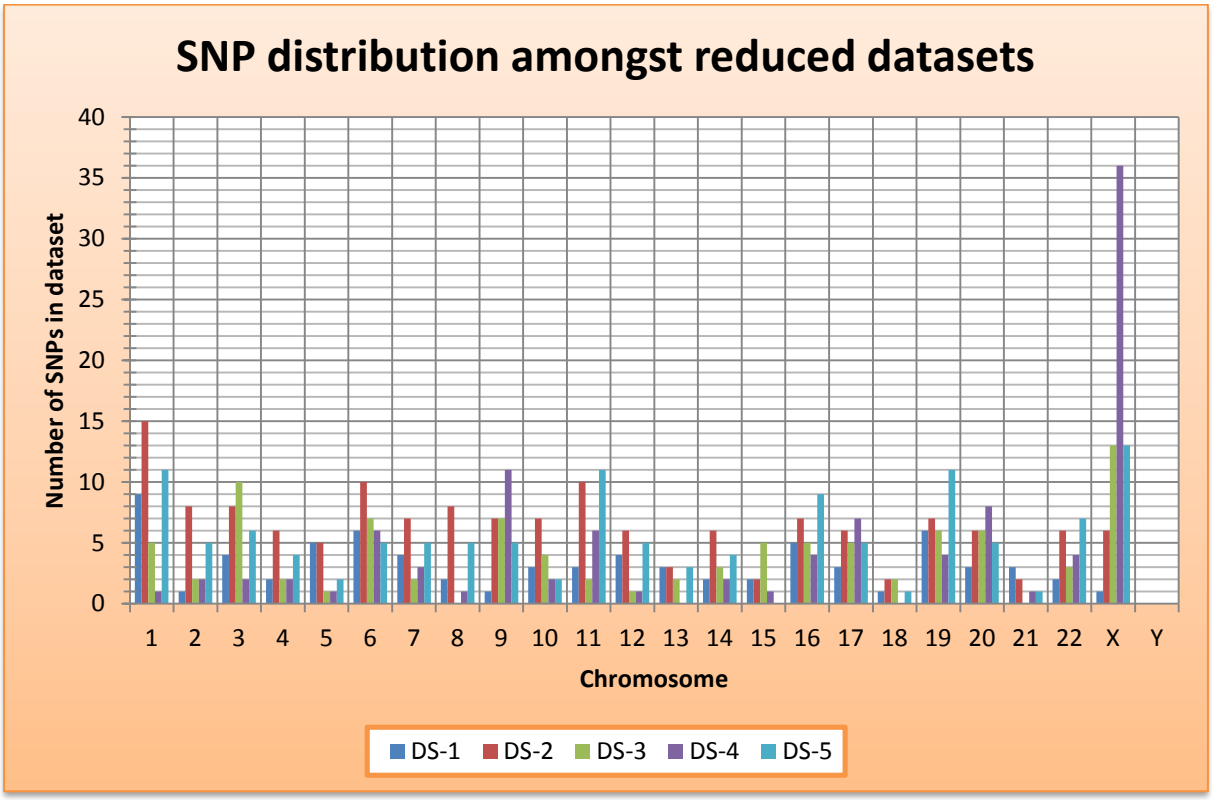


Figure 5 - 4: Chromosomal distribution of SNPs in reduced datasets

5.2 Method comparison

To compare methods accurately it must be ensured that the comparison is valid and useful for the investigation. To do so with any degree of certainty the overall and class accuracies were the only ones found to be valid across all platforms in this study.

5.2.1 Shortcomings of comparison

There are numerous ways of determining which method is the "best" one, but these are not always implemented in the same way in different packages. Some, like WEKA, are more defined in terms of what analysis statistics they provide and how they are implemented. Others, for example the code-only implementations developed at, have very little and what in comparison could be thought of as more simplistic measurements. As WEKA is a freely available program and has been used in a number of published studies (Kharbat et al., 2008; Shah & Kusack, 2007), it has an enhanced range of comparative statistics available to the researcher. With WEKA being part of an open source development environment many developers from around the world have now contributed to its development and the implementation of many of its analysis systems. This has enhanced and greatly enlarged the development community and expertise for this product.

With a limited amount of directly comparable statistics to use, decisions were made to utilise the overall, case and control accuracy measures. These could be calculated for all software environments and are well accepted as measures of relative model "success". As stated in chapter 4, the use of all three accuracy measures in combination results in an additional consideration – that of the relative weights to give each measure to a final indication of "best" method. Some of this is dependent upon the implementation environment for the optimised best method, and not just on a mathematical probability weighting. In this study the determinant of overall accuracy has been tempered by the relative merits of case (diseased) and control class accuracies. As the end implementation environment is medical it is deemed to be of greater value to have higher class accuracy for the case group as this refers to the person's susceptibility to develop the disease. An error in ranking someone with a higher risk factor for susceptibility than they may actually have is preferred to the alternative of not ranking their risk high enough. The exact weighting to put to each level of accuracy is likely to depend on the medical community more than any other group, although from the results of

the WW*k*NN method where thresholding is applied a "lean" towards the case rather than control samples yields a higher level of accuracy.

5.2.2 Comparison of methods for datasets based on samples from two regions

5.2.2.1 The datasets

The output given by the various execution environments (WEKA, NeuCom, and MATLAB), was recorded in text files and the comparative statistics transferred to a spreadsheet. The abilities of the spreadsheet software to perform conditional formatting enabled the tracking of "best performing" in each category to be done dynamically as each new result was entered. The end result of this can be found in the colour coding of results presented in the tables at the end of chapter 4. The coding system allows both a comparison of methods and datasets to be performed in one sheet and displayed in a single table. The shaded cells refer to calculations determining best and worst performing of each of the datasets with a specific method variation and are calculated horizontally. The coding relating to the formatting of the cell text refers to comparisons of performance within a dataset which are calculated vertically.

From the results in tables 4-13 to 4-16, it is easily seen that DS-4 consistently outperforms all other datasets, where only DS-5 records an accuracy level higher than that of DS-4 for any one test. The predominance of green shaded cells reveals that DS-4 gives the highest overall accuracy in a vast majority of occasions. The correspondingly high number of blue shaded cells reveals that DS-4 also achieves higher individual class accuracy, especially that of the case class. In the two occasions where DS-5 performed better than DS-4, the difference is very small. As DS-4 makes up the second highest contributing proportion of DS-5 it may be possible that the presence of so many DS-4 SNPs in DS-5 forms a major contributor to this result.

Both datasets, DS-1 and DS-2, performed the worst of all five datasets. They both consistently appear to have shaded cells of an apricot colour, indicating the worst performing

option for that particular method and variation. It is interesting to note that these two data sets were obtained using similar methods and came from the stratified group selection process. In a number of cases these data sets return accuracy levels far below 75% and in some cases less than 50% for overall accuracy. This adds further weight to the conclusion that score cannot be considered in the same way as other numerical measurements.

From these results it can be argued that performing the selection process on the complete set of original data is more beneficial, and ultimately accurate, than by attempting to reduce the data initially by other means. (DS-4 was generated from an ordered list of relative differences in mean scores for each SNP, which was the simplest of the selection processes applied). The process of applying analysis, even as simple as a correlation test, can be a laborious and time-consuming process with this amount of raw data. The raw data used in this study is by comparison small against that obtained from other microarray chips. This study uses a 15K chip and is comparable to other data within the original study that uses a 500K chip. This larger chip is more common and with the development of larger and larger chips the size of the original data pool is only going to increase. The jump in data size resulting from the larger capacity of the chips will make the process of analysing it and storing it in an efficient manner a more difficult one. Possible solution to this would be a mechanism by which error analysis tools interacted with the data within a database, allowing the data to be stored in a way that is sensible and accessible for multiple purposes and by multiple programs. Added to this is the task of aligning data from newer tests (and chips) that contain SNPs not present in previous studies, or not considered in previous studies, and how they interact with other SNPs.

