Imperial College London

NANYANG TECHNOLOGICAL UNIVERSITY

Decision Support Continuum Paradigm for Cardiovascular Disease: Towards Personalized Predictive Models

Tay Jia Xian, Darwin

Department of Bioengineering
Imperial College London

School of Chemical and Biomedical Engineering
Nanyang Technological University

A thesis submitted to Imperial College London and Nanyang Technological University

In partial fulfillment of the requirement for the degree of

Doctor of Philosophy in Bioengineering

and the Diploma of Imperial College London

2015

This thesis consists of **261** pages (including this page)

Acknowledgement

I would like to extend my utmost appreciation to the gracious extension of invaluable support, encouragement and knowledge imparted by my PhD advisors, Professor Richard I. Kitney (from Imperial College London, United Kingdom) and Assistant Professor Chueh Loo Poh (from Nanyang Technological University, Singapore). Without their consistent inspiration, guidance and insights, I would not have reaped the fruitful results for my doctoral study.

I would like to take this opportunity as well to express my sincere gratitude for their effort to make my PhD experience as wonderful, memorable and rewarding as possible. Moreover, special thanks to Christine L. Doran and Helen S. Challis, staff from the international office at Imperial College London, for making every effort to ensure my stay at London was as pleasant as possible.

Furthermore, without the technological sophistication and scientific research environment provided, I would not be able to carry out my experiments. To this end, I would like to thank the members and staff of the Department of Bioengineering, Imperial College London and Division of Bioengineering, Nanyang Technological University (Singapore).

Finally, I sincerely appreciate the School of Chemical & Biomedical Engineering, Nanyang Technological University (Singapore) and Department of Bioengineering, Imperial College London, for awarding me with the Nanyang Technological University-Imperial College London Joint PhD Scholarship. Additionally, I would like to thank The Engineering and Physical Science Research Council (EPSRC) and the Ministry of Education (Singapore) for their partial support in my work.

Abstract

Clinical decision making is a ubiquitous and frequent task physicians make in their daily clinical practice. Conventionally, physicians adopt a cognitive predictive modelling process (i.e. knowledge and experience learnt from past lecture, research, literature, patients, etc.) for anticipating or ascertaining clinical problems based on clinical risk factors that they deemed to be most salient. However, with the inundation of health data and the confounding characteristics of diseases, more effective clinical prediction approaches are required to address these challenges.

Approximately a few century ago, the first major transformation of medical practice took place as science-based approaches emerged with compelling results. Now, in the 21st century, new advances in science will once again transform healthcare. Data science has been postulated as an important component in this healthcare reform and has received escalating interests for its potential for 'personalizing' medicine. The key advantages of having personalized medicine include, but not limited to, (1) more effective methods for disease prevention, management and treatment, (2) improved accuracy for clinical diagnosis and prognosis, (3) provide patient-oriented personal health plan, and (4) cost containment.

In view of the paramount importance of personalized predictive models, this thesis proposes 2 novel learning algorithms (i.e. an immune-inspired algorithm called the Evolutionary Data-Conscious Artificial Immune Recognition System, and a neural-inspired algorithm called the Artificial Neural Cell System for classification) and 3 continuum-based paradigms (i.e. biological, time and age continuum) for enhancing clinical prediction. Cardiovascular disease has been selected as the disease under investigation as it is an epidemic and major health concern in today's world.

We believe that our work has a meaningful and significant impact to the development of future healthcare system and we look forward to the wide adoption of advanced medical technologies by all care centres in the near future.

Author’s Declaration

I hereby declare that the work presented in this thesis was carried out in accordance to the regulations of both Imperial College London and Nanyang Technological University (Singapore). This work is original except where indicated by special reference in the text. It is conducted at the Department of Bioengineering, Imperial College London and Division of Bioengineering, Nanyang Technological University, under the supervision and guidance of Professor Richard I. Kitney and Assistant Professor Poh Chueh Loo respectively.

Any views and opinion expressed in this thesis are those of the author and are in no way represent those of the aforementioned universities.

Any errors in this thesis are the responsibility of the author.

This thesis has not been presented to any other universities for examination or academic award either in Singapore, United Kingdom or any other countries.

Publications

The work presented in this thesis has resulted in the following original publications.

Peer Reviewed Conference Papers

1. **D. Tay**, C.L. Poh, R.I. Kitney, "An Evolutionary Data-Conscious Artificial Immune Recognition System", 15^{th} Annual Conference on Genetic and Evolutionary Computation (GECCO), pp. 1101-8, 2013.

Peer Reviewed Journal Papers

2. **D. Tay**, C.L. Poh, C. Goh, R.I. Kitney, "A Biological Continuum based Approach for Efficient Clinical Classification", Journal of Biomedical Informatics, vol. 47, pp. 28-38, 2014.

3. **D. Tay**, C.L. Poh, E.V. Reeth, R.I. Kitney, "The Effects of Sample Age and Prediction Resolution on Myocardial Infarction Risk Prediction", IEEE Journal on Biomedical and Health Informatics, 2014. (Accepted)

4. **D. Tay**, C.L. Poh, R.I. Kitney, "A Novel Neual-Inspired Learning Algorithm with Application to Clinical Risk Prediction", Submitted to Journal of Biomedical Informatics, 2015.

5. **D. Tay**, C.L. Poh, R.I. Kitney, "Age-related Risk Prediction for Cardiovascular Disease", Submitted to IEEE Journal on Biomedical and Health Informatics, 2015.

Content

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligent
AIS	Artificial Immune System
AIRS	Artificial Immune Recognition System
ANCSc	Artificial Neural Cell System for classification
ANN	Artificial Neural Network
AP	Angina Pectoris
ARB	Artificial Recognition Ball
AT	Affinity Threshold
ATS	Affinity Threshold Scalar
ATSR	Affinity Threshold Similarity Ratio
AUC	Area under the Receiver Operating Characteristic Curve
BA	Balanced Accuracy
BC	Biological Continuum
BCEN	Biological Continuum-based Etiological Network
BioLINCC	Biologic Specimen and Data Repository Information Coordinating Center
BMI	Body Mass Index
CART	Classification And Regression Tree
CDSS	Clinical Decision Support System
CGP	Cartesian Genetic Programming

CHD	Coronary Heart Disease
CHF	Congestive Heart Failure
CHS	Cardiovascular Heart Study
CI	Coronary Insufficiency
CM	Candidate Memory
CNS	Central Nervous System
CPM	Critical Path Method
CPOE	Computerized Physician Order Entry
CRISP-DM	CRoss Industry Standard Process for Data Mining
CV	Cross-Validation
CVD	Cardiovascular Disease
DT	Decision Tree
EDC-AIRS	Evolutionary Data-Conscious AIRS
EKG	Electrocardiography
EM	Established Memory
EHR	Electronic Health Record
FFT	Fast Fourier Transformation
FP	False Positive
FSNA	Naïve Bayes with Feature Selection
GA	Genetic Algorithm
GWAS	Genome-Wide Association Study

HGP	Human Genome Project
HHP	Honolulu Heart Program
HPC	High Performance Computer
KNN	K-Nearest Neighbour
KS	Kennard-Stone
ICP	Integrated Clinical Pathway
ID3	Iterative Dichotomizer 3
LPT	Learning Plateau Threshold
MANB	Model-Averaged Naïve Bayes
MDR	Multifactor Dimensionality Reduction
MI	Myocardial Infarction
ML	Machine Learning
MLPNN	Multi-Layer Perceptron Neural Network
NB	Naïve Bayes
ND	Neural Density
NHLBI	National Heart, Lung and Blood Institute
NPC	Neuroplastic Coefficient
NPS	Neuronal Pool Size
NPT	Neuroplastic Threshold
NR	Neurogenic Rate
NS	Neurogenic Space

PCA	Principal Component Analysis
PERT	Program Evaluation and Review Technique
PNN	Probabilistic Neural Network
RBF	Radial Basis Function
RF	Random Forest
SNP	Single Nucleotide Polymorphism
SOM	Self-Organizing Maps
SUS	Stochastic Universal Sampling
SVM	Support Vector Machine
TIA	Transient Ischemic Attack
UD	Uniform Design

Chapter 1

Introduction

1.1. Personalized Predictive Models for Cardiovascular Disease

Personalized medicine, first introduced by Hippocrates around 2400 years ago, was about the evolution and increasing precision of diagnosis and treatment (Gordon & Koslow, 2010). With advances in medicine and technologies over the years (e.g. advances in medical knowledge and devices, analytical tools, and information technologies), a paradigm shift in medical knowledge and tools have enabled diagnosis of disease from metaphysical to physical and from cellular to molecular. More recently, system-level interactions between molecular events and higher level phenomena (e.g. cognition and behaviour) have been studied. Currently, disease diagnosis with genetic, molecular and other markers of functional significance is not uncommon. This lead us to be at the verge of making accurate prediction of whether someone will develop a disease in the future, respond positively (or negatively) to a treatment or have any serious reaction to a drug. The use and proliferation of these advanced medical forecasting technologies requires other elements of healthcare system and society to co-evolve in tandem. This includes, but not limited to, laws protecting privacy, systems of payment, regulatory guidelines, physician and patient education, and ethical framework.

Personalized clinical predictions, based on the patient's unique clinical, genetic and environmental characteristics, play an essential role in healthcare decision making and planning. These predictions can lead to improved disease prevention, management and therapeutics strategies, and potentially empower patients to initiate a dialogue that can enhance the wellness plan personalized for them. Common clinical predictions include clinical diagnosis and prognosis of patient's health status. Conventionally, predictions rely on expert knowledge. However, it has become more and more difficult with the exponential increase in informative health data. This inevitably hinders, if not incapacitates, one's ability to recall and analyse the full content of complicated patient's record effectively.

Clinical researchers have invested great effort into developing and optimizing predictive instruments to identify disease status that physicians often find it difficult to define accurately (Baxt & Skora, 1996). Myocardial infarction (MI), for example, is often difficult to ascertain for patient presenting to the emergency department with anterior chest pain. Therefore, reliable predictive models capable of foretelling events of MI are highly desirable. Several risk scoring systems based on generalized linear model have been developed with the assumption of linear relationship between the risk factors and the disease (Nilsson et al., 2006). However, in most cases, the underlying cause of a disease is commonly multifactorial and subtle, with non-linear causal dynamics. On this aspect, if a linear model is used in the presence of nonlinearity, inaccurate modelling would result. This ultimately causes poor generalization and prediction performance. Therefore, a non-linear approach, like machine learning techniques, would be more appropriate to characterize and predict a disease. Machine learning (ML) is a branch of artificial intelligence that postulates a set of computer-based methods for automatic analysis of information and recognition of patterns/concept, through repeated learning from the training data (Roganb et al., 2008). It is capable of identifying the non-trivial/non-linear relationship between the predicators and the outcome, building models capable of making data-driven prediction. Data-driven predictions have the advantage of providing guidance for relatively rare clinical or sub-clinical diseases that could elude a physician, but could be elucidated by the data-driven integration of limitless experiences of many physicians and patients (Chawla & Davis, 2013). This approach may contribute to the transition of medicine from population-based evidence to one that amalgamates both population and individual-based evidence.

To this end, the employment of data mining techniques – suggested by Snyderman et. al. as a "central feature" for future healthcare system (Snyderman & Langheier, 2006) - have received escalating interests for performing diagnosis and prognosis of diseases that physicians often find it challenging to adjudicate accurately. It has been demonstrated in (Baxt & Skora, 1996; Eftekhar et al., 2005; Li et al., 2000) that medical decision support system based on machine learning techniques like artificial neural network (ANN) outperform physicians' judgment and classical statistical models such as multivariable logistic regression analysis. Hence, ML methods have been applied in several clinical domains, aspiring to leverage the performance of clinical diagnosis and prognosis. This is of paramount importance as with improved sensitivity, many lives can be saved while with improved specificity, the healthcare costs can be greatly reduced as unnecessary admission and procedures could be eradicated. Clinical diseases that have been studied with the use of ML techniques include but not limited to, cardiovascular disease (CVD) (Nilsson et al., 2006; Baxt & Skora, 1996; Eggers et al., 2007), cerebrovascular disease (Khosla et al., 2010; Yeh et al., 2011), cancer (Cruz & Wishart, 2006; Liu, 2004) and traumatic brain injury (Mushkudiani et al., 2008; Eftekhar et al., 2005; Li et al., 2000).

However, even with the use of ML techniques, several challenges still exist which prohibits the efficient development of accurate personalized and predictive models. The key challenges addressed in this thesis are summarized below:

1. Unlike manufacturing process, in which the products are standardized, patients are generally different and may not fit well within a standard prediction model. This means that if inadequate consideration was given when designing the clinical prediction models, inefficient development process and poor prediction performance would thrive.

2. Evolving medical knowledge and continual addition of new clinical information result in a large number of clinical features that need to be analysed. This situation, commonly known as the curse of dimensionality (Bellman, 1961), often jeopardizes the ability of ML techniques to learn and generalize.

3. The health status of individuals tends to change over time (e.g. as one ages). Similarly, the concept that underlies the clinical data tends to drift over different prediction scale and intervals. These, if not handled properly, often degrade the performance of the prediction models.

On this note, the ability to (1) efficiently handle large number of clinical features, (2) understand the design issues related to the development of clinical prediction models (e.g. sample peculiarity), and (3) recognize the importance of employing learning algorithms with high generalization ability are valuable for the development of patient-oriented prediction models. Achieving these aspects would ultimately leverage on the diagnostic/prognostic performance, increase efficiency and lower the cost incurred by both the hospital and the patients (e.g. cost containment through early diagnosis and eradication of unnecessary clinical tests). In this thesis, the disease we focus on is cardiovascular disease – one of the leading cause of death worldwide (World Health Organization, 2008; Go et al., 2013).

1.2. Motivations

Clinical decision making, such as disease diagnosis and prognosis, is a coveted and elusive clinical task. In the United States (U.S.), for example, missed or wrong diagnosis is not uncommon and has detrimental implications – e.g. causing preventable and permanent damage or death (Tehrani et al., 2013). This however, when carried out properly, would significantly improve the quality of healthcare and saves many lives. Further, with the clinical data deluge in today's healthcare industry, the performance of effective analysis of clinical data becomes a challenge for a healthcare provider; the large amount of data that needs to be processed concurrently is beyond the human scale of thinking and analysis. Hence in this thesis, we investigate on methods to alleviate and ameliorate the task of clinical decision making. Particularly, computational methods based on the idea of artificial intelligence (AI) for predicting cardiovascular health outcomes are delved into. It is noteworthy that these predictive models can be highly effective and efficient in providing instant clinical prediction on the likelihood of a disease when properly calibrated. Additionally, it has been demonstrated to achieve comparable, if not better, predictive accuracy as clinicians (Baxt, 1991; Harrison et al., 1991). With such system present in the clinical settings, it has been postulated to enhance clinicians' judgement.

Coronary heart disease (CHD), the narrowing or blockage of blood vessels that supply oxygen and nutrients to the heart, is the leading cause of mortality in many developed countries, such as the U.S. and the United Kingdom (U.K.) (Go et al., 2013; Wilson et al., 1998; British Heart Foundation Statistics Database, 2010); accounting for approximately 12.7% of all global deaths (in 2008). Despite considerable advances in medicine, approximately 1 in every 6 deaths in the U.S. (in 2007) is caused by CHD. Moreover, MI, a form of CHD, approximately occurs every 34 seconds in the U.S. and about 15% who experience MI will die from it (Go et al., 2013). This places a heavy burden on the healthcare systems (Leal et al., 2006; McGovern et al., 1996; Jemal et al., 2005). The complexity of MI arises from the fact that multiple subclinical and clinical diseases typically interact with each other in a complicated and often unknown manner. Moreover, it is unlikely that a single aspect of health status to be the sole predicator (Fried et al., 1998; Song et al., 2004). This results in very significant challenges in relation to the analysis and understanding of the disease. Therefore to reduce the number of MI incidences, improved methods of detection and management are necessary. Clinical decision support system (CDSS) is one such method developed to assist physicians and other healthcare professionals in making decision for clinical tasks like diagnosis. One type of CDSS uses AI, or more specifically ML methods, to learn and consolidate the knowledge required to perform the clinical tasks. The motivation for creating such intelligent computer system was to create the perfect "doctors in a box" in an attempt to improve the ability to detect, manage and treat different types of disease. The benefits of having such system include:

1. Bridging the gap between individual practitioners (through the condensation of the most up-to-date knowledge and experience) so that they can aspire to the same level of practice as the best in their field and offered copious experience to gauge the impact of diseases.

2. Serving as a second opinion for the patients, a highly recommended step for ascertaining the diagnosis and determining the course of treatment. In addition, it offers the reassurance that is much needed by the patient (i.e. whether the best possible choice of treatment is offered to the patient).

3. Providing predictive tools capable of offering personalized and preventative means to medicine. The potential advantages for this approach include early detection and intervention, more precise diagnosis and prognosis, more appropriate selection of treatment strategies, and cost containment among others.

Although CDSS based on ML techniques has shown improved clinical prediction performance over conventional methods on many clinical problems (Kim et al., 2005; Song et al., 2004; Li et al., 2000), we are still far from the goal of personalized and predictive medicine. Personalized medicine is a model capable of recommending decisions for diagnosis, treatment and prevention that are specific to an individual patient. The recommendation is often based on the patient's unique clinical, genetic and environmental characteristics (Lenfant, 2012). This differs from the traditional approach where patient care is based on generalization from randomized controlled clinical trials. The significant disadvantage associated with this traditional approach is the delivery of "average medicine" where clinical interventions are offered to a patient on the basis that they, from a statistical perspective, work well on other patients. This drawback is supported by studies in the field of pharmacogenomics where each individual (with unique genotypic makeup) has different degree of response with respect to a specific type and amount of drug (Evans & Relling, 1997). Therefore, the selection of therapy based on large randomized clinical trials may soon be replaced because of the estimated benefits it promises for individual patients. Another goal of personalized medicine is to offer healthcare professionals and even patients with predictive tools that encourage healthcare to be more proactive and preventive, allowing appropriate medical interventions to be carried out early to prevent or procrastinate the onset of the disease. This differs from the conventional approach where it is disease-oriented, reactive, episodic, and geared towards acute crisis intervention where the disease has already manifested and largely irreversible (Ginsburg & Willard, 2009).

With the completion of the Human Genome Project in 2003 (Austin, 2003), it provides clinicians and scientists with a diverse and important set of molecular information that can be used to better understand the mechanisms that underpin a disease. However, the use of gene expression profiles to define broad group distinctions has its limitation; resulting in considerable heterogeneity within the broadly defined groups and poor clinical predictions for individual patients. On this note, a more holistic analytical approach is required to improve the prediction accuracy. Beyond information at the molecular and gene levels, clinical observations (e.g. electrocardiography, blood pressure, ultrasound data, magnetic resonance image, etc.) have been used and have shown great potential in understanding the biology of a disease experienced by an individual (Hsia et al., 2003). These inevitably offer the conceptual advances necessary to drive the healthcare system to one that is predictive and patient-oriented. This growing transition in the healthcare system has been utilized to stratified risk for several diseases such as cancer, cardiovascular disease, traumatic brain injury, and diabetes (Lisboa & Taktak, 2006; Vellido et al., 2008; Nevins et al., 2003; Tsai & Watanabe, 1999; Polat et al., 2006; Li et al., 2000; Polat & Güneş, 2007; Barakat et al., 2010). These developments clearly serve as examples where CDSS and patient's unique clinical/genetic characteristics have resulted in the opportunities to better characterize diseases and at the same time redefine therapeutic strategies.

1.3. Thesis Contributions and Objectives

This thesis contributes to the development of novel machine learning algorithms, their application for solving clinical problems, and new methodologies for addressing issues unique to clinical predictions. A total of 2 new algorithms and 3 continuum paradigms (i.e. biological, time and age continuum) were developed as part of the pursuit to (1) ameliorate the task of clinical predictions (e.g. through the development of more robust and accurate learning algorithms), and (2) create clinical models that are carefully calibrated and personally effective for the individuals (e.g. by recognizing the potential advantages of incorporating the concept of continuum for clinical prediction). The following are the main contributions:

- **Biological Continuum Model for Clinical Prediction**: Clinical classification, based on machine learning techniques, provides the disease diagnosis for an individual. It is a significant task of pragmatic value in the clinical settings. However, with the exponential growth of clinical features in the healthcare industries, the efficient development of up-to-date and efficacious clinical classification models becomes a challenge. To address this issue, we propose a novel feature selection methodology for the development of clinical classification model. We believe that information about one level of concept should be, in many cases, generalized to other levels. Hence, we employed the conceptual framework of biological continuum (BC) (Kitney & Poh, 2006; Poh et al., 2007), together with the optimization capability of genetic algorithm

(GA) (Holland, 1992) and the classification ability of support vector machine (SVM) (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1999) to build clinical classification models. Evaluation of the proposed method was carried out using the cardiovascular heart study (CHS) dataset (Fried et al., 1991). Results demonstrate that with the adoption of this methodology, a significant speedup of 4.73-fold (when compared to conventional GA based wrapper approach using SVM) can be achieved for the development of clinical classification model without compromising the classification accuracy.

- **Evolutionary Data-Conscious Artificial Immune Recognition System (EDC-AIRS)**: We introduce a novel immune-inspired supervised learning algorithm that exploits on 3 human immune system's phenomena observed – namely the (1) increase in the concentration of antibodies in antigen infected regions, (2) spatial independency and distribution of lymph nodes across the human body, and (3) characteristics and specificity of surface receptors on B-Cells necessary to recognize and bind to a certain type of antigen. This algorithm, called EDC-AIRS, is an optimized version of the artificial immune recognition system version 2 (AIRS2) algorithm proposed by Andrew Watkins in 2004 (Watkins et al., 2004). The key difference between EDC-AIRS and AIRS2 algorithms is that EDC-AIRS algorithm contextualizes the immune response to the concentration, distribution and characteristics of the antigens and is no longer a global centralized response. Empirical experiments with 6 benchmark datasets showed promising results and clinches a place in the top 3 positions when compared to other state-of-the-art classification algorithms.

- **Time-related Continuum Model for Clinical Prediction**: The best practice to avoid human mortality caused by diseases is to detect them early and prevent its onset. The ability to do so in current clinical practice is highly attractive but is equally challenging. Therefore, a more holistic, sophisticated, predictive, preventative and personalized approach is required to detect diseases early. This is very important as delayed treatment could cause permanent damage where full recovery becomes impossible or even death. Therefore, we investigate on the use of SVM and EDC-AIRS algorithms as classifiers to predict the risk of individuals experiencing MI over different time scales and intervals (using baseline datasets having different sample age). The CHS observational study (Fried et al., 1991), which contains a comprehensive set of biomarkers, was analysed. Results indicate that SVM algorithm is capable of achieving high sensitivity, specificity and balanced accuracy of 95.3%, 84.8% and 90.1% respectively over a time interval of 6 years. Further, experiment results indicate that sample age, prediction scale and intervals do not have a significant impact on prediction models developed using subjects and 65 and above. This opens the opportunity for constructing prediction models capable of detecting MI early, allowing clinicians to take preventative measures promptly, improving the quality of individuals' life, and reducing avoidable mortality.

- **Artificial Neural Cell System for classification (ANCSc)**: We propose a novel neural-inspired supervised learning algorithm for solving classification problems – one of the most common and well-studied tasks in predictive data mining and knowledge discovery. It is developed based on new source of inspirations that are responsible for developing and enriching the brain – namely neurogenesis, neuroplasticity, nurturing and apoptosis. This novel algorithm, called ANCSc, capitalizes on the mechanisms associated with the brain's ability to (1) produce new neurons during both prenatal and postnatal development phases (i.e. neurogenesis), (2) refine neural pathways and synapses in support for learning and adapting to changes (i.e. neuroplasticity), (3) promote neurogenesis and neuroplasticity when knowledge are inculcated to individuals (i.e. nurturing), and (4) programmed cell death of redundant cells during an organism's lifecycle. Evaluation of ANCSc algorithm with 6 benchmark datasets demonstrated that it is a robust learning algorithm capable of achieving highly competitive classification results.

- **Age-related Continuum Model for Clinical Prediction**: The performance of prediction models not only relies on the predictive ability of the learning algorithm used, but is also highly dependent on the quality and characteristics of the data analysed. In order to continuously ameliorate the performance of prediction models, it is necessary to explore and investigate on different components that would influence the performance of prediction models. One clinical study suggests that differences in degree of severity of CVD risk factors as one age plays a crucial role in age-related excess risk for CVD. Hypertension and diabetes, for example, tend to prevail with age while total cholesterol levels and body mass index (BMI) often decline with age (Abbott et al., 2002). This indicates a confounding and evolving role CVD risk factors take. Therefore, to investigate whether this observation has an impact on prediction models developed using machine learning algorithms, we propose a (age-related) risk prediction approach that takes the effect of evolving risk factors (over a range of ages) into consideration – i.e. develop risk prediction models using only individuals from a specific age group. Three algorithms, namely ANCSc, EDC-AIRS and SVM, were employed to develop these risk prediction models. Juxtaposition of these algorithms was performed to investigate on their ability to generalize. Data from the Honolulu Heart Program (Syme et al., 1975; Marmot et al., 1975; Robertson et al., 1977) were utilized to perform this risk prediction task. Results demonstrate that age-related risk prediction outperforms unified risk prediction approach (i.e. prediction model developed using individuals of all ages). This offers the advantage of providing a spectrum of accurate prediction models suitable for individuals of all ages; enabling a continuum of high quality healthcare to be given to the patients.

1.4. Thesis Organization

The task of developing effective and efficient MI risk prediction models is very important. The ability to do so is highly desirable as it would allow early detection of MI risk, and consequently enables preventative measures to be given promptly. This would inevitably increase the opportunity of avoiding the full manifestation of the disease. Such proactive approach would ultimately improve the quality of individuals' life as they would not need to undergo the painful experience that is associated with MI. To this end, this thesis describes methods for developing accurate MI risk prediction models in an effective and efficient manner.

The structure of the thesis is illustrated in Figure 1.1 and described as follows. Chapter 2 provides background information to our studies which include cardiovascular disease, personalized predictive medicine and clinical support technologies.

Chapter 3 proposes a methodology for alleviating the computational effort needed to construct up-to-date clinical classification models. Detailed information on feature selection, data pre-processing, structure adopted (i.e. biological continuum), data used (i.e. CHS dataset), and classification model development are provided. The performance and speedup achieved for our proposed method is also presented.

Chapter 4 introduces an optimized immune-inspired supervised learning algorithm called EDC-AIRS. Background information of AIRS (version 2) algorithm, detailed description of EDC-AIRS algorithm and the classification performance of our proposed algorithm tested using 6 widely benchmarked datasets are provided.

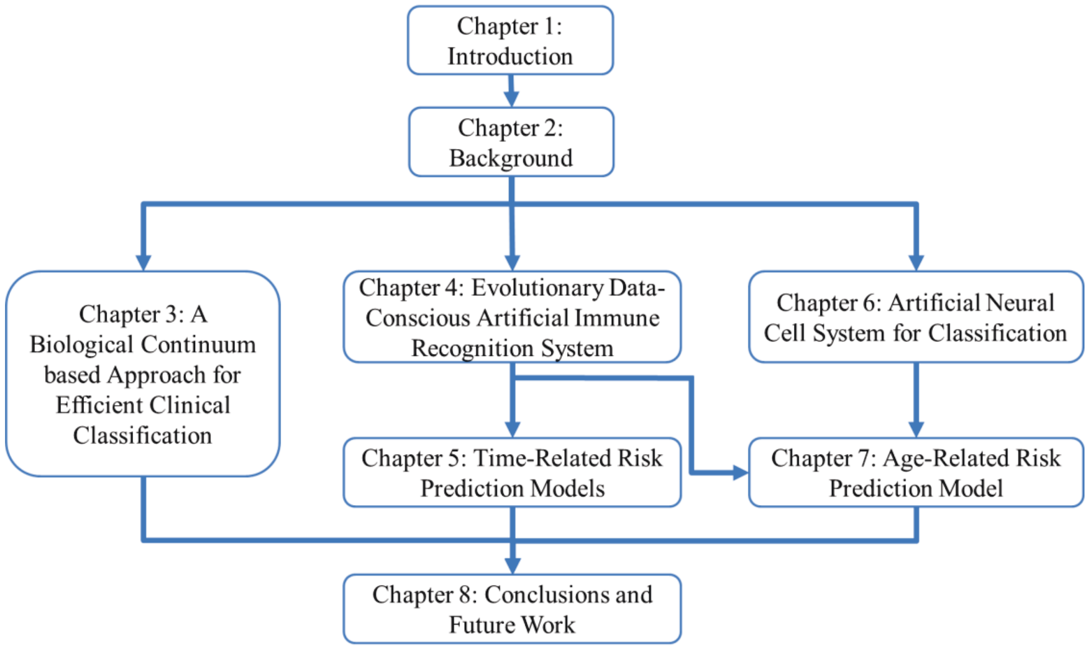


Figure 1.1: Structure of the Thesis

Chapter 5 investigates on the importance of performing MI risk prediction (1) using baseline data comprising individuals in different age range (i.e. sample age), and (2) over different time scale and interval (i.e. prediction resolution). SVM and EDC-AIRS algorithms are used as an approach to perform this task; promoting a personalized, predictive and preventative diagnostic framework capable of detecting the risk of individuals experiencing MI, improving the quality of healthcare and life of individuals.

Chapter 6 describes a novel algorithm inspired by neurogenesis (i.e. generation of new neurons), neuroplasticity (i.e. change in neuronal structure and connectivity), nurturing (i.e. inculcation of best knowledge and practice), and apoptosis (i.e. programmed cell death of redundant cells). This new algorithm, called ANCSc, achieved highly competitive results when benchmarked with 6 datasets made available in the UCI machine learning repository (C.L. Blake & C.J. Merz, 1998). Detailed information on the source of inspirations, how the algorithm works, its classification performance and the sensitivity of ANCSc's parameters are presented.

Chapter 7 delved into the significance of developing risk prediction models using clinical data of individuals stratified into different age groups; an important step in the construction of accurate prediction models suitable for individuals of all ages. This age-related risk prediction approach is hypothesized to be important as CVD risk factors tend to evolve and confound the disease as one age. This task was carried out using several algorithms (i.e. ANCSc, EDC-AIRS and SVM). Investigation on which algorithm is most capable at performing risk prediction for CVD is also conducted.

Finally, Chapter 8 summaries the achievements of this thesis and provides the possible directions for future research.

Chapter 2

Background

2.1. Cardiovascular Disease

Cardiovascular disease (CVD) is a group of diseases associated with the heart and/or blood vessels. It includes disorders that cause (1) narrowing of blood vessels supplying blood to the heart (i.e. coronary heart disease), brain (i.e. cerebrovascular disease) and limbs (i.e. peripheral arterial disease), (2) damage to the heart muscle and heart valves from rheumatic fever (i.e. rheumatic heart disease), (3) weakening of heart muscle to pump adequate blood into the blood vessels (i.e. congestive heart failure), (4) abnormal formations of heart structures at birth (i.e. congenital heart disease), and (5) formation of blood clots in leg veins which could give rise to serve pain and disability, or even life threatening complications when the clots dislodge and move to the heart and lungs (i.e. deep venous thrombosis and pulmonary embolism) (World Health Organization, 2013).

An acute event that is of particular interest is acute myocardial infarction (MI) - commonly known as heart attack. This is because it is a deleterious health issue experienced by people worldwide; causing substantial mortality (Go et al., 2013; Wilson et al., 1998; British Heart Foundation Statistics Database, 2010). MI is often linked to atherosclerosis - the formation of plaque that builds up in the walls of the blood arteries, narrowing them, and increasing the difficulties for blood to flow through. MI events usually arises when myocardial ischemia (an inadequate supply of blood to the heart) occurs for a considerable period of time, overwhelming the myocardial cellular repair mechanisms designed to support the normal operating function and homeostasis of the cardiovascular system. If this imbalanced supply and demand of blood (or more specifically, oxygen and nutrients) reaches a critical threshold and left for an extended period of time, it would result in an irreversible myocardial cell damage or necrosis. Such event is often caused by plaque rupture with thrombus formation in a coronary vessel, and in the most unfortunate case it would lead to the death of a person.

In view of the detrimental impact of MI on the society, several epidemiology studies have been carried out to better understand and characterize the disease. This includes the Cardiovascular Health Study (CHS) (Fried et al., 1991), the Honolulu Heart Program (HHP) (Robertson et al., 1977; Marmot et al., 1975; Syme et al., 1975), the Framingham Heart Study (O'Donnella & Elosua, 2008) and the INTERHEART study (Ounpuu et al., 2001). These studies have identified major risk factors associated with CVD which include age, gender, cholesterol, hypertension, obesity, diabetes, smoking, alcohol, psychosocial factors, sedentary lifestyle and unhealthy diet (Yusuf et al., 2004; Hubert et al., 1983; Psaty et al., 2001; Stokes et al., 1989; Yano et al., 1984; Anand et al., 2008).

Broadly, risk factors can be categorized into 2 groups, namely non-modifiable and modifiable risk factors. The non-modifiable risk factors include age, gender, race and family history while the modifiable risk factors include blood pressure, cholesterol, body mass index, diabetes, smoking, diet and physical activities among others. Identification of these risk factors is important as they are measurable elements or characteristics that are causally correlated to an increased risk of a disease (O'Donnella & Elosua, 2008). However, caution need to be taken when analysing risk factors as their degree of impact on individuals' health may change as one ages (Asia Pacific Cohort Studies Collaboration, 2006) (which will be addressed in this thesis). Therefore, careful monitoring, analysis and management of these risk factors could reduce mortality rate.

Several risk scoring systems and survival curves (Clayton et al., 2005; Lloyd-Jones et al., 2004; Levy et al., 2006) have been proposed for clinical risk prediction. However, these models (e.g. logistic regression models) tend to flounder as the number of interacting predictors that need to be analysed becomes large (e.g. in genome-wide studies). This is because when identifying interactions between predicators with traditional statistical methods, there is a need to specifically detail a model for the interaction (McKinney et al., 2006). For example, in logistic regression, interaction between polymorphisms A and B need to be explicitly specified in the logistic equation in order to allow for interaction between the polymorphisms. This problem becomes increasingly severe when the number of predictors, and thus the number of possible interactions, becomes large. Another caveat of a traditional statistical model is that it assumes that the predicators (e.g. genes or clinical measurements) are independent and that linear combination of these predictors can successfully describe the underlying patterns and predict the outcome (e.g. disease status). However, this conceptual phenomenon is not common in most biological systems (Cruz & Wishart, 2006). All these challenges often hinder the ability of traditional statistical models to identify and characterize the predictors' interactions and the biological pathways that underpin a disease. Therefore, caution has to be exercised when employing these models. Nevertheless, they have been a good tool (in conventional medicine) for estimating the risk of an individual experiencing or re-experiencing a disease.

Technological advances and accelerating pace of change in healthcare; including, new modalities, socio-economic needs for cost containment (through prevention and early diagnosis and prognosis), and escalating demands for personalized therapy (Vellido et al., 2008), predicates for tools capable of offering patient-specific diagnoses, prognoses and recommendations. This has been part of an effort towards a predictive, preventative and personalized (3P) approach to medicine (Snyderman & Williams, 2003). One goal of such "3P" concept to medicine is to provide clinicians with sophisticated, efficient and effective risk assessment methods, and patients with early, accurate and personalized diagnosis that could prevent the onset of MI; thus improving their quality of life and reducing preventable mortality (which will be addressed in this thesis).

2.2. Personalized Medicine

Personalized medicine, a form of medical innovation, refers to the use of genomic signatures of patients in a target population to ameliorate diagnosis accuracy, allow early interventions that could prevent or delay the onset of diseases, and promote the assignment of more effective therapies (Moon et al., 2007). It has been recognized to have major impact to human health and has been identified as one of the grand societal challenge to which engineers (e.g. biomedical engineers, biologists, computer scientists, etc.) can contribute (i.e. to the advance of personalized healthcare) over the next 20 years (College of Fellows, 2013). The prospect of offering individualized risk predictions and treatment decisions through the examination of individual's genomic details has been attractive, albeit a challenging one (Ginsburg & Willard, 2009). Personalized medicine has become possible, in part, due to the Human genome project (HGP) (Collins & Galas, 1993) and the Genome-wide association study (GWAS) (U.S. Department of Health & Human Services, 2013). The use of genetic information has been the key player in certain aspect of personalized medicine since inception. However, its scope has broadened over the years to include various types of personalized measurements like clinical data and environmental triggers - including several other objectives such as: the identification of risk factors, diagnostic features, and therapies based on large healthcare databases; remote monitoring of individual's compliance with treatment regimens; and scrutinization of relationships between an individual environment and his or her health (College of Fellows, 2013).

Personalized medicine differs from traditional medicine is several aspects. In particular, traditional clinical diagnosis and management focus on medical and family history, observable clinical signs/symptoms, laboratory results, and imaging data for diagnosing and treating patients. This often is described as a reactive approach where treatment begins only when the signs of the disease appear (e.g. when the patient is in pain or their daily life has been affected by the disease). With the increasing emphasis to improve the quality of life among individuals, there is a need for a more proactive approach where diseases are detected and treated early before they fully manifest. To do so, a more personalized approach is necessary, where patients are examined in an attempt to identify the disease signature specific to each patient at the individual level. This is important to ensure that the most appropriate, effective, and ideally a non-invasive healthcare intervention, plan and/or recommendation is given to the patient.

Despite the advantages associated with the analysis of genomic data, several challenges exist. Bioinformaticians, for example, faced the difficulties of the need to (1) process large-scale robust genomic data; (2) interpret the functional impacts of genomic variation; (3) integrate data to relate complex interactions with phenotypes; and (4) translate these discoveries into clinical practices (Fernald et al., 2011). Other challenges include the high failure rate of molecular targeted therapeutics, unexpected effects on patient outcomes cause by bypass mechanisms, and the difficulties of identifying and validating the molecular markers, homeostatic feedback loops and molecular crosstalk (Gonzalez-Angulo et al., 2010). Although these challenges impede our advancement towards the complete understanding of the biological complexity and eventually personalized cure to a disease, the ability to circumvent these issues would have far-reaching clinical ramifications. This, ultimately, would enable a more comprehensive, effective and safe (i.e. with no side effects) therapeutic strategies to be developed. Through this personalization of therapeutic interventions, it is aimed that not only years are added to individual’s life, but life is also added to those years.

2.3. Clinical Support Technologies

Several clinical support technologies are available to directly support clinical tasks and leverage on the benefits that patients, healthcare professionals and healthcare systems would eventually accrue. This includes electronic health record (EHR) system, computerized physician order entry (CPOE) system, integrated clinical pathways (ICP) and clinical decision support system (CDSS). EHR system refers to an aggregated computerized legal medical record system that allows the storage, retrieval and manipulation of patients' health and history records across multiple locations. It is an important component in current clinical settings as accurate clinical decision making is highly dependent on the amount of viewable clinical data. Clinical studies have shown that with EHR system integrated into daily clinical practice, the immediate benefits gained include improvement in quality of care, decrement in medication errors, reduction in cost, and improve availability, timeliness and accuracy of clinical data (Wang et al., 2003; Hillestad et al., 2005). In view of such advantages and guidelines from the Health Information Technology for Economic and Clinical Health (HITECH) Act (U.S. Department of Health and Human Services, 2009) – which provides incentives to healthcare providers that adopt health information technology to advance clinical processes and improve outcomes – EHR is becoming more common in (U.S.) clinical settings (Neill, 2013).

CPOE system refers to a variety of computer-based systems that enable electronic medication ordering and ensures standardized, legible and complete orders (Kaushal et al., 2003). CPOE systems alone, however, offer limited benefits without CDSS (Sittig & Ash, 2009). CDSS system encompasses a variety of tools and interventions which include: computerized alerts, reminders and recommendations; clinical guidelines; order sets; patient data reports and dashboards; documentation templates; diagnostic/prognostic support; and clinical workflow tools (Osheroff et al., 2007). Therefore, CPOE with CDSS is essential to enable medication orders to be integrated with the patient medical information (from EHR system) and automatically cross-referenced to identify potential medical errors. The impact of CPOE with CDSS on quality of care has been studied with positive results showing better adherence to clinical guidelines, decline in medication errors (such as drug dosage errors, frequency errors, route errors, drug allergies, incorrect therapy and wrong contraindications), reduction in unnecessary healthcare utilization, hospital admission and hospitalization duration (Bates et al., 1999; Kaushal et al., 2003; Eslami et al., 2008; Garg et al., 2005).