Of all the methods tested only two had the capacity to reduce the number of attributes used in the final model. They were the J48 tree which allows pruning; both for accuracy of the model produced and increase in speed of execution. Pruning these trees allows for a reduction, if not elimination, of repeatedly testing the same thing to determine the same result. The other method is the WW*k*NN method developed at KEDRI. One of the parameters that can be set in this method is the number of attributes to be used. To employ this feature correctly the attributes within the data file need to be listed in PCA order, that is from the most significant to the least significant contributing factor. When the number of

attributes to be used is entered this is counted from the beginning of the file and thus it is important that the attributes are in the correct order. As the various datasets were not of the same size a percentage reduction was used so that a more direct comparison could be made. The results of reducing the attributes in WW*k*NN did not significantly affect the accuracy levels in any way that would lead to the conclusion that the datasets could be further reduced without loss of accuracy. The tables in chapter 4 only reflect the results of testing with 100% of the attributes for two reasons. Firstly, so that a direct comparison can be made with the other methods which used the same size of dataset and; secondly, there was a noted drop off and accuracy when attribute numbers less than 95% of the original were used.

5.2.2.2 The methods

At this stage of the investigation it was not expected that any method would perform well enough to achieve much over 85% in accuracy and definitely not over 95%. It was surprising to note that a large number of tests resulted in very high accuracy levels, over 90%. There was also a group of tests that resulted in very low accuracy levels, which was what was expected. The aim of performing so many tests and variations within the chosen methods was to determine which methods could be eliminated from further investigation and which variations were worth a second look. To this end the results were very successful.

With 51 variations found to have accuracies over 95% this gives a sizable group to investigate further. How much optimisation is needed, and in what areas? Does one level of optimisation actually result in a lowered accuracy level? Can methods be used together to achieve an even better result? These are but a few of the questions further investigation would answer.

The list of 76 best variations was determined by grading against overall accuracy and including the best of each of the classification methods including results where multiple variations achieved the same results. Doing this allows a comparison amongst the methods which results in the observation that with the exception of the tree method (J48) the personalised methods consistently achieved the highest accuracy levels, for both overall and class accuracy. The level of success that J48 achieved was very surprising as it was not

expected that a global method would compete so well with the personalised ones. It is possible that the pruning of the trees developed by J48 achieves a level of optimisation that allows for the resulting high accuracy levels.

For some of the tests the overall accuracy level was good, but when the class accuracy levels are included the relative worth of the overall accuracy is downgraded. This is due to the drop in performance of a single class against the other. For example for the best overall results from the SVM localised method an overall accuracy of 96.23% was achieved, with a case accuracy of 100% and a control accuracy of 59%. This means that the method classified all of the case class (diseased) correctly, but only 59% of the control class. This result is not acceptable in terms of classification accuracy independent of the application area, and particularly in medicine which is the application area of this study. It could be argued that to correctly classify the samples (individuals) with the disease correctly is preferable, and this is so. But this should not be done at the expense of the classification of control or non-diseased samples. It is self-explanatory that the aim of high accuracy is best achieved by having the individual class accuracies as high as possible, resulting in a correspondingly high overall accuracy. Yet this can be overlooked when the reported overall accuracy is high, and a method or result is accepted when it should not be.

For the methods resulting in a low accuracy level, between 50% and 55%, it was common for all samples to be classified as a single class. This results in the corresponding accuracy levels of 100% and zero for the respective classes. This was something that each of the SVM methods tested using WEKA reported across all five datasets. Not to disregard this method entirely, the results were compared to that reported by the SVM implementation in NeuCom. (This is one of the methods that were able to be compared across implementation environments; unfortunately this list was not long.) In NeuCom the SVM variations produced between 93% and 95% overall accuracy levels. The lowest class accuracy level of these variations was 54%, which is also too low. In all three SVM variations tested in NeuCom the best results were obtained from DS-4, which matches the overall trend for all other tests.

When the methods are grouped into global, localised and personalised methods, a pattern of results emerges. The local methods performed worst, noting some of the issues with the SVM methods mentioned above. When consistence of result is taken into account the personalised methods outperformed all others. These methods may not have the highest overall accuracy levels but the accuracy of each class is split more evenly giving rise to a more stable and ultimately trustworthy result. If the methods are ranked in terms of accuracy, giving respectively lowering priorities to overall, case and control accuracies; then 9 of the top 10 are personalised methods. When the performance is viewed for each method amongst its respective variations the personalised methods achieved a higher level of consistency than methods in either of the other two groups. This has led to the conclusion that personalised methods are worth further investigation.

5.3 Random test sample

As the data used in the testing to date had been drawn from only two regions a test sample of 100 individuals was used as a comparison and verification point of the results obtained so far. This random sample was taken, as its name indicates, from a random sampling of 50 individuals from each of the case and control groups. It was possible that the data used previously for method testing could in some way not be a true representation of the entire population or, that those results were somehow skewed by region. Taking a random sample across all the regions seeks to eliminate, or at least illuminate, any variation between the results obtained by data from two regions as opposed to all regions.

The WW*k*NN method whilst having the facility to set the number of attributes at each execution was used with 100% attributes only in this test so that the results were directly compatible with earlier results and as earlier testing showed that there was no improvement to be gained at this stage in lowering the attribute numbers in each dataset. The datasets used here were based on DS-4 and DS-5 as these showed to be the best performing from earlier testing. DS-5 was included for completeness as it performed the best on a particular type of method.

It was not expected that the same high level of accuracy would be achieved with the random sample and the results bore this out. It was surprising to note that the accuracy levels although reduced did not drop as much as expected. The multiple linear regression method dropped the most in terms of overall accuracy down to 62% and 49% for the two datasets respectively. When viewed as a ranked list there was little change between the results obtained using only two regions and the random sample, at least at the top end. The top 10-15 ranked methods are the same between the two sets of results. Within the WW*k*NN method there was a difference in the performance at different threshold settings and this could easily relate to the way in which the data was ordered in the datasets. For the two regions group of datasets there was a definite order where the control samples were listed first. This was not strictly enforced in the random sample testing. The difference is a shift of 0.05 either side of 0.5 showing that a slight "lean" towards one of the classes yields the better results. From the two region results it would seem to be in the direction of the case class. This is one factor that would benefit from further investigation and is likely to be part of a fuller optimisation processes.

The relative performance of the datasets used bore out the conclusion from two region testing that the SNP attributes contained in DS-4 yielded the highest and most consistent accuracy results irrespective of method used. The relative differences in results between DS-4 and DS-5 in the random sample testing proved more pronounced where DS-5 failed to eclipse the performance of DS-4 in any method or variation. As DS-4 is based on the simplest of the methods for attribute list formulation it is somewhat surprising that it has performed so well. This said it does hold some interesting prospects for future attribute selection processes with the ever increasing amount of raw data becoming available.

5.4 Conclusion

This study has found a number of interesting results. One is that using data mining methods for attribute reduction has discovered the same SNPs as a number of previous studies, thus verifying the validity of this approach and that it is possible to apply data mining and knowledge engineering techniques to biological data and achieve at least the same results as existing methods. The difference is that in this study several SNPs were found significant that

did not appear in any previous work. Another is that simple methods can be just as effective as their more complex counterparts.

The comparison of modelling methods can be difficult due to lack of compatible measurement, and the need to set what is meant by “best” at each occasion. Many studies utilise accuracy, but only note overall accuracy and do not take into account the accuracy of individual classes. When class accuracies are considered the resultant “best” method is different on a number of occasions. Again this leads to the necessity to decide on the way in which each accuracy measure is used and what weighting each has in the final decision of “best”.

From the results found in this study it can be concluded that attribute reduction methods are as important as the modelling applied, and that personalised methods yield higher and more consistent results. Added to this is the unexpectedly high classification accuracy values obtained, most in excess of any of the previous studies findings. All these results point to one thing, further research.

Chapter 6 Conclusions and future directions of study

6.1 A brief summary of the problem and the work done in this thesis

The research presented in this thesis focuses on determining the relative merits of various pre-processing techniques, attribute reduction methods, and classification methods. The aim was to develop a prioritised list of methods that deserved further investigation alongside any added insights into the data itself discovered during the different phases of investigation.

Determining the susceptibility to MS form SNPs data has been an area of interest for some time and this study benefits from previous research done into the use of SNP data in susceptibility prediction for other diseases. It was identified that this was principally a classification problem and as such currently represents a two class situation, that of diseased or not diseased persons. In reality it is a three class problem, where there is some area of uncertainty as to the determination of class boundaries, especially with existing medical diagnostic procedures. To this end it was not expected that addressing the issue of susceptibility as a two class problem would yield results much over 85% in terms of model accuracy. The other expectation was that there would be a major reduction in SNPs from the original dataset to those representing a more appropriate or significant group for the purposes of classification. Traditionally less than 100 attributes are used for classification modelling, but the levels of selection for attribute candidates was not limited by numbers but by level of significance. This was done to accommodate the opinion of several other researchers that it is all too easy to remove attributes too early and that there is likely to be a large number of SNPs that have greater collective effect than they do individual effect.