An ICP is a multidisciplinary plan that displays goals for patients, and provides the sequence and timing of actions necessary to achieve these goals with optimal efficiency (Uzark, 2003). The concept of ICP was first demonstrated in the industrial sectors as a tool to define, organize and manage the essential tasks and rate-limiting processes. Examples of such pathways include ‘program evaluation and review technique’ (PERT) and ‘critical path method’ (CPM), which were developed to assist with the planning and scheduling of tasks (Chu & Cesnik, 1998). The success of these pathways (in terms of both cost and productivity) was quickly being realized. Subsequently, similar tools were adopted in the healthcare industries in response to the rising healthcare costs and clinical demands. In the clinical context, the primary aims of ICP are to: provide high-quality and safe patient care that is delivered in a timely, organized and cost effective manner (Kwan, 2007); promote evidence-based and guideline-based care; standardize the care processes; increase use of recommended medical therapies; decrease use of unnecessary tests; decrease the hospitalization duration; provide a framework for data collection and analysis; alleviate documentation burdens; and improve patient satisfaction (Cannon & O'Gara, 2007; Cheah, 2000). It is noteworthy that cardiovascular medicine is an area in which clinical pathways have embraced. This is due in part to the high volume and high cost associated with CVD and the related procedures (Every et al., 2000).

One key problem with current ICP is that it often addresses processes in the "ideal" patient, and in some cases do not address issues in the majority of patients who enter the pathway. Hence, placing patients within a standardize pathway may not be beneficial (Every et al., 2000) as each patient may have unique response to the given clinical interventions. Therefore, the development of patient-specific pathways is highly desirable for providing personalized care.

Generally, for any support system (that aids in decision making) to be considered useful in the clinical settings, it must possess at least one of the following characteristics (Lisboa, 2002):

1. *Attention focusing* or *alerts* that aim to notify users of any abnormalities which might otherwise be overlooked.

2. *Patient-specific assessments and advices* that provide customized medical recommendations (e.g. diagnostic and prognostic inferences) for individual patient. It must have the ability to make medical predictions with comparable, if not better, accuracy than a human physician.

3. *Interactive tools for critiquing, analysing, planning and testing clinical hypotheses*. This would allow discovery of new insights about a patient's condition or the possible effects of different treatment choices.

2.3.1. Clinical Decision Support System

CDSS refers to any electronic system designed and developed to objectively assist in clinical decision making. An aspect that is of particular interest in recent years is its capability to analyse the characteristics of individual patients to generate patient-specific assessments or recommendations; which are then presented to the clinicians for consideration (Bright et al., 2012). CDSS has emerged as one of the most important components in future healthcare due to its capability to capitalize on the wealth of clinical information reaped from day-to-day clinical practice - which otherwise are left unexploited - providing data-driven recommendations for clinical processes like diagnosis and prognosis, and even discovery of new medical insights (e.g. the underlying mechanisms of a disease). Moreover, it can integrate and offer the functionality of several systems (e.g. EHR, CPOE and ICP), making it a powerful tool whose value and usefulness in the clinical settings should not be underestimated.

Not only does CDSS help in clinical decisions, it may open the possibilities of bridging the gap between individual practitioners, allowing them to aspire to the same level of practice as the best in their field. In view of these benefits it can offer in the healthcare industry, the prospect of CDSS has become increasingly attractive with many interesting works put forward by the artificial intelligence research community (Chawla & Davis, 2013; Baxt, 1991; Wiens et al., 2012; Khosla et al., 2010; Neill, 2013) - postulating approaches that amalgamate both knowledge-driven and data-driven concepts for medical related analysis and management. With the continual advancement of CDSS, it is aimed that it would eventually enable physicians to focus on tasks where they are most needed (e.g. at the patient's bedside, listening and understanding their problems, comforting them, etc.), leaving the task of recalling, searching and analysing the "encyclopedic" aspect of medicine to CDSS.

The uptake of CDSS in current clinical settings is slow despite the critical role it plays in the emergence of personalized healthcare, and the ubiquity of computer systems in the commercial, industrial and scientific areas to enhance the accuracy, efficiency and productivity. Typical barriers that hinder the wide adoption of CDSS include (Miller & Sim, 2004; Coiera et al., 2003; Garg et al., 2005; Hillestad et al., 2005):

1. Additional time and efforts required by physicians to learn and deploy the system before it can be used effectively and efficiently for their daily tasks.

2. Concerns about physicians being overly dependent on CDSS, resulting in eroded capacity in making independent decision.
3. CDSS fit poorly into the current clinical practice, either solving issues perceived as trivial or imposing changes in the way clinicians worked.
4. Scepticism over the applicability of CDSS in terms of their explanatory ability, adaptability to the changing population and the capability of adjusting to the idiosyncratic health phenomena exhibited by population from different regions.
5. Uncertainty in the degree of proven benefits needed to be demonstrated before mass deployment should be carried out.
6. High upfront implementation cost and uncertainty in lucrative benefits.
7. High disincentive for healthcare providers to invest in these systems while the savings go to the patients. This misalignment of incentives hinders healthcare transformation.
8. Compatibility issues between heterogeneous systems and the lack of ubiquitous data exchange between different disciplines.
9. Dearth of physicians' exhortation and supports.
10. Legal considerations.

In face of these multifactorial obstacles that healthcare organizations may encounter in their efforts to wide deployment of CDSS, and the increasing demands for CDSS to be effective and adaptive under unprecedented circumstances, ten grand challenges that impede the inception of high quality, effective means of designing, developing, presenting, implementing, evaluating and maintaining all types of CDSS capabilities for clinicians, patients and consumers have been identified (Sittig et al., 2008). Proposed and listed in their order of importance, that when solved, the full potential of these systems can be realized are to:

1. Improve the human-computer interface whereby the presentation of the CDSS recommendations should support and not interrupt the clinical workflow. This requires developers of modern CDSS to take a socio-technical approach where the goals of CDSS go beyond the original focus of producing expert-level advisories and extend to encompass support for tasks like producing better documentation, retrieving relevant literature and facilitating communication among providers (Peleg & Tu, 2006).
2. Identify, describe, evaluate, collect, categorize, synthesize and disseminate the best practices for CDSS design, development, implementation, maintenance and evaluation.
3. Intelligently and automatically summarize all patient-level information, allowing 'at a glance' assessment of patient status.
4. Automatically prioritize and filter recommendations according to a multi-attribute utility model by combining both patient-specific and provider-specific data.
5. Create architecture for sharing executable CDSS modules and services so that one can implement new state of the art CDSS interventions with little or no extra effort on their part.
6. Identify and eliminate redundant, potentially discordant or mutually exclusive guideline-based recommendations for patients with co-morbid conditions or multiple medications.
7. Prioritize CDSS content development and implementation according to (1) value to patients, (2) cost to the healthcare system, (3) availability of

reliable data, (4) difficulty of implementation, and (5) acceptability to clinicians and patients, among others.

8. Create internet-accessible CDSS repositories that allow these interventions and services to be easily downloaded, maintained, locally modified and installed.
9. Extract clinical information contained in the free-text portions of electronic health record systems into a form that would drive CDSS.
10. Mine large clinical databases to create new, valuable guidelines and CDSS interventions.

Similarly, Bates et al. investigated the common factors that lead to successful implementation and stated Ten Commandments for effective CDSS (Bates et al., 2003). The Ten Commandments are: (1) speed is everything (i.e. the speed of the information system – for example, process and response time - is highly important); (2) anticipate needs and deliver in real-time (i.e. provide the appropriate information to the clinicians at the time they need it); (3) fit into the user's workflow (i.e. provision of appropriate guidelines, on the same screen, to clinicians when they are in the process of ordering); (4) little things can make a big different (i.e. usability of CDSS is very important); (5) recognize that physicians will strongly resist stopping (i.e. physicians dislike suggestions that resist the performance of an action without providing an alternative); (6) changing direction is easier than stopping (i.e. modifying clinician behaviour can be carried out more easily when the change is a single attribute of an order which the clinician does not have strong disagreement with – for example recommending changes to the dose, route or frequency of a medication); (7) simple interventions work best (i.e. substantial condensation and simplification of guidelines onto a single screen is essential); (8) ask for

Table 2.1: Possible Features Leading to an Effective CDSS

Type	Feature
General System Features	1. Support workflow integration using charting or order entry system. 2. *Generation of decision support using a computer.*
Clinician-System Interaction Features	1. *Provision of automatic decision support as part of clinician workflow.* 2. Eradication of requirement for additional clinician data entry 3. Documentation of reason for not following CDSS recommendations is required. 4. *Provision of real-time decision support.* 5. Execution of recommended orders by agreement.
Communication Content Features	1. *Provision of recommendation, not just assessment.* 2. Encourage execution of action rather than inaction. 3. Provision of reasoning for the recommended action. 4. Justification of decision support with the provision of research evidence.
Auxiliary Features	1. Involvement of local users during the development process. 2. Provision of decision support results to patients as well as providers. 3. Performance of periodic CDSS performance feedback by users. 4. Provision of conventional education on the use and features of the deployed CDSS.

Features stated in italics are strongly correlated to the implementation of effective CDSS.

additional information only when you really need it (i.e. plans must be made to handle situations when providers do not provide the piece of required information, and over time, ensure that key information are collected as part of the routine care); (9) monitor impact, get feedback and respond (i.e. track and assess suggestions and make appropriate midcourse correction); and (10) manage and maintain your knowledge-based systems (i.e. evaluate system usage pattern and ensure that it is in pace with changes in medical knowledge).

Additionally, Kawamoto et al. described numerous potentially important features that (Kawamoto et al., 2005). Fifteen of these features (categorized into four groups) are listed in Table 2.1. In particular, four of these features: (1) provision of automatic decision support as part of clinician workflow, (2) provision of could lead to the implementation of an effective CDSS recommendation rather than just an assessment, (3) provision of real-time decision support, and (4) generation of decision support using a computer, demonstrated strong correlation to the implementation of effective CDSS. Therefore, they are highly recommended to be implemented whenever possible.

It is also important to examine and understand the characteristics of the data CDSS are learning from. For instance, human medical data are known to have their own unique characteristics which may pose challenges for medical data mining. Some of these dominant characteristics include voluminous and heterogeneous raw medical data, lack of standardization in disease description, poor mathematical characterization of medical data, lack of canonical form in biomedicine, data ownership, privacy and security of human data, strong obligatory towards statistical philosophy (Cios & Moore, 2002), sparseness in outcome events, redundancy in medical records, conflict in patients' predictors and outcomes, and sequential recording of medical records (Suka et al., 2008). In spite of the challenges, human medical data are vital and rewarding to mine and analyse since human subjects can provide feedbacks (like visual and auditory sensations, perception of pain, discomfort, hallucinations and recollections). These are of great importance for both short-term and long-term disease observations. Hence, appropriate actions such as: (1) preprocessing of raw medical data, (2) analysis of sample peculiarity exhibited by the collected clinical data, (3) determination of the prediction characteristics, (4) meticulous selection of mining (e.g. machine learning) techniques, and (5) acquiring supports from healthcare providers, patients and professional organizations, should be considered during development of CDSS systems.

2.4. Introduction to Machine Learning

Machine learning (ML), a term coined by Samuel (Samuel, 1959) in the 1950s, is concerned with the creative design and development of learning procedures capable of empowering computers with the ability to autonomously learn to solve a problem without explicitly being programmed (Min, 2010). Specifically, ML is a process that aims to select, explore and acquire knowledge directly from plethora of data (with minimal human intervention); constructing a concise model capable of describing unknown patterns or relationships, and in turn solves challenging problems. This learning process is usually performed through repeated exposure to the defined (data) problem, allowing the model to self-optimize and continuously improve its ability to solve future related problems. Key differences between statistics and ML techniques include: (1) statistics use a rigorous mathematical approach while ML methods allow partial adoption of heuristics to solve the problem; (2) statistics only allows the manipulation of numerical data while ML methods often allow multiple types of data (e.g. numerical or categorical) to be handled simultaneously; and (3) statistics is of hypothetico-deductive nature (i.e. a hypothesis is postulated and subsequently, data is collected to test the hypothesis) while ML is of inductive nature (i.e. from the data collected, a knowledge or evidence-based hypothesis is deduced) (Yoo et al., 2012).

In the context of clinical classification (e.g. discriminating patients from healthy individuals), supervised learning algorithms are a typical set of ML methods used to perform this predictive modelling. Supervised classification is a ML task that reasons from labelled data instances provided externally (i.e. exemplars that consist of observation/measurement values about an item of interests and the desired output value) to generate a hypothesis model capable of making predictions (i.e. assigning the output value) about future (unseen) instances. A variety of supervised learning algorithms exists and has garnered a significant amount of attention due to their successful application to different types of real-world problems; 2 novel supervised learning algorithms developed would be described in this thesis. Some of the most commonly used algorithms are described below:

1. **Artificial Neural Network (ANN):** ANN is a type of ML model first introduced by McCulloch and Pitts in 1943 (McCulloch & Pitts, 1943). It is inspired by the neurological functions of the brain and is capable of performing many tasks like classification and regression. It consists of interconnected artificial neurons (i.e. computational nodes) that (1) accept input data, and (2) compute an output value based on the given input values (Baxt, 1991). A key advantage of ANN over conventional statistical methods is that ANN is capable of modelling complex non-linear relationships. This provides ANN the competitive advantage when modelling non-trivial tasks; allowing it to achieve good performance when applied on various challenging science and engineering problems. However, ANN has several drawbacks: (1) it is highly sensitive to its parameters' value; (2) the architecture and complexity of the network constructed play a significant role in its performance; (3) it has a high computational training cost; and (4) the resulting induction models may be difficult to interpret by humans (Bellazzi & Zupan, 2008).

2. **Support Vector Machine (SVM)**: SVM, introduced by Vapnik and Cortes in 1995 (Cortes & Vapnik, 1995), is a learning algorithm based on statistical learning theory (Vapnik, 1999). The fundamental strategy of this algorithm is to search for a (linear) hyper-plane that could maximally separate exemplars from different categories. This hyper-plane is then used to (linearly) classify new exemplars by determining which side of the hyper-plane they fall on. For non-linearly separable problems, non-linear kernels can be used to map the original feature space onto a higher dimensional space so that they can be linearly separated. Popular kernel functions include polynomial, sigmoid and radial basis functions. One key advantage

of SVM is its excellent predictive performance while its key disadvantage is the extensive computational time required (Bellazzi & Zupan, 2008).

3. **Decision Trees (DT):** DT is a type of decision tools that uses a directed acyclic graph constructed from training data (through recursive data partitioning) to perform classification. Within the tree structure, each non-leaf node is responsible for testing a feature while each leaf node corresponds to a class label. One of the pioneering (landmark) DTs is ID3 (Iterative Dichotomizer 3) developed by Quinlan in 1986 (Quinlan, 1986). Some of the currently popular DT algorithms include C4.5 (successor of ID3), See5 (successor of C4.5) (Quinlan, 1992) and CART (Classification And Regression Tree) (Breiman et al., 1984). One notable advantage of DT is its low computational complexity while the key disadvantage is that the constructed tree may become very complex when the analysed dataset contains many features (Yoo et al., 2012).

4. **Naïve Bayesian classifier:** Naïve Bayesian classifier is an efficient probabilistic classifier based on Bayesian theorem. It estimates various probabilities from the input data and assumes that the input features are conditionally independent of each other – i.e. the presence (or absence) of one feature is unrelated to the absence (or presence) of another. Despite it being a relatively simple classifier that makes unrealistic independence assumption, it is capable of achieving comparable performance in relation to other more sophisticated algorithms (Bellazzi & Zupan, 2008), and is one of the popular classifiers used for medical diagnosis (Rish, 2001). However, when biomarkers exhibit non-linear relationships, more sophisticated algorithms like ANN and SVM are capable of surpassing the performance of Naïve Bayesian classifier (Bellazzi & Zupan, 2008).

2.4.1. Standards for Model Development

Standards exist for the development of predictive models. An example is the CRISP-DM (CRoss Industry Standard Process for Data Mining) reference model defined in the CRISP-DM project (Wirth & Hipp, 2000). It is a generic process model that proposes the following 6 phases for developing predictive models - namely (1) business understanding, (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation, and (6) deployment. Figure 2.1 provides the canonical flow of the CRISP-DM process model.

In the context of clinical prediction, the business understanding phase involves the definition of clinical objectives and requirements. In the second phase, data understanding, clinical data are amassed (e.g. from paper-based forms or data warehouses) and analysed to identify data quality problems (e.g. missing data, sample size, etc.), learn insights about the data, or propose hypotheses from the detection of interesting clinical data samples. If required, one would cycle between the data understanding and business understanding phases to better design the project plan and collect insightful data.

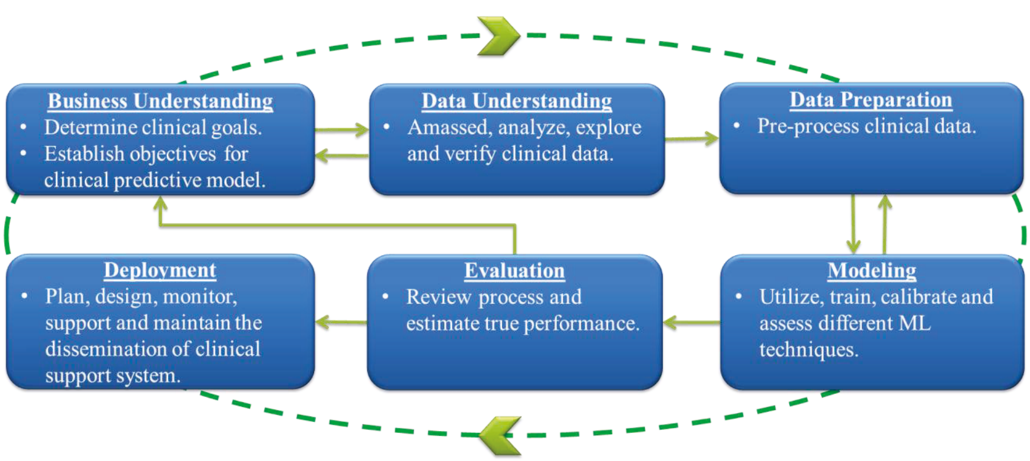


Figure 2.1: Overview of CRISP-DM Process Model
Six phases are proposed for the development, evaluation and deployment of predictive models.

The data preparation phase is responsible for constructing the final dataset that will be deployed for learning and construction of ML models. It consists of data pre-processing steps like feature ranking, feature selection, feature construction, data cleaning, data imputation and data transformation. Additionally, it is important to split the initial data into 2 mutually independent datasets (namely training dataset and validation dataset) during this phase in order to postulate a reliable approach for the estimation of the true performance of the constructed predictive models. The training dataset is used for the construction of the final predictive model while the validation dataset is used to test the constructed model developed using the training dataset (Bellazzi & Zupan, 2008). In the fourth phase, modelling, different ML algorithms are employed, trained and calibrated with the training dataset to construct the predictive model. The performance of these constructed models is compared and the best performing model is selected for evaluation and deployment. If necessary, one would cycle between modelling and data preparation phases to construct the best possible predictive model. Typically during this phase, k-fold stratified cross validation strategy is adopted to develop the predictive model. This approach divides the training dataset into k data subsets of approximately equal size and outcome distribution. Consequently, k-1 data subsets are used to develop the predictive model while the remaining (testing) subset is used to test the constructed model. This process of training and testing is repeated k times, each time using a unique testing subset. A graphical illustration of 10-fold cross validation strategy is provided in Figure 2.2.

The evaluation phase aims to evaluate whether the clinical objectives defined during the business understanding phase were satisfied and to estimate the true performance of the constructed predictive model. If the clinical objectives were not satisfied by the constructed model, one would need to repeat the entire process. To evaluate the constructed model, it is crucial to use

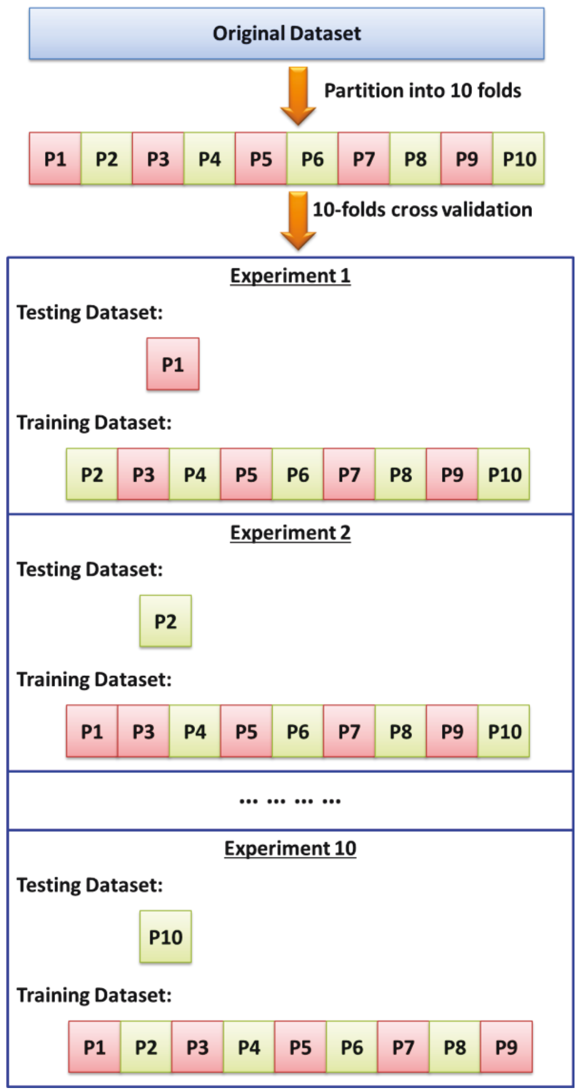


Figure 2.2: An Illustration of 10-Fold Cross Validation
The average performance achieved in the 10 experiments corresponds to the 10-fold cross validation performance.

the validation dataset to perform this task. This is because reporting results based on the training dataset may be overly optimistic and prone to over-fitting.

Finally, the deployment phase deals with the dissemination of the resulting model to the clinical environments. This includes: (1) development of dedicated user interfaces that aims to ease the usage of the predictive model by physicians and healthcare professionals; (2) generation of insightful clinical reports; (3)

provision of a platform for updating the predictive models; (4) support of standards for the exportation, importation and communication of predictive models across different decision support systems; and (5) provision of either web or mobile based decision support shells (Wirth & Hipp, 2000; Bellazzi & Zupan, 2008).

2.5. Machine Learning based Clinical Decision Support System

Medical diagnosis and prognosis conducted by physicians today tend to be highly subjective and vary based on their personal intuition, experiences, judgement, emotions, and knowledge. Exacerbated by the fact that medical history, clinical biomarkers and symptoms seldom follow a linear relationship, and the expected outcome at individual level does not always abide to the rules of epidemiology; it is necessary for the healthcare industry to adopt a more objective approach (Chattopadhyay, 2013). One method that has been postulated is the use of computational machine learning techniques that allow the extraction of interesting, meaningful and predictive information from clinical data. This approach has the potential to: (1) eradicate some degree of physician's subjectivity; (2) allow the epidemiology to work more precisely at the patient level; (3) enable more comprehensive set of data to be analysed simultaneously; and (4) ensure a more objective output to be generated. However, it is noteworthy that the final clinical decision should be made by the physicians as humans are more flexible and capable at identifying outlying details that CDSS is unable to account for (e.g. due to the lack of certain information). Hence, CDSS should serve as guidelines aiming to leverage on the overall standard of healthcare and should not be used as a replacement for physicians. An ideal scenario is to capitalize on the highly accurate prediction that machine learning based CDSS can offer while allowing physicians to have full flexibility and responsibility in making good clinical judgement (Snyderman & Langheier, 2006).

Albeit the challenges for the wide deployment of CDSS in clinical practices, efforts have continually been invested to improve and enhance the capabilities of CDSS. This is, in part, because of the growing body of literature that demonstrates the potential benefits of adopting CDSS as part of the clinical routine (Neill, 2013; Wiens et al., 2012; Levin et al., 2012); a necessity to gain greater appreciation and eventual adoption from the clinicians. Moreover, it has been realized that CDSS do offer significant advantages (e.g. improved patient safety, quality of care, and efficiency in healthcare delivery) when deployed appropriately (Coiera et al., 2003) . Hence, a vital task is to accurately identify those aspects of clinical practice that are best suited for their introduction. These promises have anticipated the current confluence of interests on the employment of artificial intelligence (AI) and statistical modelling as computational reasoning tools to support clinical decision. These approaches have the distinct advantages of performing non-linear inference, exploratory data analysis, tolerating noise, circumventing the difficulties of acquiring expert knowledge and the ability to accommodate and model new manifestations of disease (Lisboa, 2002).

In view of these promising benefits, a plethora of CDSS have been developed in recent years using ML methods. This approach empowers users to automatically discover the underlying medical knowledge (from large medical databases that could be stored in different sources) through the process of learning from experiences. This process of learning allows the performance of certain tasks to improve over time with experience; here, experience refers to the data that is used for training the ML inference model. In other words, the algorithm will search through the possible hypotheses (within the boundaries of the selected mathematical or computational model) to identify the one that best suit the observed data and any prior knowledge possessed by the learning algorithm. The nature of the data, in this case, can be described by nominal or numerical information called attributes (e.g. gender, age, family history, etc.) and/or time-series information (e.g. electrocardiogram, blood pressure, etc).

Clearly, if an algorithm is allowed to learn from more data, it will gain more experience. Similarly, if high quality data is presented to a classification algorithm, good experience will be gained. This would result in good discriminative ability reaped by the algorithm.

Table 2.2: List of Algorithms Used for the Development of Clinical Decision Support Technique

Algorithm	Area of Concern	Reference
Artificial immune recognition system	Atherosclerosis	(Latifoğlu et al., 2008)
Artificial immune recognition system	Heart disease	(Polat et al., 2006)
Artificial immune recognition system	Thyroid disease	(Polat et al., 2007)
Artificial neural network	Nosocomial infection	(Suka et al., 2008)
Artificial neural network	Cardiovascular disease	(Ohlsson, 2004)
Artificial neural network	Diabetic retinopathy	(Schaefer & Leung, 2007)
Self-organizing maps Probabilistic neural network Multi-layer perceptron neural network	Kidney dysfunction	(AlTimemy & Naima, 2010)
Artificial neural network Support vector machine K-nearest neighbour	Soft tissue tumor	(García-Gómez et al., 2004)
Artificial neural network Decision tree	Coronary artery disease	(Kurt et al., 2008)
Bayesian classifier	Pyloric stenosis	(Alvarez et al., 2006)
Bayesian learning	Pulmonary gas exchange	(Murley et al., 2005)
Model-averaged Naïve Bayes Naïve Bayes Naïve Bayes with feature selection	Alzheimer's disease	(Wei et al., 2011)
Naïve Bayes Decision tree	Diabetes	(Huang et al., 2007)
Decision tree	Dengue	(Tanner et al., 2008)
Decision tree	Pancreatic cancer	(Yu et al., 2005)
Decision tree	Acute myocardial infarction	(Mair et al., 1995)
Random forest	Coronary Artery Disease	(Kelm et al., 2011)
Random forest	Cardiac arrhythmia	(Özçift, 2011)
Support vector machine	Tracheal intubation	(Yan et al., 2009)
Support vector machine	Clostridium difficile	(Wiens et al., 2012)
Support vector machine	Breast cancer	(Daemen et al., 2007)
Support vector machine	Lung cancer	(Nguyen et al., 2007)

Common ML techniques that have been employed to build CDSS include Artificial Neural Network (ANN), Decision Tree (DT), Random Forest (RF), Bayesian Network, Naïve Bayes, Support Vector Machine (SVM) (Geert et al., 2009), and Artificial Immune Recognition System (AIRS). The use of these algorithms for extracting insights from large medical databases is invaluable as medicine is a domain that is complex and difficult to model by humans. These techniques are capable of handling large amount of data from different sources, incorporate expert knowledge into the analysis, offer data-driven predictions that can assist clinicians in making their decision. A list of examples that employs popular machine learning techniques as an approach to enhance clinical decision making is shown in Table 2.2. Selected examples are succinctly described below:

1. Suka et al. (Suka et al., 2008) proposed a multiple ANNs approach to estimate the probability of nosocomial infection. Multiple ANNs was constructed by connecting individual ANNs (that predicts the probability of nosocomial infection at different time period) sequentially, where the output of an ANN is connected to the input of the next ANN. Experimental results show that with multiple ANNs, it outperforms multivariate regression models in predicting the risk of nosocomial infection.

2. Ohlsson (Ohlsson, 2004) presented a technique using ANN to automate the interpretation of heart images. ANN was compared with logistic discrimination and K-nearest neighbour (KNN). Results indicate an advantage of using ANN over the other 2 methods evaluated.

3. Schaefer et al. (Schaefer & Leung, 2007) proposed a neural network-based approach for automatic detection of exudates in retina images – an early indicator for diabetic retinopathy (a common eye disease that is directly associated with diabetes, which can eventually result in

blindness). Experimental results demonstrated high sensitivity and specificity of 94.78% and 94.29% respectively.

4. AlTimemy et al. (AlTimemy & Naima, 2010) compared the accuracy of using self-organizing maps (SOM), probabilistic neural network (PNN) and multi-layer perceptron neural network (MLPNN) for the prediction of kidney dysfunction. Over 600 analytical laboratory tests have been collected and evaluated with the 3 aforementioned types of neural networks. Their results indicate that PNN offers faster and more accurate prediction for kidney dysfunction.

5. Yan et al. (Yan et al., 2009) employed support vector machine (SVM) with polynomial kernel for the prediction of whether tracheal intubation would be easy or difficult before anesthesia is carried out. A total of 264 medical cases and 13 physical features were analysed in this study. The use of 13 basic and anthropometrical features has a significant advantage over the approach taken by some anaesthetists where a single feature is examined ahead of anaesthesia. This is because most specialists agree that full consideration of multiple features would improve the prediction accuracy of airway physical examination. Based on 4-fold cross-validation, an average classification accuracy of 90.53% was achieved in the study.
6. Wiens et al. (Wiens et al., 2012) compared the use of SVM and HMM for predicting patient risk of clostridium difficile. The problem is formulated as a time-series problem, which is of particular importance as the nature and timing of diagnostic and therapeutic activities, and the overall evolution of the patient's pathophysiology over time have a significant impact towards the patient's risk for adverse events. A total of 8166 unique patients were analysed in this study. It was found that classifiers that consider the temporal aspect of patient health outperform classifiers that only consider a patient's current state ($p<0.05$).

7. Murley et al. (Murley et al., 2005) used Bayesian learning to determine the desirable physiological model parameters for pulmonary gas exchange. It aims to support and improve the selection of inspired oxygen fraction. The model was tested with 16 post-operative cardiac patients. Results demonstrate that it is both accurate and safe to use the prediction model to support clinicians.

8. Wei et al. (Wei et al., 2011) applied model-averaged naïve Bayes (MANB) method to predict late onset of Alzheimer's disease. A total of 1,411 individuals who each had 312,318 SNP measurements available were analysed. MANB performance was compared with Naïve Bayes (NB) and Naïve Bayes with feature selection (FSNB). The area under the receiver operating characteristic curve (AUC) achieved for MANB, NB and FSNB were 0.72, 0.59 and 0.71 respectively. Although the performance of MANB and FSNB was statistically not significant, the training time required by MANB is significantly faster than FSNB (~104-fold faster).

9. Latifoglu et al. (Latifoğlu et al., 2008) performed diagnosis of atherosclerosis from Carotid Artery Doppler Signals using Artificial Immune Recognition System (AIRS) as the classification algorithm. Prior to classification, features were first extracted using Fast Fourier Transformation (FFT) modelling and calculation of maximum frequency envelope of sonograms. Subsequently, Principal Component Analysis (PCA) was used to reduce the number of features which are then weighted using K nearest neighbour (KNN). A total of 60 cases and 54 controls were studied. Based on the method proposed, a classification accuracy of 100% was obtainable using 10-fold cross-validation.

10. Polat et al. (Polat et al., 2007) employed AIRS and fuzzy theory to perform thyroid disease diagnosis. The thyroid dataset, available at the

UCI machine learning repository, consists of 215 instances and 3 classes. The 10-fold cross-validation classification accuracy achieved was 85%. This performance, when compared to previous work, is the highest.

Although the current role of ML based CDSS revolves around patient diagnosis, prognosis and image analysis, it is postulated that it has great potential to improve copious aspects of clinical healthcare in the future. Examples include (1) personalization of therapeutic strategies that maximizes efficacy and safety, (2) recommendation of the most appropriate and cost-efficient diagnostic process, (3) real-time and transparent monitoring of patients' health, and (4) discovery of new medical knowledge that has a direct and profound impact to the quality of patients' health and care (Neill, 2013).

Chapter 3

A Biological Continuum based Approach for Efficient Clinical Classification[1]

Clinical diagnosis is a significant task of pragmatic value. The conventional approach to this task is based on expert knowledge and judgement (i.e. analysis of patient's clinical data by a physician and based on his knowledge and experience, determine the health status and treatment for the patient). However, with the inundation of clinical data/features in current healthcare industries, this approach is becoming increasingly challenging. Therefore, computer-aided techniques, like data mining, have been proposed to alleviate this challenge.

In this chapter, we introduce a novel clinical feature selection methodology for efficient development of clinical classification model. It is an approach that adopts the conceptual framework of biological continuum (BC) (Kitney & Poh, 2006; Poh et al., 2007), the optimization capability of genetic algorithm (GA) (Holland, 1992) and the classification ability of support vector machine (SVM) (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1999). Together, a network of associated clinical risk factors (from different biological levels) was constructed. We call this network the Biological Continuum based Etiological Network (BCEN). Evaluation of our proposed methodology was carried out using the CHS dataset (Fried et al., 1991). Results demonstrate that our methodology, when compared with the conventional approach, is capable of achieving a significant speedup of 4.73-fold without compromising classification accuracy. The key advantage of our approach is the provision of a

[1] The work presented in this chapter has been published in the Journal of Biomedical Informatics and reprinted from "Journal of Biomedical Informatics, Tay, Poh, Goh & Kitney, "A biological continuum based approach for efficient clinical classification", 2014 (doi:10.1016/j.jbi.2013.09.002)", with permission from Elsevier. This paper can be found in Appendix C.

reusable (feature subset) paradigm for efficient development of up-to-date and efficacious clinical classification models.

3.1. Introduction

The efficient development of accurate clinical classification models has been a challenge for many reasons. One problem that is commonly encountered is the 'curse of dimensionality' (Bellman, 1961), where the linear growth of clinical features (i.e. predicators) results in an exponential growth in the search space. This inevitably hinders the development of classification models as it becomes computationally expensive to investigate a plethora of clinical features simultaneously using search heuristics that analyse features in combinations (particularly, when performing multivariate analysis based on wrapper approach). This situation is exacerbated by the fact that up-to-date and sophisticated clinical classification models need to be constantly developed in order to continually improve the quality of clinical diagnosis. Specifically, the clinical classification models need to be rebuilt whenever new clinical risk factors that could potentially ameliorate the performance of the classification model are introduced. An example of such clinical effort is the perpetual studies of different types of clinical risk factors and approaches that could improve the ability to identify events of myocardial infarction (MI) (Baxt & Skora, 1996; Menown et al., 2000). This is of paramount importance as MI is a leading cause of morbidity and mortality in many developed countries, such as the United States (U.S.) and the United Kingdom (U.K.) (Go et al., 2013; Wilson et al., 1998; British Heart Foundation Statistics Database, 2010). Despite considerable advances in medicine, MI approximately occurs every 34 seconds in the U.S. and about 15% who experience MI will die from it (Go et al., 2013). Moreover, MI is difficult to ascertain in patients presenting to the emergency department with anterior chest pain (Baxt & Skora, 1996). This advocates for the need of an efficient approach to develop up-to-date MI classification models for performing accurate diagnosis.

Furthermore, investigation of the association between a range of clinical observations (e.g. medical history, chemotherapy, stage of disease, gene, etc.) and the disease at the human population level is important as it has demonstrated promising potential for improving disease classification performance (Hsia et al., 2003; Pittman et al., 2004). However, when such an investigation is carried out on a larger scale, this would involve a large amount of clinical features, making analysis challenging and even computationally infeasible. Additionally, it also hinders the ability for any machine learning method to perform accurate disease classification. One approach to mitigate the aforementioned problems is through dimensionality reduction - where significant clinical risk factors are identified, reducing the total number of predicators that need to be analysed.

In this chapter, we introduce a novel clinical feature selection methodology for the development of MI classification model. This approach utilizes on the conceptual framework of biological continuum (BC) (Kitney & Poh, 2006; Poh et al., 2007), the optimization capability of genetic algorithm[2] (GA) (Holland, 1992) for performing feature selection and the classification ability of support vector machine[3] (SVM) (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1999) for dichotomizing patients experiencing a phenotypic manifestation from healthy individuals. The BC is the hierarchy of the human organism comprising body, systems, viscera, tissue, cells, proteins and genes. Detailed analysis of the biology of a disease at different levels along the BC is very important. As an example, one significant advantage postulated by molecular medicine is the ability to prevent a disease at the molecular or cellular level. This is highly attractive as less damage would have occurred and the likelihood of full recovery is much higher.

[2] A brief description of how GA works can be found in Appendix A.
[3] A brief description of how SVM works can be found in Appendix B.

In this study, the BC provides the biological paradigm necessary for segregating a range of available clinical features; offering the advantage of reducing the number of clinical features that needs to be analysed concurrently. A GA based wrapper approach using SVM, which selects significant clinical features capable of dichotomizing patients experiencing a phenotypic manifestation from healthy individuals, was implemented. This hybrid algorithm (called GA-SVM) was used to identify important clinical features at each level of the BC and incrementally built a network of clinical risk factors, called the biological continuum based etiological network (BCEN). The primary advantage of BCEN used for the construction of up-to-date clinical classification model is that it allows new clinical features to be considered for incorporation into the classification model without the need for a total reanalysis from scratch.

The reliability of the constructed BCEN was assessed by comparing the set of identified risk factors found in the (obesity-system) sub-network, with the risk factors found in previous clinical studies. Promising results were obtained from this analysis. An MI classification model was subsequently developed based on the clinical features identified and present in the BCEN. Significant reduction in the computational time required to develop the classification model was achieved. It is noteworthy that comparable classification accuracy was obtained between the proposed method (i.e. pre-selection of clinical features using BCEN) and the baseline approach (i.e. no pre-selection was performed). The Cardiovascular Health Study (CHS) (Fried et al., 1991) dataset was analysed in this study.

The rest of the chapter is organized as follows. Section 3.2 provides the background information on feature selection. In Section 3.3, the experimental methodology involved in the development of the clinical feature selection technique and the clinical classification model is presented. The experimental results are presented in Section 3.4 and discussed in Section 3.5. Finally, conclusions are drawn in Section 3.6.

3.2. Background

Conventionally, clinical predictions which provide the disease diagnosis for an individual are based on expert knowledge. However, with the exponential growth of clinical data generated in healthcare industries, this approach has become more and more difficult and costly. An approach to mitigate this challenge is to process and analyse the large amount of clinical data, extracting knowledge that enables support for cost-containment and decision making (Bhatla & Jyoti, 2012). Machine learning is one method that has been proposed to address this issue. It provides the techniques necessary for the analysis of the data, discovery of hidden patterns and provides healthcare professionals with an additional source of knowledge for decision making. In the parlance of literature, machine learning is defined as a branch of artificial intelligence that postulates a set of computer-based methods for automatic analysis of information and recognition of patterns through repeated learning from the training data (Roganb et al., 2008), and is a more powerful and sophisticated descendant of traditional statistical models. It is generally model-free and is capable of efficiently detecting and modelling the non-linear interactions in high dimensional datasets. Additionally, the associations or patterns detected by machine learning methods tend to be logical and can be identified by human experts if they analyse the problem carefully enough (Baxt & Skora, 1996). Clearly, this entails that machine learning is capable of saving both the time and effort necessary for the discovery of underlying patterns.