The constructivist approach to research lends itself well to addressing a number of issues related to this study. Of particular importance is that it allows for on-going development of the artefact under investigation.

6.2 Findings – expected and unexpected

The investigation into the raw data proved that there remains much to be done in terms of storage and handling of large datasets. This was expected and the resulting hours of manual processing undertaken with the knowledge that this would not always be the case.

From the attribute selection/reduction process it was most surprising that the simplest method, of ranking the difference between the average score for each SNP between the case and control samples, proved to provide the best performing dataset. When the same group of SNPs were used for the random sample dataset it continued the trend. All five contributing datasets had PCA reduction also applied where it was designated as a cut off position of explaining 95% of the variation in the data to retain as many SNPs as possible whilst not keeping those that were not needed. A further unexpected result when comparing the final reduced datasets with exiting published data was that 7 SNPs (technically 6 SNPs plus gender) were present in the data sets and not in the published lists. Of these data sets, DS-4, the dataset which came from the reduction mentioned above, had the highest number of new SNPs (5 in total). Finding these new informative SNPs for the accurate prediction of susceptibility to MS on a personalised level is new information/knowledge discovered that needs to be published and discussed with experts in MS.

Model comparison across implementation platforms can be fraught with danger in terms of having applicable measures available that are consistently implemented throughout. This resulted in the use of classification accuracy, both overall and individual class accuracy, as the sole measure of each model's best performance.

It was anticipated that the modelling methods that were classed as personalised methods would be the better performing techniques in this situation. The results obtained here were both expected and unexpected. It was expected that the personalised methods would perform well and the local methods would perform poorly and this was borne out in the results. What was unexpected was the performance of the decision tree method and that accuracy levels could be achieved that were so high, many over 90%, by approximately ⅓ of all variations tested. When all three levels of accuracy were considered together the number of candidate methods was reduced, but still remained unexpectedly high. Testing with the random sample dataset produced similar results, with the personalised methods suffering the least reduction

in accuracy and showing the greatest level of consistence amongst variations. It was expected that the accuracy of the random sample dataset would be lower than that obtained with the two region datasets which they did do, but the results showed that the accuracy levels remained unexpectedly high.

Two expected outcomes from this study come in the form of limitations. The limitation was that the experimental design and modelling for this study had to be optimised to cope with the computing resources available. This meant that the approach had to address the two key bottlenecks namely the amount of RAM and the processor speed in the PC used. The consequence of this is that the time required to complete the testing increased proportionally compared with the resources used in similar studies.

Time constraints were anther issue in this study. The principle one being the amount of time needed to process the data initially as well as perform the reduction and model testing. This resulted in very little time to do further model enhancements, which are being carried out in subsequent work.

6.3 Open questions for a future work

Much has been learnt from this research that will fuel further research. These are best posed in terms of questions for a further study.

How can the process of data transformation, storage etc. be automated?

How does the effect of SNP interactions affect the models and the resulting accuracy levels?

Is accuracy the best method to compare methods?

Does gender matter?

- comes from the finding of significant SNPs on chromosome X and none on Y

How do computational costs impact of the development of future integrated systems using SNPs data for disease susceptibility predictions?

How does the optimisation of models increase the accuracy and can the combination of methods make it even better?

What roll is there for decision trees (as the method performed exceptionally well) in the overall process, can they be one of the combined methods?

How can the results from newer SNPs chips producing much more data points be incorporated into existing models and how does this change the importance of the SNPs that have been previously included or excluded?

How can other related data be incorporated? Data that relates to the disease, gene expression, protein encoding etc., along with patient data and other possible sources may or may not enhance the performance of the model in terms of susceptibility. These factors are likely to be a part of an overall model personalised for each patient incorporating both prediction and treatment.

Reference

Abdleazeem, S., & El-Sherif, E. (2008). Arabic handwritten digit recognition *International Journal of Document Analysis and Recognition (IJDAR), 11*(3), 127 - 141. doi:10.1007/s10032-008-0073-5

Alpaydin, E., & Kaynak, C. (1998). Cascading classifiers *KYBERNETIKA, 34*(4), 369 - 374.

Alpaydin, E., & Kaynak, C. (1998). Data Set Information. Retrieved2011, from http://archive.ics.uci.edu/ml/machine-learning-databases/optdigits/optdigits.names

Arshadi, N., Chang, B., & Kustra, R. (2009). Predictive modeling in case-control single-nucleotide polymorphism studies in the presence of population stratification: a case study using Genetic Analysis Workshop 16 Problem 1 dataset. *BMC proceedings, 3*(suppl 7).

Bakir-Gungor, B., & Sezerman, O. U. (2011). A new methodology to associate snps with human diseases according to their pathway related context. *PLoS ONE, 6*(10).

Ban, H.-J., Heo, J. Y., Oh, K.-S., & Park, K.-J. (2010). Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine. *BMC genetics, 11*(1). doi:10.1186/1471-2156-11-26

Ban, M., Goris, A., Lorentzen, Å. R., Baker, A., Mihalova, T., Ingram, G., ... Compston, A. (2009). Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. *European Journal of Human Genetics, 17*, 1309-1313. doi:10.1038/ejhg.2009.41

Bull, L. (2008). Learning Classifier Systems in Data Mining Learning Classifier Systems in Data Mining: An Introduction. In *Studies in Computational Intelligence* (Vol. 125, pp. 1-15). Berlin, Heidelberg: Springer Singapore Pte. Limited. Retrieved from http://www.springerlink.com.ezproxy.aut.ac.nz/content/m6j026t56g514343/. doi:10.1007/978-3-540-78979-6_1

Bull, L., & Kovacs, T. (2005). Foundations of Learning Classifier Systems: An Introduction. In *Studies in Fuzziness and Soft Computing* (Vol. 183/2005, pp. 1-17). Berlin Heidelberg: Springer-Verlag. Retrieved from http://www.springerlink.com.ezproxy.aut.ac.nz/content/1ek31c71797lhfjt/. doi:10.1007/11319122_1

Calcagno, G., Staiano, A., Fortunato, G., Brescia-Morra, V., Salvatore, E., Liguori, R., ... Sacchetti, L. (2010). A multilayer perceptron neural network-based approach for the identification of responsiveness to interferon therapy in multiple sclerosis patients. *Information sciences, 180*(21).

Cao, C., Leong, T.-Y., Leong, A. P. K., & Seow, F. C. (1998). Dynamic decision analysis in medicine: a data-driven approach. *International Journal of Medical Informatics, 51*(1), 13-28. doi:10.1016/S1386-5056(98)00085-9

Chen, J. J. (2007). Key aspects of analyzing microarray gene-expression data. *Pharmacogenomics, 8*(5).

Cortes, A., & Brown, M. A. (2010). Promise and pitfalls of the Immunochip. *Arthritis Research and Therapy, 13*(1).

D'Addabbo, A., Palmieri, O., Latiano, A., Annese, V., Mukherjee, S., & Ancona, N. (2011). RS-SNP: A random-set method for genome-wide association studies. *BMC Genomics, 12*.

Duin, R., van Breukelen, M., Tax, D., & den Hartog, J. (1998). Handwritten digit recognition by combined classifiers *KYBERNETIKA, 34*(4), 381 - 386.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data *Association for Computing Machinery. Communications of the ACM, 11*(9).

Feng, T., Elston, R. C., & Zhu, X. (2011). Detecting rare and common variants for complex traits: Sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genetic Epidemiology, 35*(5), 398-409.