Clinical prediction (e.g. diagnosis of cardiovascular disease) based on machine learning approaches has gained popularity over the years (Baxt & Skora, 1996; Eggers et al., 2007; Palaniappan & Awang, 2008; Bhatla & Jyoti, 2012; Hossain et al., 2013; Latifoğlu et al., 2008; Ohlsson, 2004; Nilsson et al., 2006) and shown to be an extremely useful tool in medical innovation (Hossain et al., 2013). It is often based on the patient's unique clinical, genetic and environmental characteristics and plays a significant role in healthcare decision making and planning. Since each clinical feature collected is associated with a different financial cost, diagnostic value and risk (Yang & Honavar, 1998), it is highly desirable to reduce the number of clinical tests that need to be taken by a patient. This would inevitably reduce the financial cost, and the time incurred on both the analysts and patients. One approach commonly adopted by machine learning techniques to reduce the number of clinical features while improving the diagnostic/classification accuracy is feature selection.

Feature selection is the process of selecting a subset of relevant features for model construction and provides better insights into the target concept of a real-world problem (Kohavi & Sommerfield, 1995). It differs from other dimensionality reduction techniques like project and compression where their original representation of the variables is modified. Therefore, feature selection has the advantage of preserving the original semantics of the features which enables domain experts to interpret the selected features. Furthermore, it has shifted from being an illustrative example to one of real prerequisite for developing classification models (Saeys et al., 2007). This is, in part, because of the exponential increase in the dimensionality of the data (e.g. in clinical and bioinformatics domains), the fact that most classifiers were originally not designed to handle plethora of irrelevant features, and the need to generate more accurate classifiers efficiently. In general, feature selection aims to identify a parsimonious subset of useful features (from a large set of features) that (1) does not decrease the classification accuracy, (2) reduces the computational time needed to learn a sufficiently accurate classification model, (3) does not acutely changes the class distribution while adequately representative for descripting the target concept, and (4) reduces the amount of examples that need to be collected in order to develop a classification model with the desired accuracy (Dash & Liu, 1997).

Feature selection algorithms typically fall under 4 categories depending on how it is performed in relation to the classification algorithm. They include (1) selection based on expert knowledge, (2) filter approach, (3) wrapper approach, and (4) embedded approach. Each has its own competitive advantages and drawbacks. Selection based on expert knowledge (e.g. human domain expert or referencing the scientific literature) offers a set of features with high interpretability in relation to the target concept. However, its major drawbacks are that it can be time consuming and human expert is required to perform the task. An illustration of this approach is demonstrated in (Emily et al., 2009), where the number of interaction tests that need to be performed can be limited with the use of experimental knowledge of the biological network. More specifically, knowledge extracted from protein interaction databases reduces the number of interaction tests from $1.25x10^{11}$ to $7.1x10^{4}$, allowing more efficient analysis of genome-wide studies to be carried out.

Filter methods, on the other hand, evaluate the relevance of each feature by assessing only the intrinsic characteristics of the data. Although this approach does not need a domain expert to intervene, is simple, efficient and can easily scale to very high-dimensional datasets, it does not always guarantee improved performance (Chu et al., 2012) as it ignores the inductive bias associated with the classifier (Yang & Honavar, 1998). Examples of filter techniques include chi-square test, t-test, information gain, correlation-based feature selection and Markov blanket filter.

Wrapper methods embed the inductive bias associated with the classifier within the feature selection process. In this case, subsets of features are generated and their performance is assessed by training and testing them on a specific classification algorithm. The advantages of this approach are: (1) the freedom to choose the desired classification algorithm, (2) allowing interaction between feature selection and model selection, and (3) ensuring that feature dependencies are taken into consideration (i.e. the need to add or remove more than 1 feature at the same time in order to improve the performance (Guyon & Elisseeff, 2003; Yang & Honavar, 1998)). Consideration of feature dependencies is important, especially in the medical field, as it has become evident that multiple genes collectively contribute to the etiology and clinical manifestation of human diseases (Li & Agarwal, 2009). Hence, important genotypic factors might be missed if they have been examined in isolation or in a linear fashion - without allowing for potential interactions. This situation would be exacerbated when performing genome-wide association studies where hundreds of thousands of single nucleotide polymorphisms (SNPs) need to be analysed. Wrapper approach, on the downside, becomes computationally intensive when the number of features grows exponentially. This is because every feature subsets generated need to be executed on the selected learning algorithm. Moreover, it has a higher risk of over-fitting the classifier than filter approach. Examples of this technique include sequential forward selection, sequential backward selection, simulated annealing, genetic algorithm and estimation of distribution algorithm.

Finally, embedded approach integrates the process of identifying the optimal subset of features within the learning algorithm. Based on this mechanism, it has the advantage of being more computationally efficient (compared to wrapper approach) while maintaining interaction with the classifier. Examples include decision trees and weighted naïve Bayes.

3.3. Material and Methods

In Section 3.3.1, a description of the CHS dataset used is provided. Section 3.3.2 lists the steps taken in constructing BCEN – the proposed framework for efficient and repetitive development of up-to-date clinical classification models. Specifically, data imputation (Section 3.3.2.1) was first conducted on the CHS dataset as it contains a significant amount of missing data. Subsequently, data class balancing (i.e. selecting a similar number of cases and controls) is performed on the imputed dataset (see Section 3.3.2.2). Through these 2 steps, we aim to improve the quality of the data to be analysed. In Section 3.3.2.3, how clinical features amassed from the CHS observational study are segregated

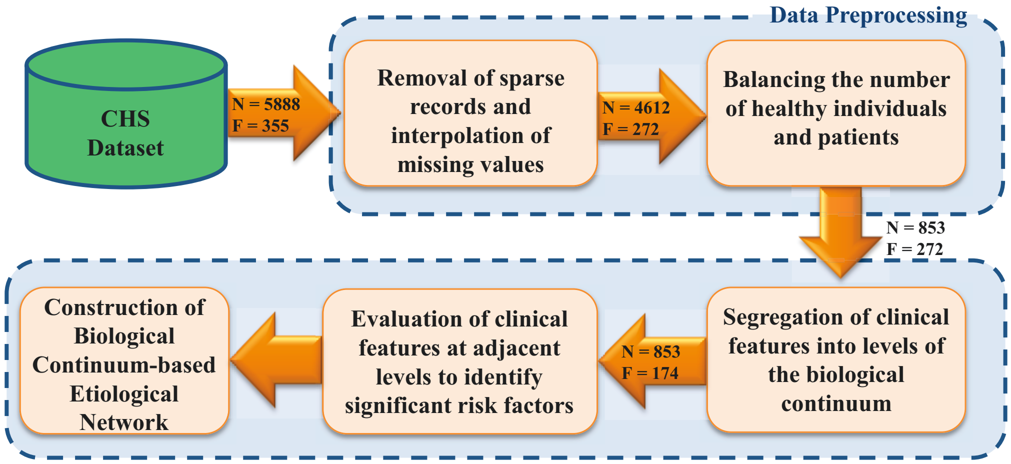


Figure 3.1: Canonical Flow of the Methods Adopted to Construct BCEN
'N' denotes the number of instances and 'F' represents the number of features present in the dataset at different stages.

along the BC is explained. The algorithm (i.e. GA-SVM) used to perform the classification task is detailed in Section 3.3.2.4. Through the use of GA-SVM algorithm and the clinical features segregated at each level along the BC, the BCEN paradigm is constructed (see Section 3.3.2.5). Finally, in Section 3.3.3, the steps taken to evaluate the classification performance of the developed BCEN is described.

3.3.1. Dataset

The CHS dataset, as described in (Fried et al., 1991), is an epidemiology study of the elderly (defined as adults aged 65 and older). It comprises of elderly subjects from four U.S. communities, namely Forsyth County, North Carolina; Sacramento County, California; Washington County, Maryland; and Pittsburgh, Pennsylvania. A total of 5888 individuals from urban and rural areas form the baseline cohort of CHS. Eligible individuals were sampled from Medicare eligibility lists in each area. Eligible participants included all individuals sampled from the Health Care Financing Administration (HCFA) sampling frame - they were 65 years or older at the time of examination, non-institutionalized, expected to remain in the area for the next 3 years, able to give informed consent and do not require a proxy respondent at baseline. Individuals who were wheelchair-bound at home at baseline, receiving hospice treatment, radiation therapy or chemotherapy for cancer were excluded. The eligible individuals were examined yearly from 1989 to 1999. Extensive physical and laboratory evaluations were carried out to identify the presence and severity of cardiovascular disease (CVD) risk factors - such as hypertension; hypercholesterolemia and glucose intolerance; subclinical disease, such as carotid artery atherosclerosis; left ventricular enlargement; and transient ischemia. Criteria for identification of MI events include: observation of evolving Q-wave, cardiac pain and abnormal enzymes together with an evolving ST-T pattern or new left bundle branch block. A total of 355 clinical features related to the individual's health status were selected from the CHS dataset for this study.

The dataset was chosen because of (1) the relatively high prevalence of coronary heart disease (CHD) among the elderly, (2) worldwide demographic aging, (3) paucity of information regarding risk factors for CHD among elderly, and (4) the changing clinical characteristics of CHD with advancing age (Fried et al., 1991; Wiener & Tilly, 2002; Go et al., 2013; Abbott et al., 2002).

3.3.2. Biological Continuum based Etiological Network (BCEN)

Several steps were taken to construct the BCEN for MI with the canonical flow illustrated in Figure 3.1. A succinct description of the key steps taken is given below while we dedicate separate sections for the discussion of the details:

1. Sparse records were removed and missing entries in the dataset were imputed to ensure good quality data is used to model the risk factors associated with MI. This was performed with the K-Nearest Neighbour (KNN) algorithm (Cover & Hart, 1967) - it calculates the missing value

by taking the K nearest training set vectors (based on Euclidean distance) into consideration.

2. Healthy individuals, forming a large proportion of the dataset in relation to the number of patient records, were sampled to avoid jeopardizing the ability of SVM to learn and generalize. This is carried out with Kohonen Self-Organizing Map (SOM) (Kohonen, 1990), where a representative subset of the majority class (i.e. healthy individuals) present in the CHS dataset was selected, a process known as under-sampling.

3. Clinical features, such as blood pressure, electrocardiography (EKG) readings, ultrasound data, hematology data, etc, were segregated along the BC - the hierarchy of the human organism. It comprises 7 levels, namely the body, system, viscera, tissue, cell, protein and gene.

4. GA-SVM, a hybrid algorithm used to identify significant clinical features, was implemented. It is used repeatedly at each level of the BC to identify significant risk factors that are related to the different phenotypic manifestations, and ultimately MI.

5. With the significant risk factors identified at the different levels of the BC, they were consolidated to construct a consensus network, known as the BCEN in this work. These risk factors, in turn, were used to perform MI classification using the GA-SVM algorithm.

3.3.2.1. Data Imputation

As with many datasets collected from real subjects and patients, missing data is unavoidable. This may be due to various factors, e.g. the refusal of respondents, malfunction of equipment, data not entered correctly and the death of patients (Batista & Monard, 2003). Moreover, since the quality of the results is largely determined by the quality of the data used in the analysis, detailed consideration was given before using the CHS dataset. It was found that the CHS dataset contains a significant percentage of missing information. Hence, data imputation was first conducted.

Data imputation, the process of substituting missing values in a dataset with plausible values, was performed using KNN. KNN imputation was used because of its excellent performance in estimating missing values (Troyanskaya et al., 2001; Acuña & Rodriguez, 2004; Batista & Monard, 2002; Jerez et al., 2010) and its ability to estimate both qualitative and quantitative attributes. This makes it highly suitable for extrapolating the missing entries in the CHS dataset.

Firstly, individuals with unknown MI status were removed from the analysis. Next, to foster more accurate data imputation, individuals and clinical features with high percentage of missing entries were removed. It is important to have low percentage of missing values because the accuracy of the imputed result would suffer if too little complete entries were available for KNN to reference when estimating the missing values (Garcia-Laencina et al., 2008; Troyanskaya et al., 2001; Jerez et al., 2010). Hence, individuals and clinical features with more than 20% and 4.5% missing entries, respectively, were removed. Consequently, the resultant dataset was normalized to unit variance before data imputation was performed using KNN. This is important as it ensures that variables with large scale do not dominate the (Euclidean) distance measure (Minaei-Bidgoli et al., 2003).

The optimal value of K for each clinical feature was determined by 10-fold cross-validation. After the value of K for each clinical feature had been determined, data imputation for each missing attribute was performed. The type of replacement method used depends on the type of data present in each clinical feature. For instance, if the data is categorical, a reliable choice is to use the mode of the K nearest neighbours to assign the value for the missing entries (Acuña & Rodriguez, 2004; Cover & Hart, 1967). On the other hand, if the data is continuous, the weighted-mean of the K nearest neighbour is used instead to calculate the missing value. Weighted-mean estimation has been demonstrated in (Dudani, 1976; Troyanskaya et al., 2001) to be robust and accurate.

3.3.2.2. Class Imbalance Data Problem

The class imbalance data problem is not uncommon in medical datasets where the data is predominated by the healthy subjects (i.e. controls), with only a small number of disease-affected subjects (i.e. cases). Consequently, this limited the effectiveness ability of standard machine learning algorithms - where the algorithms tend to be overwhelmed by the major class and ignore the minor one. This, in turn, hinders performance (Japkowicz, 2000; Li et al., 2010). This class imbalance data problem prevails in the CHS dataset as well. Therefore, data balancing was performed before deploying the data to GA-SVM.

SOM, an unsupervised (neural network) learning algorithm, was employed to under-sample the major class. This algorithm was chosen because it is capable of generating high quality samples that are representative of the original dataset (Kohonen, 1990) and it has been shown in (Wu et al., 1996) that SOM outperforms random selection. Once the imputed dataset was obtained, the SOM was trained in two phases; namely, the ordering phase and the tuning phase. Two key adaptive parameters, neighbourhood size and learning rate, were used when training the SOM. Neighbourhood size defines the number of neurons that surround the winning neuron (i.e. most stimulated neuron) at each epoch, while the learning rate controls the degree of change for the adapting neurons.

During the ordering phase, large initial neighbourhood size (i.e. 10) and learning rates (i.e. 0.9) were used. Conversely, small neighbourhood size (i.e. 1) and learning rates (i.e. 0.02) were used during the tuning phase - where the neighbourhood size will shrink progressively to 1. This is to allow the SOM to adjust quickly to the input pattern during the ordering phase and to stabilize the feature map during the tuning phase (Kohonen, 1990). The following value for the SOM parameters was determined experimentally and used in this study: number of neurons: 21 by 21; topology function: hexagon; distance function: Euclidean; epoch: 1000; ordering phase learning rate: 0.9; tuning phase learning rate: 0.02; initial neighbourhood size: 10; final neighbourhood size: 1. The reason for using these values is because they have shown to provide reasonable performance.

3.3.2.3. Segregation of Clinical Features

The Biological Continuum was central to the development of the BCEN. It was utilized in this case to provide the necessary biological paradigm to relate the disease mechanisms to the clinical manifestations at various levels of the biological continuum. Upon analysing the clinical features, it was found that these features fall under 4 key levels along the BC, namely: body, system, viscera and protein level. Clinical features related to medication were removed from the study as it was difficult to adjudicate to which level of the BC they belong. Categorization of the rest of the clinical features, in relation to the levels of the BC, was undertaken using the following guidelines:

- Body level – Contains clinical features related to individuals' personal statistics (e.g. age, weight), lifestyle (e.g. smoking status, exercise intensity) and cardiovascular events which that individual is experiencing.

- System level – Consists of clinical features related to individuals' medical history (e.g. arthritis, diabetes), symptoms (e.g. hearing/vision problems) that the individual is experiencing and blood pressure measurements.

- Visceral level - Clinical measurements, e.g. EKG, ultrasound data and treatment specific to an organ were classified under this level.

- Protein level – Clinical features related to hematology were grouped under this level.

3.3.2.4. GA-SVM

GA-SVM, a hybrid algorithm that comprises of (1) SVM that models the statistical properties necessary to distinguish healthy individuals from patients experiencing a clinical phenotype, and (2) GA that selects the significant features that contribute to the construction of an accurate SVM model, was implemented. The reason for combining GA and SVM to identify significant clinical features and to perform clinical classification is because (1) GA - given a reasonable time to perform computation - is capable of providing a good approximate solution to problems that cannot be easily solved if other (conventional) techniques were to be used, and (2)

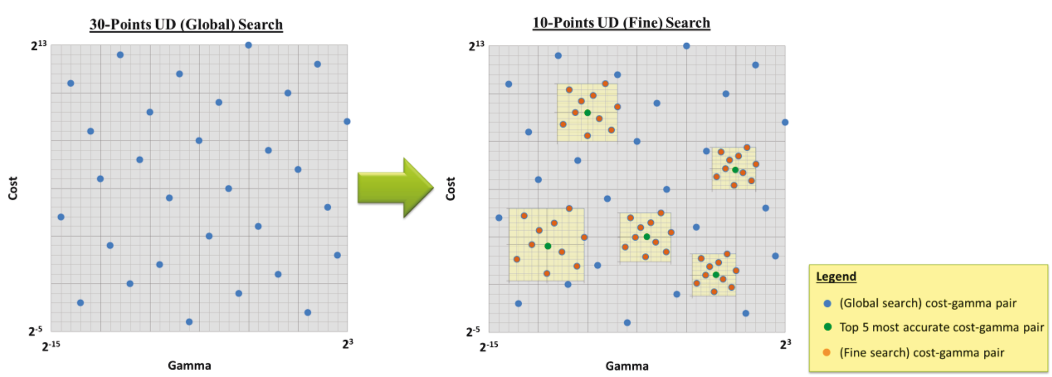


Figure 3.2: Graphical Illustration of SVM Parameter Optimization Using UD Technique
30-point UD (Global) search is first performed to determine regions with cost-gamma combinations that would produce the optimal SVM model. Subsequently, 10-point UD (fine) search is carried out to determine the optimal parameter set.

SVM has demonstrated excellent classification performance on a plethora of diverse problems. Therefore, we believe that this combination would be a very good method for our investigation.

In this work, SVM uses radial basis function (RBF) as its kernel function and is defined as:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \quad (1)$$

where γ is a variable used to adjust the width of the Gaussian functions of the kernel. RBF is used due to its ability to solve non-linearly separable problems, low complexity involved during model selection and excellent performance. Two parameters, namely the regularization cost and gamma (used in RBF) parameters, were tuned over the recommended range [2^{-5}, 2^{13}] and [2^{-15}, 2^{3}] respectively (Chang & Lin, 2001). Optimization of SVM parameters were performed by evaluating a set of cost-gamma combinations defined using uniform design (UD) method (Fang et al., 2000). UD is a technique that scatters

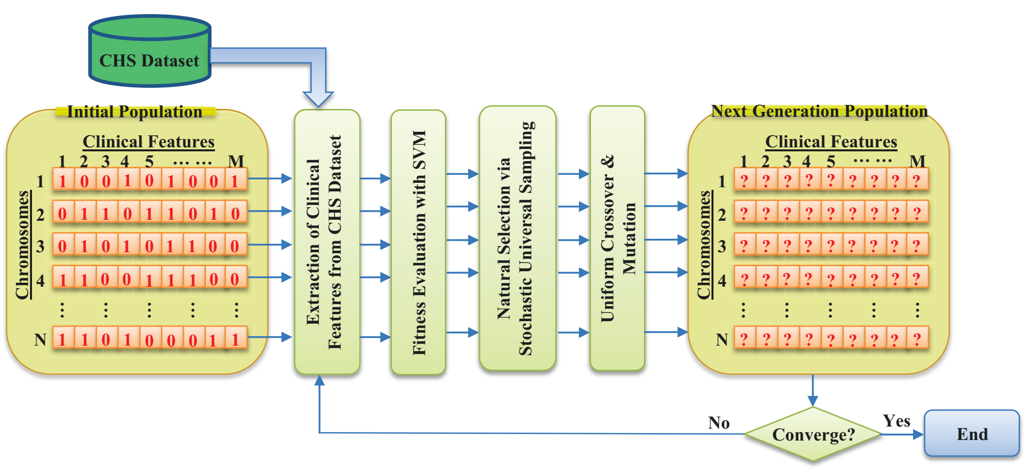


Figure 3.3: A Schematic Illustration of Clinical Feature Selection based on GA-SVM
A string of binary value in each chromosome (of size M) represent the present (value of 1) or absent (value of 0) of a risk factor during evaluation. This set of chromosomes (of size N) is randomly generated at onset and undergo the selection process postulated by GA to find the set of risk factors that produce the highest fitness value. Fitness evaluation is performed with SVM.

a set of points uniformly across the cost-gamma landscape, proposed to alleviate the computational loads associated with the search for the optimal cost-gamma pair (Chow et al., 2008). This search process begins by initializing a 30-points UD (global) search across the defined cost-gamma landscape. Next, it identifies the top 5 most accurate (global) cost-gamma pairs, where they form the centroid for 10-points UD (fine) search. If improved accuracy was achieved, the points will form the centroid for another 10-points UD search. This process repeats until no further improvement is achieved. Figure 3.2 provides an illustration of this method.

Figure 3.3 provides the schematic illustration of GA-SVM algorithm. The flow of the algorithm is as follow: GA first (randomly) initializes a pool of clinical feature subsets (Figure 3.3 - chromosome 1 to N) from the CHS dataset (consisting of M clinical features). Each bit in the chromosome is assigned with a value of either ‘1’ or ‘0’, indicating whether that feature is selected or eliminated from consideration by the classifier, respectively. This produces a pool of chromosomes representing different input features. Consequently, each chromosome was evaluated by SVM (where optimization of SVM parameters was performed independently for each chromosome) in an attempt to determine how informative and discriminative the clinical features are in relation to the associated clinical or subclinical manifestation. This evaluation is conducted by performing a 10-fold stratified cross-validation. Subsequently, these subsets of clinical features undergo natural selection, crossover and mutation phases postulated by GA. The process repeats until GA converges or the maximum number of generations has been reached. GA is considered to have converged if the maximum fitness value (i.e. balanced accuracy – the average of sensitivity and specificity) does not improve after 20 consecutive generations. Upon termination, the subset of clinical features that yielded the highest balanced accuracy will be selected and considered as significant risk factors. A consensus network was constructed if several combinations of clinical feature subset yielded the same fitness performance. The reason for doing this is to build a parsimonious model that maximizes the likelihood of the clinical features that are most influential to the development of the phenotypic manifestation. It was derived by identifying clinical features that existed in more than 75% of the highest-performing clinical feature combinations. The parameters value used by GA are as follow: population size: 250; maximum generation: 300; natural selection: stochastic universal sampling; crossover type: uniform crossover; crossover probability: 0.8; mutation probability: 0.01. These values were used as they perform reasonably well over a range of values when evaluated experimentally. The algorithm was written in Matlab (MathWorks Inc., Natick, MA) and executed in parallel using a high performance computer (HPC) cluster.

3.3.2.5. Construction of BCEN

The underlying cause of MI is multifactorial and subtle, with nonlinear causal dynamics. Moreover, with the plethora of clinical predicators available, analysis of all of them becomes computationally impractical. In view of such challenges, GA-SVM, together with the conceptual framework of the BC, were used to construct the BCEN for MI.

Firstly, by segregating the clinical features into various levels along the BC, the number of clinical features to be analysed is effectively reduced to the number of clinical features present at each level (i.e. dimensionality reduction). Secondly, with the employment of GA, which is capable of performing global heuristic searches both effectively and efficiently, the computational burden of discovering significant risk factors is alleviated. Finally, facilitated by SVM, which outperforms popular technique like multifactor dimensionality reduction (MDR) (Chen et al., 2008), it ensures that accurate estimation of the association between the clinical features at adjacent levels of the BC is being carried out.

At onset, clinical features grouped under the “body level” of the BC were input into GA-SVM for investigation. This step aims to identify clinical features that contribute significantly to the development of an accurate inference model for MI. Consequently, significant risk factors, defined in this work as risk factors that can potentially contribute to the manifestation of a clinical or subclinical risk, were identified - forming the top level of the BCEN. If any of these identified risk factors are continuous, it is discretized based on the extended χ^2 algorithm (Su & Hsu, 2005). The reason for performing this step was to alleviate the associated computational complexity when analysis was performed with SVM.

Next, clinical features categorized under the "system level" of the BC were input into GA-SVM for investigation. This, similar to the earlier step, aims to identify clinical features that have a significant impact to the inference of the phenotypic manifestation previously identified at the "body level". The

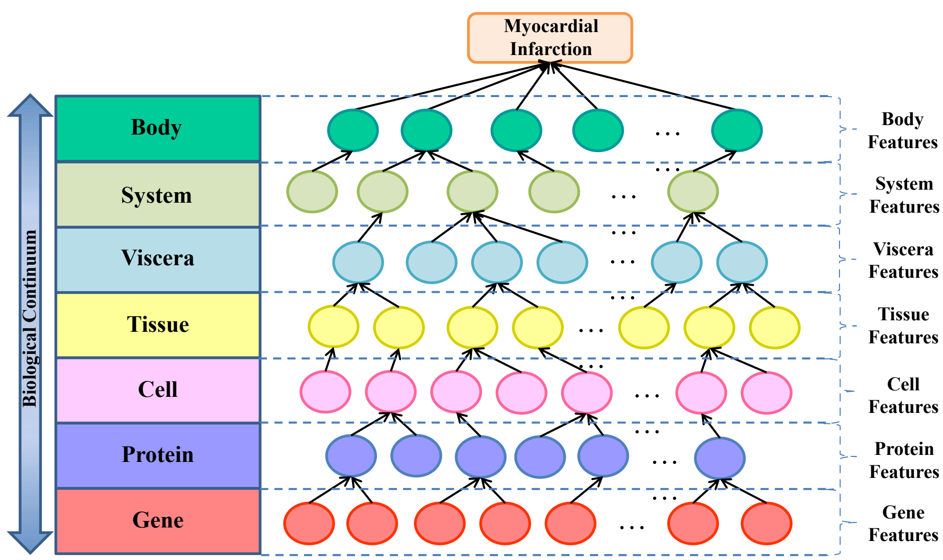


Figure 3.4: Graphical Illustration of BCEN
The circles represent clinical feature that belong to the respective levels of the BC. The arrows linking the clinical features indicate that a significant correlation was found between them.

resultant output from this step forms the "system level" of BCEN. This procedure is repeated for the rest of the levels along the BC, constructing a probabilistic tree-structured BCEN at the end of this propagation. The resultant BCEN is capable of scrutinizing how, for instance, clinical features at the visceral level are associated with those at the system level and, in turn, how these features at the system level are associated with those at the body level. This concept is graphically illustrated in Figure 3.4.

3.3.3. MI Classification with BCEN

After the construction of BCEN for MI, the distinct risk factors present in the network were used to develop an MI classification model. The performance (both classification accuracy and computational time) yielded with this approach was compared with an MI classification model that uses all clinical features present in the CHS dataset. GA-SVM was used as the classification algorithm for both the postulated approaches; hence, any benefits or drawbacks of using this classifier would prevail in both approaches.

3.4. Experimental Results

3.4.1. Data Preprocessing

Records and clinical features with considerable missing entries were removed. In addition, only records with known MI status were selected. This resulted in a dataset comprising of 4612 instances and 272 clinical features, with less than 1% of missing values (with respect to the entire dataset) and 40.8% of records with complete entries. The training and query datasets thus have 1881 and 2731 instances (both with 272 features), respectively. Subsequently, the K neighbour value for each clinical feature was determined based on the normalized training dataset. This yielded an average K value of 9.80, with

Table 3.1: Details of Best-Performing Clinical Feature Subsets

	Parent Node	Child Nodes	# Inner Nodes	# Leaf Nodes	Total Nodes	ACC	SN	SP	PR	FM	BA
	MI Status	Clinical Features at Body Level	5	6	11	0.828	0.786	0.866	0.846	0.815	0.826
Body Level	ANGBASE	Clinical Features at System Level	4	19	23	0.814	0.741	0.929	0.416	0.428	0.835
	CHFBASE		2	16	18	0.958	0.559	0.855	0.596	0.575	0.707
	STRKBASE		0	12	12	0.958	0.701	0.905	0.878	0.672	0.803
	CBD		3	18	21	0.955	0.734	0.983	0.841	0.784	0.858
	OVRWT120		2	32	34	0.737	0.737	0.738	0.704	0.720	0.737
System	ANBLMOD	Clinical Features at Viscera Level	0	25	25	0.785	0.562	0.931	0.841	0.673	0.746
	CLBLMOD		0	18	18	0.955	0.426	0.991	0.767	0.548	0.709
	SUPPUL16		0	9	9	0.828	0.865	0.749	0.834	0.849	0.807
	CHSTPN		0	16	16	0.717	0.794	0.609	0.695	0.741	0.702
	VISPROB		0	21	21	0.829	0.667	0.981	0.568	0.609	0.824

Column 1 provides the best-performing clinical features at different levels of the BC: ANGBASE = angina status at baseline; CHFBASE = congestive heart failure at baseline; STRKBASE = stroke status at baseline; CBD = self-reported stroke, transient ischemic attack and cardiac endarterectomy; OVRWT120 = obesity > 120% ideal; ANBLMOD = angina modified at baseline status; CLBLMOD = claudication modified baseline status; SUPPUL16 = supine reading: 30 second heart rate; CHSTPN = chest pain; VISPROB = vision problem.

Columns 6 to 11 represent the various performance measurements: ACC = Accuracy; SN = Sensitivity; SP = Specificity; PR = Precision; FM = F-Measure; BA = Balanced Accuracy.

standard deviation of 9.38. Data imputation was next performed to impute the missing entries found in the query dataset.

The imputed dataset obtained has a high fraction of controls (i.e. without MI - 4200 instances) and a relatively small portion of cases (i.e. with MI - 412 instances). SOM was thus employed to resolve this class data imbalanced problem. Under-sampling was performed on the major class (i.e. controls), yielding 441 instances. The final dataset produced has 853 instances and 272 clinical features.

3.4.2. Segregation of Clinical Features

The construction of a BCEN involved the segregation of the clinical features (173 diagnostic measurements and 1 MI status) along the BC. These 173 clinical features (after excluding medication) satisfied the characteristics of only 4 levels of the BC; namely, body, system, viscera and protein. Among these clinical features, 38, 74, 41 and 20 belong to the body, system, viscera and protein levels, respectively. A description of the segregated clinical features is provided online as an Appendix at http://www.bg.ic.ac.uk/jtay/web/chs_appendix.html. Readers may refer to the CHS data dictionary made available at the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) website for more information (https://biolincc.nhlbi.nih.gov/studies/chs/).

3.4.3. Construction of BCEN and Classification of MI

Clinical features at the body level were first deployed to GA-SVM to determine the set of risk factors that were highly correlated to MI (root node). A total of 11 risk factors, namely ANGBASE (angina status at baseline), CHFBASE (congestive heart failure at baseline), STRKBASE (stroke status at baseline), CBD (self-reported stroke, transient ischemic attack (TIA) and cardiac endarterectomy), SCORE03 (social support score), AMOUNT (cigarettes smoked per day), WGTEEN (teenage weight category), OVRWT120 (obesity > 120% ideal), EDUC (education level), WAIST (waist circumference – cm) and ALCOH (number of alcoholic beverages per week) were identified at the body level (note that these modifiable risk factors are also identified in earlier reported clinical studies (Yusuf et al., 2004; Rosengren et al., 2009)).

When extending the network, only clinical feature subsets (child nodes) that yielded a balanced accuracy of at least 0.7 were considered. This threshold was imposed to reflect only child nodes that are highly correlated to their parent node. This resulted in 5 inner nodes at the body level - namely ANGBASE, CHFBASE, STRKBASE, CBD and OVRWT120. This criterion was applied to the rest of the levels of the BC.

Figure 3.5: Sub-Network of BCEN for MI

Eleven clinical features at the body level were found to be potential etiological factors of MI. Obesity, one of the risk factor of MI, consists of 34 highly correlated clinical features at the system level.

Table 3.2: Performance of Classification with and without BCEN

Experiment	#Features Considered	#Gen	Time Taken, Hours (Mean±SD)	ACC	SN	SP	PR	FM	BA
Baseline Method: Classification with Original Set of Risk Factors	173	73	69.6±0.136	0.941	0.993	0.893	0.897	0.942	0.943
Proposed Method: Classification with Risk Factors Present in BCEN	111	21	14.7±0.005	0.931	0.995	0.871	0.878	0.933	0.933

These experiments were executed in parallel over an 8-core computer server. The best-performing clinical feature subset is the same for the different runs. '#Gen' denotes the number of generations taken by GA before it converges.

The resultant inner nodes identified at the system level include ANBLMOD (angina modified at baseline status), CLBLMOD (claudication modified baseline status), SUPPUL16 (supine reading: 30 second heart rate), CHSTPN (chest pain) and VISPROB (vision problem). Table 3.1 provides the details of the best- performing clinical feature subsets that satisfy the aforementioned criteria. Note that none of the clinical features at the protein level correlated well with those at the visceral level. The authors believe that this could be due to the discontinuity in continuum along the BC (i.e. missing data at the tissue and cell levels) when estimating the association between the clinical features and phenotypic manifestation that resulted in the low performance.

The resultant BCEN consists of 111 distinct nodes (Body level: 11; System Level: 63; Viscera Level: 37) in total, accounting for 64.1% of the original number of clinical features analysed. The complete BCEN for MI (created using prefuse toolkit (Heer et al., 2005)) is illustrated in Web Figure 1 - available at http://www.bg.ic.ac.uk/jtay/web/chsBCENFull.html. The BCEN provides a visual and interactive etiological network for the user to visualize and comprehend the relationship among the different risk factors along the BC for MI. For our discussion here, a sub-network of the BCEN was analysed because of its complexity and numerous interrelated risk factors present in the complete network. This sub-network is presented in Figure 3.5.

Referring to Figure 3.5, it can be seen that obesity (OVRWT120), a risk factor of MI, has 34 risk factors at the system level that are highly correlated with it. These risk factors are related to rheumatology, physical function, oncology, pulmonology, thromboembolism, sleep disorder, ophthalmology, otolaryngology, cognitive function and endocrinology. They account for 45.9% of the clinical features analysed at the system level. This suggests that not all clinical features at the system level are good predictors of obesity and it could be more fruitful to focus investigations on significantly contributing clinical features.

MI classification, with GA-SVM algorithm, was next performed with the 111 clinical features that were present in BCEN. Baseline comparison was made with the original set of 173 clinical features present in the imputed CHS dataset. Results, as shown in Table 3.2, were obtained from averaging 3 runs of GA-SVM. For each method, the best-performing clinical feature subset for the different runs is the same. Comparable classification performance was achieved for both the methods. However, the computational time required by the proposed method (i.e. deploying only risk factors present in the BCEN to GA-SVM algorithm) to develop the MI classification model was much lower (approximately 14.7 hours).

3.5. Discussion

To develop MI classification models efficiently in high dimensional datasets, we introduced a novel methodology for the reduction of clinical features to be analysed without compromising the performance of the classification model. Classification (without feature selection) conducted on a large number of clinical risk factors often produced low-performing classification models, as the performance is often jeopardized by the present of irrelevant or redundant predicators. On the other hand, the development of classification models with feature selection (e.g. the baseline method used in this work) conducted on a large number of clinical risk factors is usually computationally expensive. Therefore, pre-selection of clinical risk factors is vital to mitigate this problem contributed by the 'curse of dimensionality'. This was performed by segregating the clinical features along the various levels of the BC. The segregation process effectively reduces the data dimension, where its size is dependent on the number of clinical features categorized under each level of the BC. In this study, for example, analysis performed at the "body level" requires only 38 clinical features to be considered at a time. This, in contrast to the initial 173 clinical features, offers a reduction of 4.55-fold in the data dimension. Having to analyse a smaller number of clinical features inevitably reduces the amount of computational time required to develop the classification model. Moreover, if prior knowledge is available the data dimension can be further restricted. For instance, Emily et. al. (Emily et al., 2009) utilized knowledge from protein databases to reduce the search of SNPs to gene pairs that are known to interact and reference. A similar concept can be applied to other levels of the BC to alleviate the search effort required.

Although effort is required to construct the BCEN, the resultant network has several advantages. Firstly, with the introduction of new clinical risk factors the entire BCEN need not be reconstructed. It provides a reusable framework where only the level of the BC, at which the new clinical risk factor belong to, need to be redeveloped. If the newly introduced clinical risk factor is identified as an etiological factor (i.e. risk factor contributing to the cause of the disease), then starting with that clinical risk factor as the root node, the network is extended for levels of the BC that is below that of the newly inserted etiological factor. This approach thus provides a significant reduction in the time and effort required to build up-to-date clinical classification models. Secondly, the BCEN provides an excellent paradigm for the illustration of the potential biological pathways that underpin the different phenotypic manifestations and has the significant advantage of analysing only clinical risk factors that are biologically

Table 3.3: Obesity-System Level Risk Factors

Variable	Description
ARTH01*	Arthritis
DIAG01*	Ever diagnosed with cancer
BRONCH*	Bronchitis confirmed by doctor
PNEUMON*	Pneumonia detected by doctor
EMPHYSEM*	Emphysema detected by doctor
THROMB*	Deep vein thrombosis
ROSEIC*	Intermittent claudication by rose questionnaire
GROGGY[Ɨ]	Groggy in morning
TRSLEEP*	Trouble falling asleep
WKERLY*	Wake up far too early
RECOGN*	See enough to recognize person
TELE*	Hear enough to use phone
CONVER*	Hear enough to converse
ADL*	Activities of daily living (ADL)
IADL*	Instrumental ADL score
UES[Ɨ]	Upper extremity score
BLEED12[Ɨ]	Bleed or bruise easily
CLOT12*	Disorder related to blood clotting
LTAAI*	Left ankle-arm index (%)
SUPPUL16[Ɨ]	Supine reading: 30 second heart rate
BIORES21*	Bioelectric impedance – resistance
BAL22[Ɨ]	Dizziness, loss of balance screen
LOSBAL22*	Loss of balance
DIZZY22[Ɨ]	Dizzy/light-headed when stand up quickly
TIMEWLK*	15 feet walk time-sec
DIGCOR*	Digit symbol score
SCOR3510*	Mini-mental score (35pt)
SCORE30*	Mini-mental score (30pt)
ANBLMOD*	Angina modified baseline status
CHBLMOD*	CHF modified baseline status
STBLMOD*	Stroke modified baseline status
TIBLMOD*	TIA modified baseline status
HYPER*	Calculated hypertension status
DIABADA*	ADA guidelines diabetic status

* Risk factors found in previous work
[Ɨ] Potential risk factors not found in previous studies (to the best of our knowledge)

plausible. This not only allows the identification of significant risk factors that can be used for efficient development of accurate classification models, but, also, (1) reveals relationships that are not readily apparent from the study of individual disorders, (2) provide a global perspective of the different risk factors and etiologic pathways associated with the disease, and (3) identify new risk factors that could pave the way to the development of novel diagnostic, preventive or therapeutic strategies. Therefore, BCEN may be a simple etiological network, but it has the potential to provide significant insights into the mechanisms of a disease.

The constructed BCEN was validated by comparing the identified inter-relationship among different risk factors with those reported in previous clinical studies. All risk factors found at the body level of BCEN were also identified in previous clinical studies. Further, comparisons of a sub-network of BCEN (i.e. obesity-system sub-network) have shown that there is a large overlap (of 82.4%) between the identified relationships and those found in previous work. A possible reason for the identification of the additional inter-relationships is the employment of machine learning techniques. Since previous clinical studies tend to use linear statistical models to perform the analysis, non-trivial and non-linear relationships may go undetected. Therefore, the use of machine learning techniques in this work could potentially identify the non-trivial, non-linear and interacting etiological factors. This enables one to better understand the underlying causes of the disease, allowing more appropriate and focus interventions to be recommended to the patients. Table 3.3 lists the risk factors found to be highly associated with obesity and their presence in the clinical literature.

Arthritis, for instance, has been reported previously to be more prevalent among obese patients (Holliday et al., 2011; Park & Lee, 2011). This is primarily due to the presence of excess biomechanical stress, inducing deleterious effect on the joints. Similarly, obese individuals have a higher risk of cancer related to endometrium, prostate, colon, esophagus and stomach (Kane et al., 2005; Yang et al., 2009). Previously reported investigations have also shown association between obesity and bronchitis, pneumonia, emphysema, deep vein thrombosis, intermittent claudication, duration of sleep, blindness, hearing impairment, activities of daily living, pulmonary embolism, ankle-arm index, loss of balance, walking capacity, cognitive function, unstable angina, stroke, transient ischemic attack, hypertension and diabetes (Guerra et al., 2002;

Corrales-Medina et al., 2011; Samama, 2000; Golledge et al., 2007; Patel et al., 2008; Patterson et al., 2004; Habot-Wilner & Belkin, 2005; Fransen et al., 2008; Himes, 2000; Stein et al., 2005; Tison et al., 2011; Gray et al., 1989; Corbeil et al., 2001; Hulens et al., 2003; Elias et al., 2003; Wolk et al., 2003; Winter et al., 2008).

This suggests that the BCEN is feasible and effective in characterizing a disease and identifying the possible etiological factors. It is noteworthy that analysis of the obesity-system sub-network identified 6 new clinical features that were not previously identified in previous work. This could indicate that these clinical features are potential etiological factors of MI where further investigations could improve the understanding and treatment of the disease. We hypothesize that the reconstruction of the etiologic pathways is of major importance in healthcare as it would allow a more proactive approach for providing medical interventions to eradicate or delay the onset of a disease. This differs from the traditional reactive approach where individuals visit a physician only when they are sick or in pain, which sometimes results in a situation where treatment is too late to achieve complete recovery. Early medical interventions can be realized with BCEN by monitoring and controlling the risk factors (especially at the lower levels of the BC) that contribute to the development of a disease (e.g. MI).

The employment of BCEN to reduce the number of clinical features to be analysed significantly alleviated the computational demands. Without acutely compromising the classification performance, a speedup of approximately 4.73-fold was achieved. This was possible due to the earlier convergence of GA, suggesting that significant risk factors are already identified and present in BCEN. This facilitates the identification of risk factors that contribute significantly to the modelling of accurate MI classification model.

This study has a few limitations. Firstly, only a single dataset (i.e. CHS dataset) was used to build the etiological network for MI. This inevitably limits the power to detect all the associated risks and conclusively state that the BCEN has described the complete etiology of MI. Additionally, it limits the ability to state that the proposed method provides efficiency for all clinical classification problems. Nonetheless, it does shed some light to a novel approach for investigating the etiology of MI and efficient clinical classification. Secondly, only a single classification algorithm (i.e. SVM) has been used to identify the association between the clinical features and for developing MI classification model. This may hinder the discovery of the underlying associations and the performance of the classification model, as no single machine learning technique or statistical model is optimal for every problem. The reason for this is because each method would have its own inductive bias (Freitas & Timmis, 2007). Hence, it is suggested in (Cruz & Wishart, 2006) that comparison between multiple machine learning techniques, traditional statistical models and expert-based schemes should be conducted in order to assess the suitability of each method for a particular problem. Finally, the CHS dataset only contains risk factors that fall under the body, system, visceral and protein levels. This hinders the construction of a complete BCEN, limiting the ability to provide a more comprehensive illustration of the underlying etiology of a disease and the development of a more accurate classification model.

Nevertheless, the constructed BCEN is potentially capable of presenting the etiology of a disease in a biologically-structured manner that could facilitate the understanding and management of a disease. Moreover, it offers an effective and efficient approach for the development of MI classification model.

3.6. Summary

In view of the high prevalence of MI worldwide, better ability to characterize and classify the disease is both appropriate and necessary. In this chapter we have presented an integrated approach to build a single probabilistic network (i.e. BCEN which identifies and relates the etiological factors associated with MI) that aims to provide an efficient approach for the development of MI classification model.

Validation of the constructed BCEN was conducted and our results indicate that the network is reliable and capable of identifying significant etiological factors. There is a large overlap between the relationships identified by our approach and those found in previous work. Out of the 34 clinical features identified at the obesity-system level, 28 (82.4%) of them were found in the previous clinical studies. However, 6 new clinical features, that had not been identified previously, were found to be associated with obesity in this study. These new clinical features could be probable risk factors for MI. They indicate the need for further clinical investigations to improve the understanding and treatment of the disease.

Based on the distinct risk factors identified and present in BCEN, a classification model for MI was developed. The classification model obtained demonstrated high balanced accuracy of 0.933. It was developed at a rate of 4.73-fold faster than its counterpart that does not adopt any pre-selection strategy. This suggests that BCEN may be a desirable approach for developing clinical classification models when a large number of clinical features need to be considered.

Although further validation of this methodology is necessary, this approach may be valuable in exploring and identifying risk factors that underpin a disease. To conclude, the BCEN is an etiological network that is simply built but profoundly useful. It has the potential to provide insights, from a novel perspective, into the characteristics of (current/new) diseases - allowing more efficient and effective understanding, analysis, management and classification to be undertaken. We look forward to a more comprehensive understanding of the disease etiology and eventually, towards personalized medicine.

Disclaimer

The CHS dataset described in this chapter is provided by the National Heart, Lung and Blood Institute (NHLBI).

Chapter 4

Evolutionary Data-Conscious Artificial Immune Recognition System[4]

Artificial Immune Recognition System (AIRS) algorithm (version 2), introduced by Andrew Watkins (Freitas & Timmis, 2007) in 2004, offers a promising methodology for supervised data classification. It is an immune-inspired learning algorithm that works efficiently and has shown comparable performance with respect to other classifier algorithms. For this reason, it has received escalating interests in recent years. However, the full potential of the algorithm was yet unleashed.

In this chapter, a novel supervised classification algorithm further inspired by the natural immune system is presented. This algorithm, called the evolutionary data-conscious artificial immune recognition system (EDC-AIRS), is an improvised version of artificial immune recognition algorithm version 2 (AIRS2). It exploits 3 additional immune metaphors which empowers the algorithm with the ability to robustly adapt to the different density, distribution and characteristics exhibited by each data class. Promising results have been achieved when evaluated with six widely used benchmarking datasets.

[4] The work presented in this chapter has been published in the ‘Proceeding of the fifteenth annual conference on Genetic and Evolutionary Computation Conference (GECCO)’ and reprinted with permission: Tay, Poh & Kitney, “An evolutionary data-conscious artificial immune recognition system”, GECOO’13, © 2013 ACM, Inc. http://doi.acm.org/10.1145/2463372.2463499 (ISBN: 978-1-4503-1963-8).
This paper can be found in Appendix D.

4.1. Introduction

The human immune system is a highly sophisticated, distributed, complex and powerful natural defense mechanism that comprises of several functional mechanisms, positioned in strategic locations, conferring resistance against viruses and foreign pathogens. It has the ability to learn the characteristics of the foreign antigens and contrive a defense strategy to detect and neutralize them. Specifically, the immune system possesses properties such as the capability of recognition, memory acquisition, diversity and self-regulation, making it highly suitable for learning patterns that underlie a data. On this note, it has inspired the development of the artificial immune system capable of solving many problems related to computer science and engineering (e.g. computer security, anomaly detection, optimization, machine learning, etc.) (Freitas & Timmis, 2007; Castro & Timmis, 2002). One such algorithm that has received escalating interests is the Artificial Immune Recognition System version 2 (AIRS2) (Watkins et al., 2004).

Although AIRS2 algorithm has shown to be an effective classification algorithm, some useful immune mechanisms are yet to be exploited by the algorithm. For instance, artificial recognition balls (ARBs) are used in AIRS2 algorithm to denote a representative subset of B-Cells. They would compete for survival based on the idea of resource limited system (Timmis & Neal, 2001). However, the creation and elimination of the ARBs do not correspond to the density of the data in which they cover (i.e. a larger number of ARBs do not survive in regions that are more densely populated with data). This contradicts with the natural immune system where macrophages would flood the extracellular space of the infected regions (attempting to eliminate the harmful agents) and B-Cells would proliferate and secrete antibodies profoundly in response to pathogenic agents. In other words, a larger concentration of defense agents would be present in regions that has received intense invasion from harmful antigens. Another area that the original AIRS2 algorithm did not explore and exploit is the distributed diversity exhibited by the lymph nodes found in the natural immune system. The AIRS2 algorithm uses a common parameter set to model the distribution of different data classes. This is undesirable in cases where the distribution of different data classes differ by a considerable degree. Observation of the strategic positioning of the lymph nodes in human bodies (which promotes better immune defense) advocates for the need of a more specific and distinct parameter set (e.g. affinity threshold scalar, density and total resources parameters) to model each data class (i.e. instances that belong to a specific class). Finally, it is important to generate B-Cells that can affiliate/bind well with the antigens. This is realized biologically through the production of highly specific surface receptors on the B-Cells which facilitates the detection and eradication of the foreign antigens. To mimic this concept computationally, feature selection can be performed where highly informative features that can describe the underlying association were identified and used for classification.

This chapter presents a novel algorithm called the evolutionary data-conscious AIRS (EDC-AIRS) algorithm, which extends the existing AIRS2 algorithm by contextualizing the immune response to the concentration, distribution and characteristics of the antigens and is no longer a global centralized response. When evaluated using 6 widely used benchmarking datasets, our method has exhibited improved learning ability and classification accuracy.

The rest of the chapter is organized as follows. Section 4.2 provides a brief introduction to the AIRS2 algorithm. A description of the proposed EDC-AIRS algorithm is presented in Section 4.3. The experimental results are offered in Section 4.4 and discussed in Section 4.5. Finally, Section 4.6 concludes this chapter.

4.2. Artificial Immune Recognition System

The natural immune system is a highly rapid and efficient biological self-defense mechanism that protects a given host against infections, for example, from foreign antigens or pathogens. The immune system functions by detecting a wide variety of agents and distinguish the foreign antigens (e.g. viruses) from the organism's own healthy cells or molecules (also known as self-antigens). The immune system consists of a number of components. Two examples are macrophages and lymphocytes (e.g. B-cell ad T-Cell) which are responsible for the recognition and elimination of the determined infectious agents. The lymphocytes have highly specific surface antigenic receptors to a given antigenic determinant, in which they would only proliferate in response to a specific infection. Therefore, the type of antibodies present in an individual could reflect the infections to which they are infected with.

The antibody's polypeptide chains composed of a highly variable amino-terminal region (V-region) and a carboxy-terminal region (C-region) that can be of a few types. The V-region is responsible for the antigenic detection while the C-region is responsible for a variety of effector functions. The polypeptide chain of an antibody is formed through the genetic recombination and somatic hyper-mutation of multiple gene segments scattered along the chromosome of the genome. Such formation mechanism used to generate antibodies introduces diversity into the underlying immune defense (Castro & Timmis, 2002), ameliorating the ability of the antibodies to recognize/bind to the antigens.

Inspired by the robustness exhibited by the natural immune system, the AIRS2 algorithm (Watkins et al., 2004) - a novel one-shot incremental supervised learning algorithm was developed and applied to solve classification problems. It has several attractive characteristics such as the ability to (1) adaptively develop an appropriate architecture during the learning process, (2) achieve competitive accuracy compared to other classification algorithms, (3)

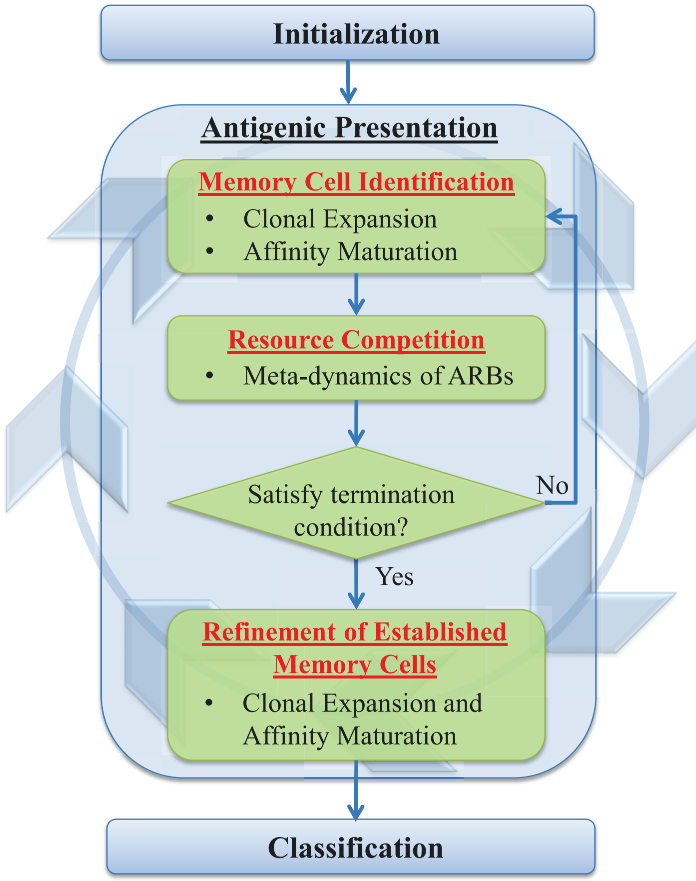


Figure 4.1: Canonical Flow of the AIRS2 Algorithm
The AIRS2 algorithm consists of 4 stages during the learning phase, namely the initialization, memory cell identification, resource competition and refinement of established memory cells stages. The last 3 stages will repeat for every data instance presented. After learning the underlying pattern of the data, classification is performed on the unseen data using K-nearest neighbour.

develop a generalized model by generating a representative set of memory cells and (4) achieve accuracy comparable to those obtained with the optimal parameter set when experimented over a wide range of parameter values (Watkins & Boggess, 2002).

The AIRS2 algorithm consists of 4 stages during the process of learning the underlying patterns of the data. They are the initialization, memory cell identification, resource competition and refinement of established memory cells stages. The canonical flow of the AIRS2 algorithm is presented in Figure 4.1.

The initialization stage is responsible for normalization of the data, parameter discovery and seeding of memory cells. The data items found in the dataset is first normalized so that the Euclidean distance between the feature vectors of any 2 data items is in the range [0, 1]. Affinity threshold, the average Euclidean distance between each data item in the training dataset, is then calculated. This value controls the quality of the memory cells maintained and utilized for classification. The mathematical expression for computing the affinity threshold is as follow:

$$\text{affinity threshold} = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\text{affinity}(ag_i,ag_j)}{\frac{n(n-1)}{2}} \quad (2)$$

where n is the number of training instances (antigens), ag_i and ag_j are the ith and jth training antigens in the training data, and affinity(ag_i,ag_j) returns the Euclidean distance between the two antigens. The initial memory cell pool (MCP), a collection of classifier cells that will be used for classification at the end of the training lifecycle, is then seeded by randomly selecting data item(s) from the training dataset.

A process known as the antigenic presentation is then undertaken where each training instance is subsequently presented to the AIRS2 algorithm. For each training instance presented, it first undergoes the memory cell identification stage where its affinity with the memory cells in MCP (that reside in the same class) is computed. The most stimulated memory cell (also known as matched memory cell) is then selected and cloned in proportion to its stimulation value (i.e. clonal expansion phase). This value is calculated based on the following equation:

$$\text{stimulation}(x, y) = 1 - \text{affinity}(x, y) \quad (3)$$

where x is the presented training instance and y is the memory cell. These cloned memory cells forms the artificial recognition ball (ARB) pool where an ARB (Timmis & Neal, 2001) is a single representation for a number of similar memory cells. This allows a reduction in duplication and manages the survival of classifier cells within the population. The cloned ARBs are then mutated at a rate inversely proportional to the antigenic affinity, introducing diversity into the system (i.e. affinity maturation phase). The range of the maturated value assigned to a selected attribute is centered at the attribute's initial value and spanned over the difference between 1 and the ARB's stimulation value. In other words, mutated ARB offspring of highly stimulated cells are only allowed to explore and mutate to a value near its initial state while less stimulated ARB offspring are allowed to mutate over a lager range.

Next, the ARBs will compete for survival based on the concept of resource allocation mechanism (Timmis & Neal, 2001), where the ARBs are allocated a number of resources proportional to their normalized stimulation values. The resulting ARBs with insufficient resources are subsequently pruned (i.e. meta-dynamic phase). The average simulation level for the ARBs is then computed based on the following equation:

$$\text{avg\_stimulation}_{\text{i}} = \frac{\sum_{j=1}^{|AB_i|} ab_j.stimulation}{|AB_i|}\ ,\ ab_j \in AB_i \qquad (4)$$

where AB refers to the ARB pool, ab ε AB; $|AB_i|$ is the number of ARBs in class i. The average stimulation is then compared with the user-defined stimulation threshold. If it is greater than the user-defined threshold, the training cycle stops for that training instance. Otherwise, the training cycle repeats.

Once the termination condition is satisfied, the most stimulated ARB is selected as the candidate memory (CM) cell. If this CM cell's stimulation level is higher than all the memory cells in the established memory (EM) set (i.e. collection of ARBs that have survived the resource competition stage), then it is added into the EM set. Otherwise, this CM cell is discarded. Finally, replacement of the EM cells is carried out first by computing the memory cell replacement cutoff value as defined as:

$$\text{Cutoff} = \text{AT} * \text{ATS} \qquad (5)$$

where AT refers to affinity threshold and ATS denotes affinity threshold scalar. If the affinity between this CM cell and the best affiliated memory cell found previously (i.e. EM cell) is below the cutoff value, the EM cell will be removed and replaced with the CM cell. Consequently, the next training instance is deployed to the AIRS2 algorithm until all the training instances are presented. This process ultimately identifies a set of representative memory cells that provides a generalized representation of the pattern that underlies the data, which will then be used for classification. The classification algorithm employed is K-nearest neighbour (KNN) where the classification outcome for each unseen data instance is determined by taking the majority vote of the k most stimulated EM cells. For a more detailed description of the algorithm, readers can refer to (Watkins et al., 2004; Brownlee, 2005).

```
CandStim ← stimulation(ag, mc_candidate)
MatchStim ← stimulation(ag, mc_match)
CellAff ← affinity(mc_candidate, mc_match)
if (CandStim > MatchStim)
    if (CellAff < AT * ATS)
        MC ← MC - mc_match
    end
    MC ← MC ∪ mc_candidate
end
```

Figure 4.2: Pseudo-code for Memory Cell Introduction used in AIRS2 Algorithm – adopted from (Watkins et al., 2004)

CandStim (and MatchSim) denotes the stimulation level between the presented antigen and the candidate (and matched) memory cell. CellAff refers to the affinity between the candidate and matched memory cell. MC represents the memory cell pool.

4.3. Material and Methods

4.3.1. Evolutionary Data-Conscious AIRS (EDC-AIRS) Algorithm

This study formulates a novel immune-inspired (EDC-AIRS) algorithm that employs several natural immune mechanisms. In particular, how antibodies evolve and adapt to the different concentration, location and type of foreign antigens are being mimicked in addition to those proposed by the AIRS2 algorithm. This, when implemented as a high fidelity computational technique, empowers the algorithm with the ability to independently adapt to the distinct (1) density, (2) distribution and (3) characteristics of each data class.

Firstly, the ability to adapt to the different (local) density present in the data was addressed by the observation of the rapid growth of macrophages and

```
CellAff ← affinity(mc_candidate, mc_match)
Density_count ← 0
foreach (ag_i in AG)
do
    AntigenAff ← affinity(ag_i, mc_candidate)
    if (AntigenAff < AT*Radius_density)
        Density_count ← Density_count + 1
    end
done
Density_ratio ← Density_count / Density_max
if (CellAff < (1- Density_ratio) * AT * ATS)
    MC ← MC - mc_match
end
MC ← MC ∪ mc_candidate
```

Figure 4.3: Pseudo-code for Memory Cell Introduction used in EDC-AIRS Algorithm

$Density_{count}$ represents the number of antigens that is proximal to the candidate memory cell. $Density_{max}$ denotes the maximum number of antigen present in the training data.

B-Cells in response to the invasion of foreign antigens (particularly, at the regions of infection). More specifically, a relative proportion of antibodies to antigens were necessary to neutralize the harmful agents. This mechanism was incorporated in the EDC-AIRS algorithm by allowing a relatively larger number of ARBs to survive in regions that are more densely populated with training data. Implementation was carried out by removing and modifying some of the criteria present in the original AIRS2 algorithm. In particular, the way the memory cells are introduced into the system is modified. The original pseudo-code for memory cell introduction (Watkins et al., 2004) used in AIRS2 algorithm is shown in Figure 4.2. In this (original) implementation, the candidate memory cell ($mc_{candidate}$) is first identified by determining which memory cell generated has the highest stimulation level (Figure 4.2 line 1) to the training antigen (ag) presented. Next, this new $mc_{candidate}$ is introduced into the existing memory cells (MC) pool (Figure 4.2 line 8) if it is more stimulated to the training antigen presented than the most stimulated memory cell (mc_{match}) in the MC pool (Figure 4.2 line 4). $mc_{candidate}$ will replace mc_{match} (Figure 4.2 line 6) if the affinity between them is less than the product of affinity threshold and affinity threshold scalar (Figure 4.2 line 5).

In our proposed implementation, the criterion that requires the candidate memory cell ($mc_{candidate}$) to be more stimulated (by the training antigen, ag) than the matched memory cell (mc_{match}) before it was added to the memory cell pool was first removed. The reason for doing so is to encourage new ARBs that are highly stimulated (ensured by the high stimulation threshold adopted) to survive within the system. Secondly, computation of the density ($Density_{count}$) proximal to $mc_{candidate}$, based on the initial set of training antigens (AG), was implemented in the algorithm. The degree of proximity was determined by a user-defined parameter, $Radius_{density}$ (Figure 4.3 line 6). This step aims to determine the density of the training antigens surrounding $mc_{candidate}$ that is no more than '$Radius_{density}$' distance away (Figure 4.3 line 3-9). Additionally, the maximum density ($Density_{max}$) present in AG was also computed based on the same approach. Another user-defined parameter ($Radius_{max}$) was used instead to determine the size of the region to be considered. Finally, a weighting coefficient ($Density_{ratio}$) was derived based on the following equation:

$$Density_{ratio} = \frac{Density_{count}}{Density_{max}} \tag{6}$$

This weighting coefficient (Figure 4.2 line 10) was subsequently multiplied with the product of affinity threshold (AT) and the affinity threshold scalar (ATS) to determine whether the new memory cell should be introduced into the memory cell pool (Figure 4.3 line 11-14). Specifically, if the affinity between $mc_{candidate}$ and mc_{match} is less than the product, mc_{match} will be replaced by

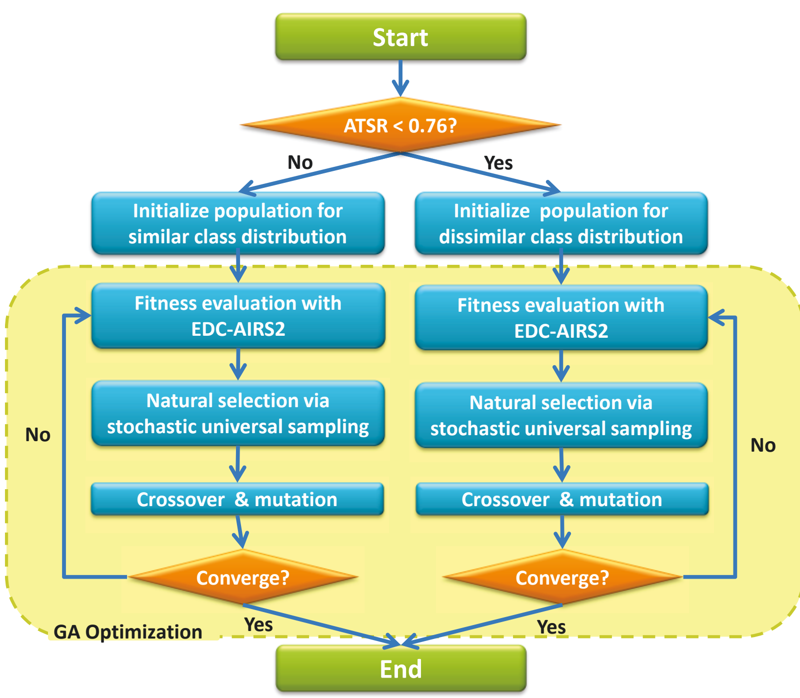


Figure 4.4: Proposed Methodology for Optimization of Parameter Set for Binary Class Classification Problems
The ATSR of the dataset was first computed. If it was smaller than the empirically derived threshold (0.76), an independent set of parameters for each data class was created. Otherwise, a common parameter set was used for all data classes. These parameters were then optimized (with respect to classification accuracy) using GA.

$mc_{candidate}$ (Figure 4.3 line 12). Otherwise, $mc_{candidate}$ will be added to MC pool (Figure 4.3 line 14). The revised pseudo-code for memory cell introduction used in EDC-AIRS algorithm is depicted in Figure 4.3.

The strategy that was delved into next is associated with the distribution characteristic of different data classes. This is vital according to the mechanism observed in the natural immune system, where lymph nodes are located in strategic positions – producing antibodies that could detect and eradicate the foreign antigens more efficiently. The spatial independency of the lymph nodes (Moses & Banerjee, 2011) and the circulatory networks in the immune system is of significant importance as it enables decentralized immune defense while protecting the human body in a global fashion. Therefore, if the distribution of different data classes differs by too much, this could indicate that the location at which the antibodies are produced (i.e. the position of the lymph nodes) would need to be adjusted so that the antibodies produced could detect and eradicate the antigens found in each data class in a more efficient manner. On the contrary, if the distribution of different data classes is near symmetry, this could indicate that no additional lymph nodes are required for more efficient neutralization of antigens; mitigating the required search effort as a result. This was empowered within the EDC-AIRS algorithm by having an independent set of parameters for evolving the memory cells if the distribution similarity between the data classes was below an empirically derived threshold, known as the Affinity Threshold Similarity Ratio (ATSR). Otherwise, a common set of parameters was used for all data classes. The ATSR was calculated by first computing the affinity threshold (i.e. the average affinity value over all training data) associated with each data class. After which, the minimum affinity threshold found among the different data classes was divided by the maximum affinity threshold found. The mathematical expression is given

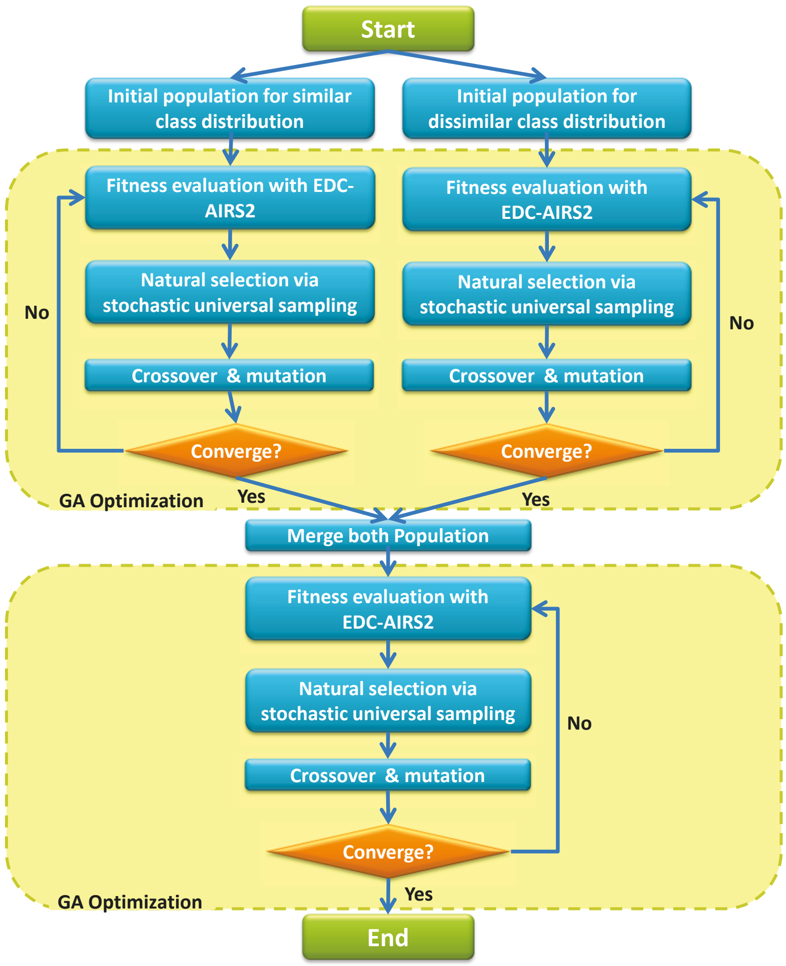


Figure 4.5: Proposed Methodology for Optimization of Parameter Set for Multiclass Classification Problems
Both similar and dissimilar populations were initialized and optimized concurrently. Upon convergence, these populations were merged and re-optimized by GA.

below:

$$ATSR = \frac{\boldsymbol{min_{i=1,\dots,C}(affinity_{threshold_i})}}{\boldsymbol{max_{j=1,\dots,C}(affinity_{threshold_j})}} \tag{7}$$

where C is the total number of classes. This provides a yardstick to determine how similar the distributions of different data classes were. The parameters that orchestrate the evolution, survival and existence of the memory cells include the total resources, affinity threshold scalar (ATS), $\text{Radius}_{\text{density}}$ and $\text{Radius}_{\text{max}}$

parameters. Therefore, these parameters were duplicated and optimized independently for each data class if the ATSR computed for the dataset was below the pre-defined threshold value. All parameters were optimized using Genetic Algorithm (GA) (Holland, 1992), a search heuristic that imitates the process of natural evolution. The optimization algorithm was developed using MATLAB GA toolbox (Chipperfield & Fleming, 1995) and was executed in parallel over a high performance computer (HPC) cluster.

Figure 4.4 and 4.5 illustrate the canonical flow of the proposed methodology used for binary and multiclass classification problems respectively. For binary classification problem, we assume either a similar or dissimilar distribution based on the calculated ATSR and optimize that parameter set only (i.e. distinct parameter set for each data class if the computed ATSR is below the pre-defined threshold and a common parameter set if ATSR is above the pre-defined threshold). In contrast, both similar and dissimilar distributions were assumed for multiclass classification problems. In other words, both set of parameters were optimized concurrently by GA for multiclass datasets. Upon convergence of both runs (i.e. no improvement after 10 generations or the maximum number of generations has been reached), both the populations were merged and re-optimized by GA once again.

Finally, the ability to adapt to the characteristics of the data was performed by mimicking the genetic recombination and somatic hyper-mutation of gene segments scattered along the chromosome of the genome when forming a natural antibody. This process produces highly specific surface receptors of B-Cell necessary to recognize and bind to a certain type of antigen (that possess distinct structure). From a computational perspective, this was achieved through feature selection where a subset of informative features, that could capture the true patterns underlying the particular dataset, was selected for the learning process. GA was selected to perform this feature selection task as it has the

		EDC-AIRS	
		Misclassification	Correct Classification
AIRS2	Misclassification	**a**	**b**
	Correct Classification	**c**	**d**

Figure 4.6: Contingency Table for McNemar's Test(EDC-AIRS vs AIRS2)
'a' indicates the number of data items misclassified by both EDC-AIRS and AIRS2 algorithms; 'b' represents the number of data items misclassified by AIRS2 algorithm but correctly classified by EDC-AIRS algorithm; 'c' denotes the number of data items misclassified by EDC-AIRS algorithms but correctly classified by AIRS2 algorithm; 'd' dictates the number of data items correctly classified by both EDC-AIRS and AIRS2 algorithms.

potential to generate the optimal feature subset (Huanga & Wangb, 2006). The GA parameters were determined experimentally and kept constant between benchmarks. The setup details of GA are as follow: population size: 100; maximum generation: 100; natural selection: stochastic universal sampling; crossover type: discrete recombination; crossover probability: 0.8; mutation rate: 1/P, where P is the number of parameters. The value of the EDC-AIRS parameters that was either assigned (i.e. given as a constant value) or tuned by GA (i.e. given as a range of value) are as follow: seed: 1; clonal rate: 10; hyper-mutation rate: 2; stimulation threshold: 0.9; initial memory pool size: [0, 200]; K-nearest neighbour value: [1, 15]; affinity threshold scalar: [0, 1]; total resource: [150, 300]; $Radius_{density}$ = [0, 3]; $Radius_{max}$ = [0, 3].

The performance of EDC-AIRS algorithm was evaluated with 4 benchmarking datasets, namely the Fisher's Iris, Ionosphere, Pima Indians Diabetes and Sonar Datasets. Hold-out validation was performed on the Ionosphere dataset while cross-validation was performed on the remaining 3 datasets. More specifically, the first 200 data items of the Ionosphere dataset was selected as the training data and was tested on the remaining 151 data items. As for the Iris, Pima Indians Diabetes and Sonar datasets, 5, 10 and 13-fold cross-validation was carried out respectively. The reason for choosing such validation strategy was to remain comparable to other experiments reported in the literature. Further details about the validation procedures applied on these benchmarking datasets can be found in (Watkins, 2001).

The performance yielded by EDC-ARIS algorithm was (statistically) compared with those obtained by the AIRS2 algorithm. We have chosen McNemar's test to determine whether the performance of the 2 aforementioned supervised algorithms are statistically different as it had been demonstrated to have low type 1 error (Dietterich, 1998). To perform the test, both EDC-AIRS and AIRS2 algorithms were first trained with the training data and tested with the testing data. The predicted outcome for each data item in the testing data was recorded and used to construct the contingency table shown in Figure 4.3. If the sum of 'b' and 'c' is greater than 25, chi-square test with 1 degree of freedom is used for performing McNemar's test. Otherwise, to provide a better estimation of the small sample (i.e. $b + c \leq 25$), binomial distribution is used for (exact) McNemar's test. The 2 algorithms are considered to be statistically different if the p-value computed with McNemar's test is smaller than 0.05.

4.3.2. Dataset

Four benchmarking datasets obtained from (C.L. Blake & C.J. Merz, 1998) were used to evaluate the performance of the novel EDC-AIRS algorithm. A brief description of these datasets is as follow:

1. Fisher's Iris Dataset – Consists of 4 features that describe the length and width of the sepal and petal. Three classes exist which represent the type of the iris plant (i.e. Iris Sentosa, Iris Vericolour and Iris Virginica). It has a sample size of 150 with 50 instances per class. The Iris Sentosa class is linearly separable from the other 2 classes while the Iris Vericolour and Iris Virginica classes are not linearly separable from each other.

2. Ionosphere Dataset – A binary class classification problem that contains 351 instances and 34 features. The 2 classes represent “good” or “bad” radar returns. The “good” radar returns refer to those that show some types of structure in the ionosphere while “bad” radar returns have their signals passed through the ionosphere.

3. Pima Indians Diabetes Dataset – Patients in this dataset are all females who are at least 21 years of age and are of Pima Indian heritage. It is a binary class classification problem that aims to distinguish between patients tested positive for diabetes and those who are not. It contains 768 instances and 8 features.

4. Sonar Dataset – The objective of this experiment is to determine whether an object is a mine (metal) or rock by bouncing sonar signal off the object at various angles and conditions. It contains 208 instances and 60 features.

In order to investigate on how different data class distribution affects the performance of the classification algorithm, several additional benchmarking datasets were acquired from (C.L. Blake & C.J. Merz, 1998). Both similar and dissimilar distributions among the data classes were assumed for these datasets. Experiments were then conducted using these 10 datasets, with different degree of data class distribution (as determined by the computed ATSR value), to determine the impact of data class distribution on the algorithm’s classification performance. A succinct description of these datasets is as follow:

1. Wine Dataset – Contains results obtained from the chemical analysis of 3 different cultivars grown in the same region in Italy. It is a tri-nary classification problem that consists of 178 instances and 13 features.

2. Magic Dataset – This dataset, obtained from the Major Atmospheric Gamma Imaging Cherenkov (MAGIC) Telescope project, is a Monte Carlo generated data that aims to simulate the registration of high energy gamma

particles in a ground-based atmospheric Cherenkov gamma telescope. It is a binary class classification problem which contains 19020 instances and 10 features.

3. Hill-Valley Dataset – This dataset consists of 606 instances, 100 features and 2 classes. Each instance represents 100 data points. When plotted (in the given order) on a 2-dimensional graph, the resultant plot would represent either a hill (a "bump" in the terrain) or a valley (a "dip" in the terrain).

4. Bupa Liver Disorder Dataset – This dataset contains examination results (e.g. quantity of alcoholic beverages consumed per day and blood tests) of males which are used to investigate liver disorders. It has a total of 345 instances and 6 features.

5. Statlog Heart Dataset – Investigation of the presence or absence of heart disease in an individual is carried out based on various medical diagnoses. This result is dictated in this dataset, which contains 270 instances and 13 features.

6. Cardiovascular Health Study (CHS) Dataset – This dataset, as described in (Fried et al., 1991), is an epidemiology study of risk factors for cardiovascular diseases in elderly aged 65 and above. The cohort consists of elderly subjects from four U.S. communities, namely Forsyth County, North Carolina; Sacramento County, California; Washington County, Maryland; and Pittsburgh, Pennsylvania. Data collected in year 5 of the CHS study was utilized. The balanced case-control sample size consists of 270 instances and 253 features. It is a binary class classification problem (i.e. with or without myocardial infarction).

4.4. Experimental Results

EDC-AIRS algorithm was developed by extending AIRS2 algorithm. Three areas of optimization were carried out, each addressing an aspect of the phenomenon observed in the natural immune system (i.e. the concentration, distribution and characteristics of the antigens). In order to better generate a set of representative memory cells, it is necessary to empirically determine the ATSR threshold first. To perform this investigation, 10 datasets with different degree of data class distribution were evaluated. The ATSR value of these datasets ranges from 0.609 to 0.957, where a lower value indicates that the distribution of the data classes differs by a larger degree. Both classification with a common set of parameter (assuming similar distribution among data classes) and a distinct set of parameters for each data class (assuming dissimilar distribution among data classes) were performed. Based on the results shown in

Table 4.1: Empirical Experiments with ATSR based on Datasets with Different Data Class Distribution

Measurement	Ionosphere	Iris	Wine	ks_yr50611	MAGIC	Pima Indians Diabetes	Hill-Valley	Bupa-Liver Disorder	Sonar	Statlog Heart
#Instances	200	150	178	270	19020	768	606	345	208	270
#Attributes	34	4	13	253	10	8	100	6	60	13
#Classes	2	3	3	2	2	2	2	2	2	2
#Class1 Instances	99	50	59	135	12332	268	305	145	97	120
#Class2 Instances	101	50	71	135	6688	500	301	200	111	150
#Class3 Instances	-	50	48	-	-	-	-	-	-	-
Validation Type	Holdout	5-CV	LOO	10-CV	5-CV	10-CV	Holdout	10-CV	13-CV	10-CV
Class 1 AT	0.437	0.106	0.162	0.308	0.160	0.217	0.121	0.157	0.271	0.427
Class 2 AT	0.266	0.129	0.223	0.408	0.209	0.183	0.107	0.167	0.283	0.408
Class 3 AT	-	0.152	0.185	-	-	-	-	-	-	-
Overall AT	0.371	0.288	0.266	0.366	0.187	0.202	0.114	0.164	0.283	0.448
ATSR	**0.609**	**0.698**	**0.727**	**0.756**	**0.764**	**0.842**	**0.885**	**0.937**	**0.957**	**0.957**
Acc. for Similar Distribution	96.7%	99.0%	98.9%	65.9%	**83.1%**	**77.3%**	**56.3%**	**69.9%**	**88.5%**	**84.8%**
Acc. for Dissimilar Distribution	**97.4%**	**99.6%**	**99.6%**	**67.0%**	82.8%	77.1%	55.7%	69.6%	87.0%	83.7%

Accuracy (Acc.) was used to evaluate how datasets with varying degree of data class distribution affects the performance of the algorithm. The dataset 'ks_yr50611', which uses the CHS dataset, predicts the occurrence of MI (from year 6 to 11) based on a balanced case-control sample obtained in year 5.
AT means affinity threshold, CV denotes cross-validation and LOO refers to leave-one-out cross-validation.

Table 4.2: Classification Performance of the Benchmarking Datasets with Different Issues Addressed

Experiment	Description	Iris	Ionosphere	Pima Indians Diabetes	Sonar
1	AIRS2	96.0%	95.6%	74.2%	84.9%
2	GA-AIRS2	98.7%	97.4%	**77.3%**	86.5%
3	Density	98.7%	96.7%	**77.3%**	88.5%
4	Density & Distribution	**99.6%**	97.4%	**77.3%**	88.5%
5	Density, Distribution and Characteristics (EDC-AIRS)	**99.6%**	**98.0%**	**77.3%**	**90.9%**
	McNeamar's Test (p-value)	**0.008**	0.126	**0.020**	**0.042**

Using GA-AIRS2 as the base algorithm, the techniques described in experiments 3, 4 and 5 are implemented respectively.