Freitas, A. A. (2003). A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery [TH-013]. In *Advances in Evolutionary Computing* (Vol. Natural Computing Series, pp. 819-845): Springer. Retrieved from http://www.springerlink.com.ezproxy.aut.ac.nz/content/j07734581405q133/fulltext.pdf. Retrieved 2012-02-27. doi:10.1007/978-3-642-18965-4_33

Gamberger, D., Lavrac, N., & Dzeroski, S. (2000). Noise Detection and Elimination in Data Preprocessing: Experiments in Medical Domains. *Applied Artificial Intelligence, 14*(2), 205-223. doi:10.1080/088395100117124

Genes shed new light on cause of MS. Retrieved Augsut 26, 2011, from http://www.skynews.com.au/health/article.aspx?id=648911

Goertzel, B., Pennachin, C., Coelho, L., Shikida, L., & Queiroz, M. (2007). Biomind ArrayGenius and GeneGenius: Web Services Offering Microarray and SNP Data Analysis via Novel Machine Learning Methods*The AAAI Press, Menlo Park, California*. Symposium conducted at the meeting of the Twenty-Second AAAI Conference on Artificial Intelligence, Vancouver, British Columbia, Canada. Retrieved from https://www.aaai.org/Papers/AAAI/2007/AAAI07-275.pdf

GWAS Central. Retrieved from https://www.gwascentral.org/

Hoh, J., & Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature reviews. Genetics, 4*(9).

Kalatzis, F. G., Exarchos, T. P., Giannakeas, N., Rizos, P., Fotiadis, D. I., Markoula, S., ... Georgiou, I. (2010, 9–11 October 2009). *Point-of-Care Monitoring and Diagnostics for Rheumatoid Arthritis and Multiple Sclerosis*. presented at the meeting of the ADVANCED TOPICS IN SCATTERING THEORY AND BIOMEDICAL ENGINEERING Patras, Greece. Retrieved from http://eproceedings.worldscinet.com/9789814322034/9789814322034_0010.html doi:10.1142/9789814322034_0010

Kasabov, N. (2006). Global, local and personalised modeling and pattern discovery in bioinformatics: An integrated approach *Pattern Recognition Letters, 28*, 673 - 685. doi:10.1016/j.patrec.2006.08.007

Kasabov, N. (2007). Evolving Connectionist Systems: Methods & Applications in Bioinformatics, Brain Study & Intelligent Machines (2nd ed.). London: Springer Verlag.

Kharbat, F., Odeh, M., & Bull, L. (2008). Knowledge Discovery from Medical Data: An Empirical Study with XCS. In *Studies in Computational Intelligence* (Vol. 125, pp. 93-121). Retrieved from http://www.springerlink.com.ezproxy.aut.ac.nz/content/k244831p100j2817/. doi:10.1007/978-3-540-78979-6_5

Kim, J. H. (2002). Bioinformatics and genomic Medicine. *Genetics in Medicine, 6*(6, Supplememnt), 62S-65S.

Kwon, M.-S., Kim, K., Lee, S., Chung, W., Yi, S.-G., Namkung, J., & Park, T. (2011. GWAS-GMDR: A program package for genome-wide scan of gene-gene interactions with covariate adjustment based on multifactor dimensionality reduction. presented at the meeting of the IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2011,

Leong, T. Y. (1998). Multiple perspective dynamic decision making. *Artificial Intelligence, 105*, 209-261.

Li, Y., & Schwartz, C. E. (2011). Data mining for response shift patterns in multiple sclerosis patients using recursive partitioning tree analysis. *Quality of Life Research 20*(10), 1543-1553. doi:10.1007/s11136-011-0004-7

Lin, J.-H., & Haug, P. J. (2006). Data Preparation Framework for Preprocessing Clinical Data in Data Mining. *AMIA Annu Symp Proc. 2006*, 489–493.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics, 39*(7), 906-913. doi:10.1038/ng2088

Marchiondo, K. (2010). Multiple sclerosis *MedSurg Nursing, 19*(5).

MATLAB.

Mitha, F., Herodotou, H., Borisov, N., Jiang, C., Yoder, J., & Owzar, K. (2011). SNPpy - Database management for SNP data from genome wide association studies. *PLoS ONE, 6*(1).

Multiple Sclerosis New Zealand. (2011). What is MS? Retrieved August 26, 2011, from http://www.msnz.org.nz/Page.aspx?pid=276

NeuCom. Retrieved from www.theneucom.com

Online Mendelian Inheritance in Man.

Pappa, G. L., & Freitas, A. A. (2008). Discovering New Rule Induction Algorithms with Grammar-based Genetic Programming In O. Maimon & L. Rokach (Eds.), *Soft Computing for Knowledge Discovery and Data Mining* (pp. Part II, 133-152). Retrieved from http://www.springerlink.com.ezproxy.aut.ac.nz/content/v6365361ww852l26/. doi:10.1007/978-0-387-69935-6_6

Piatetsky-Shapiro, G. (2003). Microarray data mining facing the challenges. *SIGKDD explorations, 5*(2). doi:10.1145/980972.980974

Quevedo, J. R., Bahamonde, A. b., Perez-Enciso, M., & Luaces, O. (2012). Disease liability prediction from large scale genotyping data using classifiers with a reject option. *IEEE/ACM transactions on computational biology and bioinformatics, 9*(1), 88-97.

Sawcer, S. (2010). What role for genetics in the prediction of multiple sclerosis? *Annals of neurology, 67*(1). doi:10.1002/ana.21911

Schwartz, C. E., Sprangers, M. A. G., Oort, F. J., Ahmed, S., Bode, R., Li, Y., & Vollmer, T. (2011). Response shift in patients with multiple sclerosis: an application of three statistical techniques *Quality of life research, 20*(10), 1561-1572. doi:10.1007/s11136-011-0056-8

Shah, S., & Kusack, A. (2007). Cancer gene search with data-mining and genetic algorithms. *Computers in biology and medicine, 37*(2).

Sumathi, S., & Sivanandam, S. N. (2006). Data Mining & KDD. In *Studies in Computational Intelligence* (Vol. 29, pp. 231-241). Retrieved from http://www.springerlink.com.ezproxy.aut.ac.nz/content/x47n1111r324622l/fulltext.pdf

The Welcome Trust Case Control Consortium, & The Australo-Anglo-American Spondytitis Consortium. (2007). Association scan of 14,500 nsSNPs in four common diseases

identifies variants involved in autoimmunity. *Nature Genetics*, *39*(11), 1329-1337. doi:10.1038/ng.2007.17

The Wellcome Trust Case Control Consortium. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *nature*, *464*. doi:10.1038/nature08979

Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., ... Clayton, D. G. (2007). Robust association of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature genetics*, *39*(7), 857-864. doi:10.1038/ng2068

Tomida, S., Hanai, T., Koma, N., Suzuki, Y., Kobayashi, T., & Honda, H. (2002). Artificial neural network predictive model for allergic disease using single nucleotide polymorphisms data. *Journal of bioscience and bioengineering*, *93*(5), 470-478.

Wagholikar, K. B., Sundararajan, V., & Deshpande, A. W. (2011). Modeling Paradigms for Medical Diagnostic Decision Support: A Survey and Future Directions. *Journal of Medical Systems*. doi:10.1007/s10916-011-9780-4

Wang, X., Prins, B. P., Sober, S., Laan, M., & Snieder, H. (2011). Beyond genome-wide association studies: New strategies for identifying genetic determinants of hypertension. *Current hypertension reports*, *13*(6), 442-451. doi:10.1007/s11906-011-0230-y

WEKA. Retrieved from http://www.cs.waikato.ac.nz/ml/index.html

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Margan Kauffman.

Wu, X., Yu, P. S., Piatetsky-Shapiro, G., Cercone, N., Lin, T. Y., Kotagiri, R., & Wah, B. W. (2003). Data Mining: How Research Meets Practical Development? *Knowledge and Information Systems*, *5*(2), 248 - 261. doi:10.1007/s10115-003-0101-1

Yang, C., Wan, X., Yang, Q., Xue, H., Tang, N. L. S., & Yu, W. (2011). A hidden two-locus disease association pattern in genome-wide association studies. *BMC Bioinformatics*, *12*.

Yoo, J., Lee, Y., Kim, Y., Rha, S. Y., & Kim, Y. (2008). SNPAnalyzer 2.0: A web-based integrated workbench for linkage disequilibrium analysis and association analysis. *BMC bioinformatics*, *9*(1). doi:10.1186/1471-2105-9-290

Yu, Z., Wong, H.-S., Wang, D., & Wei, M. (2010). Neighborhood Knowledge-Based Evolutionary Algorithm for Multiobjective Optimization Problems *Evolutionary Computation, IEEE Transactions on 15*(6), 812 - 831. doi:10.1109/TEVC.2010.2051444

Yung, L. S., Yang, C., Wan, X., & Yu, W. (2011). GBOOST: A GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. *Bioinformatics*, *27*(9), 1309-1310.