Table 4.1, it is indicative that with an ATSR value of 0.756 and below, a distinct parameter set for each data class is capable of achieving a higher accuracy. Therefore, an ATSR threshold of 0.76 was used for the rest of the experiments.

The performance of the proposed EDC-AIRS algorithm (when compared with AIRS2 algorithm) was evaluated using 4 benchmarking datasets. The algorithm was evaluated 3 times with consistent classification result obtained each time (i.e. standard deviation of 0). The classification accuracy for the incremental implementation of the 3 aforementioned mechanisms is given in Table 4.2. Baseline comparison was made with GA-AIRS2 algorithm - an AIRS2 algorithm with its parameters tuned via GA. It is noteworthy that GA-AIRS2 performs better than AIRS2 for all 4 benchmarking datasets.

With the implementation to address the density issue (Table 4.2 – experiment 3), ameliorated performance was observed for the Sonar dataset. However, the performance on the Ionosphere dataset exacerbates while the performance for the rest of the datasets remains comparable. With the additional implementation to amortize the impact of different distribution exhibited by each data class (Table 4.2 – experiment 4), the deterioration in performance observed previously on the Ionosphere dataset vanished. Moreover,

Table 4.3: Performance Comparison of Different Classification Algorithms

	Iris		Ionosphere		Pima Indians Diabetes		Sonar		Wine		Statlog Heart	
Rank	Algorithm	Acc	Algorithm	Acc	Algorithm	Acc	Algorithm	Acc	Algorithm	Acc	Algorithm	Acc
1	Grobian (rough)	100%	3-NN + Simplex	98.7%	Logdisc	77.7%	TAP MFT Bayesian	92.3%	**EDC-AIRS**	**99.6%**	Lin. SVM 2D QCP	85.9%
2	**EDC-AIRS**	**99.6%**	**EDC-AIRS**	**98.0%**	IncNet DIPOL92	77.6% 77.6%	**EDC-AIRS**	**90.9%**	kNN, Manh, auto k=1-10 IncNet, Gauss	98.9% 98.9%	**EDC-AIRS**	**84.8%**
3	SSV C-MLP2LN PVM 2 rules	98.0% 98.0% 98.0%	3-NN IB3	96.7% 96.7%	**EDC-AIRS** Linear Disc. Analysis	**77.3%** 77.5 – 77.2%	Nave MFT Bayesian SVM Best 2-layer MLP + BP, 12 hidden	90.4% 90.4% 90.4%	SSV	98.3%	Naive-Bayes	84.5%
4	PVM 1 rule	97.3%	MLP + BP	96.0%	SMART GTO DT (5xCV)	76.8% 76.8%	AIRS2	84.9%	kNN, Euclidean, k=1	97.8%	K*	76.7%
5	AIRS FuNe-I NEFCLASS	96.7% 96.7% 96.7%	AIRS2	95.6%	ASI	76.6%	MLP+BP, 12 hidden	84.7%	FSM	96.1%	IB1c	74.0%
6	AIRS2 CART	96.0% 96.0%	AIRS C4.5	94.9% 94.9%	Fischer Disc. Analysis	76.5%	MLP+BP, 24 hidden	84.5%			1R	71.4%
7	FUNN	95.7%	RIAC	94.6%	MLP+BP	76.4%	1-NN, Manhantta n	84.2%			T2	68.1%
8			SVM	93.2%	LVQ LFC	75.8% 75.8%	AIRS	84.0%			MLP + BP	65.6%
9			FSM + rotation	92.8%	RBF	75.7%	FSM	83.6%			FOIL	64.0%
10			1-NN	92.1%	kNN, k=22, Manh MML NB	75.5% 75.5% 75.5 – 73.8%					RBF	60.0%
…					… …							
n					AIRS2	74.2%						
n+1					AIRS	74.1%						

'Acc' denotes the classification accuracy.

the accuracy obtained for the Iris dataset improved while the accuracy for both Pima Indians Diabetes and Sonar datasets remain the same. Finally, when the characteristic of the dataset was delved into (Table 4.2 – experiment 5), further improvement in accuracy for Ionosphere and Sonar datasets was obtained. Accuracy for Iris and Pima Indians Diabetes datasets remains unchanged, probably due to the limited features available for selection (i.e. 4 and 8 features respectively).

Statistical comparison of EDC-AIRS and AIRS2 algorithms indicate that EDC-AIRS algorithm achieved comparable, if not better, performance than AIRS2 algorithm. Specifically, EDC-AIRS algorithm outperforms AIRS2 algorithm for 3 out of 4 datasets (i.e. Fisher's Iris, Pima Indian Diabetes and Sonar datasets) while comparable performance was achieved for Ionosphere dataset. Six benchmarking datasets were used to compare the performance of EDC-AIRS algorithm with other well-known classifiers(Duch, 2000; Watkins et al., 2004; Duch, 2000) is provided in Table 4.3. The EDC-AIRS algorithm has shown promising results, clinching a place in the top 3 positions for all the datasets evaluated.

4.5. Discussion

We have developed an immune-inspired supervised classification algorithm called EDC-AIRS that have shown improved learning and classification capability. The success of the algorithm is primarily due to the recognition of the importance of additional immune metaphors, namely the ability to adapt to the different concentration, distribution and characteristics of the antigens. However, the EDC-AIRS algorithm did not achieve ameliorated performance for all classification problems investigated in this study (e.g. Pima Indian Diabetes dataset). This is not surprising as every learning algorithm has an inductive bias that would work reasonably well for some, but not all, datasets or application domains (Freitas & Timmis, 2007). This phenomenon has been described as the selective superiority problem (Brodley, 1993).

The AIRS2 parameters reported in (Watkins et al., 2004) has been tuned manually. This apparently hinders the true potential of the AIRS2 algorithm. As demonstrated, the employment of GA to optimize the AIRS2 parameters (i.e.

GA-AIRS2) improved the classification accuracy (ranging from 1.6% to 3.1% improvement) for all the 4 benchmarking datasets investigated. Clearly, this indicates that optimization of parameters with an evolutionary computing algorithm (e.g. GA) that is capable of dynamically searching through the defined search space is invaluable in discovering the optimal parameter setting. This is especially so when dealing with datasets from various application domains where the patterns that underlie these data would be very different, causing exhaustive manual tuning of the parameters to flounder as it would be very time consuming to carry out this task.

The EDC-AIRS algorithm, when juxtaposed with the AIRS2 algorithm, has several distinctive strengths when learning the underlying patterns within the data. Firstly, by adopting a mechanism to handle the different data density exhibited at different regions, it is capable of producing representative memory cells that could better characterize and capture the real data pattern. As a result, it is at an advantage when applied on datasets (such as Sonar dataset) that have data density which tends to fluctuate at different regions. Secondly, the EDC-AIRS algorithm is more capable at dealing with difference in distribution among data classes, generating representative memory cells for each data class. The ability to do so is important because it is unlikely for different data classes to have the same distribution and even more unlikely for a classifier to recognize and robustly adapt to such deviation without explicitly allowing for it. Efforts were therefore taken in this work to calculate the ATSR value and to determine whether to optimize a common or distinct parameter set. Ten datasets from diverse domains with different characteristics were used to evaluate the importance of implementing this technique. Results shown that for datasets with ATSR value lesser than 0.76 (e.g. Iris and Ionosphere datasets), it is more desirable to have a distinct parameter set for each data class. The need to compute the ATSR value and differentiate them into similar or dissimilar distribution is not an essential step but is advantageous to do so. This is because it is theoretically possible for GA to tune the parameter set meant for dissimilar distribution to one suitable for similar distribution. However, it is computationally intensive to do so. Therefore, by performing this simple step of differentiation, it can help to alleviate the complexity involved when tuning the parameters with GA. This complexity is introduced by the (linear) increase in the number of parameters that needs to be tuned, which in turn contributed to an exponential increase in the search space. This makes the task of discovering the optimal value for the parameters very challenging. This problem is commonly referred to as the ‘curse of dimensionality’ (Bellman, 1961).

Finally, the EDC-AIRS algorithm is capable of selecting features that are highly informative and relevant. This avoids some of the difficulties when dealing with datasets (e.g. Ionosphere and Sonar datasets) that have irrelevant or redundant features which often jeopardize the algorithm’s ability to learn and generalize. Moreover, it has the crucial advantage of identifying important features that best associate with an outcome, building a parsimonious classification model as a result. This property is highly desirable in accordance to the law of parsimony (Occam’s razor principle (Blumer et al., 1987)) where a simpler model with minimal complexity is preferred.

When EDC-AIRS algorithm was benchmarked with 4 datasets, promising results were obtained consistently. It outperforms AIRS2 algorithm in all the 4 cases. The increase in classification accuracy is 3.6%, 2.4%, 3.1% and 6% for Iris, Ionosphere, Pima Indians Diabetes and Sonar dataset respectively. This suggests that EDC-AIRS algorithm is a robust learner that is capable of adapting to different profound data patterns and structures.

4.6. Summary

Further inspired by the characteristics of the natural immune system, we have developed an adaptive and robust supervised classification algorithm called the EDC-AIRS algorithm. The performance of the proposed algorithm was evaluated with 6 benchmarking datasets. When ranked with other classifiers, the classification performance of EDC-AIRS algorithm is in the top 3 positions for all the datasets evaluated. Ameliorated performance achieved by the algorithm signifies the importance of empowering an algorithm with the ability to independently adapt to the distinct density, distribution and characteristics of each data class. However, this approach does not guarantee improved performance for all classification problems in face of the selective superiority problem.

Disclaimer

Figure 4.2 was adopted from (Watkins et al., 2004) with kind permission from Springer Science and Business Media.

Chapter 5

Time-Related Risk Prediction Models[5]

Myocardial infarction (MI) is one of the leading causes of death in many developed countries. Hence, early detection of MI events is critical for effective preventative therapies. One approach for early disease prediction is the use of prediction models developed using machine learning techniques. These models, we hypothesize, could be better achieved through detailed consideration of (1) sample age of clinical data amassed from routine medical examination, and (2) prediction resolution (i.e. prediction scales and intervals) used. In this chapter, we investigated on the effects of the aforementioned 2 factors on the performance of MI risk prediction models developed using Support Vector Machine (SVM) and Evolutionary Data-Conscious Artificial Immune Recognition System (EDC-AIRS) algorithms. The cardiovascular health study (CHS) dataset was used in this study. Results indicate that SVM algorithm is capable of achieving high sensitivity, specificity and balanced accuracy of 95.3%, 84.8% and 90.1% respectively over a time span of 6 years. Further, both sample age and prediction resolution were found not to have a significant impact on the performance of MI risk prediction models developed using subjects aged 65 and above. This implies that risk prediction models developed using different sample age and prediction resolution is a feasible approach and could offer patients with a more comprehensive estimation of their health risk.

[5] The work presented in this chapter has been accepted by IEEE Transactions on Biomedical and Health Informatics. This paper can be found in Appendix E.

5.1. Introduction

The best practice to avoid human mortality caused by life threatening diseases like myocardial infarction (MI) is to detect them early and prevent its onset. One approach is to devise computational methods that capitalize on clinical biomarkers to better screen the possible risk of (future) MI so that the most effective, personalized and preventive measures can be offered promptly. This ultimately would result in a reduction in avoidable mortality. However, the development of reliable and accurate clinical risk prediction models for MI remains a challenge.

The current approaches for assessing the risk of individuals experiencing MI include risk scoring system and survival curves (Clayton et al., 2005; Lloyd-Jones et al., 2004; Levy et al., 2006). These, however, have limitations like the inability to substantially identify minority of individuals with subsequent risk of experiencing MI (Alty et al., 2007). Moreover, clinical biomarkers and symptoms seldom follow a linear relationship and the expected outcome at individual level does not always abide to the rules of epidemiology (Chattopadhyay, 2013). As a result, conventional risk scoring systems – which model relationships in a linear manner - often flounder in view of these challenges (Song et al., 2004; Kim et al., 2005).

In recent years, there is an exponential increase in the amount of clinical and molecular data collected from routine medical examination. To overcome the challenges associated with human scale of thinking and analysis, data mining techniques – which have been postulated as a "central feature" for future healthcare system (Snyderman & Langheier, 2006) – became a popular method for extracting insights from this data deluge. Advantages of using data mining techniques include the capability of dealing with plethora of information, solving non-trivial problems, producing data-driven prediction models, and handling non-linear relationships among biomarkers. Examples of data mining techniques used to estimate disease risk include works from: (1) Wiens et al. (Wiens et al., 2012) who employed support vector machine (SVM) to identify patients who are at high risk of experiencing hospital acquired Clostridium difficile (C. diff); and (2) Khan et al. (Khan et al., 2001) who used artificial neural network (ANN) for discriminating small, round blue-cell tumors (SRBCTs).

Investigation from (Asia Pacific Cohort Studies Collaboration, 2006) suggests that differences in severity of cardiovascular disease (CVD) risk factors could contribute to age-related excess risk for CVD – i.e. the impact of a risk factor on one's health could change as one ages. These, from the perspective of preventive medicine and clinical risk prediction, motivated us to hypothesize the importance of sample age and prediction resolution – 2 aspects that are not commonly examined in the literature – in relation to clinical risk prediction models. Here, sample age refers to the average age of individuals found in the baseline (i.e. input) dataset used to construct the clinical risk prediction model while prediction resolution refers to the prediction scale (i.e. number of years into the future where prediction of MI occurrence begins) and interval (i.e. time duration, in years, that marks the start and end of MI outcomes to be considered) employed by the clinical risk prediction model.

This chapter presents the development of MI risk prediction models constructed using Support Vector Machine (SVM) (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1999) and Evolutionary Data-Conscious Artificial Immune Recognition System (EDC-AIRS) (Tay et al., 2013) algorithms. Additionally, the effects of sample age and prediction resolution (using subjects aged 65 and above) on the performance of the developed models were examined. Participants amassed from the Cardiovascular Health Study (CHS) (Fried et al., 1991) were analysed. We have chosen CHS dataset in this work because of the wide range of clinical measurements and risk factors accrued during the CHS observational study.

The rest of the chapter is organized as follows. Section 5.2 provides details of CHS dataset, and delineates the methodology involved in developing the predictive models. Section 5.3 provides the experimental results achieved by the risk prediction models developed using different combinations of sample age and prediction resolution. Key results are discussed in Section 5.4 and conclusions are drawn in Section 5.5.

5.2. Material and Methods

In Section 5.2.1, details of CHS dataset are provided. This dataset, however, consists of a significant percentage of missing data and a highly skewed data distribution (commonly known as the class imbalanced data problem). Hence, for effective analysis, data imputation and class data balancing are performed and described in Section 5.2.2 and 5.2.3 respectively. Section 5.2.4 explains how the various MI risk prediction models based on different combinations of baseline data and prediction resolution were developed and validated.

5.2.1. Dataset

The CHS dataset (Fried et al., 1991), an observational study of cardiovascular risk factors associated with the elderly, was analysed. Further details of this dataset can be found in Section 3.3.1.

5.2.2. Data Imputation

Data imputation is the process of substituting missing entries in a dataset with plausible values and aims to improve the quality of the data. It was performed using weighted K-nearest neighbour (KNN) because of its excellent performance in estimating missing values (Troyanskaya et al., 2001; Jerez et al., 2010). Moreover, it has the capability to estimate both qualitative and quantitative attributes. Hence, it is highly suitable for interpolating the missing values in the CHS dataset.

Individuals with unknown MI status and clinical features that were uninformative (i.e. features with consistent value throughout) were first removed from the analysis. Individuals and clinical features with high percentage of missing entries were also removed. This is to ensure that there is an adequate supply of complete entries for weighted KNN to reference when estimating the missing values, which in turn promotes a more accurate data imputation process (Garcia-Laencina et al., 2008; Troyanskaya et al., 2001; Jerez et al., 2010). The resulting dataset was next normalized to unit variance to ensure that the attributes with large scale do not dominate the (Euclidean) distance measure (Minaei-Bidgoli et al., 2003). Subsequently, the optimal value of K for each clinical feature was determined by 10-fold cross-validation and used for the data imputation process. The type of replacement method used by weighted KNN depends on the data type. For instance, if categorical (continuous) data were encountered, the weighted-mode (weighted-mean) of the K nearest neighbours was used to assign the value for the missing entries. The use of weighted KNN estimation has been demonstrated in (Dudani, 1976; Troyanskaya et al., 2001) to be robust and accurate.

5.2.3. Class Imbalanced Data Problem

In order to create an unbiased dataset for SVM and EDC-AIRS algorithms to learn from, under-sampling of the majority class is necessary. The Kennard-Stone (KS) algorithm (Kennard & Stone, 1969) was employed to perform this task because of its excellent performance as demonstrated in a comparative study (Wu et al., 1996). This algorithm sequentially selects representative data that are uniformly scattered across the data domain space. This is carried out by first selecting a data object that is closest to the mean of the dataset and is included as the first data candidate. Subsequently, the data object that is most distant from the first one (based on Euclidean distance) is included as the second data candidate. The next data object is chosen by identifying the one farthest away from the previously selected data candidates. This process repeats until the desired number of candidates has been identified (Wu et al., 1996; Shahlaeiab et al., 2012).

In this study, the KS algorithm was used to under-sample the majority class found in the imputed CHS dataset. The number of candidates to select is equivalent to the number of samples in the minority class. In other words, after this process, the number of controls and cases would be identical.

5.2.4. MI Risk Prediction Models

Risk prediction of MI events is a highly alluring task as it would allow early detection and better management of the disease, and ultimately improve the individuals' quality of life. To develop such risk prediction models, 2 algorithms (SVM and EDC-AIRS) were employed in this study. SVM algorithm is a robust supervised learning algorithm that is capable of yielding excellent generalization performance on an extensive area of problems (Chen et al., 2005; Osuna et al., 1997; Listgarten et al., 2004). It is derived from statistical learning theory and is capable of solving linearly and non-linearly separable problems. Fundamentally, SVM performs classification through the construction of an N-dimensional hyper-plane that optimally separates the data into two or more categories whereby the margin of separation between the different categories is maximized.

EDC-AIRS algorithm (Tay et al., 2013) is a supervised classification algorithm inspired by the principles and processes associated with the human immune system. Adscititious to the typical mechanisms adopted by artificial immune system – like clonal expansion, somatic hyper-mutation, resource competition and memory cell formation – EDC-AIRS algorithm proposed strategies for robustly adapting memory cells to the different density,

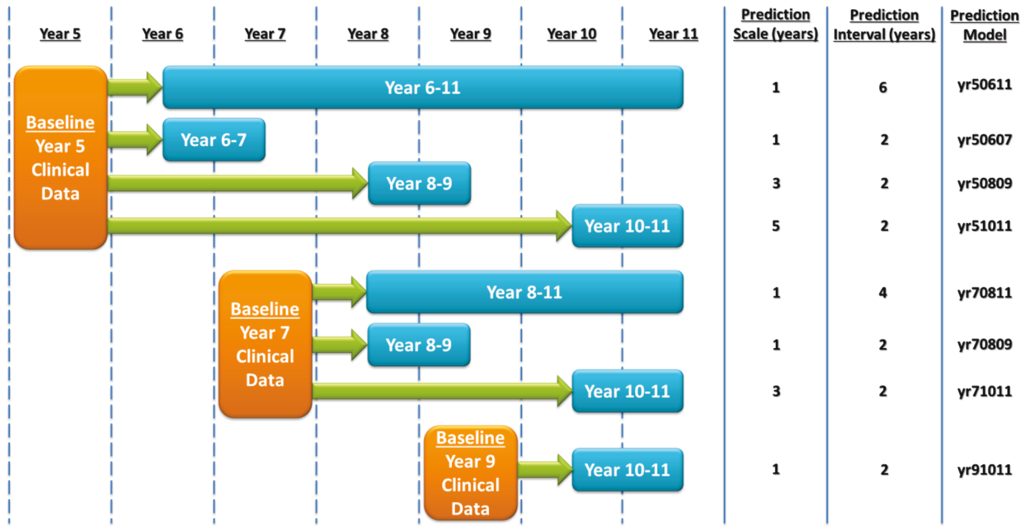


Figure 5.1: MI Risk Prediction of Various Prediction Scale and Interval
MI risk prediction at various time scale and interval using the CHS dataset was performed. Prediction scale refers to the number of years into the future where prediction of MI occurrence begins while prediction interval refers to the time duration (in years) that marks the start and end of MI outcomes to be considered.

distribution and characteristics exhibited by each data class. This algorithm, when tested on several widely benchmarked datasets, has demonstrated highly competitive classification performance (Tay et al., 2013). To adopt a ceteris paribus experimental design, the parameters for both algorithms were first tuned using Genetic Algorithm (GA) and subsequently, feature selection was conducted (using GA) to identify predictive biomarkers.

Clinical data - recorded during the 5th to 11th year in which the CHS clinical study was undertaken - were utilized. The reason for using clinical data recorded from year 5 onwards was because clinical examinations taken by the two different cohorts recruited at different phases synchronized from that year onward. The reason for ending the prediction at year 11 is because from year 12 onwards, participants were only monitored annually via phone calls and no clinical examinations were conducted.

		EDC-AIRS	
		Misclassification	Correct Classification
SVM	Misclassification	**a**	**b**
	Correct Classification	**c**	**d**

Figure 5.2: Contingency Table for McNemar's Test
'a' indicates the number of data items misclassified by both SVM and EDC-AIRS algorithms; 'b' represents the number of data items misclassified by SVM algorithm but correctly classified by EDC-AIRS algorithm; 'c' denotes the number of data items misclassified by EDC-AIRS algorithm but correctly classified by SVM algorithm; 'd' dictates the number of data items correctly classified by both SVM and EDC-AIRS algorithms.

To test the hypothesis, prediction models - using different baseline datasets (with different sample age) - capable of predicting the risk of experiencing MI at various prediction scales and intervals were developed. As illustrated in Figure 5.1, 8 different prediction models were designed to investigate how time factor in relation to the onset of MI would affect the performance of the prediction model. Three different baseline datasets were used. These datasets contain clinical examination results recorded in year 5, year 7 and year 9 of the CHS study. Each of these datasets was used to predict whether an individual would experience MI in the near future. Three different prediction scales (1, 3 and 5 years) and 3 different prediction intervals (2, 4 and 6 years) were investigated. Specifically, healthy individuals present in year 5 of the CHS dataset were used as the baseline data to predict whether one would experience MI from year 6 to 11 (prediction scale: 1 year; prediction interval: 6 years), year 6 to 7 (prediction scale: 1 year; prediction interval: 2 years), year 8 to 9 (prediction scale: 3 years; prediction interval: 2 years) and year 10 to 11 (prediction scale: 5 years; prediction interval: 2 years). Similarly, clinical examination results of healthy participants in year 7 was initialized as the baseline data, where prediction of whether one would suffer from MI from year

Table 5.1: Details of the Imputed CHS Dataset

Prediction Model	Sample Size* (cases/controls)	#Features	Age (Mean±SD)
yr50611	3102 (6.2%/93.8%)	237	75.7 ± 5.34
yr50607	3102 (2.4%/97.6%)	237	75.7 ± 5.34
yr50809	3034 (2.1%/97.9%)	237	75.7 ± 5.34
yr51011	2978 (2.1%/97.9%)	237	75.7 ± 5.36
yr70811	2407 (2.1%/97.9%)	233	77.2 ± 5.40
yr70809	2407 (2.1%/97.9%)	233	77.2 ± 5.40
yr71011	2362 (2.0%/98.0%)	233	77.2 ± 5.40
yr91011	1909 (1.9%/98.1%)	242	78.8 ± 5.09

*This sample size refers to the number of individuals that remain in the CHS dataset after removal of records with significant missing entries.
'yrXYYZZ' denotes that the prediction model uses clinical measurements observed in year X to make prediction of whether one would experience MI from year YY to ZZ.

8 to 11, year 8 to 9 and year 10 to 11 were conducted. Likewise, clinical data recorded in year 9 was utilized to perform prediction of MI occurrence from year 10 to 11.

Each baseline dataset was randomly split into two subsets having balanced class distribution. The first subset contains 70% of the initial data. Using this subset, the prediction model was trained and tuned based on 10-fold cross-validation. The second subset, which contains the remaining 30% of the data, was used to validate the developed model. This splitting process was repeated 3 times and independently used to develop and test the respective prediction model. It is highly encouraged to do so to avoid the developed model from capturing not only the true associations, but, also, idiosyncratic features of the training data, which often produces an overly optimistic model (Taylor et al., 2008). When developing and testing each prediction model, each algorithm is executed 3 times. Three commonly used performance measurements were employed to evaluate the prediction models developed - namely sensitivity, specificity, and balanced accuracy (i.e. average between sensitivity and specificity).

Table 5.2: Details of Datasets Used to Build the Prediction Models

Prediction Model	#Training Instances	#Validation Instances
yr50611	270	114
yr50607	104	42
yr50809	92	38
yr51011	88	36
yr70811	136	58
yr70809	70	30
yr71011	66	28
yr91011	52	20

All training and testing datasets contain equal number of cases and controls.

Finally, to determine whether the prediction models developed using SVM and EDC-AIRS algorithms are statistically different from each other, McNemar's test was conducted. This statistical test was chosen as it has been demonstrated to have low type 1 error (Dietterich, 1998). For each prediction model, this test was carried out by first recording the prediction outcomes obtained (by each algorithm) when tested using each validation dataset. The results obtained from each algorithm were then used to construct the contingency table shown in Figure 5.2. Referring to the figure, if the sum of 'b' and 'c' is greater than 25, chi-square test with 1 degree of freedom is used for performing McNemar's test. Otherwise, to provide a better estimation of the small sample (i.e. $b + c \leq 25$), binomial distribution is used for (exact) McNemar's test. The prediction model is considered to be statistically different from the ground truth if the p-value computed using McNemar's test is smaller than 0.05.

Table 5.3: Classification Performance of SVM and EDC-AIRS Algorithms (Cross-Validated)

Prediction Model	SVM				EDC-AIRS			
	#Features Selected	Sensitivity	Specificity	Balanced Accuracy	#Features Selected	Sensitivity	Specificity	Balanced Accuracy
yr50611	130±18	0.943±0.015	0.975±0.015	0.959±0.006	113±9	0.896±0.032	0.731±0.088	0.814±0.030
yr50607	120±17	0.974±0.011	0.974±0.011	0.974±0.011	107±10	0.974±0.011	0.750±0.084	0.862±0.045
yr50809	109±2	1.000±0	0.978±0	0.989±0	103±8	0.971±0.013	0.899±0.070	0.935±0.029
yr51011	111±4	1.000±0	0.970±0.013	0.985±0.007	112±8	1.000±0	0.720±0.164	0.860±0.082
yr70811	109±16	0.900±0.136	0.967±0.016	0.951±0.040	109±9	0.966±0.022	0.838±0.039	0.902±0.022
yr70809	105±3	1.000±0	0.990±0.016	0.995±0.008	103±9	0.981±0.033	0.848±0.044	0.914±0.014
yr71011	104±4	1.000±0	0.970±0	0.985±0	107±10	0.980±0.017	0.828±0.046	0.904±0.032
yr91011	111±2	0.962±0	0.962±0	0.962±0	114±6	0.962±0	0.872±0.059	0.917±0.029

The number of feature selected refers to the number of biomarkers identified by GA as predictive towards the prediction of MI. All performance measurements range between 0 and 1.

5.3. Experimental Results

5.3.1. Data Preprocessing

Efforts were taken to ensure the quality of the data. Firstly, removal of records and clinical features with significant missing entries were performed. Table 5.1 presents the details of the resulting CHS datasets.

Subsequently, these datasets went through the weighted KNN data imputation process where missing values were estimated. Finally, under-

Table 5.4: Classification Performance of SVM and EDC-AIRS Algorithms (Tested with Validation Dataset)

Prediction Model	SVM			EDC-AIRS		
	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy
yr50611	0.953±0.037	0.848±0.054	0.901±0.022	0.924±0.020	0.649±0.110	0.786±0.058
yr50607	0.921±0.055	0.873±0.055	0.897±0.055	0.841±0.099	0.587±0.0550	0.714±0.024
yr50809	0.947±0	0.772±0.030	0.860±0.015	0.772±0.219	0.667±0.122	0.719±0.080
yr51011	0.944±0	0.852±0.085	0.898±0.042	0.926±0.064	0.574±0.032	0.750±0.048
yr70811	0.747±0.293	0.874±0.053	0.810±0.121	0.828±0.120	0.713±0.173	0.770±0.040
yr70809	0.844±0.102	0.800±0	0.822±0.051	0.867±0	0.556±0.077	0.711±0.038
yr71011	0.905±0.041	0.857±0.124	0.881±0.055	0.833±0.109	0.595±0.149	0.714±0.036
yr91011	0.967±0.058	0.700±0.173	0.833±0.115	0.967±0.058	0.467±0.252	0.717±0.126

These performance measurements were obtained by evaluating each developed prediction model with their respective test dataset.

Table 5.5: Statistical Evaluation of Developed Prediction Models

Prediction Model	McNemar's Test# (p-value) SVM vs EDC-AIRS
yr50611	**<0.0001**
yr50607	**0.0001**
yr50809	**0.0052**
yr51011	**0.0009**
yr70811	0.3072
yr70809	**0.0414**
yr71011	**0.0013**
yr91011	0.0654

#The p-value of McNemar's test is presented examining whether the performance of the developed prediction model is statistically different from the ground truth.

sampling was performed with the KS algorithm to obtain a balanced number of cases and controls.

These data preprocessing steps taken affected the overall size of the dataset used to build each prediction model. Details of the resultant datasets are summarized in Table 5.2. The training datasets were used to develop the prediction models while the validation datasets were used to evaluate the robustness of the developed models.

5.3.2. MI Risk Prediction Models

Prediction models - using baseline dataset with different sample age - at various time scales and intervals were developed using the training datasets. Cross-validation was carried out to evaluate the performance of each prediction model. For all prediction models developed, results (as shown in Table 5.3) indicate consistently high predictive performance was achieved by both SVM and EDC-AIRS algorithms. For example, a balanced accuracy of at least 0.95 and 0.81 was achieved by SVM and EDC-AIRS algorithms respectively.

To assess whether the prediction models developed generalize well, validation was performed using the validation datasets. Results, as presented in

Table 5.4, demonstrate that a balanced accuracy of at least 0.81 and 0.71 was achieved by SVM and EDC-AIRS algorithms respectively.

McNemar's test was conducted to determine whether the performance of SVM and EDC-AIRS algorithms are statistically different from each other. Results (as shown in Table 5.5) indicate that for most of the prediction models (i.e. except prediction models 'yr70811' and 'yr91011'), the performance of SVM and EDC-AIRS algorithms are statistically different.

5.4. Discussion

MI risk prediction models developed using baseline datasets with different sample age, and based on different prediction resolution combinations were analysed. Cross-validation was utilized during the training phase as an approach to evaluate and develop potent MI risk prediction models. The resultant prediction models developed by both algorithms achieved a relatively high sensitivity, specificity and balanced accuracy (for SVM algorithm, the respective performance achieved is at least 0.90, 0.96 and 0.95; while for EDC-AIRS algorithm, the respective performance achieved is at least 0.89, 0.72 and 0.81). Investigation on whether the prediction models developed were over-trained was conducted by validating each developed model with an unseen dataset (i.e. not used to develop the prediction model). The aim of this step was to assess the generalizability of the developed models. Results indicate that SVM algorithm (and EDC-AIRS algorithm) – across all prediction models tested - achieved a sensitivity, specificity and balanced accuracy of at least 0.74, 0.70 and 0.81 (and 0.77, 0.46 and 0.71) respectively. Furthermore, it can be observed that in general there is a drop in the validation sensitivity (SVM: 0.071±0.059; EDC-AIRS: 0.097±0.078), specificity (SVM: 0.151±0.061; EDC-AIRS: 0.210±0.104) and balanced accuracy (SVM: 0.112±0.038; EDC-AIRS: 0.153±0.063) among all the prediction models developed. It is noteworthy that the drop in performance is less severe for SVM algorithm (when compared to EDC-AIRS algorithm). This portends that SVM algorithm tends to perform better on noisy data (in contrast to EDC-AIRS algorithm) even after data imputation was conducted. This observation is supported by the results obtained from the performance of McNemar's test. From this statistical evaluation, it was demonstrated that SVM algorithm outperforms EDC-AIRS algorithm for 6 out of 8 prediction models tested.

Prediction models developed (with SVM algorithm) using baseline dataset from year 5 (and year 7), and tested using their respective validation datasets have shown comparable sensitivity, specificity and balanced accuracy. Analysis of variance (ANOVA) test was conducted on the respective group of prediction models (i.e. developed using either year 5 or 7 as baseline dataset) that has a prediction interval of 2 years. Results demonstrate that they are statistically comparable - with p-value of 0.473 for prediction models using baseline dataset from year 5 (and 0.245 for prediction models using baseline dataset from year 7). This signifies that predication scale does not have a significant impact on the performance of (SVM-based) prediction models developed and tested using subjects aged 65 and above. Similar analysis was performed on prediction models developed based on different prediction interval. Results indicate that these models are statistically comparable – with p-value of 0.918 and 0.883 for prediction models developed using baseline dataset from year 5 and 7 respectively. This means that prediction interval does not have a significant impact on the performance of prediction models developed using SVM algorithm.

As for prediction models developed using EDC-AIRS algorithm, similar analysis was conducted. For prediction models developed using baseline dataset from year 5 (and year 7) that are based on 2-year prediction interval, and tested using their respective validation datasets, ANOVA test was conducted. Results indicate that the prediction models in their respective group are statistically comparable – having a p-value of 0.712 (for prediction model using year 5 baseline dataset) and 0.926 (for prediction model using year 7 baseline dataset).

This indicates that predication scale does not have a significant impact on prediction models developed using EDC-AIRS algorithm as well. Likewise, prediction models developed based on different prediction interval were analysed. Results show that these models are statistically comparable – having a p-value of 0.118 and 0.139 for prediction models developed using baseline dataset from year 5 and 7 respectively. This suggests that prediction interval does not have a significant impact on the performance of prediction models developed using EDC-AIRS algorithm as well. In view of these observations, we aim to investigate the effects of prediction resolution on subjects in younger age groups as part of our future work. A summary of the p-values discussed is provided in Table 5.6.

Analysis of prediction models that aim to predict the likelihood of MI occurrence in individuals' subsequent 2 years (i.e. 'yr50607', 'yr70809' and 'yr91011') indicate comparable performance – with p-value of 0.504 and 0.996 for SVM and EDC-AIRS algorithms respectively. Comparison of age among individuals belonging to different baseline datasets indicates that they are statistically different (p-value < 0.0001). This portends that sample age does not have a significant impact on the performance of prediction models.

Table 5.6: Statistical Evaluation of Prediction Resolution

Prediction Models Compared	ANOVA Test# (p-value)	
	SVM	EDC-AIRS
Prediction Scale		
yr50607; yr50809; yr51011	0.473	0.712
yr70809; yr71011	0.245	0.926
Prediction Interval		
yr50611; yr50607	0.918	0.118
yr70811; yr70809;	0.883	0.139

#The p-value of ANOVA test is presented examining the significance of prediction scale and interval for both SVM and EDC-AIRS algorithms.

One benefit of performing risk prediction using different prediction resolution and sample age is that it allows more refined and progressive risk prediction to be conducted (without compromising accuracy). This provides the advantage of estimating the seriousness of a disease one is experiencing; enabling clinicians to offer a more personalized management and/or therapeutic strategy to the patient.

The limitation of this investigation includes the use of a single dataset to evaluate the effects of sample age and prediction resolution in relation to the performance of MI risk prediction. This limits the power to conclusively state how each factor influences the performance of the prediction model. Nevertheless, it does provide some insights on whether sample age and prediction resolution have an impact on the performance of clinical risk prediction model.

5.5. Summary

Early detection of individuals with high risk of experiencing MI through the use of prediction models that are simple to use and provide instant prediction has been a coveted and elusive clinical task. To this end, we investigated on the effects of sample age and prediction resolution in relation to the development of accurate clinical risk prediction model. Our experiments indicate that both sample age and prediction resolution do not have a significant impact on prediction models developed using subjects aged 65 and above. In view of this observation, the decision of which combination of sample age and prediction resolution to use in clinical practice – in our opinion – would depend on the availability of appropriate treatment/management plans for the patients. This is very important as we do not want to burden the patients with unnecessary emotional stress (which might implicitly exacerbate their health) if we do not have a solution for them. Such consideration is critical in order to provide high quality biological, psychological and sociological care for the patients.

Overall, high validation sensitivity, specificity and balanced accuracy were achieved by SVM algorithm. This opens the opportunity for constructing predictive models capable of detecting MI early, allowing clinicians to take preventative measures promptly, improving the quality of individuals' life, and reducing avoidable mortality.

In view of these results, we suggest the use of different prediction resolution to provide a more detailed health screening of elderly subjects so that more appropriate preventative measurements - in relation to the individual's risk level - can be taken.

Disclaimer

The CHS dataset described in this chapter is provided by the NHLBI.

Chapter 6

Artificial Neural Cell System for Classification[6]

The human brain has always been looked upon with great interests and studied by many researchers from multiple disciplines. It has been considered as the organ responsible for functions like information processing/storage and recall, decision making and initiating actions on external environment. The mechanisms that develop the brain and empowering it with such capabilities have strong similarity to machine learning and classification, and have inspired us to develop a learning algorithm for problem solving and optimization.

Exploiting on 3 natural mechanisms responsible for developing and enriching the brain (i.e. neurogenesis, neuroplasticity via nurturing and apoptosis), we introduce a novel learning algorithm for solving classification problems. We call this new neural-inspired classification algorithm as the Artificial Neural Cell System for classification (ANCSc). Benchmark testing on ANCSc algorithm was conducted and highly competitive classification results were achieved. Through this work, we aim to suggest new approaches that might be of value to the construction of learning systems.

[6] The work presented in this chapter has been submitted to Journal of Biomedical Informatics and is currently under review.

6.1. Introduction

This chapter introduces a novel supervised learning algorithm for solving classification problems – one of the most common and well-studied tasks in predictive data mining and knowledge discovery with a wide range of applications such as medical diagnosis, genomic analysis, pattern recognition, digital security, among others. It is inspired by the characteristics exhibited by 3 natural phenomena responsible for developing and enriching brain function - namely (1) neurogenesis, (2) neuroplasticity as a result of the dynamic interplay between nature and nurture, and (3) apoptosis. These mechanisms (among others) enable human to learn, identify, differentiate and organize objects, patterns, sounds, concepts, etc. This model of neural operations has many features in common, generally in the field of machine learning, to the task of classification – the problem of identifying which category an observation belongs to, on the basis of a pre-specified set of data containing observations with known category membership. Hence, these neural processes - which to our knowledge have not been exploited for the development of machine learning algorithms - become an ideal candidate for the study and modeling of learning systems.

Neurogenesis, in neuroscience, is the process by which new neurons are generated in the nervous system from neural stem/progenitor cells (Wiskott et al., 2006). The generated neurons are not stagnant throughout the life of a species and can be stimulated by behavioral and environmental factors (Lillard & Erisir, 2011). This is vital and necessary for adapting the brain to any changing elements it encounters; refining the neural pathways and synapses essential for learning and adapting to changes, and circumvent any undesirable side effects. This process of molding and reshaping the brain in face of changes in behavior, environment and neural processes is often referred to as neuroplasticity (Taupin, 2006). Intentional exposure to new environments and (supervised/guided) inculcation of desirable information/behavior to a human (e.g. taught by an instructor) may trigger neuroplastic changes as well. This process, considered as nurturing, capitalizes on what the nature can provide (i.e.

the individuals' innate qualities), enriches and leverages on the individuals' ability so that they can perform at their greatest potential.