Zhao, F., & Leong, T.-Y. (2000). A data preprocessing framework for supporting probability-learning in dynamic decision modeling in medicine. *Proc AMIA Symp. 2000*, 933–937.

Appendix A

This appendix contains the full-scale tabled results of the methods of classification outlined in chapter 4. They are the originals form which the images in chapter 4 have been taken.

	Method	DS-1			DS-2			DS-3			DS-4			DS-5		
		Overall	Case	Control	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control
Rules																
	OneR	85.14%	93.75%	75.50%	85.14%	93.75%	75.50%	85.14%	93.75%	75.50%	91.27%	95.09%	87.00%	88.92%	94.64%	82.50%
	Decision Table – Best First	91.51%	84.54%	75.50%	91.51%	86.08%	93.50%	91.51%	93.30%	89.50%	95.28%	99.11%	91.00%	89.39%	90.18%	88.50%
Trees																
	J48 – confidence factor = 0.25	93.40%	95.88%	91.00%	93.40%	95.88%	91.00%	95.75%	96.88%	94.50%	96.23%	95.98%	96.50%	**97.64%**	97.32%	98.00%
	J48 – Confidence factor = 0.5	93.15%	94.85%	91.50%	93.15%	94.85%	91.50%	**95.99%**	96.88%	95.00%	**96.70%**	95.98%	97.50%	97.41%	96.88%	98.00%
	J48 – Confidence factor = 0.1	**93.65%**	95.88%	91.50%	**93.65%**	95.88%	91.50%	95.75%	96.88%	94.50%	96.23%	95.98%	96.50%	**97.64%**	97.32%	98.00%
Bayesian																
	Naive Bayes	61.17%	29.90%	91.50%	61.17%	29.90%	91.50%	65.80%	40.63%	94.00%	74.06%	54.46%	96.00%	68.63%	46.88%	93.00%
	Bayesian Logistic Regression	70.05%	70.10%	70.00%	70.05%	70.10%	70.00%	79.72%	80.80%	78.50%	88.44%	95.98%	80.00%	88.68%	97.32%	79.00%
Nearest Neighbour																
No distance weighting																
	K = 1	77.16%	85.05%	69.50%	77.16%	85.05%	69.50%	87.26%	87.05%	87.50%	94.81%	94.64%	95.00%	87.26%	89.29%	85.00%
	K = 3	75.38%	87.63%	63.50%	75.38%	87.63%	63.50%	87.26%	88.39%	86.00%	93.63%	94.64%	92.50%	88.92%	93.75%	83.50%
	K = 5	75.89%	88.14%	64.00%	75.89%	88.14%	64.00%	88.68%	92.86%	84.00%	93.63%	95.54%	91.50%	87.74%	94.20%	80.50%
	K = 7	73.10%	89.69%	57.00%	73.10%	89.69%	57.00%	88.68%	94.64%	82.00%	93.16%	96.43%	89.50%	87.03%	95.09%	78.00%
	K = 9	70.81%	89.69%	52.50%	70.81%	89.69%	52.50%	89.39%	95.09%	83.00%	92.69%	96.88%	88.00%	87.97%	97.77%	77.00%
	K = 11	70.30%	91.24%	50.00%	70.30%	91.24%	50.00%	88.21%	95.54%	80.00%	92.22%	96.88%	87.00%	86.79%	98.21%	74.00%
	K = 15	68.02%	92.27%	44.50%	68.02%	92.27%	44.50%	87.97%	96.43%	78.50%	90.80%	96.88%	84.00%	84.43%	98.21%	69.00%
	K = 20	69.04%	91.24%	47.50%	69.04%	91.24%	47.50%	88.68%	96.43%	80.00%	90.57%	95.98%	84.50%	84.67%	97.77%	70.00%
	K = 30	69.29%	91.24%	48.00%	69.29%	100.00%	48.00%	83.96%	95.09%	71.50%	88.92%	97.77%	79.00%	80.42%	98.21%	60.50%
	K = 50	65.48%	96.39%	35.50%	65.48%	96.39%	35.50%	83.73%	93.30%	73.00%	84.20%	100.00%	66.50%	78.07%	98.66%	55.00%

Weight by 1/distance																
K = 1	77.16%	85.05%	69.50%	77.16%	85.05%	69.50%	87.26%	87.05%	87.50%	94.81%	94.64%	95.00%	87.26%	89.29%	85.00%	
K = 3	75.63%	87.63%	64.00%	75.63%	87.63%	64.00%	87.97%	88.84%	87.00%	94.58%	95.54%	93.50%	89.15%	93.75%	84.00%	
K = 5	76.14%	87.63%	65.00%	76.14%	87.63%	65.00%	90.33%	93.30%	87.00%	94.34%	95.54%	93.00%	88.92%	95.54%	81.50%	
K = 7	73.86%	89.69%	58.50%	73.86%	89.69%	58.50%	89.62%	94.20%	84.50%	94.58%	96.88%	92.00%	87.50%	95.98%	78.00%	
K = 9	71.32%	89.69%	53.50%	71.32%	89.69%	53.50%	90.09%	94.20%	85.50%	93.87%	97.32%	90.00%	88.44%	97.77%	78.00%	
K = 11	71.07%	91.75%	51.00%	71.07%	91.75%	51.00%	88.68%	94.64%	82.00%	93.40%	97.32%	89.00%	88.44%	97.77%	78.00%	
K = 15	70.30%	92.78%	48.50%	70.30%	92.78%	48.50%	90.57%	96.43%	84.00%	91.75%	96.43%	86.50%	86.32%	98.21%	73.00%	
K = 20	70.56%	93.30%	48.50%	70.56%	93.30%	48.50%	90.33%	96.43%	83.50%	91.51%	96.88%	85.50%	86.32%	98.21%	73.00%	
K = 30	70.56%	95.88%	46.00%	70.56%	95.88%	46.00%	89.15%	97.32%	80.00%	90.57%	98.21%	82.00%	82.55%	98.66%	64.50%	
K = 50	67.51%	95.88%	40.00%	67.51%	95.88%	40.00%	87.50%	95.09%	79.00%	87.03%	100.00%	72.50%	81.37%	98.66%	62.00%	
Weight by 1-distance																
K = 1	77.16%	85.05%	69.50%	77.16%	85.05%	69.50%	87.26%	87.05%	87.50%	94.81%	94.64%	95.00%	87.26%	89.29%	85.00%	
K = 3	75.38%	87.63%	63.50%	75.38%	87.63%	63.50%	87.26%	88.39%	86.00%	93.63%	94.64%	92.50%	88.92%	93.75%	83.50%	
K = 5	75.89%	88.14%	64.00%	75.89%	88.14%	64.00%	88.68%	92.86%	84.00%	93.63%	95.54%	91.50%	87.74%	94.20%	80.50%	
K = 7	73.10%	89.69%	57.00%	73.10%	89.69%	57.00%	88.68%	94.64%	82.00%	93.16%	96.43%	89.50%	87.03%	95.09%	78.00%	
K = 9	70.81%	89.69%	52.50%	70.81%	89.69%	52.50%	89.39%	95.09%	83.00%	92.69%	96.88%	88.00%	87.97%	97.77%	77.00%	
K = 11	70.30%	91.24%	50.00%	70.30%	91.24%	50.00%	88.21%	95.54%	80.00%	92.22%	96.88%	87.00%	86.79%	98.21%	74.00%	
K = 15	68.02%	92.27%	44.50%	68.02%	92.27%	44.50%	87.97%	96.43%	78.50%	90.80%	96.88%	84.00%	84.43%	98.21%	69.00%	
K = 20	69.29%	93.30%	46.00%	69.29%	93.30%	46.00%	88.44%	96.88%	79.00%	91.04%	96.88%	84.50%	84.43%	98.21%	69.00%	
K = 30	68.78%	94.33%	44.00%	68.78%	94.33%	44.00%	85.14%	97.32%	71.50%	89.15%	98.21%	79.00%	80.42%	98.66%	60.00%	
K = 50	65.23%	96.39%	35.00%	65.23%	96.39%	35.00%	84.43%	95.09%	72.50%	84.20%	100.00%	66.50%	78.07%	98.66%	55.00%	
Support Vector Machine																
Polynomial	50.76%	0.00%	100.00 %	50.76%	87.63%	63.00%	52.83%	100.00%	0.00%	52.83%	100.00%	0.00%	52.83%	100.00 %	0.00%	
Linear	64.97%	75.77%	54.50%	75.13%	0.00%	100.00%	86.56%	88.84%	84.00%	86.56%	95.54%	89.00%	91.51%	97.77%	54.50%	
Radial Based Function	76.14%	37.63%	75.00%	76.14%	85.57%	67.00%	89.62%	100.00%	0.00%	96.23%	100.00%	59.00%	89.39%	100.00 %	0.00%	
Multilayer Perceptron	76.14%	85.57%	67.00%	76.14%	85.57%	67.00%	89.62%	91.52%	87.50%	96.23%	96.43%	96.00%	89.39%	92.41%	86.00%	