Motivated by the profound significance of the aforementioned mechanisms in human learning process, and the ability to autonomously trim off non-essential cells during human development (commonly known as apoptosis), we introduce a novel supervised classification called the Artificial Neural Cell System for classification (ANCSc). In a nutshell, ANCSc algorithm bio-mimics the mechanisms underlying the neuronal behavior associated with the process of learning and interaction with the external environment. It allows artificial neurons (i.e. candidate solutions) to (1) proliferate in the solution space (i.e. bio-mimicking neurogenesis), (2) progressively and independently refine and adapt to the (data) environment presented (i.e. bio-mimicking neuroplasticity as a result of nurturing), and (3) survive or undergo programmed cell death as part of an effort to construct a concise and efficacious classification model (i.e. bio-mimicking apoptosis). The utilization of these learning mechanisms is a novel contribution towards the development of neural-inspired learning algorithms, and in our opinion, would promote the development of robust classification models that are less complex to design; for example, the ANCSc algorithm, in contrast to artificial neural network (ANN), does not require the network architecture (i.e. number of neurons and layers) to be defined.

The classification performance of the ANCSc algorithm, when evaluated with 6 benchmark datasets, demonstrates that it is a robust learning algorithm capable of achieving highly competitive classification results. This novel learning method is an important contribution as the capability to better learn profound data structures and make accurate prediction of new observations are beneficial for many classification problems.

The rest of the chapter is organized as follows. Section 6.2 provides a brief overview of neural processes. A detailed description of the proposed ANCSc algorithm is presented in Section 6.3. Materials and methods used in this study are delineated in Section 6.4. Performance of ANCSc algorithm and its corresponding sensitivity analysis are offered in Section 6.5. Section 6.6 discusses the key results and properties associated with the algorithm. Finally, conclusions are drawn in Section 6.7.

6.2. Overview of Neural Processes

Neurons, a group of specialized impulse-conducting cells that process and transmit information through electrical and chemical signals, form the core components of the nervous system (e.g. the brain). The human brain contains on average 86.1 billion neurons (Azevedo et al., 2009), connected to each other to form neural networks. Communication among the neurons occurs via synapses – specialized connections between neurons that allow electrical and chemical signals to be transmitted. This interaction among neurons is the cellular basis for tasks like thinking and decision making. In particular, neurons are interconnected in smaller groups – called neuronal pools – defined on the basis of function (i.e. each neuronal pool is responsible for enabling a specific function to be carried out) (Martini et al., 2011).

New neurons are generated in the human brain from neural stem/progenitor cells – a process called neurogenesis. It is most active during prenatal development and declines sharply over the adolescence period (Wiskott et al., 2006). Neurogenesis in the adult brain occurs primarily in two discrete areas - namely the dentate gyrus of the hippocampus and the subventricular zone, along the lateral ventricles. The number of new neurons added to an adult brain is dependent on the rate of cell generation and the probability of cell survival (i.e. generated cells might undergo programmed cell death after a period of time – a phenomenon known as apoptosis) (Wiskott et al., 2006). As demonstrated in several studies, the rate at which neurogenesis occurs is modulated by several intrinsic and environmental stimuli. Intrinsic regulators include age (Kuhn et al., 1996), gender (Tanapat et al., 1999) and genetic factors (Kempermann et al., 1997) while environmental stimuli comprise of environmental enrichment (Nilsson et al., 1999), physical (Praag et al., 1999) and social (Fowler et al.,

2002) activities, stress (Gould & Tanapat, 1999), smell (Tanapat et al., 2001) and diet (Stangl & Thuret, 2009). It is noteworthy that adult neurogenesis, in any cases, occurs (during most part of the life) at a very low rate (Wiskott et al., 2006; Taupin, 2006). Further, there is also growing evidence suggesting an association between adult hippocampal neurogenesis to several processes like neuro-inflammation, learning and memory. It has been demonstrated that neuro-inflammation inhibits neurogenesis in adult hippocampus (Ekdahl et al., 2003) while increased hippocampal neurogenesis is potentially involved in ameliorated learning and memory (Neves et al., 2008; Gould et al., 1999; Shors et al., 2001). Long-surviving neurons in the brain have been postulated to be more stable and preserve the encoding of the learned environment, whereas newly generated neurons are more plastic – which allows the brain to adapt itself to the new environment (i.e. occurrence of neuroplasticity as a consequence of learning) (Wiskott et al., 2006).

In the parlance of literature, neuroplasticity refers to the malleability of the brain - usually observable as changes in neuronal structure (e.g. changes in the position of the neurons) and connectivity, functional changes in the brain and neurogenesis. This typically occurs as a result of learning (e.g. taught/nurtured by an instructor), training (e.g. practicing to improve the ability to perform a task) and experience (e.g. exposure to certain event or environment), rendering the brain capable of adapting to environmental dynamics (Taupin, 2006). It is noteworthy that it has become increasingly evident that both neurogenesis and neuroplasticity occur in the human brain throughout life; instead of during prenatal development or juvenile period only (Gage, 2002; Lillard & Erisir, 2011).

Apoptosis, the process of controlled cell death, is an important feature that offers significant advantages during an organism's lifecycle. It promotes healthy (e.g. nervous system) development where defective apoptotic processes would be detrimental – leading to diseases like cancer (as a result of inadequate apoptosis) or atrophy (as a consequence of excesses apoptosis).

6.3. Artificial Neural Cell System for Classification (ANCSc) Algorithm

ANCSc, a novel neural-inspired learning algorithm, will be presented in this section. It is a supervised classification algorithm that bio-mimics how new neurons are populated, refined and maintained in the mammalian brain. Through this process, it aims to "educate" the ANCSc classification model (in an incremental manner) key patterns that underlie the training data.

To provide a comprehensive description of ANCSc algorithm, Section 6.3.1 describes the key terms and parameters vital for the understanding of the algorithm while Section 6.3.2 provides a tour of the training routine associated with the algorithm.

6.3.1. Key Concept and Parameters

This subsection describes the definitions for the key terms and parameters used in relation to the ANCSc algorithm.

Key Terms

- **Affinity**: The Euclidean distance between two neurons (feature vectors). In this implementation, this distance is between 0 and 1 (where 0 represents high affinity while 1 indicates low affinity).

- **Apoptosis**: The removal of neurons from the artificial cognitive system that mimics the naturally occurring and genetically determined process of self-destruction of unwanted cells. It is a regulated process that offers the advantage of producing a parsimonious yet accurate artificial cognitive system for performing classification at the end of the training routine.

- **Artificial Cognitive System**: A collection of representative neurons (which evolve during the training process of ANCSc) capable of describing the training data presented. Given that communication among neurons (e.g.

within a neuronal pool) enables human to think or recognize objects, we propose the use of KNN algorithm (Cover & Hart, 1967) to perform classification (at the end of each training cycle) due to their metaphorical similarity – i.e. both defines a pool of elements for conducting a task of interest.

- **Artificial Neuron**: In neuroscience, neuroplastic changes (i.e. slight changes in the position of the neurons) have been proposed as the consequence of learning and memory formation in species like human (Gage, 2002). To bio-mimic this phenomenon, we propose the artificial neurons developed in ANCSc algorithm as feature vectors (with its associated class) that contribute to the formation of the artificial cognitive system. Synaptic connections between neurons are not considered in order to simplify the construction of the learning model. Artificial neurons can be added, modified or removed from the postulated artificial cognitive system during ANCSc training cycle.

- **Artificial Neuronal Pool**: A group of proximal neurons that describe a specific pattern determined within the data problem presented. Its formation is regulated by the associated classification performance and defined on the basis of cell proliferation, adaptation and survival. The size of the artificial neuronal pool determines the number of neurons (i.e. k value) to be used for classification by KNN.

- **Class**: The category assigned to a given feature vector. For binary classification problems, each feature vector is assigned to one of the 2 pre-defined categories.

- **Feature Vector**: An n-dimensional vector of categorical/numerical features that describe the characteristics of an object/observation.

- **Neuroplasticity**: The adaptation of neurons (i.e. modification of the feature vectors) in the artificial cognitive system triggered by the process of learning and generalization. This procedure aims to promote the generation of highly representative artificial neurons capable of describing the given data environment.

- **Testing Data**: A collection of data items, that represent observations/measurements of a subject of interests, used to estimate the performance of the classification model trained with the training data. It is a distinct set of data that is used in an iterative process to evaluate and improve the performance of the trained model.

- **Training Data**: A collection of data, similar to the testing data, used to develop a classification model. Training data are commonly used in various areas of information science for the discovery of predictive relationship between the feature vector and the class. In this particular context, they serve as the data environment that promotes proliferation, adaptation and survival of neural cells.

Key Parameters

- **Learning Plateau Threshold (LPT)**: A termination criterion which defines the number of learning cycles that the ANCSc algorithm would iterate for before termination. Improvement in classification accuracy (during a learning cycle) would reset this (integer) parameter.

- **Neural Density (ND)**: This value, which ranges between 0 and 1, aims to spread neurons with high affinity. This offers the potential advantage of generating a set of representative neurons.

- **Neurogenic Space (NS)**: A parameter, used during the prenatal development phase, which determines the size of the region at which artificial neurons would develop in the fetal artificial cognitive system. The value of this parameter ranges between 0 and 1.

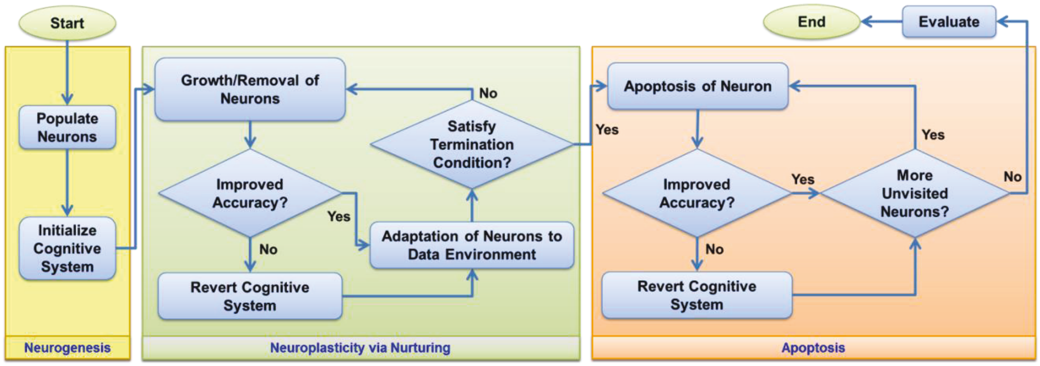


Figure 6.1: Canonical Flow of ANCSc Algorithm
The ANCSc algorithm consists of 3 key phases: Neurogenesis, neuroplasticity via nurturing and apoptosis. During the neurogenesis phase, the initial set of artificial neurons is created. These artificial neurons then evolve (through cell proliferation, adaptation and survival) in the subsequent 2 phase, generating a set of representative artificial neurons capable of describing the underlying patterns of the data presented.

- **Neurogenic Rate (NR)**: The rate at which neurons are generated during prenatal development phase. The value of this parameter ranges between 0 and 1.

- **Neuronal Pool Size (NPS)**: The number of neurons that should be used (by KNN classifier) to determine the classification of a given test data item. This integer value is used as the k value parameter required in KNN algorithm.

- **Neuroplastic Coefficient (NPC)**: This parameter specifies the degree to which the generated neurons migrate in the artificial cognitive system. This offers an opportunity for the neurons to generalize and circumvent situation like overfitting (i.e. modelling the idiosyncratic features of the data under study, which often results in poor classification performance on unseen data). The value of this parameter ranges between 0 and 1.

Algorithm 6.1: Overview of ANCSc Algorithm

Input: **D** (training data)
T (testing data)
Output: **O** (class label prediction)

Initialization

Step 1: Set t = 1. Normalize **D** and **T** to the range [0,1].

Neurogenesis Phase

Step 2: Populate a pool of artificial neurons $\mathbf{P_1}$ to form the initial cognitive system **C**.
$\mathbf{P_1}$ is generated by searching for representative data items in **D**, $\mathbf{P_1} \subseteq \mathbf{D}$.

Step 3: A_1 = accuracy of classification model $\mathbf{P_1}$ when evaluated with **D**.
Set t = t + 1.

Neuroplasticity via nurturing Phase

Step 4: Identify $p_i \in \mathbf{P_t}$ that resulted in largest number of misclassification.
If class label of p_i contradicts with NPS artificial neurons at its neighborhood, removed p_i from $\mathbf{P_t}$.
Otherwise, generate centroid artificial neuron p_j among the NPS artificial neurons (with same class label).
Add p_j to $\mathbf{P_t}$.

Step 5: A_t = accuracy of classification model $\mathbf{P_t}$ when evaluated with **D**.
If A_t is greater or equals to best accuracy achieved thus far, update C to $\mathbf{P_t}$. Otherwise, discard $\mathbf{P_t}$.
Set t = t + 1.

Step 6: Scatter closely clustered $p_i \in \mathbf{C}$. Resulting model forms $\mathbf{P_t}$.

Step 7: A_t = accuracy of classification model $\mathbf{P_t}$ when evaluated with **D**.
If A_t is greater or equals to best accuracy achieved thus far, update **C** to $\mathbf{P_t}$. Otherwise, discard $\mathbf{P_t}$.
Set t = t + 1.

Step 8: If termination criteria are satisfied, proceed to Step 9. Otherwise, go to Step 4.

Apoptosis Phase

Step 9: If eradication of $p_i \in \mathbf{C}$ does not deteriorate classification performance when evaluated with **D**, remove p_i from **C**.
Otherwise, keep p_i.

Evaluation

Step 10: Evaluate performance of classification model **C** on **T**. Generated class labels of **T** are assigned to **O**.

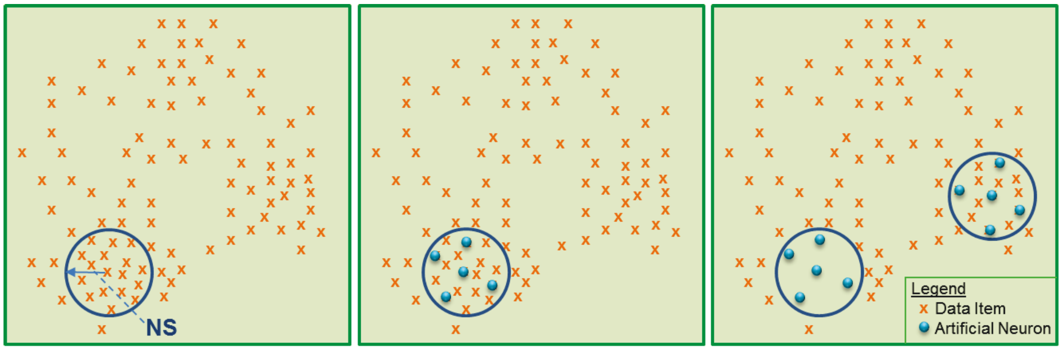


Figure 6.2: Graphical Illustration of Neurogenesis Phase
From the initial training data, the region (defined by NS) that is most populated with data items is identified. (b) From the identified region, artificial neurons are created using KS algorithm (in this example, numNeurons=5). (c) Upon selection, data items within that region are removed. This process is repeated.

- **Neuroplastic Threshold (NPT)**: The number of cycles allowed for ANCSc algorithm to generalize the artificial cognitive system before termination. This (integer) parameter resets if there is an improvement to the classification accuracy.

6.3.2. Training Routine of ANCSc

This subsection provides a detailed description of the key routines, methods and equations proposed in ANCSc algorithm. The canonical flow of the algorithm is illustrated in Figure 6.1 while Algorithm 6.1 provides the corresponding pseudocode. In this implementation, all data are normalized such that the Euclidean distance between any 2 feature vectors is between 0 and 1.

The ANCSc algorithm consists of 3 key development phases – namely neurogenesis, neuroplasticity via nurturing and apoptosis phases. All steps proposed in ANCSc algorithm to develop the classification methodology are explained independently below.

Neurogenesis Phase

The primary objective of this phase is to generate a reduced set of representative artificial neurons (or data items) from the training dataset. This establishes the fetal artificial cognitive system that would be refined and enhanced in the later phases. It begins the process of populating new artificial neurons by requiring the specification of 2 parameters – namely neurogenic space (NS) and neurogenic rate (NR). It proceeds by searching for the region (radius defined by NS) that is most populated with data items (within the training dataset). Upon finding it, a uniformly distributed subset of data items from that region is selected. This selection technique of uniformly distributed data item is similar to the Kennard-Stone (KS) algorithm (Kennard & Stone, 1969). However, unlike KS algorithm, we proposed that the number of data items (numNeurons) to be selected be dynamically determined by the following equation:

$$\text{numNeurons} = ||NS|| * NR \tag{8}$$

where $||NS||$ is the number of data items found within the defined region, and NR is a user-defined probability parameter that determines the proportion of data items that would be selected as artificial neurons for the development of the fetal artificial cognitive system. This NR parameter is tantamount to the intrinsic and environmental stimuli (described in Section 6.2) that regulate the rate of neurogenesis in human brain.

Subsequently, all data items within the previously defined region are removed and the aforementioned process repeats to create the fetal artificial cognitive system. At the end of this phase, a set of representative artificial neurons would form the artificial cognitive system. An illustration of this process is given in Figure 6.2. Finally, the classification performance of the constructed fetal artificial cognitive system is evaluated with the initial training data.

Neuroplasticity via Nurturing Phase

Neuroplasticity, as a consequence of nurturing, plays a significant role in promoting the construction of a robust classification model that promises enhanced performance over one that regurgitates memorized patterns learned during the neurogenesis phase. This phase was inspired by observation of how neuronal structures change (i.e. change in the position of the neurons) in tandem with healthy brain development, learning and memory formation. Changes in connectivity among the neurons (i.e. synaptic connection) are not considered in order to postulate a simple and efficient learning model.

The primary objective of this phase is to (1) grow artificial neurons at locations that would contribute to better classification performance, (2) remove existing artificial neurons that exacerbate the classification performance, and (3) adapt engendered artificial neurons to the input data environment to promote better classification performance. This phase begins by identifying the artificial neuron that resulted in the largest number of misclassifications. If the class of this artificial neuron (for example, it is class 1) contradicts with most of the other artificial neurons (i.e. they are of class 0) at its proximity, it is removed from the artificial cognitive system. Otherwise, a new artificial neuron with the same class (as those at its proximity) is generated at the centroid of those artificial neurons, and added to the artificial cognitive system. A condition that must be satisfied for this addition is that the class of the artificial neuron to be added must belong to the minority data class. This is to encourage a balanced number of artificial neurons (i.e. similar number of artificial neurons with class 0 and 1 labels) to thrive in the developed artificial cognitive system. We hypothesize that this would potentially deliver a solution that could generalize better.

An aging mechanism is implemented, "aging" the newly added artificial neurons. This is to allow "younger" artificial neurons to have an opportunity to be involved in the learning process (i.e. mimicking the concept - in neuroscience - that younger neurons in human brain are more plastic (Wiskott et al., 2006)). If this phase resulted in an artificial cognitive system that shows improved performance, it would be kept for future development. Otherwise, it would be discarded.

To better adapt the engendered artificial neurons to the input data environment, closely clustered artificial neurons are scattered apart if it does not compromise the resulting classification performance. This adaptation step begins by searching for the artificial neuron (dNeuron) – within a region whose radius is defined by the neural density (ND) parameter - that is most populated with other artificial neurons. Upon finding this artificial neuron, the closest artificial neuron (cNeuron) affiliated to it (i.e. with highest affinity to dNeuron) is modified so that they are more distributed apart. The degree of spread is determined by the neuroplastic coefficient (NPC) parameter and defined with the following equation:

$$\text{cNeuron}_i = \text{cNeuron}_i + \text{NPC}*(\text{cNeuron}_i - \text{dNeuron}_i) \qquad (9)$$

where cNeuron_i and dNeuron_i are the ith attribute of cNeuron and dNeuron, respectively. Through modicum adjustment of the artificial neurons in the artificial cognitive system, we aim to promote the construction of a more diverse set of representative artificial neurons; sequella for mitigating the risk of overfitting. Similar to the previous step, a (separate) aging mechanism is implemented. This is to ensure that different artificial neurons that are densely clustered together have a chance to deviate and generalize. Likewise, if this newly developed artificial cognitive system constructed in this phase demonstrates ameliorated performance, it would be saved. Otherwise, it would be removed from further consideration.

Termination of Neuroplasticity via Nurturing Phase

The stopping criterion for neuroplasticity via nurturing phase is reached if there is no improvement in the classification performance after LTP (a user-defined value) iterations or the same classification performance is achieved consecutively after NPT (a user-defined value) iterations. Otherwise, neuroplasticity via nurturing phase repeats, inculcating the artificial cognitive system with key patterns that underlie the training data.

Apoptosis Phase

Naturally occurring apoptotic processes are very important in healthy development of organism. For example, apoptosis occurs between the fingers and toes of a human during the embryonic stage (which initially appears like duck's webbed feet), giving them the freedom to maneuver individually. Metaphorically, this feature may offer ANCSc algorithm the ability to trim away redundant neurons, delivering a concise and efficacious classification model.

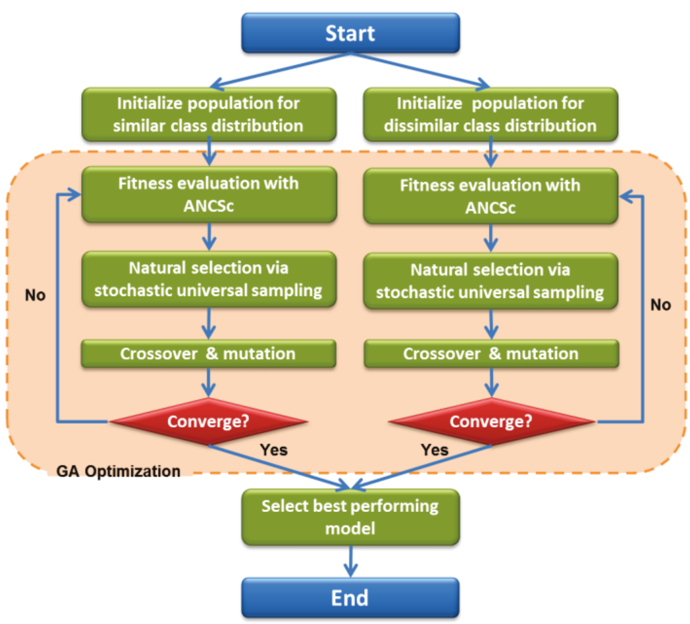


Figure 6.3: Proposed Strategy for Optimization of Parameter Set for Binary Class Classification Problems
Optimization using GA was conducted on ANCSc algorithm that has a common and independent set of parameters. The best performing model computed from these two experiments was selected as the resulting classification model.

This process of removing redundant artificial neurons is carried out upon termination of the neuroplasticity via nurturing phase. It aims to eradicate redundant artificial neurons that do not contribute to the construction of an efficacious and concise artificial cognitive system, but instead exacerbate the overall performance. The determination of which artificial neuron to apoptosize is governed by 2 questions. First, whether the Euclidean distance of the artificial neuron under examination and another artificial neuron in the postulated artificial cognitive system is smaller than the product of NS and NR? Second, whether removal of the neuron under examination would contribute to an improved artificial cognitive system? If the answer is 'yes' to both these questions then that artificial neuron is removed. Otherwise, it remains in the artificial cognitive system.

Evaluation

At the end of the training cycle described above, KNN algorithm is used to predict the class value of unseen data items. It works by determining the k (defined by NPS parameter) artificial neurons closest to an unseen data item and adopting a majority vote scheme to suggest the class value. This is similar to activating the neurons in the corresponding neuronal pool – in human brain – when one recall an event or object.

6.3.3. Data Class-specific ANCSc Parameters

Neurogenesis has been shown to occur in 2 distinct areas of the brain, namely the dentate gyrus of the hippocampus and the anterior part of the subventricular zone. Each area harbors a population of neural stem/progenitor cells that divide and proliferate independently. Moreover, each area is responsible for different function - the hippocampus is claimed to be the putative area for information storage while the subventricular zone is associated with the development of the olfactory bulb. The occurrence of autonomous

		ANCSc	
		Misclassification	Correct Classification
EDC-AIRS	Misclassification	**a**	**b**
	Correct Classification	**c**	**d**

Figure 6.4: Contingency Table for McNemar's Test (ANCSc vs EDC-AIRS)
'a' indicates the number of data items misclassified by both EDC-AIRS and ANCSc; 'b' represents the number of data items misclassified by EDC-AIRS but correctly classified by ANCSc; 'c' denotes the number of data items misclassified by ANCSc but correctly classified by EDC-AIRS; 'd' dictates the number of data items correctly classified by both EDC-AIRS and ANCSc.

neurogenesis in areas of the brain responsible for different function suggests that decentralized development may be the strategy that nature adopts.

These observations underscore the importance of locality and task specific regulation. One approach to bio-mimic this computationally is to independently analyze and model each data class (i.e. having an independent parameter set for each data class). This, when applied to an immune-inspired algorithm (Tay et al., 2013), demonstrated improved performance. Therefore a similar technique was implemented in ANCSc algorithm. The parameters that orchestrate the proliferation, adaptation and survival of the neural cells include NS, NR, ND and NPC. Hence, these parameters were duplicated and optimized independently for each data class. Genetic algorithm (GA) (Holland, 1992) - a search heuristic inspired by natural evolution - was employed to optimize these parameters.

Figure 6.3 illustrates the canonical flow of the strategy used to solve binary classification problems. Two sets of parameters were initialized and optimized in parallel - namely a common set (i.e. a single set of parameters used to model both data classes) consisting of 7 parameters and an independent set comprising 11 parameters. Upon termination of the optimization process carried out by GA, the best performing classification model obtained is used to predict future unseen data items.

6.4. Material and Methods

6.4.1. Performance Evaluation of ANCSc Algorithm

ANCSc algorithm was evaluated with a number of widely used benchmark datasets to assess its learning capability and classification performance. A total of 9 datasets, used in (Tay et al., 2013), were employed to evaluate whether having independent parameter set influence the classification performance of ANCSc algorithm. These 9 datasets used include: Ionosphere, Fisher's Iris, Wine,

Cardiovascular Health Study (CHS), Pima Indians Diabetes, Hill Valley, Bupa Liver Disorder, Sonar, and Statlog Heart datasets. The performance yielded by ANCSc algorithm was (statistically) compared with those obtained by the evolutionary data-conscious artificial immune recognition system (EDC-AIRS) (Tay et al., 2013). We have chosen McNemar's test to determine whether the performance of the 2 supervised algorithms described are statistically different as it has been demonstrated to have low type 1 error (Dietterich, 1998). To perform the test, both EDC-AIRS and ANCSc algorithms were first trained with the training data and tested with the testing data. The predicted outcome for each data item in the testing data was recorded and used to construct the contingency table shown in Figure 6.4. Referring to the figure, if the sum of 'b' and 'c' is greater than 25, chi-square test with 1 degree of freedom is used for performing McNemar's test. Otherwise, to provide a better estimation of the small sample (i.e. $b + c \leq 25$), binomial distribution is used for

Table 6.1: Cross-Validation Scheme Employed for Each Dataset

Dataset	Number of CV Fold
Fisher's Iris	5
Pima Indians Diabetes	10
Sonar	13
Wine	10
Statlog Heart	10

These cross-validation schemes were selected to remain comparable to other experiments reported in the literature.

(exact) McNemar's test. The 2 algorithms are considered to be statistically different if the p-value computed with McNemar's test is smaller than 0.05.

Further investigation was conducted to empirically evaluate (1) the classification performance of ANCSc algorithm when compared to other state-of-the-art classification algorithms, and (2) the performance and number of postulated artificial neurons at the end of each phase of ANCSc implementation. A total of 6 datasets, obtained from the data repository at the University of California (Irvine) (C.L. Blake & C.J. Merz, 1998), were used. These datasets include: Fisher's Iris, Ionosphere, Pima Indians Diabetes, Sonar, Wine and Statlog Heart datasets. Hold-out validation was carried out for Ionosphere dataset, while cross-validation (CV) was performed on the remaining 5 datasets. In particular, the first 200 data items of the Ionosphere dataset were selected as the training data and the remaining 151 data items were chosen as the testing dataset. As for the rest of the datasets, the cross-validation

Table 6.2: Empirical Experimental Results for Using Common and Independent Parameter Sets

Measurement	Bupa Liver Disorder	ks_yr50611	Statlog Heart	Hill-Valley	Ionosphere	Iris	Pima Indians Diabetes	Sonar	Wine
#Instances	345	270	270	606	200	150	768	208	178
#Attributes	6	253	13	100	34	4	8	60	13
#Classes	2	2	2	2	2	3	2	2	3
#Class1 Instances	145	135	120	305	99	50	268	97	59
#Class2 Instances	200	135	150	301	101	50	500	111	71
#Class3 Instances	-	-	-	-	-	50	-	-	48
Validation Type	10-CV	10-CV	10-CV	Holdout	Holdout	5-CV	10-CV	13-CV	10-CV
Acc. Obtained with Common Parameter Set	**72.8%**	**80.4%**	**86.3%**	62.7%	96.7%	98.9%	75.9%	89.9%	98.9%
Acc. Obtained with Independent Parameter Set	70.1%	79.6%	85.6%	**63.0%**	**98.0%**	**99.1%**	**77.5%**	**91.8%**	**99.3%**

Accuracy (Acc.) was used as the metric to evaluate how common and independent parameter sets influence the performance of ANCSc algorithm. The dataset 'ks_yr50611', which uses the CHS dataset, predicts the occurrence of MI (from year 6 to 11) based on a balanced case-control sample obtained in year 5. CV denotes cross-validation.

scheme used is described in Table 6.1. The reason for choosing these validation strategies was to remain comparable to other experiments reported in the literature.

Experiments on each dataset were conducted 3 times to evaluate its consistency. It was optimized with GA with the following setup details: population size: 100; maximum generation: 100; natural selection: stochastic universal sampling; crossover type: discrete recombination; crossover probability: 0.8; mutation rate: 1/P, where P is the number of parameters. The value of the ANCSc parameters that was either assigned (i.e. given as a constant value) or tuned with GA (i.e. given as a range of value) are as follow: Seed: 1; NPS: [1, 15]; LPT: [0, 10]; NPT: [0, 100]; NS: [0, 0.5]; NR: [0, 1]; ND = [0, 0.5]; NPC = [0, 0.5]. These parameter values were determined experimentally and kept constant between benchmarks.

6.4.2. Dataset

Several standard benchmark datasets were used in this investigation. A succinct description of the datasets used can be found in section 4.3.2.

6.5. Experimental Results

Several experiments were conducted to investigate the properties and classification ability of ANCSc algorithm. Four key experiments were carried out and presented in this section. Their objectives are to determine: (1) the significance of implementing independent parameter set for each data class; (2) the performance of ANCSc algorithm when compared to other state-of-the-art algorithms; (3) the performance of ANCSc algorithm when juxtaposed with EDC-AIRS algorithm - one of the top performing classification algorithm compared; and (4) the sensitivity of each parameter that orchestrates and influences the development of the neurons and its impact on the classification

Table 6.3: Performance Comparison of Different Classification Algorithm

	Iris		Ionosphere		Pima Indians Diabetes		Sonar		Wine		Statlog Heart	
Rank	**Algorithm**	**Acc**	**Algorithm**	**Acc**	**Algorithm**	**Acc**	**Algorithm**	**Acc**	**Algorithm**	**Acc**	**Algorithm**	**Acc**
1	Grobian (rough)	100%	3-NN + Simplex	98.7%	Logdisc	77.7%	TAP MFT Bayesian	92.3%	EDC-AIRS	99.6%	**ANCSc**	**86.3%**
2	EDC-AIRS	99.6%	**ANCSc**	**98.0%**	IncNet	77.6%	**ANCSc**	**91.8%**	**ANCSc**	**99.3%**	Lin. SVM 2D QCP	85.9%
					DIPOL92	77.6%						
3	**ANCSc**	**99.1%**	EDC-AIRS	97.4%	**ANCSc**	**77.5%**	Nave MFT Bayesian	90.4%	kNN, Manh, auto k=1-10	98.9%	EDC-AIRS	84.8%
					EDC-AIRS	77.3%	SVM	90.4%	IncNet, Gauss	98.9%		
					Linear Disc. Analysis	77.5 – 77.2%	Best 2-layer MLP + BP, 12 hidden	90.4%				
4	SSV	98.0%	3-NN	96.7%	SMART	76.8%	EDC-AIRS	88.5%	SSV	98.3%	Naive-Bayes	84.5%
	C-MLP2LN	98.0%	IB3	96.7%	GTO DT (5xCV)	76.8%						
	PVM 2 rules	98.0%										
5	PVM 1 rule	97.3%	MLP + BP	96.0%	ASI	76.6%	AIRS2	84.9%	kNN, Euclidean, k=1	97.8%	K*	76.7%
6	AIRS	96.7%										
	FuNe-I	96.7%	AIRS2	95.6%	Fischer Disc. Analysis	76.5%	MLP+BP, 12 hidden	84.7%	FSM	96.1%	IB1c	74.0%
	NEFCLASS	96.7%										
7	AIRS2	96.0%	AIRS	94.9%	MLP+BP	76.4%	MLP+BP, 24 hidden	84.5%			1R	71.4%
	CART	96.0%	C4.5	94.9%								
8	FUNN	95.7%	RIAC	94.6%	LVQ	75.8%	1-NN, Manhanttan	84.2%			T2	68.1%
					LFC	75.8%						
9			SVM	93.2%	RBF	75.7%	AIRS	84.0%			MLP + BP	65.6%
10			FSM + rotation	92.8%	kNN, k=22, Manh	75.5%	FSM	83.6%			FOIL	64.0%
					MML	75.5%						
					NB	75.5 – 73.8%						
…					… …							
n					AIRS2	74.2%						
n+1					AIRS	74.1%						

'Acc' denotes the classification accuracy. The performance of EDC-AIRS algorithm without feature selection is shown in this comparison (as feature selection was not performed by other algorithms compared in this table).

performance of ANCSc algorithm. For each experiment the algorithm was executed 3 times. Consistent classification results were obtained for all the runs (i.e. standard deviation of 0). This signifies that ANCSc algorithm exhibits deterministic learning capability.

Table 6.4: Performance Comparison of ANCSc and EDC-AIRS Algorithms using McNemar's Test

Dataset	McNemar's Test# (p-value)
Bupa Liver Disorder	**0.016**
ks_yr50611	**0.008**
Statlog Heart	0.290
Hill-Valley	**0.016**
Ionosphere	1.000
Iris	1.000
Pima Indians Diabetes	0.924
Sonar	0.144
Wine	1.000

#The p-value of McNemar's test is presented, examining whether the performance of ANCSc algorithm is statistically different from EDC-AIRS algorithm.

Table 6.5: Performance of ANCSc Algorithm at Each Phase of Implementation

	Neurogenesis Phase		Neuroplasticity via Nurturing Phase		Apoptosis Phase		
	Accuracy	#Neurons	Accuracy	#Neurons	Accuracy	#Neurons	%Neurons Eradicated
Iris	97.8%	109.2±0.4	97.8%	105.8±3.0	99.1%	28.8±5.1	72.8%
Ionosphere	94.7%	153±0	95.4%	146±0	98.0%	113±0	22.6%
Diabetes	73.8%	555.2±1.2	74.1%	554.1±1.4	77.5%	215.6±9.6	61.1%
Sonar	86.1%	178.5±0.9	87.5%	172.7±2.8	91.8%	105.4±1.0	39.0%
Wine	98.5%	116.1±1.6	98.5%	116.1±1.6	99.3%	89.2±0.9	23.2%
Heart	85.2%	92.7±1.8	85.6%	88.8±2.9	86.3%	84.8±12.2	4.5%

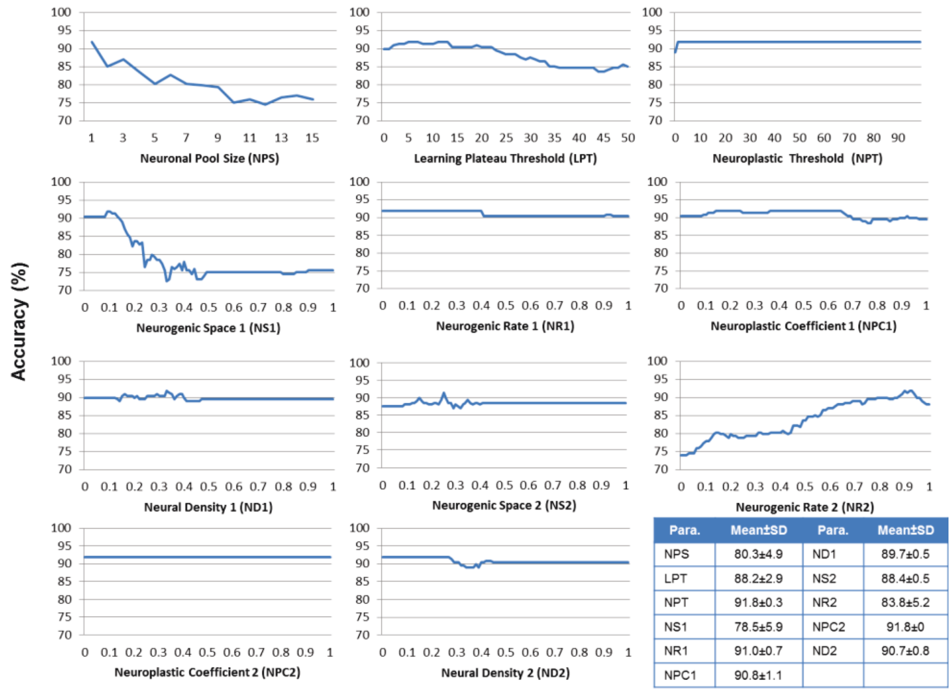
'#Neurons' refers to the average number of artificial neurons generated in the artificial cognitive system after executing each phase. '%Neurons Eradicated' denotes the percentage reduction in the number of artificial neurons after the apoptosis phase is conducted. The best performing model obtained for each dataset was used to perform this analysis.

6.5.1. Performance of ANCSc Algorithm

The importance of having independent parameter set for ANCSc algorithm was evaluated using 9 benchmark datasets. The corresponding classification performance is provided in Table 6.2. From the results, it can be observed that 6 out of 9 datasets evaluated benefited from this implementation. Comparison of ANCSc algorithm with other well-known classifiers (Duch, 2000; Duch, 2000) is given in Table 6.3. The ANCSc algorithm has shown promising results – achieving highly competitive performance for all the datasets evaluated. To assess how ANCSc algorithm fare when juxtaposed with other top performing algorithms evaluated, we statistically compare the classification performance yielded by ANCSc algorithm and EDC-AIRS algorithm (one of the top performing algorithms evaluated) using McNemar's test. The results, as shown in Table 6.4, indicate that ANCSc algorithm achieved comparable, if not better, performance than EDC-AIRS algorithm. Specifically, ANCSc algorithm outperforms EDC-AIRS algorithm (with statistically significant improvement) for 'Bupa Liver Disorder', 'ks_yr50611' and 'Hill-Valley' datasets while comparable performance was achieved for the remaining datasets.

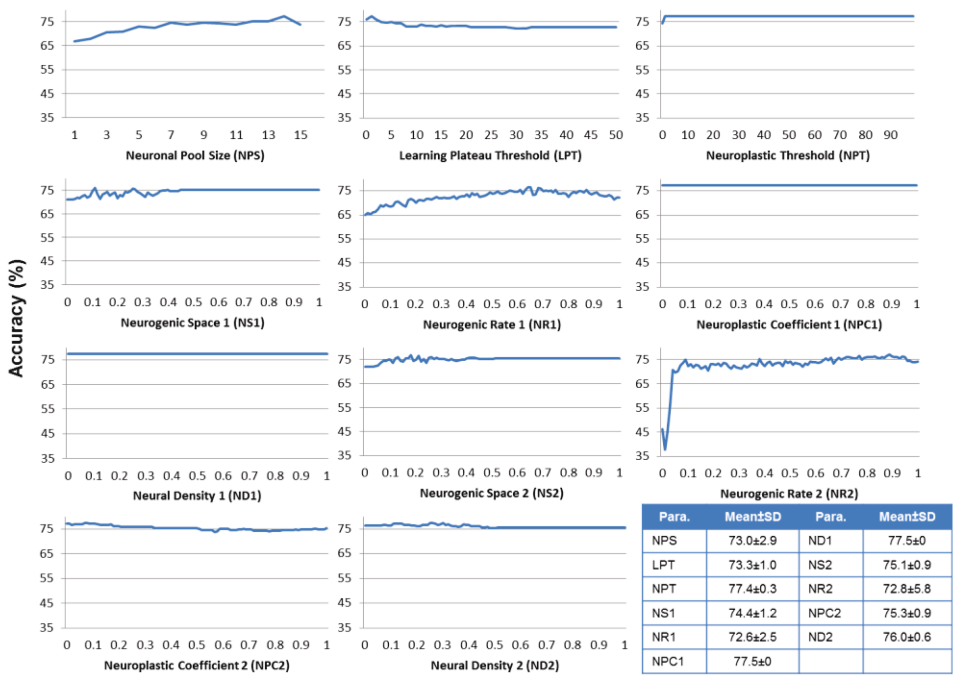
The classification performance and number of artificial neurons engendered at each phase were scrutinized using 6 datasets (i.e. Iris, Ionosphere, Pima Indians Diabetes, Sonar, Wine and Statlog Heart datasets). Results, as given in Table 6.5, demonstrate that after the execution of each phase, improved classification accuracy was achieved. Moreover, significant number of redundant artificial neurons engendered (during the neurogenesis, and neuroplasticity via nurturing phases) was pruned away during the apoptosis phase. This resulted in the formation of a concise (i.e. memory efficient) classification models with ameliorated performance (having an improvement of up to 4.9%).

The average computational time required by ANCSc algorithm to develop the best performing model (i.e. executed with parameter values that produce the highest classification accuracy) for the 9 benchmark datasets used was analyzed. For each dataset, the algorithm is executed 10 times on an Intel Xeon 2.66 GHz (18 GB RAM) server. The average computational time required ranges from 4.22±0.15 seconds (for Ionosphere dataset) to 348.6±0.7 seconds (for Pima Indians Diabetes dataset). As part of our future work, we aim to ameliorate the computational efficiency of the algorithm.



Para.	Mean±SD	Para.	Mean±SD
NPS	80.3±4.9	ND1	89.7±0.5
LPT	88.2±2.9	NS2	88.4±0.5
NPT	91.8±0.3	NR2	83.8±5.2
NS1	78.5±5.9	NPC2	91.8±0
NR1	91.0±0.7	ND2	90.7±0.8
NPC1	90.8±1.1		

(a) Sonar Dataset



Para.	Mean±SD	Para.	Mean±SD
NPS	73.0±2.9	ND1	77.5±0
LPT	73.3±1.0	NS2	75.1±0.9
NPT	77.4±0.3	NR2	72.8±5.8
NS1	74.4±1.2	NPC2	75.3±0.9
NR1	72.6±2.5	ND2	76.0±0.6
NPC1	77.5±0		

(b) Pima Indians Diabetes Dataset

Figure 6.5: ANCSc Sensitivity Analysis Performed on 11 Parameters for Binary Classification Problems

Sensitivity analysis performed on (a) Sonar and (b) Pima Indians Diabetes datasets. The best performing model achieved for each dataset was used to perform the sensitivity analysis (i.e. varying the value of the parameter under study while fixing the remaining parameters' value). The number suffixed to NS, NR, NPC and ND indicates whether it is used to model the first or second data class.

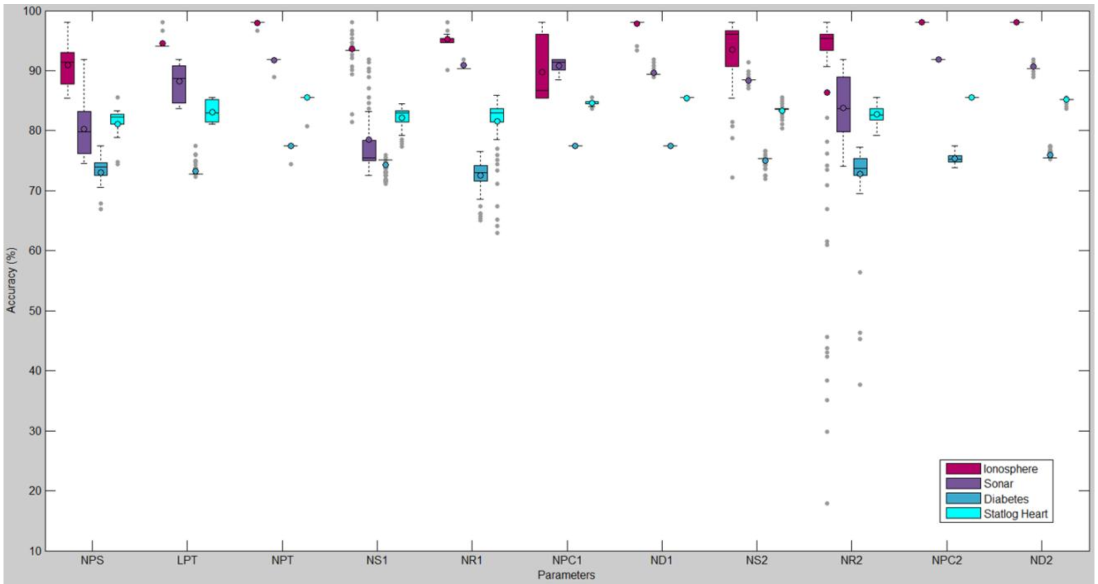


Figure 6.6: ANCSc Sensitivity Analysis Performed on 4 Datasets
Sensitivity analysis was performed on 4 (binary class) datasets. The effect of 11 parameters on ANCSc classification performance was investigated. The number suffixed to NS, NR, NPC and ND indicates whether it is used to model the first or second data class.

6.5.2. Sensitivity Analysis

The proposed ANCSc algorithm performs proliferation, adaptation and survival of artificial neurons based on 7 key user-defined parameters, namely NS, NR, ND, NPC, NPT, LPT and NPS. These parameters influence (1) the initial neuronal map developed in the fetal artificial cognitive system (i.e. parameters NS and NR); (2) the proliferation, eradication and adaptation of artificial neurons in response to learning the data environment (i.e. parameters ND and NPC); (3) the apoptosis of artificial neurons in support for the development of an artificial cognitive system that can generalize better (i.e. parameters NS and NR); (4) the duration of learning allowed (i.e. parameters LTP and NPT); and (5) the number of proximal artificial neurons deemed responsible for describing a specific pattern (i.e. parameter NPS).

Sensitivity analysis, the study of how the uncertainty of input parameters would affect the output of the inference model, was carried out by varying the parameter under study while fixing the remaining parameters. The parameter values related to the best performing model (optimized and yielded with GA) was used as the base model to perform this analysis. Experiments on 4 datasets (i.e. Ionosphere, Pima Indians Diabetes, Sonar and Statlog Heart datasets) demonstrated that each of these parameters have an effect on the classification performance of ANCSc algorithm. The experimental results for 2 datasets (Sonar and Pima Indians Diabetes) are provided in Figure 6.5. From the figure, NPT (for values greater than 1) seems to be ostensibly redundant as it did not affect the classification performance. However, on further investigation with arbitrary values for all the 11 parameters, it was found that NPT has an impact on the resulting classification performance. A summarize view, given as box plot, of the (sensitivity analysis) results for the 4 datasets investigated is shown in Figure 6.6. From the boxplot, it can be observed that the interquartile range for all the experiments conducted tends to be small. This is highly desirable as it signifies that ANCSc algorithm is robust enough to postulate the optimal model over a wide range of parameter values.

6.6. Discussion

We have developed a novel algorithm called ANCSc. It is a supervised classification algorithm inspired by the importance and robustness of several mechanisms (i.e. neurogenesis, neuroplasticity, nurturing and apoptosis) that occur during the development of the brain. These mechanisms empower individuals with the capability and creativity to interact and solve environmental problems in an innovative, effective and efficient manner.

During neurogenesis phase proposed in ANCSc algorithm, the fetal artificial cognitive system begins by developing artificial neurons and taking shape. It subsequently advances to the neuroplasticity via nurturing phases, whereby the initially grown artificial neurons were stimulated and “nurtured” by the data environment it is presented with. In this regard, the artificial neurons in the artificial cognitive system evolved during each learning cycle by growing new artificial neurons, performing niche refinement to existing ones and/or eradicating artificial neurons that hinder the inculcation process. Through this repeated learning process, the aim is to "educate" the artificial cognitive system with key patterns found within the training data; enabling the artificial neurons to develop further and collectively realize their full potential. Experimental results (see Table 6.5) demonstrate that this inculcation process (i.e. neuroplasticity via nurturing) incrementally ameliorate the classification performance of the developing artificial cognitive system. Termination of this learning algorithm proceeds with the removal of redundant artificial neurons (i.e. apoptosis phase) that potentially exacerbate the resulting classification performance. With this implementation, it is worth noting that (based on the 6 datasets evaluated) on average, 37.2% of neurons were removed from the postulated classification models while improving the classification accuracy by 2.5%. Hence, evolution, cooperation and altruism among the neuronal cultures are the most important factors that resulted in the success of ANCSc algorithm.

Further enhancement to the algorithm was carried out by empowering the ANCSc algorithm with the ability to model each data class autonomously. Locality/task specific regulation of changes in the neurons was mimicked and implemented by introducing an independent parameter set for each data class involved. Experiments on 9 benchmark datasets showed a small improvement in classification performance (ranges from 0.2% to 1.9%). Nevertheless, we believe that if the idiosyncratic characteristics of each data class under study differ significantly, the advantage of having such independent parameter set would become more prominent.

The ANCSc algorithm has achieved promising results and outperformed several state-of-the-art classification algorithms. To objectively assess the performance of ANCSc algorithm, we have employed McNemar's test to statistically compare the classification performance of ANCSc and EDC-AIRS (one of the top performing algorithm evaluated) algorithms. From the results, it was demonstrated that ANCSc classification performance is comparable, if not better, than EDC-AIRS algorithm.

Sensitivity analysis was conducted on 4 binary class datasets. The average standard deviation for the 11 parameters analyzed ranges from 0.002 to 0.081.

Results indicate that NR, NPS and NS play a significant role in producing an optimal classification model. The average standard deviations of the top 3 most sensitive parameters (i.e. NR2, NPS and NS1) are 0.081, 0.030 and 0.028 respectively. High NR and NS sensitivity signify that the fetal artificial cognitive system constructed during the neurogenesis phase has a significant impact on the performance of the final classification model generated. High NPS sensitivity suggests that complex and discrete patterns are ubiquitous within the data problem under examination where changes in the proposed artificial neuronal pools have a significant impact on the classification result. In other words, changing the value of NPS might cause distinct artificial neuronal pool to overlap, jeopardizing the ability of the classification model to generalize and predict the correct class for unseen data items.

Although only 3 parameters were accentuated in this section, the other 8 parameters (used during neuroplasticity via nurturing and apoptosis phases) do contribute to the success of the algorithm – results are as demonstrated by sensitivity analysis and performance at different phases of ANCSc implementation (see Figure 6.5 and Table 6.5). To this end, proper optimization of all the 11 parameters is highly recommended for the production of an accurate and robust classification model.

To summarize, ANCSc algorithm has several attractive features as a supervised learning algorithm. These include, but are not limited to, the ability to: (1) autonomously develop an appropriate, representative, and concise cognitive architecture during the learning process; (2) incrementally learn and model each data class independently; (3) achieve highly competitive classification performance when juxtaposed with other state-of-the-art classification algorithms; and (4) generate an optimal model over a wide range of parameter values. The results, to date, show that the ANCSc algorithm is a robust learner that is capable of adapting to different profound data patterns and structures.

It is noteworthy that every classification algorithm has its own inductive bias that work reasonably well for some, but not all, datasets or application domains

– an observation commonly referred to as the selective superiority problem (Brodley, 1993) in the literature. Therefore, ANCSc algorithm does not guarantee improved performance for all classification problems or outperform all other classification algorithms in view of this problem.

We believe that future work to advance the algorithm can be carried out along 3 main research directions: (1) the extension of ANCSc algorithm for unsupervised learning and time-series analysis; (2) the study of the implication and possible application of ANCSc algorithm in various research fields (e.g. pattern recognition, bioinformatics, optimization, etc.); and (3) the exploration of techniques for solving large-scale problems effectively and efficiently (e.g. parallelism, storage efficiency, incrementally learning, etc.).

6.7. Summary

We have presented a novel supervised learning algorithm inspired by natural phenomena related to neurogenesis, neuroplasticity, nurturing and apoptosis. Leveraging on the fetal artificial cognitive system developed from the input data environment, ANCSc algorithm "nurture" it in an attempt to unleash its greatest potential. Application of ANCSc algorithm to classical classification problems have been performed with promising results.

The learning approach postulated by ANCSc algorithm, in our opinion, has great potential for learning profound data structures and producing a concise model capable of descripting the problem. Additionally, it offers a novel learning methodology in which classification problems can be solved by approaching them from a different perspective.

Chapter 7

Age-Related Risk Prediction Model[7]

Cardiovascular disease (CVD) is currently the leading cause of mortality in many developed countries. One reason for this phenomenon is the poor understanding of the disease etiology. This, in part, is due to the confounding and evolving effect of risk factors associated with CVD. This, we believe, has an impact on computational-based risk prediction for CVD as well. To investigate this impact, we present in this chapter a (age-related) risk prediction approach that takes the effect of evolving risk factors (over a range of ages) into consideration. Three algorithms - namely ANCSc, EDC-AIRS and SVM - were employed to develop these risk prediction models. Juxtaposition of these algorithms was performed to investigate on their ability to generalize. Data from the Honolulu Heart Program (Syme et al., 1975; Marmot et al., 1975; Robertson et al., 1977) were utilized to perform this risk prediction task. Results demonstrate that age-related risk prediction outperforms unified risk prediction approach.

[7] The work presented in this chapter has been submitted to IEEE Transactions on Biomedical and Health Informatics and is currently under review.

7.1. Introduction

Cardiovascular disease (CVD) is an epidemic and major health concern in today's world. It is the leading cause of mortality in many developed countries, such as the United States (US) and the United Kingdom (UK) (Go et al., 2013; British Heart Foundation Statistics Database, 2010). The risk of CVD death has been demonstrated to increase considerably with age for both genders (Tunstall-Pedoe et al., 1994). This has been postulated to be associated with differences in levels of CVD risk factors; which contribute to age-related excess risk for CVD (Asia Pacific Cohort Studies Collaboration, 2006). For example, hypertension and diabetes tend to prevail with age while total cholesterol levels and body mass index (BMI) often decline with age (Abbott et al., 2002). This, inevitably, suggests an accentuated role for hypertension and diabetes in the development of CVD in older individuals and an evolving role for total cholesterol levels and BMI in relation to age. Such evolution in risk factors, which may not be clinical overt to date, suggests that further research, discovery and development would have a beneficial impact to CVD healthcare – for example, early detection, ameliorated diagnostic precision, better understanding of risk factors evolution, effective treatments (e.g. recommendation of appropriate drug dosage in accordance to patient's age), and cost containment.

To this end, we aim to develop age-related risk prediction models capable of determining the first CVD event (over a 2-year period) experienced by individuals belonging to different age group – i.e. age 46 to 65, 46 to 55 and 56 to 65. This allows the effect of age-related risk factors in relation to risk prediction (developed using 3 learning algorithms) to be evaluated. Further, we hypothesize that the performance of the prediction models could be improved as predictive performance not only relies on the predictive ability of the learning algorithms used but also on the quality and characteristics of the input data presented.

The 3 learning algorithms used to develop the prediction models include

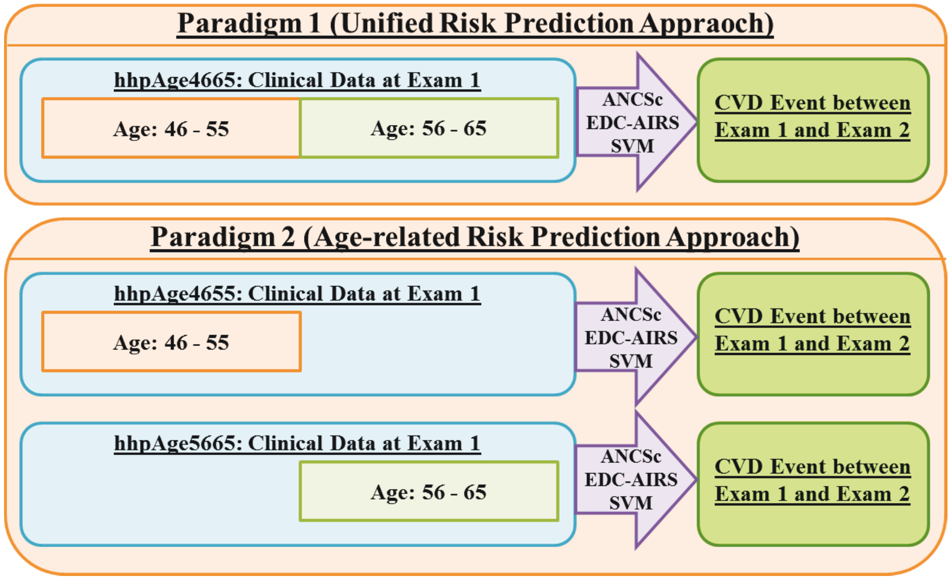


Figure 7.1: Age-related Risk Prediction Models for CVD
Age-related predictions employing 10-year and 20-year age models were performed to determine the impact of risk factors, associated with individuals in different age category, when modelled with different machine learning techniques.

Artificial Neural Cell System for classification (ANCSc) (Tay et al., 2014), Evolutionary Data-Conscious Artificial Immune Recognition System (EDC-AIRS) (Tay et al., 2013), and Support Vector Machine (SVM) (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1999). The employment of these algorithms enable us to investigate the classification capability of each algorithm when tested using the Honolulu Heart Program dataset (Syme et al., 1975; Marmot et al., 1975; Robertson et al., 1977). These 3 algorithms were examined as they have been shown in (Tay et al., 2014) to yield the highest performance when tested with a widely benchmarked heart disease dataset (i.e. Statlog Heart).

The rest of the chapter is organized as follows. Section 7.2 delineates the methodology used to construct the age-related prediction models and provides details of the Honolulu Heart Program dataset. Section 7.3 provides the experimental results achieved by each age-related prediction model developed using ANCSc, EDC-AIRS and SVM algorithms. Key results are discussed in Section 7.4 and conclusions are drawn in Section 7.5.

7.2. Material and Methods

7.2.1. Age-related Risk Prediction

Two age-related risk prediction paradigms for CVD were proposed to evaluate whether age-related risk factors have an impact on the performance of prediction models built using machine learning techniques. In the first paradigm, it encompasses a 20-year age model that consists of participants aged between 46 and 65 (denoted as 'hhpAge4665'). We call this the unified risk prediction

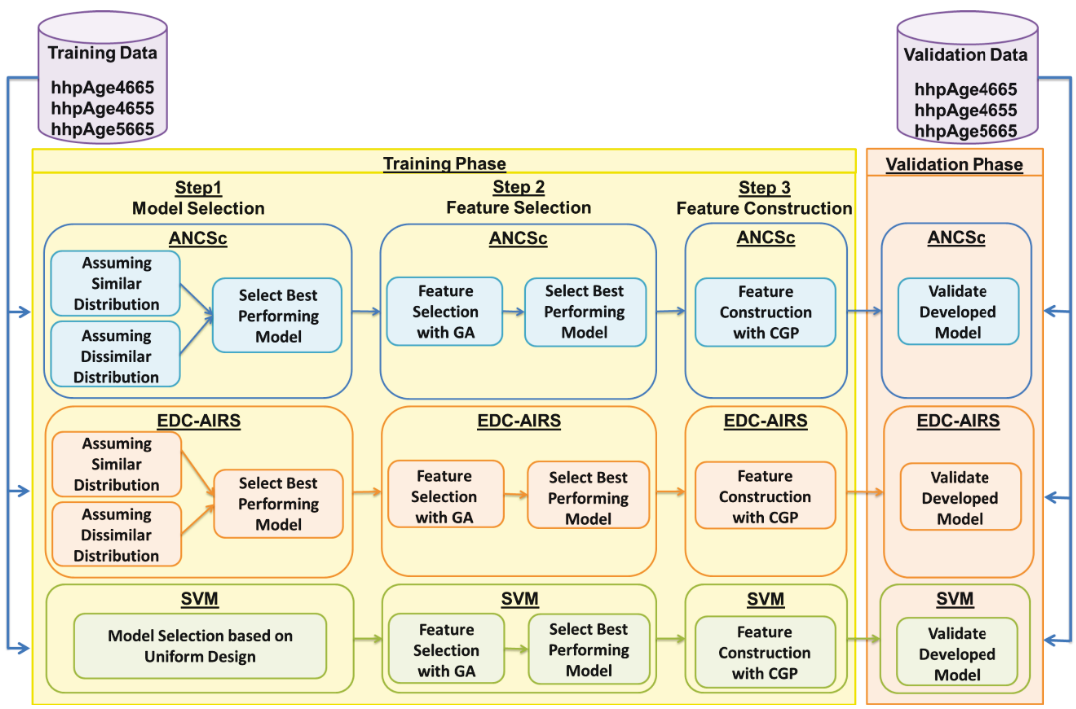


Figure 7.2: Methodology Employed to Develop the Age-related Prediction Models
The training phase, which uses the training data, is responsible for developing the prediction models. It consists of 3 distinct steps which include model selection, feature selection and feature construction. The developed prediction model is then validated for generalizability, during the validation phase, using the validation data.

approach as available participants of all ages were consolidated and used to build the prediction model. In the second paradigm, it comprises two 10-year age models consisting of participants aged between 46 and 55 (denoted as 'hhpAge4655'), and between 56 and 65 (denoted as 'hhpAge5665'). We call this paradigm as the age-related risk prediction approach as available participants were stratified into different age category before being used to build the corresponding risk prediction models. The risk of experiencing CVD between exam 1 and exam 2 (as reported in the Honolulu Heart Program) was modelled using 3 algorithms, namely (1) Artificial Neural Cell System for classification (ANCSc) (Tay et al., 2014) – a novel supervised classification algorithm inspired by neurogenesis, neuroplasticity and nurturing; (2) Evolutionary Data-Conscious Artificial Immune Recognition System (EDC-AIRS) (Tay et al., 2013) – an immune-inspired supervised classification algorithm; and (3) Support Vector Machine (SVM) (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1999) – a supervised classification algorithm based on statistical learning theory. Specifically, we aim to determine (1) whether age-related risk prediction approach outperforms unified risk prediction approach over a 2-year period (i.e. between exam 1 and exam 2), and (2) which algorithm is most capable at performing CVD risk prediction (i.e. which has the greatest generalization ability). The conceptual proposition for the investigation of the aforementioned objectives is illustrated graphically in Figure 7.1.

For each of the age model (i.e. 'hhpAge4665', 'hhpAge4655' and 'hhpAge5665'), prediction models were developed based on 3 consecutive optimization steps - namely model selection, feature selection and feature construction (see Figure 7.2). This enables ceteris paribus experiments to be conducted by all algorithms under scrutinization. Genetic algorithm (GA), unless otherwise stated, was utilized in this study to optimize the parameters. The choice of parameter settings for GA was experimentally determined. The details are as follow: population size: 100; maximum generation: 100; natural selection: stochastic universal sampling; crossover type: discrete recombination;

crossover probability: 0.8; mutation rate: 1/P, where P is the number of parameters.

First, model selection was conducted using 10-fold cross-validation. For ANCSc and EDC-AIRS algorithms, it is postulated in (Tay et al., 2014; Tay et al., 2013) that data class distribution plays an important role in the development of accurate classification models. Hence, both similar and dissimilar data class distribution were assumed and evaluated for these algorithms. The best performing model procured would be used in subsequent optimization steps. As for SVM, uniform design (Fang et al., 2000) method was used to determine the cost and gamma parameters required by SVM kernel (i.e. radial basis function). This approach was adopted as it has been shown to produce promising results, and at the same time alleviate the computational loads associated with the search for the optimal cost-gamma pair (Chow et al., 2008; Tay et al., 2014).

Next, feature selection using GA was carried out independently for each algorithm to select informative and predictive features that could enhance the process of dichotomization (i.e. separating cases from controls). This process is capable of removing redundant and/or irrelevant features that contribute to potential sources of noise and ambiguity; producing more efficacious prediction models as a result. The set of features that yielded the highest performance would be delivered to the next optimization step (i.e. feature construction) to construct new features that have the potential to ameliorate the prediction performance of the algorithm.

Feature construction is the process of discovering unknown relationship between features and augments the existing feature space with new composite features (Liu & Motoda, 1998). Cartesian Genetic Programming (CGP) (Miller & Thomson, 2000), a highly effective form of genetic programming that has demonstrated success in garnering parsimony (i.e. more human-comprehensible) (Kowaliw & Banzhaf, 2012), was employed to construct new features. It is noteworthy that we gave preference to the usage of the reduced feature set (i.e.

features selected after performing feature selection) to construct new features as we aim to build parsimonious prediction model (i.e. one that uses the least number of features). This is important, especially in the clinical settings, as each clinical test is associated with a different financial cost and risk for obtaining them. Therefore, it is highly desirable to develop prediction models that require the least number of clinical tests while not compromising predictive accuracy. Following an informal parameter search, the following CGP parameters were chosen: #inputs = feature dimension; #output = 1; #rows = 1; #columns = 10; arity = 2; levels back = 10; functions = {addition, subtraction, multiplication, division}.

Finally, the developed prediction models were validated for generalizability using the validation data – a dataset which is distinct and separate from the training data. To determine whether a single algorithm statistically outperforms the others, McNemar's test was conducted. This statistical test was chosen as it has been demonstrated to have low type 1 error (Dietterich, 1998). This test has been carried out by first training each algorithm with the training data and tested with the testing data. The predicted outcome for each data item in the testing data was recorded. Each time, results from 2 algorithms were used to construct the contingency table shown in Figure 7.3. If the sum of 'b' and 'c' is greater than 25, chi-square test with 1 degree of freedom is used for

		Algorithm 2	
		Misclassification	Correct Classification
Algorithm 1	Misclassification	**a**	**b**
	Correct Classification	**c**	**d**

Figure 7.3: Contingency Table for McNemar's Test (ANCSc vs EDC-AIRS/SVM)
'a' indicates the number of data items misclassified by both algorithm 1 and algorithm 2; 'b' represents the number of data items misclassified by algorithm 2 but correctly classified by algorithm 1; 'c' denotes the number of data items misclassified by algorithm 2 but correctly classified by algorithm 1; 'd' dictates the number of data items correctly classified by both algorithm 1 and algorithm 2.

performing McNemar's test. Otherwise, to provide a better estimation of the small sample (i.e. $b + c \leq 25$), binomial distribution is used for (exact) McNemar's test. One algorithm is considered to be statistically better than the other if the p-value computed with McNemar's test is smaller than 0.05.

7.2.2. Dataset & Data Pre-processing

The Honolulu Heart Program (Syme et al., 1975; Marmot et al., 1975; Robertson et al., 1977), initiated in 1965 by the National Heart, Lung and Blood Institute (NHLBI) as a prospective study of environmental and biological causes of CVD among Japanese Americans living in Hawaii, was analysed in this study. Subjects, followed for the development of CVD, collected between 1965 and 1968 (exam 1) were utilized as the baseline data. It consists of 8006 Japanese-American men living on the island of Oahu, Hawaii. At the time of study, participants received a comprehensive examination (e.g. physical measures, medical history/lifestyle, dietary, anthropometric measures, etc.) when aged between 45 and 68 (54.4±5.60). This resulted in 412 clinical features being collected. Out of these participants, only individuals (a total of 7383) who were free from angina pectoris (AP), coronary insufficiency (CI) and myocardial infarction (MI) were considered.

Cardiovascular events that occurred after the baseline examination (i.e. exam 1) were monitored through surveillance of hospital discharges, subsequent examinations, death certificates and autopsy records. A total of 392 individuals were found to experience cardiovascular diseases between exam 1 and exam 2 (which occurred between 1968 and 1970). Cardiovascular diseases, in this study, include AP, CI, MI, transient ischemic attack (TIA), stroke and congestive heart failure (CHF). To establish the age-related risk prediction models, participants' record was matched between exam 1 and exam 2 (i.e.

Table 7.1: Number of Instances used for Training and Validation

Prediction Model	Training Instances	Validation Instance
hhpAge4665	326	136
hhpAge4655	172	72
hhpAge5665	154	64

Instances used in both training and validation phases have the equal number of cases and controls.

follow-up examination of participants not conducted in exam 2 were removed). Finally, to mitigate class imbalance data problem (i.e. the tendency of the algorithm overwhelmed by the major class and ignores the minor one) (Japkowicz, 2000; Li et al., 2010), a balanced number of cases and controls were randomly selected. In addition, uninformative features (i.e. features with constant value for all participants) were removed, resulting in a total of 370 clinical features and 326 instances.

For each prediction model, 70% of the baseline data was used to develop/train the model (commonly referred to as the training instances) while the remaining (common known as the validation instances) was used to validate the developed model. Details of the datasets used for the different prediction models are given in Table 7.1.

7.3. Experimental Results

Several experiments were conducted to investigate on the significance of age-related risk prediction models and the classification capability of ANCSc, EDC-AIRS and SVM when applied to CVD prediction task. A total of 2 prediction paradigms, postulating either a 10-year or 20-year age model, were analysed. Development of the respective age-related risk prediction model was carried out by executing each algorithm 3 times. Consistent classification results were obtained for all the runs (i.e. standard deviation of 0).

Section 7.3.1 to 7.3.3 describes the performance obtained during the training phase for ANCSc, EDC-AIRS and SVM algorithms respectively. Section 7.3.4 provides the results achieved when the developed models were tested with their corresponding validation dataset (i.e. results obtained during validation phase). Additionally, McNemar's test results are provided to illustrate the statistical significance of the prediction outcomes.

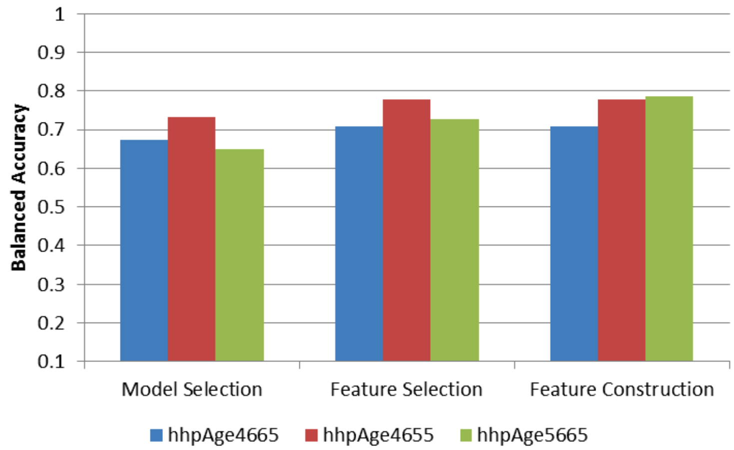
7.3.1. Age-related Risk Prediction with ANCSc algorithm

ANCSc algorithm (Tay et al., 2014) - a novel neural-inspired algorithm developed recently – was employed to perform CVD prediction. This algorithm bio-mimics the neuronal behaviour associated with the process of learning and interaction with the external environment; embracing it with the mechanisms necessary for the evolution of the neurons (i.e. candidate solution). Through this process, it promotes the generation of a set of representative neurons capable of accurately describing the underlying patterns within the data problem presented. Results, as demonstrated in (Tay et al., 2014), portend that ANCSc algorithm is a highly effective classification algorithm and has outperformed several state-of-the-art algorithms.

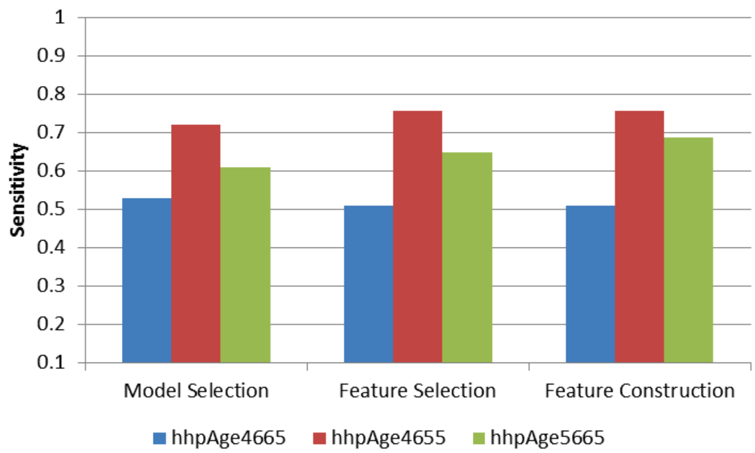
Table 7.2: Performance of ANCSc Algorithm (Training Phase)

Experiment	#Features	Sensitivity	Specificity	Balanced Accuracy
Step 1: Model Selection				
hhpAge4665	370	0.528	0.816	0.672
hhpAge4655	370	0.721	0.744	0.733
hhpAge5665	370	0.610	0.688	0.649
Step 2: Feature Selection				
hhpAge4665	184	0.509	0.908	0.709
hhpAge4655	179	0.756	0.802	0.779
hhpAge5665	192	0.649	0.805	0.727
Step 3: Feature Construction				
hhpAge4665	184	0.509	0.908	0.709
hhpAge4655	179	0.756	0.802	0.779
hhpAge5665	198	0.688	0.883	0.786

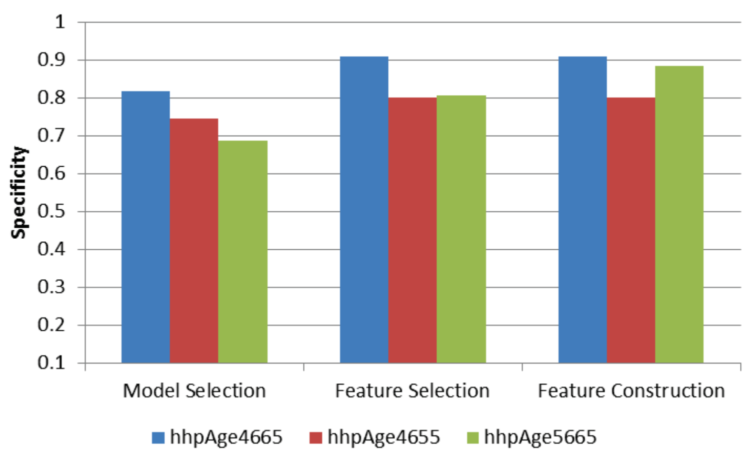
For each step, 10-fold cross-validation was conducted (with ANCSc algorithm) on each age model to build the prediction model for CVD.



(a) Balanced Accuracy Performance Metric



(b) Sensitivity Performance Metric



(c) Specificity Performance Metric

Figure 7.4: Classification Performance of ANCSc (Training Phase)
These performance measurements were obtained by performing 10-fold cross validation on each age model.

The performance of ANCSc algorithm at different optimization steps during model development is presented in Table 7.2 and Figure 7.4. During the model selection phase, the full feature set (i.e. 370 features) was used to develop the respective prediction models. The best prediction model computed for age model 'hhpAge4665' and 'hhpAge4655' is based on dissimilar data class distribution while age model 'hhpAge5665' yielded the best performing prediction model under the assumption of similar data class distribution. Results demonstrate that for age model 'hhpAge4665', the sensitivity performance achieved by the prediction model is relatively poor (0.528) in contrast to the other 2 age models (although it has achieved relatively good specificity performance – 0.816). This is not desirable as many patients who might experience MI would go undetected and in turn early preventive measures could not be offered to these patients; potentially leading to many avoidable deaths as a result. On other hand, age model 'hhpAge4655' achieved relatively good sensitivity (0.721) and specificity (0.744). This is much more desirable as patients who are likely to experience MI would have a higher chance of being detected whereby appropriate management strategies can be given early. Additionally, individuals who are healthy would be more likely to be detected as well - avoiding the need to undergo unnecessary tests, reducing the financial burden and anxiety on the patients.

Using the best prediction model obtained, feature selection was performed. A total of 184, 179 and 192 features were considered to be informative for age model 'hhpAge4665', 'hhpAge4655' and 'hhpAge5665' respectively. This corresponds to a reduction of 50.3%, 51.6% and 48.1% in feature dimensionality. Performance wise, an improvement in balanced accuracy (i.e. average between sensitivity and specificity) of 5.51%, 6.28% and 12.0% was obtained for age model 'hhpAge4665', 'hhpAge4655' and 'hhpAge5665' respectively. This reduction in the number of features (with increased performance) is highly desirable as it reduces the number of clinical tests that need to be conducted on the patients – reducing any risk, cost or emotional stress that patients might experience.

After which, feature construction was performed. No informative features could be inferred for age model 'hhpAge4665' and 'hhpAge4655'. However, 6 new features were constructed for age model 'hhpAge5665', ameliorating the balanced accuracy by 8.12%. This is a very useful technique as it ameliorates the performance of the prediction models without the need to conduct any clinical tests on the patient. It is noteworthy that the age-related prediction models outperform the unified model by approximately 10% (for balanced accuracy).

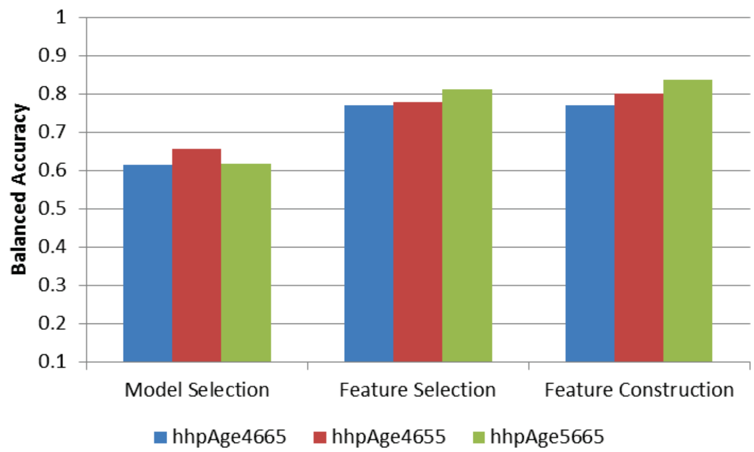
7.3.2. Age-related Risk Prediction with EDC-AIRS algorithm

EDC-AIRS algorithm is an optimized version of Artificial Immune Recognition System (AIRS2) algorithm (Watkins et al., 2004) – an immune-inspired supervised learning algorithm. EDC-AIRS algorithm extends AIRS2 algorithm by contextualizing the immune response to the concentration, distribution and characteristics of the antigens. Results, as demonstrated in (Tay et al., 2013), indicate that EDC-AIRS algorithm is a highly competitive classification algorithm that shows high fidelity to the natural immune system.

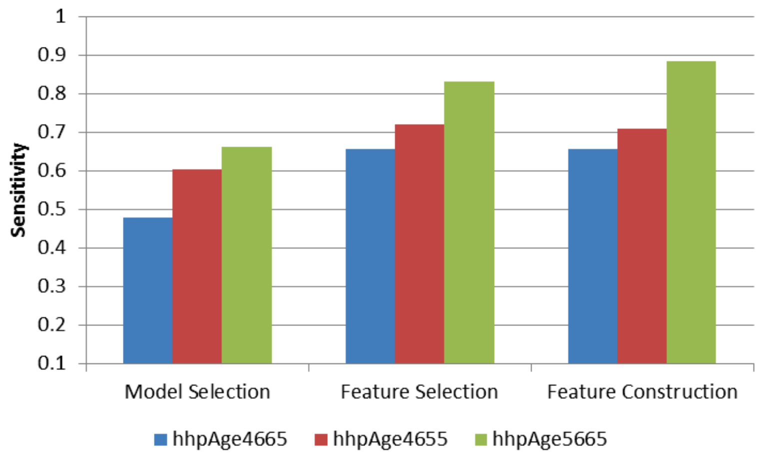
Table 7.3: Performance of EDC-AIRS Algorithm (Training Phase)

Experiment	#Features	Sensitivity	Specificity	Balanced Accuracy
Step 1: Model Selection				
hhpAge4665	370	0.479	0.748	0.614
hhpAge4655	370	0.605	0.709	0.657
hhpAge5665	370	0.662	0.571	0.617
Step 2: Feature Selection				
hhpAge4665	180	0.656	0.877	0.769
hhpAge4655	174	0.721	0.837	0.779
hhpAge5665	194	0.831	0.792	0.812
Step 3: Feature Construction				
hhpAge4665	180	0.656	0.877	0.769
hhpAge4655	175	0.709	0.895	0.802
hhpAge5665	196	0.883	0.792	0.838

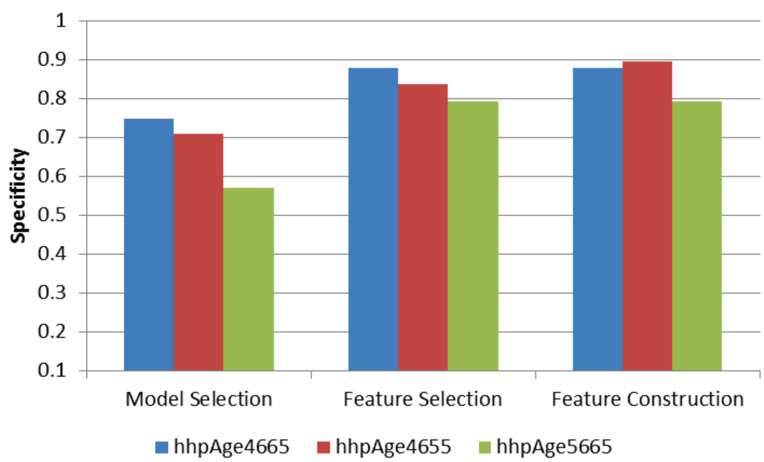
For each step, 10-fold cross-validation was conducted (with EDC-AIRS algorithm) on each age model to build the prediction model for CVD.