Table A-1: Classification testing methods implemented in WEKA

Method	DS-1			DS-2			DS-3			DS-4			DS-5		
	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control
Support Vector Machine															
Polynomial	55.55%	73.20%	75.50%	52.02%	84.02%	71.00%	59.30%	87.95%	84.00%	70.77%	95.54%	94.50%	54.06%	95.54%	88.00%
Linear	**75.11%**	74.74%	75.50%	**77.38%**	82.99%	72.00%	89.16%	90.63%	87.50%	95.04%	97.32%	92.50%	**92.69%**	98.21%	86.50%
Radial Based Function	74.87%	73.20%	76.50%	76.69%	78.35%	75.00%	86.31%	89.73%	82.50%	93.17%	96.43%	89.50%	91.50%	95.54%	87.00%
Multilayer Perceptron	74.85%	76.17%	73.50%	76.09%	81.96%	70.50%	**89.87%**	91.52%	88.00%	**96.21%**	95.98%	96.50%	91.51%	93.75%	89.00%
Evolving clustering method for classification	59.21%	60.32%	57.00%	69.58%	77.84%	61.50%	77.81%	75.00%	81.00%	94.09%	95.54%	92.50%	85.64%	90.18%	80.50%
Multiple Linear Regression	70.33%	69.59%	71.00%	69.29%	70.90%	67.00%	82.33%	82.14%	82.50%	92.44%	92.41%	92.50%	86.09%	89.29%	82.50%
Evolving Classification Function	55.55%	76.80%	87.00%	52.02%	4.64%	98.00%	59.30%	37.05%	84.00%	70.77%	52.68%	91.00%	54.06%	16.52%	96.00%

Table A-2: Classification testing methods implemented in NeuCom

Method		DS-1			DS-2			DS-3			DS-4			DS-5		
		Overall	Case	Control	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control
W kNN																
Threshold = 0.45	*K=1*	64.72%	59.28%	70.00%	**79.44%**	86.60%	72.50%	**89.15%**	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=3	69.29%	63.92%	74.50%	78.43%	87.11%	70.00%	87.26%	88.39%	86.00%	93.63%	95.09%	92.00%	88.44%	93.57%	82.50%
	K=5	70.30%	67.01%	73.50%	78.68%	89.18%	68.50%	88.21%	91.07%	85.00%	93.87%	96.43%	91.00%	87.74%	94.64%	80.00%
	K=7	68.27%	64.43%	72.00%	77.66%	89.69%	66.00%	87.26%	91.52%	82.50%	92.92%	97.32%	88.00%	87.50%	95.09%	79.00%
	K=9	68.53%	64.43%	72.50%	76.14%	90.72%	62.00%	86.56%	90.18%	82.50%	91.98%	96.43%	87.00%	85.85%	94.64%	76.00%
	K=11	70.81%	77.32%	64.50%	71.07%	97.94%	45.00%	87.50%	94.64%	79.50%	90.09%	98.21%	81.00%	82.55%	96.88%	66.50%
	K=15	70.56%	73.71%	67.50%	70.05%	97.94%	43.00%	87.74%	94.64%	80.00%	89.39%	98.66%	79.00%	82.55%	98.66%	64.50%
	K=20	72.84%	71.65%	74.00%	72.84%	99.48%	47.00%	72.84%	97.32%	81.00%	91.04%	98.66%	82.50%	82.08%	98.66%	63.50%
	K=30	72.59%	72.16%	73.00%	68.02%	99.48%	37.50%	86.08%	95.98%	75.00%	87.03%	99.11%	73.50%	78.30%	99.11%	55.00%
	K=50	71.07%	71.13%	71.00%	61.68%	98.45%	26.00%	84.43%	92.86%	75.00%	82.78%	100.00%	63.50%	74.06%	99.11%	46.00%
Threshold = 0.50	*K=1*	64.72%	59.28%	70.00%	**79.44%**	86.60%	72.50%	**89.15%**	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=3	69.29%	63.92%	74.50%	78.43%	87.11%	70.00%	87.26%	88.39%	86.00%	93.63%	95.09%	92.00%	88.44%	93.57%	82.50%
	K=5	70.30%	67.01%	73.50%	78.68%	89.18%	68.50%	88.21%	91.07%	85.00%	93.87%	96.43%	91.00%	87.74%	94.63%	80.00%
	K=7	68.27%	64.43%	72.00%	77.66%	89.69%	66.00%	87.26%	91.52%	82.50%	92.92%	97.32%	88.00%	87.50%	95.09%	79.00%
	K=9	68.53%	64.43%	72.50%	76.14%	90.72%	62.00%	86.56%	90.18%	82.50%	91.98%	96.43%	87.00%	85.85%	94.64%	76.00%
	K=11	70.05%	65.46%	74.50%	74.37%	91.24%	58.00%	87.26%	91.52%	82.50%	91.98%	96.43%	87.00%	85.61%	95.98%	74.00%
	K=15	72.08%	67.01%	77.00%	74.37%	94.85%	54.50%	87.97%	93.30%	82.00%	91.27%	96.88%	85.00%	84.20%	97.32%	69.50%
	K=20	72.59%	67.53%	77.50%	74.87%	97.42%	53.00%	88.92%	93.75%	83.50%	91.98%	96.43%	87.00%	84.43%	98.21%	69.00%
	K=30	**74.62%**	67.01%	82.00%	71.83%	98.45%	46.00%	83.96%	88.84%	78.50%	89.39%	98.21%	79.50%	80.19%	97.77%	60.50%
	K=50	70.30%	66.49%	74.00%	66.24%	94.85%	38.50%	83.96%	87.50%	80.00%	86.56%	99.11%	72.50%	79.01%	98.21%	57.50%
Threshold = 0.55	*K=1*	64.72%	59.28%	70.00%	**79.44%**	86.60%	72.50%	**89.15%**	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=3	69.29%	63.92%	74.50%	78.43%	87.11%	70.00%	87.26%	88.39%	86.00%	93.63%	95.09%	92.00%	88.44%	93.57%	82.50%
	K=5	70.30%	67.01%	73.50%	78.68%	89.18%	68.50%	88.21%	91.07%	85.00%	93.87%	96.43%	91.00%	87.74%	94.63%	80.00%
	K=7	68.27%	64.43%	72.00%	77.66%	89.69%	66.00%	87.26%	91.52%	82.50%	92.92%	97.32%	88.00%	87.50%	95.09%	79.00%
	K=9	68.27%	63.92%	72.50%	76.14%	90.72%	62.00%	86.56%	90.18%	82.50%	92.69%	96.43%	88.50%	85.85%	94.64%	76.00%
	K=11	70.05%	56.70%	83.00%	73.10%	78.35%	68.00%	84.43%	83.93%	85.00%	92.22%	93.75%	90.50%	87.26%	92.86%	81.00%
	K=15	70.56%	60.31%	80.50%	72.08%	83.51%	61.00%	87.26%	88.39%	86.00%	91.89%	94.64%	89.00%	85.85%	94.20%	76.50%
	K=20	72.34%	63.92%	80.50%	76.65%	91.24%	62.50%	86.32%	87.50%	85.00%	91.57%	98.21%	84.50%	85.61%	97.32%	72.50%
	K=30	73.60%	63.92%	83.00%	73.86%	92.78%	55.50%	82.55%	83.04%	82.00%	90.57%	96.88%	83.50%	82.08%	96.88%	65.50%
	K=50	73.35%	63.40%	83.00%	70.05%	91.75%	49.00%	81.60%	77.23%	86.50%	88.44%	98.66%	77.00%	82.55%	96.43%	67.00%

Threshold = 0.60	***K=1***	64.72%	59.28%	70.00%	**79.44%**	86.60%	72.50%	**89.15%**	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=3	69.29%	63.92%	74.50%	78.43%	87.11%	70.00%	87.26%	88.39%	86.00%	93.63%	95.09%	92.00%	88.44%	93.57%	82.50%
	K=5	67.51%	54.64%	80.00%	79.19%	82.99%	75.50%	86.56%	83.93%	89.50%	**94.58%**	94.64%	94.50%	89.86%	91.52%	88.00%
	K=7	69.80%	64.43%	87.00%	76.90%	89.69%	76.00%	82.31%	91.52%	87.50%	93.40%	97.32%	94.50%	89.62%	95.09%	88.00%
	K=9	71.57%	55.67%	87.00%	77.16%	82.47%	72.00%	84.43%	83.93%	85.00%	92.92%	92.86%	93.00%	**90.33%**	92.41%	88.00%
	K=11	70.05%	56.70%	83.00%	73.10%	78.35%	68.00%	84.43%	83.93%	85.00%	92.22%	93.75%	90.50%	87.26%	92.86%	81.00%
	K=15	70.81%	57.73%	83.50%	74.62%	79.90%	69.50%	85.61%	84.38%	87.00%	93.16%	93.75%	92.50%	87.03%	92.96%	80.50%
	K=20	70.05%	57.22%	82.50%	76.14%	85.05%	67.50%	86.56%	85.71%	87.50%	91.98%	92.41%	91.50%	85.14%	92.86%	76.50%
	K=30	72.59%	60.82%	84.00%	77.66%	88.66%	67.00%	84.20%	79.46%	89.50%	91.04%	94.20%	87.50%	83.96%	92.86%	74.00%
	K=50	73.86%	62.37%	85.00%	72.34%	86.08%	59.00%	80.42%	72.32%	89.50%	90.80%	95.54%	85.50%	83.02%	94.20%	70.50%
Threshold = 0.65	***K=1***	64.72%	59.28%	70.00%	**79.44%**	86.60%	72.50%	**89.15%**	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=3	69.29%	63.92%	74.50%	78.43%	87.11%	70.00%	87.26%	88.39%	86.00%	94.10%	95.09%	93.00%	88.44%	93.57%	82.50%
	K=5	62.18%	37.63%	86.00%	78.43%	75.26%	81.50%	82.31%	73.66%	92.00%	93.16%	91.07%	95.50%	89.39%	87.05%	92.00%
	K=7	69.80%	52.06%	87.00%	76.90%	77.84%	76.00%	82.31%	77.68%	87.50%	93.40%	92.41%	94.50%	89.62%	91.07%	88.00%
	K=9	71.57%	55.67%	87.00%	77.16%	82.47%	72.00%	84.43%	83.93%	85.00%	92.92%	92.86%	93.00%	**90.33%**	92.41%	88.00%
	K=11	69.29%	48.97%	89.00%	71.07%	61.34%	80.50%	79.25%	71.88%	87.50%	91.04%	87.50%	95.00%	88.44%	88.39%	88.50%
	K=15	70.30%	55.15%	85.00%	73.60%	72.68%	74.50%	82.55%	78.13%	87.50%	92.22%	91.07%	93.50%	86.56%	91.07%	81.50%
	K=20	68.02%	51.03%	84.50%	74.11%	73.71%	74.50%	82.31%	75.89%	89.50%	91.95%	88.84%	94.50%	86.08%	88.84%	83.00%
	K=30	70.81%	55.15%	86.00%	73.60%	72.68%	74.50%	79.25%	69.20%	90.50%	88.92%	86.16%	92.00%	83.73%	88.39%	78.50%
	K=50	73.10%	60.31%	85.50%	72.59%	77.84%	67.50%	79.01%	68.30%	91.00%	88.44%	87.50%	89.50%	80.90%	83.93%	77.50%

Table A-3: Classification testing methods implemented in MATLAB - WkNN method

Method		DS-1			DS-2			DS-3			DS-4			DS-5		
		Overall	Case	Control	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control	Overall	Case	Control
WWkNN – 100% Features																
Threshold = 0.35	*K=3*	66.50%	66.49%	66.50%	79.70%	90.21%	69.50%	89.15%	91.07%	87.00%	94.34%	95.54%	93.00%	88.21%	92.86%	83.00%
	K=5	68.27%	73.20%	63.50%	80.46%	73.20%	71.00%	91.98%	95.09%	88.50%	96.46%	97.32%	95.50%	89.62%	94.64%	84.00%
	K=7	69.54%	75.26%	64.00%	78.17%	75.26%	66.00%	90.33%	92.86%	87.50%	95.99%	96.43%	95.50%	90.33%	95.54%	84.50%
	K=9	71.57%	76.29%	67.00%	80.71%	76.29%	67.00%	**92.45%**	95.98%	88.50%	95.97%	97.77%	93.50%	89.39%	95.54%	82.50%
	K=11	70.05%	75.77%	64.50%	80.46%	95.88%	65.50%	91.75%	95.98%	87.00%	95.99%	97.77%	94.00%	92.57%	97.77%	82.50%
	K=15	68.02%	76.80%	59.50%	77.41%	76.80%	59.00%	91.27%	96.43%	85.50%	95.99%	98.66%	93.00%	90.09%	98.21%	81.00%
	K=20	70.56%	79.90%	61.50%	75.38%	79.90%	53.50%	91.51%	97.77%	84.50%	94.58%	97.21%	90.50%	88.44%	98.21%	77.50%
	K=30	71.32%	81.44%	61.50%	72.08%	98.45%	46.50%	91.27%	98.66%	83.00%	92.45%	99.11%	85.00%	86.08%	97.77%	73.00%
	K=50	72.08%	84.02%	60.50%	68.78%	98.97%	39.58%	88.21%	98.66%	76.50%	90.09%	100.00%	79.00%	83.02%	98.21%	66.00%
Threshold = 0.40	*K=3*	65.74%	63.92%	67.50%	79.70%	88.66%	71.00%	89.39%	90.63%	88.00%	94.34%	95.54%	93.00%	88.68%	92.86%	84.00%
	K=5	69.29%	72.16%	66.50%	79.95%	89.18%	71.00%	91.75%	94.20%	89.00%	96.70%	97.32%	96.00%	90.09%	94.20%	85.50%
	K=7	71.07%	74.74%	67.50%	79.19%	89.69%	69.00%	90.80%	92.41%	89.00%	96.23%	96.43%	96.00%	91.04%	94.20%	87.50%
	K=9	71.32%	74.74%	68.00%	81.73%	92.78%	71.00%	**92.45%**	94.20%	90.50%	96.46%	97.77%	95.00%	90.57%	95.54%	85.00%
	K=11	71.05%	74.23%	68.00%	80.71%	93.81%	68.00%	91.75%	94.64%	88.50%	96.70%	97.32%	96.00%	90.80%	96.43%	84.50%
	K=15	70.05%	73.71%	66.50%	78.17%	95.36%	61.50%	91.51%	95.98%	86.50%	96.93%	98.66%	95.00%	90.57%	97.32%	83.00%
	K=20	70.81%	75.26%	66.50%	76.90%	96.91%	57.50%	91.51%	96.88%	85.50%	95.28%	97.77%	92.50%	88.92%	97.77%	79.00%
	K=30	71.83%	76.29%	67.50%	76.65%	97.94%	56.00%	90.57%	97.32%	83.00%	92.92%	98.66%	86.50%	88.44%	97.77%	78.00%
	K=50	71.83%	78.57%	68.00%	70.56%	98.97%	43.00%	88.92%	98.21%	78.50%	91.51%	99.11%	83.00%	85.85%	97.77%	72.50%
Threshold = 0.45	*K=3*	65.23%	91.34%	69.00%	79.44%	87.11%	72.00%	89.62%	89.73%	89.50%	94.32%	95.09%	93.50%	89.15%	92.41%	85.50%
	K=5	68.78%	67.53%	70.00%	79.70%	87.63%	72.00%	91.27%	93.30%	89.00%	96.46%	96.88%	96.00%	90.33%	93.75%	86.50%
	K=7	71.07%	73.20%	69.00%	79.19%	87.63%	71.00%	90.57%	91.52%	89.50%	96.46%	96.43%	96.50%	91.27%	93.75%	88.50%
	K=9	72.34%	74.23%	70.50%	80.96%	90.21%	72.00%	91.98%	92.86%	91.00%	96.70%	97.77%	95.50%	91.04%	95.09%	86.50%
	K=11	72.34%	72.68%	72.00%	80.20%	91.24%	69.50%	91.56%	92.86%	90.00%	96.93%	97.32%	96.50%	91.04%	95.09%	86.50%
	K=15	72.59%	70.10%	75.00%	78.68%	93.30%	64.50%	91.98%	94.64%	89.00%	**97.41%**	98.66%	96.00%	90.80%	95.54%	85.50%
	K=20	73.10%	70.62%	75.50%	79.44%	95.36%	64.00%	91.98%	95.54%	88.00%	95.28%	96.88%	93.50%	90.09%	96.43%	83.00%
	K=30	71.57%	70.10%	73.00%	78.17%	97.94%	59.00%	91.04%	95.98%	85.50%	94.10%	97.77%	90.00%	89.39%	96.88%	81.00%
	K=50	73.35%	68.04%	78.50%	75.81%	98.45%	52.50%	91.04%	98.21%	83.00%	92.45%	98.66%	85.50%	86.32%	96.88%	74.50%
Threshold = 0.50	*K=3*	61.97%	59.79%	70.00%	79.44%	86.60%	75.50%	89.15%	88.84%	89.50%	94.34%	95.09%	93.50%	88.92%	91.52%	86.00%
	K=5	69.04%	64.95%	73.00%	80.71%	87.63%	74.00%	90.80%	91.52%	90.00%	96.46%	96.88%	96.00%	90.80%	93.75%	87.50%
	K=7	70.56%	68.56%	72.50%	78.68%	84.54%	73.00%	89.86%	89.73%	90.00%	96.46%	96.43%	96.50%	91.98%	93.75%	90.00%
	K=9	72.34%	71.13%	73.50%	80.96%	86.08%	76.00%	91.98%	91.96%	92.00%	96.70%	97.77%	95.50%	91.98%	94.20%	89.50%
	K=11	71.07%	67.53%	74.50%	80.46%	88.66%	72.50%	91.51%	91.96%	91.00%	96.93%	97.32%	96.50%	91.98%	94.64%	89.00%
	K=15	71.07%	65.46%	76.00%	80.20%	91.24%	69.50%	90.80%	92.41%	89.00%	96.93%	97.77%	96.00%	91.51%	95.09%	87.50%
	K=20	71.57%	65.46%	77.50%	80.71%	93.30%	68.50%	92.22%	94.64%	89.50%	96.23%	96.88%	95.50%	91.51%	95.98%	86.50%
	K=30	73.60%	64.95%	82.00%	78.68%	94.85%	63.00%	91.98%	95.54%	88.00%	94.58%	97.32%	91.50%	90.80%	96.43%	84.50%
	K=50	**74.11%**	65.98%	80.22%	79.44%	97.94%	61.50%	91.51%	95.09%	87.50%	92.92%	98.66%	86.50%	88.44%	96.88%	79.00%

Threshold = 0.55	*K=3*	64.72%	56.70%	72.50%	78.68%	85.05%	72.50%	89.39%	87.95%	91.00%	94.58%	94.64%	94.50%	88.68%	91.07%	86.00%
	K=5	67.77%	61.34%	74.00%	80.96%	86.60%	75.50%	90.33%	90.18%	90.50%	96.46%	96.88%	96.00%	90.80%	92.41%	89.00%
	K=7	69.80%	64.43%	75.00%	78.68%	82.47%	75.00%	90.57%	89.73%	91.50%	96.70%	96.43%	97.00%	91.75%	92.41%	91.00%
	K=9	72.08%	66.49%	77.50%	80.96%	84.54%	77.50%	91.98%	91.52%	92.50%	96.46%	96.88%	96.00%	91.98%	93.75%	90.00%
	K=11	71.32%	65.98%	76.50%	80.71%	86.08%	75.50%	91.04%	90.18%	92.00%	96.93%	96.88%	97.00%	92.69%	93.75%	91.50%
	K=15	70.81%	62.37%	79.00%	81.73%	90.21%	73.50%	90.33%	90.63%	90.00%	97.17%	96.88%	97.50%	91.98%	94.64%	89.00%
	K=20	71.07%	61.86%	80.00%	80.46%	90.72%	70.50%	91.51%	91.96%	91.00%	96.23%	95.98%	96.50%	92.45%	95.54%	89.00%
	K=30	72.59%	61.34%	83.50%	82.23%	93.30%	71.50%	91.98%	94.20%	89.50%	94.81%	95.54%	94.00%	91.98%	95.98%	87.50%
	K=50	73.60%	64.43%	82.50%	80.96%	94.33%	68.00%	89.62%	89.73%	89.50%	93.63%	98.21%	88.50%	91.51%	96.88%	85.50%
Threshold = 0.60	*K=3*	64.21%	54.12%	74.00%	78.68%	84.54%	73.00%	89.39%	87.05%	92.00%	94.34%	93.75%	95.00%	88.21%	90.18%	86.00%
	K=5	67.01%	57.73%	76.00%	80.71%	84.02%	77.50%	90.09%	88.84%	91.50%	96.26%	96.43%	96.00%	90.57%	91.52%	89.50%
	K=7	69.54%	60.82%	78.00%	79.19%	80.93%	77.50%	89.39%	87.05%	92.00%	96.46%	95.98%	97.00%	91.98%	91.96%	92.00%
	K=9	71.57%	61.86%	81.00%	81.22%	84.02%	78.50%	90.33%	87.95%	93.00%	96.70%	96.43%	97.00%	91.27%	91.96%	90.50%
	K=11	70.30%	60.82%	79.50%	80.96%	83.52%	78.50%	89.62%	87.05%	92.50%	97.17%	96.43%	98.00%	92.69%	93.30%	92.00%
	K=15	71.32%	59.28%	83.00%	81.22%	85.57%	77.00%	90.09%	88.39%	92.00%	96.70%	95.98%	97.50%	92.92%	92.41%	83.50%
	K=20	70.05%	57.73%	82.00%	80.46%	86.60%	74.50%	90.33%	89.29%	91.50%	95.99%	95.09%	97.00%	92.45%	93.75%	91.00%
	K=30	71.57%	56.70%	86.00%	81.47%	88.66%	74.50%	90.09%	89.29%	91.00%	95.75%	94.64%	97.00%	92.22%	94.20%	90.00%
	K=50	72.59%	61.34%	83.50%	80.96%	90.21%	72.00%	86.32%	82.59%	90.50%	93.63%	96.43%	90.50%	90.80%	94.20%	87.00%
Threshold = 0.65	*K=3*	64.72%	53.61%	75.50%	78.68%	82.99%	74.50%	89.15%	86.16%	92.50%	94.58%	93.75%	95.50%	88.68%	89.73%	87.05%
	K=5	66.50%	53.61%	79.00%	79.95%	81.96%	78.00%	90.33%	88.39%	92.50%	95.75%	95.54%	96.00%	90.33%	90.63%	90.00%
	K=7	65.58%	51.03%	79.50%	78.93%	78.35%	79.50%	88.92%	84.82%	93.50%	96.46%	95.98%	97.00%	91.75%	90.63%	93.00%
	K=9	68.02%	52.58%	83.00%	81.71%	81.44%	80.00%	88.68%	84.38%	93.50%	96.93%	96.43%	97.50%	90.57%	90.18%	91.00%
	K=11	69.04%	54.64%	83.00%	**82.49%**	82.99%	82.00%	88.44%	84.38%	93.00%	96.93%	95.95%	98.00%	91.98%	91.52%	92.50%
	K=15	69.04%	53.09%	84.50%	81.98%	81.96%	82.00%	89.62%	86.16%	93.50%	96.46%	95.54%	97.50%	**93.63%**	92.41%	95.00%
	K=20	69.04%	52.58%	85.00%	78.43%	78.35%	78.50%	89.86%	87.05%	93.00%	95.99%	94.64%	97.50%	92.69%	91.96%	93.50%
	K=30	69.80%	52.06%	87.00%	80.20%	81.96%	78.50%	89.15%	87.05%	91.50%	94.34%	91.07%	98.00%	92.17%	91.52%	91.00%
	K=50	70.30%	55.67%	84.50%	81.47%	84.54%	78.50%	83.49%	76.79%	91.00%	93.40%	91.52%	95.50%	89.15%	84.82%	94.00%

Table A-4: Classification testing methods implemented in MATLAB - WWkNN with no feature reduction