(a) Balanced Accuracy Performance Metric



(b) Sensitivity Performance Metric



(c) Specificity Performance Metric

Figure 7.5: Classification Performance of EDC-AIRS (Training Phase)
These performance measurements were obtained by performing 10-fold cross validation on each age model.

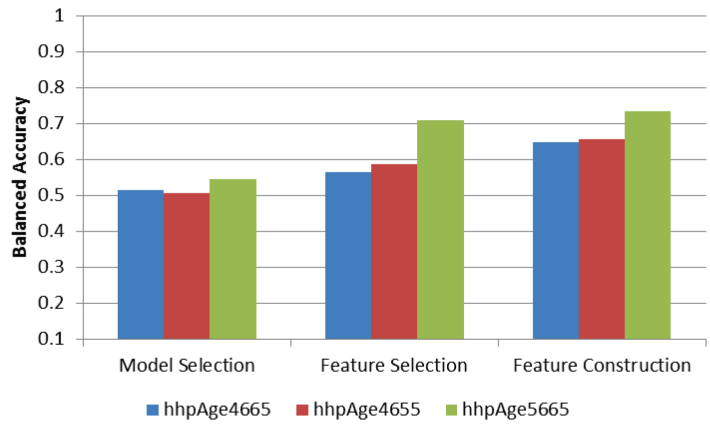
Table 7.3 and Figure 7.5 provide the performance of EDC-AIRS algorithm at different optimization steps proposed for developing the prediction models. Similar to ANCSc algorithm, model selection was first carried out with 370 features. In this case, the best performing prediction model was yielded under the assumption of dissimilar data class distribution for all age models. Results indicate that prediction model for 'hhpAge4665' achieved the worst sensitivity (0.479 - as compared to 'hhpAge4655' and 'hhpAge5665'). This is undesirable despite the much higher specificity (0.748) achieved as many patients who are likely to experience MI will go undetected. On other hand, the age-related prediction models perform relatively better. Particularly, prediction model for 'hhpAge4655' achieved relatively good balance of sensitivity (0.605) and specificity (0.709).

The performance of feature selection resulted in a reduction in feature dimensionality by 51.4%, 53.0% and 47.6% and an improvement in balanced accuracy by 25.2%, 18.6% and 31.6% for age model 'hhpAge4665', 'hhpAge4655' and 'hhpAge5665' respectively. This is highly attractive as it increases the chance of making the right diagnosis while eradicating the need to conduct a range of different tests. It is noteworthy that the age-related prediction models in general outperform the unified model.

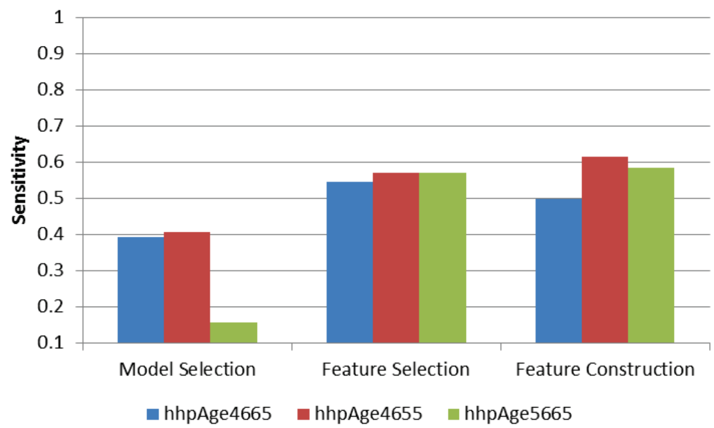
Table 7.4: Performance of SVM Algorithm (Training Phase)

Experiment	#Features	Sensitivity	Specificity	Balanced Accuracy
Step 1: Model Selection				
hhpAge4665	370	0.393	0.626	0.515
hhpAge4655	370	0.407	0.605	0.506
hhpAge5665	370	0.156	0.935	0.545
Step 2: Feature Selection				
hhpAge4665	177	0.546	0.583	0.564
hhpAge4655	168	0.570	0.605	0.587
hhpAge5665	180	0.571	0.844	0.708
Step 3: Feature Construction				
hhpAge4665	186	0.497	0.798	0.647
hhpAge4655	176	0.616	0.698	0.657
hhpAge5665	184	0.584	0.883	0.734

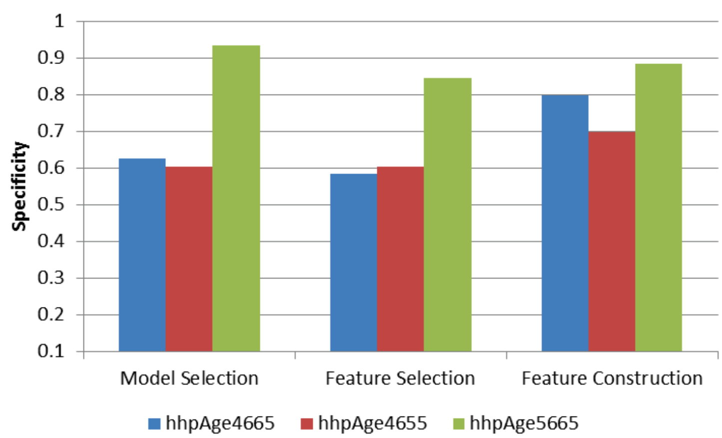
For each step, 10-fold cross-validation was conducted (with SVM algorithm) on each age model to build the prediction model for CVD.



(a) Balanced Accuracy Performance Metric



(b) Sensitivity Performance Metric



(c) Specificity Performance Metric

Figure 7.6: Classification Performance of SVM (Training Phase)

These performance measurements were obtained by performing 10-fold cross validation on each age model.

Feature construction was subsequently conducted on the reduced feature set. No new features can be inferred for age model ‘hhpAge4665’ while 1 and 2 new features were generated for age model ‘hhpAge4655’ and ‘hhpAge5665’ respectively. This resulted in an improvement in balanced accuracy by 2.95% and 3.20% for age model ‘hhpAge4655’ and ‘hhpAge5665’ respectively. With the performance of this step, the age-related prediction models remains to outperform the unified model by 4.3%-9.0% (for balanced accuracy).

7.3.3. Age-related Risk Prediction with SVM algorithm

SVM algorithm, a robust supervised learning algorithm that is capable of yielding excellent generalization performance on an extensive area of problems (Chen et al., 2005; Osuna et al., 1997; Listgarten et al., 2004), was employed. It is derived from statistical learning theory and is capable of solving linearly and non-linearly separable problems. Fundamentally, SVM performs classification through the construction of an N-dimensional hyper-plane that optimally separates the data into two or more categories whereby the margin of separation between the different categories is maximized.

Table 7.4 and Figure 7.6 provide the results achieved when SVM was trained along the 3 optimization steps postulated in Figure 7.2. In the first step, model selection, an average balanced accuracy of 0.522 was achieved for all 3 age models - such performance is near to random guess which is highly undesirable. A possible reason for SVM poor performance, in contrast to ANCSc and EDC-AIRS algorithms, is that the data contain multiple exceptional cases which SVM is bad at handling.

Next, feature selection was conducted. Improvement in balanced accuracy was seen across all 3 age models with dimensionality reduction of 52.2%, 54.6% and 51.4% for age model ‘hhpAge4665’, ‘hhpAge4655’ and ‘hhpAge5665’

Table 7.5: Performance of Developed Prediction Models (Validation Phase)

Experiment	#Features	Sensitivity	Specificity	Balanced Accuracy
ANCSc Algorithm				
hhpAge4665	184	0.456	0.794	0.625
hhpAge4655	179	0.611	0.861	0.736
hhpAge5665	198	0.531	0.781	0.656
EDC-AIRS Algorithm				
hhpAge4665	180	0.529	0.618	0.574
hhpAge4655	175	0.667	0.500	0.583
hhpAge5665	196	0.531	0.531	0.531
SVM Algorithm				
hhpAge4665	177	0.25	0.632	0.441
hhpAge4655	176	0.472	0.667	0.569
hhpAge5665	184	0.281	0.688	0.484

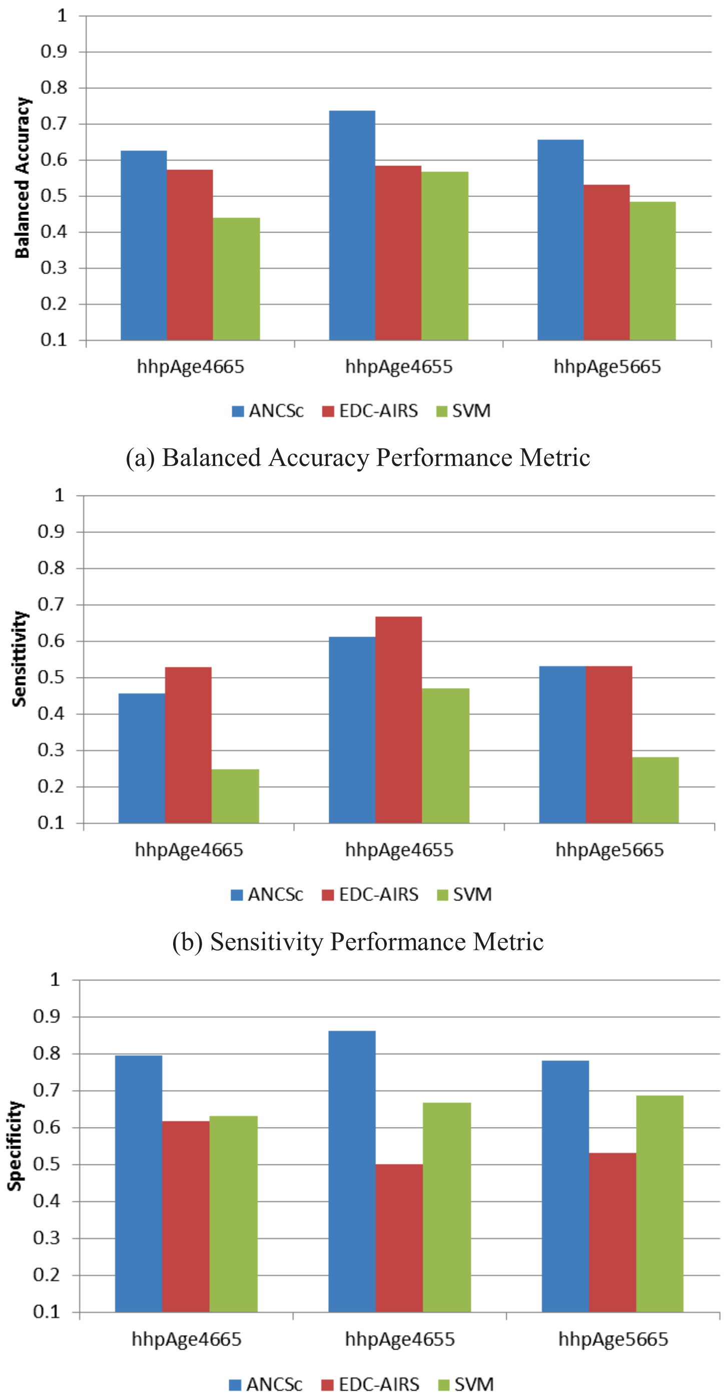
For each developed prediction model, it is validated (for generalizability) with a distinct and separate dataset.

respectively. However, the balanced accuracy is still relatively poor (around 0.57); except for age model 'hhpAge5665'. Despite the improvement in balanced accuracy (0.708) for age model 'hhpAge5665', it sensitivity is still relatively poor (0.571) – making it a less than ideal prediction model for deployment in clinical settings.

In the final step, feature construction, 9, 8 and 4 new features were inferred for age model 'hhpAge4665', 'hhpAge4655' and 'hhpAge5665' respectively. An improvement of 14.7%, 11.9% and 3.67% in balanced accuracy was obtained for age model 'hhpAge4665', 'hhpAge4655' and 'hhpAge5665' respectively. It is noteworthy that the sensitivity for all prediction models remains poor and the overall performance achieved by SVM is the worst compared to ANCSc and EDC-AIRS algorithms.

7.3.4. Validation of Developed Prediction Models

For each of the age model developed independently by the corresponding learning algorithm, validation of model generalizability (an estimate of how well the prediction models would perform when deployed in real clinical



(a) Balanced Accuracy Performance Metric

(b) Sensitivity Performance Metric

(c) Specificity Performance Metric

Figure 7.7: Classification Performance of ANCSc, EDC-AIRS and SVM (Validation Phase)

These performance measurements were obtained by evaluating each developed prediction model with their respective validation dataset.

Table 7.6: Statistical Evaluation of the Developed Prediction Models

Dataset	McNemar's Test# (p-value)	
	ANCSc vs EDC-AIRS	ANCSc vs SVM
hhpAge4665	0.262	**0.002**
hhpAge4655	**0.022**	**0.019**
hhpAge5665	0.131	**0.041**

#The p-value of McNemar's test is presented, examining whether the performance of ANCSc algorithm (statistically) outperformed EDC-AIRS and SVM algorithms.

settings) was conducted. Results, presented in Table 7.5 and Figure 7.7, indicate that ANCSc algorithm outperformed the other 2 algorithms for all age models. Improvement was at least 8.89%, 26.2% and 23.5% for age model 'hhpAge4665', 'hhpAge4655' and 'hhpAge5665' respectively.

McNemar's test, a statistical test used to compare 2 paired binomial samples, was performed to determine whether one algorithm (i.e. ANCSc) outperformed another (i.e. EDC-AIRS or SVM). It is conducted for all (validated) age models and the p-values obtained are given in Table 7.6. From the results, it can be observed that ANCSc algorithm outperformed EDC-AIRS algorithm for age model 'hhpAge4655', and SVM algorithm for all 3 age models.

It is noteworthy that age model 'hhpAge4655' developed using the 3 different algorithms in general performs the best when compared to the other 2 age models while age model 'hhpAge5665' performs comparably with the unified prediction mode (i.e. 'hhpAge4665'). This suggests that it is advantageous for us to build prediction models that are age specific.

7.4. Discussion

The ability to predict the first age-related CVD event experienced by individuals stratified to different age group was investigated. Three learning

Table 7.7: Clinical Features Unique to Modelling Age Model 'hhpAge4655' and 'hhpAge5665'

Age model 'hhpAge4655'	Age model 'hhpAge5665'
Fish intake	Past weight
Sausage intake	Chest pain
Fruit intake	Serum cholesterol
Beverage intake	Blood pressure
Total carbohydrate intake	Cancer prevalence
Percentage calories protein intake	Tiffeneau-Pinelli index

Age models 'hhpAge4655' and 'hhpAge5665' each has 6 unique clinical features.

algorithms (i.e. ANCSc, EDC-AIRS and SVM algorithms) were utilized to develop these prediction models. A total of 3 optimization steps were postulated to develop the prediction models – i.e. model selection, feature selection and feature construction. In the first step (model selection), results indicate that ANCSc algorithm, compared to other algorithms evaluated, yielded higher performance for most of the performance metrics computed – ranging from 5.19% to 44.9% higher for balanced accuracy.

In the second step, feature selection was conducted. An overall improvement in the performance of all prediction models (when evaluated via 10-fold cross-validation) was observed. The algorithm that accrued the most benefits from this optimization step is EDC-AIRS algorithm - achieving a minimum improvement of 216.3%, 50.6% and 19.6% in balanced accuracy (compared to the other 2 algorithms) for age model 'hhpAge4665', 'hhpAge4655' and 'hhpAge5665' respectively. ANCSc algorithm, on the other hand, yielded the smallest amount of improvement. This suggests that ANCSc algorithm, which achieved similar results to EDC-AIRS algorithm, is robust to high data dimensionality. Broadly, EDC-AIRS algorithm achieved the best performance for all age models in this step.

The percentage of features reduced (upon performing feature selection) ranges from 47.6% to 54.6% for all age models and algorithms. Clearly, this signifies that there are several redundant and irrelevant features present in the original clinical feature set; hindering the construction of accurate prediction models. It is noteworthy that the number of common features (after feature selection) presents between age model 'hhpAge4655' and 'hppAge5665' is 95, 102 and 84 for ANCSc, EDC-AIRS and SVM algorithms respectively. Some of the key common features identified among all 3 algorithms include: age, place of birth, medication (particularly anti-hypertension medicine), history of CVD manifestation, and amount of milk, kamoboko, safflower oil, alcohol, caffeine, complex and simple carbohydrate intake. Concerns over alcohol intake and hypertension for individuals aged between 46 and 65 dovetail with the results stated in (Abbott et al., 2002). These common features identified overlap approximately 50% of the features selected for age model 'hhpAge4655' and 'hppAge5665'. This potentially portends that clinical features having statistical properties dissimilar between age models 'hhpAge4655' and 'hppAge5665' are required to better model the characteristics of individuals in different age group.

Among the 3 algorithms analysed, key clinical features unique to describing age model 'hhpAge4655' and 'hhpAge5665' are listed in Table 7.7. Upon examining these results, it is suggestive that clinical features unique to age model 'hhpAge4655' could be potential precursors to subsequent health risk for individuals aged between 56 and 65.

In the third step, feature construction, improvement can be observed for all 3 algorithms and is most observable for age models 'hhpAge4655' and 'hhpAge5665'. This signifies that age-related risk prediction models are more sensitive to the health characteristics of individuals; resulting in increased capability of discriminating the cases from the controls. Although the number of new features added is relatively small (ranges from 1 to 8), the improvement garnered by the respective algorithm is considerable (between 2.95% to 11.9%). It is notable that EDC-AIRS algorithm outperforms the other 2 algorithms after performing feature construction.

Finally, the generalizability of the developed prediction models was determined. Clearly from the results, ANCSc algorithm yielded better overall performance than either EDC-AIRS or SVM algorithms (an improvement of at least 8.89%, 26.2% and 23.5% in balanced accuracy for age model 'hhpAge4665', 'hhpAge4655' and 'hhpAge5665' respectively). This entails that ANCSc algorithm is a robust and versatile learning algorithm more capable at performing risk prediction task for CVD and less likely to be over-trained. Moreover, it can be observed that age-related risk prediction models (i.e. age models 'hhpAge4655' and 'hhpAge5665') developed using any of the 3 algorithms are capable of achieving comparable, if not better, performance than a unified risk prediction model (i.e. age model 'hhpAge4665').

One limitation of this study is that investigation is only restricted to individuals aged between 46 and 65. This constrained our ability to conduct a more comprehensive analysis (i.e. over a wider age range) and to determine the full impact of age-related risk prediction.

7.5. Summary

We have investigated on the capability of ANCSc, EDC-AIRS and SVM algorithms to develop age-related risk prediction models for CVD. Model selection, feature selection and feature construction were performed in sequence for all algorithms in order to adopt a ceteris paribus experimental design. Results indicate that both feature selection and feature construction contribute significantly to the development of more accurate prediction models. Validation of the developed risk prediction models demonstrated that ANCSc algorithm is capable of generalizing better than EDC-AIRS and SVM algorithms. Furthermore, age-related risk prediction approach was shown to perform better than unified risk prediction approach for all algorithms investigated.

In terms of clinical impact, we believe that it has a significant contribution as it provides an easy to use prediction tool that could allow more precise and early diagnosis to be carried out, promote better understanding of risk factor evolution and disease's etiology, among others.

Disclaimer

The Honolulu Heart Program dataset described in this chapter is provided by the NHLBI.

Chapter 8

Conclusions and Future Work

8.1. Summary of Thesis Achievements

Currently, prediction models in medicine tend to be restricted to a specific domain, set of clinical data, and instant in time. Further, efficient development of effective prediction models from a deluge of complex clinical data is also a challenge. This ultimately limits our ability to provide accurate personalized prediction and offer strategies for continuum of care. The capability to do so is very important as it would (1) decrease the rate of misdiagnosis, (2) reduce avoidable mortality, (3) provide the highest quality of continuous care, (4) minimise the discomfort, pain, or anxiety that is associate with a disease through early detect, management and treatment, and (5) improve the life of individuals. Therefore, in this thesis, new approaches for efficient development of accurate clinical prediction models are presented; aiming to promote the advancement towards personalized, preventative and predictive medicine.

In chapter 3, we have demonstrated that with the employment of the biological continuum, up-to-date clinical classification models can be developed efficiently. Compared with the conventional approach, our method achieved a speedup of approximately 5-fold. Efficient development of clinical classification models is highly desirable as new biomarkers are constantly being introduced; this entails that analyse of the new biomarkers with the plethora of existing ones are necessary for the development of more accurate classification models. Hence, with our approach of analysing the deluge of biomarkers, continuous development of up-to-date clinical classification models would be better embraced by the clinical research community. Moreover, the etiological network (i.e. BCEN) constructed from the study has the potential to illustrate significant risk factors and provide the classification model for each subclinical manifestation identified. This, we believe, is a crucial step to monitor one's health along the continuum of care.

Chapter 4 introduced an optimized immune-inspired supervised classification algorithm called EDC-AIRS. The development of robust and accurate learning algorithm is critical for many tasks, including the development of personalized predictive models. Therefore, through the observation of how natural immune system works to protect us from foreign antigens, we bio-mimic the mechanisms postulated by the nature and improved the existing AIRS2 algorithm. Results portend that inspiration from the natural immune system could leverage our insights and enhance our ability to solve computational problems in a creative, effective and efficient manner.

Chapter 5 employed the SVM and EDC-AIRS algorithms for performing time-related risk prediction for MI. Detailed considerations were given to risk prediction over different prediction resolution (i.e. prediction time scale and interval), and the use of different sample age (i.e. baseline data comprising of individuals in different age range). Results indicate that both prediction resolution and sample age do not have a significant impact on the performance of MI risk prediction models developed using subjects aged 65 and above. This portends that risk prediction models developed using different sample age and prediction resolution is a feasible approach and could offer patients with a more comprehensive estimation of their health risk.

In chapter 6, we described a novel neural-inspired supervised classification algorithm called ANCSc. This algorithm bio-mimics the mechanisms responsible for the development and enrichment of the human brain. The key mechanisms include neurogenesis, neuroplasticity, nurturing and apoptosis. Benchmark testing results show that ANCSc algorithm is capable of achieving highly competitive classification performance. This portends that ANCSc algorithm is a robust algorithm capable of adapting to different profound data patterns and structures. From this study, we have again demonstrated that the nature is a wonderful source for inspiration where researchers can learn and develop techniques to solve many engineering and science problems.

Finally, in chapter 7, the effect of evolving CVD risk factors on the performance of risk prediction models (built using machine learning techniques) was taken into consideration. Results indicate that the performance of risk prediction models can be improved when they are constructed with data consisting of individuals stratified to different age group.

To summarize, we have developed 2 new algorithms and demonstrated the importance of 3 continuum models – namely biological, time and age continuum models. We hypothesize that analysis of health characteristics along continuum models is of major importance and has the significant advantage of leveraging the quality of continuous healthcare an individual can benefit from.

Table 8.1: Classification Performance Achieved on Different Datasets

	Dataset	EDC-AIRS	ANCSc	SVM
UCI Benchmarking Datasets	Iris	**99.6%**	99.1%	98.7%
	Ionosphere	97.4%	**98.0%**	**98.0%**
	Diabetes	77.3%	**77.5%**	76.7%
	Sonar	88.5%	**91.8%**	88.5%
	Wine	**99.6%**	99.3%	82.0%
	Heart	84.8%	**86.3%**	77.4%
CHS Dataset	yr50611	78.6%	78.9%	**90.1%**
	yr50607	71.4%	78.6%	**89.7%**
	yr50809	71.9%	78.9%	**86.0%**
	yr51011	75.0%	80.6%	**89.8%**
	yr70811	77.0%	74.1%	**81.0%**
	yr70809	71.1%	80.0%	**82.2%**
	yr71011	71.4%	85.7%	**88.1%**
	yr91011	71.7%	70.0%	**83.3%**
HHP dataset	hhpAge4665	57.4%	**62.5%**	44.1%
	hhpAge4655	58.3%	**73.6%**	56.9%
	hhpAge5665	53.1%	**65.6%**	48.4%

The SVM classification performance presented in this table is based on the version used in Chapter 7 of this thesis.

Furthermore, we believe that disease prevention should be the ethos of future healthcare and treatment/surgery should not dominate the clinical practice. In this regard, healthcare professionals and researchers should recognize the need for this transition and invest efforts into the realization of this healthcare transformation.

A summary of the prediction performance achieved by SVM, EDC-AIRS and ANCSc algorithms tested on different datasets is given in Table 8.1. From the table, it can be observed that EDC-AIRS and ANCSc algorithms (a type of instance-based classifier) tend to achieve similar performance when tested on the CHS and HHP datasets while SVM algorithm (a type of discriminative classifier) tends to achieve predictive performance of its kind. One possible reason for this phenomenon is that SVM tends to perform poorly on problems with exceptional cases while EDC-AIRS and ANCSc algorithms tend to be vulnerable to noisy and irrelevant features. This potentially suggests that CHS dataset comprises of noisy and irrelevant clinical features while HHP dataset contains multiple exceptional clinical cases. Similar explanation can be extrapolated for results achieved on the UCI benchmarking datasets.

8.2. Future Work

The final consideration is into future directions of research for personalized predictive models. Whilst the results demonstrated in this thesis have shown some approaches for enhancing the quality of predictive models, there is still much work that needs to be done. Some of the potential future researches that can be explored include:

1. Analysis of a more comprehensive set of biomarkers across the biological continuum (e.g. proteomic and genomic data). This would allow the discovery of highly relevant and predictive biomarkers that can better anticipate the progression or events of a disease.

2. Investigation and derivation of methods for translating risk factors that is statistically important to one that can fill the puzzle of the disease's pathology and use as part of current clinical practice. This would require risk factors that are identified through computational means to be validated by clinical experts using approaches like prospective clinical study.

3. Exploration of methods for the development of accurate prediction models that will become an important and indispensable component in clinical practice. This can potentially be achieved through the (1) collection of more predictive, relevant and specific biomarkers, and (2) development of more accurate and robust learning algorithms that can be used to perform baseline risk assessment and selection of appropriate therapeutic strategies.

4. Exploration of methods to seamlessly incorporate CDSS into routine clinical practice in an attempt to improve diagnosis, change patients' behaviour and subsequent healthcare outcome.

5. Investigation of the feasibility to monitor, detect and manage patients' well-being along the continuum of health (i.e. prevent, detect or treat subclinical manifestation before they are of clinical significance; causing damages that are irreversible).

6. Development of support system capable of offering real-time assistance. This is important as currently there is limited support at the patients' bedside to assist healthcare professionals to deliver the best standard of care. Hence, development of accurate and robust online learning algorithms is necessary for (1) monitoring and detecting anomalies in real-time clinical data, and (2) providing reliable recommendation instantaneously.

7. Translation of research into clinical practice by integrating strategies that promotes personalized, predictive and preventative medicine into the current state-of-the-art CDSS.

References

Abbott, R., Curb, J., Rodriguez, B., Masaki, K., Yano, K., Schatz, I., Ross, G., & Petrovitch, H. (2002). Age-related Changes in Risk Factor Effects on the Incidence of Coronary Heart Disease. *Ann Epidemiol, 12*(3), 173-181.

Acuña, E., & Rodriguez, C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. *Studies in Classification, Data Analysis, and Knowledge Organization* pp. 639-647).

AlTimemy, A. H. A., & Naima, F. M. A. (2010). Comparison of Different Neural Network Approaches for the Prediction of Kidney Dysfunction. *International Journal of Biological and Life Science, 2*

Alty, S., Angarita-Jaimes, N., Millasseau, S., & Chowienczyk, P. (2007). Predicting Arterial Stiffness from the Digital Volume Pulse Waveform. *IEEE Trans Biomed Eng., 54*(12), 2268-2275.

Alvarez, S. M., Poelstra, B. A., & Burd, R. S. (2006). Evaluation of a Bayesian Decision Network for Diagnosing Pyloric Stenosis. *J Pediatr Surg, 41*(1), 155-161.

Anand, S. S., Islam, S., Rosengren, A., Franzosi, M. G., Steyn, K., Yusufali, A. H., Keltai, M., Diaz, R., Rangarajan, S., & Yusuf, S. (2008). Risk factors for myocardial infarction in women and men: insights from the INTERHEART study. *Eur Heart J, 29*(7), 932-940.

Asia Pacific Cohort Studies Collaboration (2006). The Impact of Cardiovascular Risk Factors on the Age-related Excess Risk of Coronary Heart Disease. *Int J Epidemiol, 35*(4), 1025-1033.

Austin, C. (2003). The Impact of the Completed Human Genome Sequence on the Development of Novel Therapeutics for Human Disease. *Annu Rev Med, 55*, 1-13.

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., Filho, W. J., Lent, R., & Herculano-Houzel, S. (2009). Equal Numbers of Neuronal and Nonneuronal Cells make the Human Brain an Isometrically Scaled-up Primate Brain. *Journal of Comparative Neurology, 513*(5), 532-541.

Barakat, N., Bradley, A., & Barakat, M. (2010). Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. *IEEE Trans Inf Technol Biomed., 14*(4), 1114-1120.

Bates, D. W., Teich, J. M., Lee, J., Seger, D., Kuperman, G. J., Ma'Luf, N., Boyle, D., & Leape, L. (1999). The impact of computerized physician order entry on medication error prevention. *J Am Med Inform Assoc, 6*(4), 313-321.

Bates, D. W., Kuperman, G. J., Wang, S., Gandhi, T., Kittler, A., Volk, L., Spurr, C., Khorasani, R., Tanasijevic, M., & Middleton, B. (2003). Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-based Medicine a Reality. *J Am Med Inform Assoc., 10*(6), 523-530.

Batista, G., & Monard, M. (2002). *A Study of K-Nearest Neighbor as an Imputation Method.* Paper presented at the Second International Conference on Hybrid Intelligent Systems (pp. 251-260). Santiago, Chile: Soft Computing Systems: Design, Management and Applications.

Batista, G., & Monard, M. C. (2003). *A Study of K-Nearest Neighbour as an Imputation Method.* Paper presented at the In HIS

Baxt, W. (1991). Use of an Artificial Neural Network for the Diagnosis of Myocardial Infarction. *Ann Intern Med., 115*(11), 843-848.

Baxt, W., & Skora, J. (1996). Prospective Validation of Artificial Neural Network Trained to Identify Acute Myocardial Infarction. *Lancet., 347*(8993), 12-15.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *Int J Med Inform., 77*(2), 81-97.

Bellman, R. (1961). *Adaptive Control Processes.* Princeton New Jersey: Princeton University Press.

Bhatla, N., & Jyoti, K. (2012). An Analysis of Heart Disease Prediction using Different Data Mining Techniques. *International Journal of Engineering Research & Technology, 1*(8), 1-4.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's Razor. *Information Processing Letters, 24*(6), 377-380.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A Training Algorithm for Optimal Margin Classifiers.* Paper presented at the Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152). Pittsburgh, Pennsylvania, United States: ACM New York, NY, USA.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. (1984). *Classification and Regression Trees.* (1 ed.). Chapman and Hall/CRC.

Bright, T. J., Wong, A., Dhurjati, R., Bristow, E., Bastian, L., Coeytaux, R. R., Samsa, G., Hasselblad, V., Williams, J. W., Musty, M. D., Wing, L., Kendrick, A. S., Sanders, G. D., & Lobach, D. (2012). Effect of Clinical Decision-Support Systems: A Systematic Review. *Ann Intern Med, 157*(1), 29-43.

British Heart Foundation Statistics Database (2010). *Coronary Heart Disease.* Retrieved August 8, 2013, from http://www.bhf.org.uk/publications/view-publication.aspx?ps=1001546

Brodley, C. (1993). *Addressing the Selective Superiority Problem: Automatic Algorithm/Model Class Selection.* Paper presented at the In Proc. 10th Machine Learning Conf. (pp. 17-24).

Brownlee, J. *"artificial Immune Recognition System (airs). a Review and Analysis", Technical Report No. 1-02, Centre for Intelligent Systems and Complex Processes (ciscp), Faculty of Information and Communication Technologies (ict), Swinburne University of Technology (sut), Victoria, Australia: 2005.*

C.L. Blake, & C.J. Merz (1998). *UCI Repository of Machine Learning Databases.* Retrieved 8/11/2012, from http://www.ics.uci.edu/~mlearn/MLRepository.html

Cannon, C. P., & O'Gara, P. T. Cannon, C. P., & O'Gara, P. T. (Ed.). (2007). *Critical pathways in cardiovascular medicine.* (2 ed.). Lippincott WilliamsWilkins.

Castro, L. N. D., & Timmis, J. (2002). *Artificial Immune Systems: A New Computational Intelligence Approach.* (1st ed.). Springer.

Chang, C. C., & Lin, C. (2001). *LIBSVM: A Library for Support Vector Machines.* from http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Chattopadhyay, S. (2013). Mining the Risk of Heart Attack: A Comprehensive Study. *International Journal of Biomedical Engineering and Technology, 11*(4)

Chawla, N., & Davis, D. (2013). Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework. *J Gen Intern Med.,*

Cheah, J. (2000). Clinical pathways : An Evaluation of its Impact on the Quality of Care in an Acute Care General Hospital in Singapore. *Singapore Med J, 41*(7), 335-346.

Chen, S. H., Sun, J., Dimitrov, L., Turner, A. R., Adams, T. S., Meyers, D. A., Chang, B. L., Zheng, S. L., Grönberg, H., Xu, J., & Hsu, F. C. (2008). A Support Vector Machine Approach for Detecting Gene-Gene Interaction. *Genet Epidemiol, 32*(2), 152-167.

Chen, W. H., Hsu, S. H., & Shen, H. P. (2005). Application of SVM and ANN for Intrusion Detection. *Computers & Operations Research, 32*(10), 2617-2634.

Chipperfield, A., & Fleming, P. (1995). *The MATLAB Genetic Algorithm Toolbox.* Paper presented at the IEE Colloquium on Applied Control Techniques Using MATLAB

Chow, R., Zhong, W., Blackmon, M., Stolz, R., & Marsha Dowell (2008). *An Efficient SVM-GA Feature Selection Model for Large Healthcare Databases.* Paper presented at the Proceedings of the 10th annual conference on Genetic and evolutionary computation (GECCO) (pp. 1373-1380).

Chu, C., Hsu, A., Chou, K., Bandettini, P., & Lin, C. (2012). Does Feature Selection Improve Classification Accuracy? Impact of Sample Size and Feature Selection on Classification Using Anatomical Magnetic Resonance Images. *Neuroimage., 60*(1), 59-70.

Chu, S., & Cesnik, B. (1998). Improving clinical pathway design: lessons learned from a computerised prototype. *Int J Med Inform, 51*(1), 1-11.

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artif Intell Med, 26*(1), 1-24.

Clayton, T., Lubsen, J., Pocock, S., Vokó, Z., Kirwan, B., Fox, K., & Poole-Wilson, P. (2005). Risk Score for Predicting Death, Myocardial Infarction, and Stroke in Patients with Stable Angina, based on a Large Randomised Trial Cohort of Patients. *BMJ, 331*, 869-873.

Coiera, E., Magrabi, F., & Sintchenko, V. (2003). *Guide to Health Informatics.* (2 ed.). CRC Press.

College of Fellows, A. I. F. M. A. B. E. (2013). Medical and Biological Engineering in the Next 20 Years: The Promise and the Challenges. *IEEE Trans Biomed Eng., 60*(7), 1767-1775.

Collins, F., & Galas, D. (1993). A New Five Year Plan for the US Human Genome Project. *Science, 262*, 43-46.

Corbeil, P., Simoneau, M., Rancourt, D., Tremblay, A., & Teasdale, N. (2001). Increased Risk for Falling Associated with Obesity: Mathematical Modeling of Postural Control. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on,* (2), 126-136.

Corrales-Medina, V., Valayam, J., Serpa, J., Rueda, A., & Musher, D. (2011). The Obesity Paradox in Community-Acquired Bacterial Pneumonia. *International Journal of Infectious Diseases, 15*(1)

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*(3), 273-297.

Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. *Information Theory, IEEE Transactions on,* (1), 21-27.

Cruz, J., & Wishart, D. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics, 2*, 59-77.

Daemen, A., Gevaert, O., & Moor, B. D. (2007). *Integration of Clinical and Microarray Data with Kernel Methods.* Paper presented at the IEEE EMBS (pp. 5411-5415).

Dash, M., & Liu, H. (1997). Feature Selection for Classification. *Intelligent Data Analysis, 1*(3), 131-156.

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput, 10*(7), 1895-1924.

Duch, W. (2000). *Datasets used for Classification: Comparison of Results.* Retrieved 07/06/2013, from http://www.is.umk.pl/projects/datasets.html

Duch, W. (2000). *Logical Rules Extracted from Data.* Retrieved 07/06/2013, from http://www.is.umk.pl/projects/rules.html

Dudani, S. (1976). The Distance-Weighted k-Nearest-Neighbor Rule. *Systems, Man and Cybernetics, IEEE Transactions on, 6*(4), 325-327.

Eftekhar, B., Mohammad, K., Ardebili, H., Mohammad, G., & Ketabchi, E. (2005). Comparison of Artificial Neural Network and Logistic Regression Models for Prediction of Mortality in Head Trauma based on Initial Clinical Data. *BMC Med Inform Decis Mak, 5*(3)

Eggers, K., Ellenius, J., Dellborg, M., Groth, T., Oldgren, J., Swahn, E., & Lindahl, B. (2007). Artificial Neural Network Algorithms for Early Diagnosis of Acute Myocardial Infarction and Prediction of Infarct Size in Chest Pain Patients. *Int J Cardiol., 114*(3), 366-374.

Ekdahl, C. T., Claasen, J. H., Bonde, S., Kokaia, Z., & Lindvall, O. (2003). *Inflammation is Detrimental for Neurogenesis in Adult Brain.* Paper presented at the Proceedings of the National Academy of Sciences (pp. 13632-13637).

Elias, M. F., Elias, P. K., Sullivan, L. M., Wolf, P. A., & D'Agostino, R. B. (2003). Lower Cognitive Function in the Presence of Obesity and Hypertension: The Framingham Heart Study. *Int J Obes, 27*(2), 260-268.

Emily, M., Mailund, T., Hein, J., Schauser, L., & Schierup, M. (2009). Using Biological Networks to Search for Interacting Loci in Genome-wide Association Studies. *Eur J Hum Genet., 17*(10), 1231-1240.

Eslami, S., Keizer, N. F. D., & Abu-Hanna, A. (2008). The impact of computerized physician medication order entry in hospitalized patients—A systematic review. *Int J Med Inform, 77*(6), 365-376.

Evans, W., & Relling, M. (1997). Pharmacogenomics: Translating Functional Genomics into Rational Therapeutics. *Science, 286*(5439), 487-491.

Every, N. R., Hochman, J., Becker, R., Kopecky, S., & Cannon, C. P. (2000). Critical Pathways: A Review. *American Heart Association, 101*, 461-465.

Fang, K. T., Lin, D. K. J., Winker, P., & Zhang, Y. (2000). Uniform Design: Theory and Application. *Technometrics, 42*(3), 237-248.

Fernald, G., Capriotti, E., Daneshjou, R., Karczewski, K., & Altman, R. (2011). Bioinformatics Challenges for Personalized Medicine. *Bioinformatics., 27*(13), 1741-1748.

Fowler, C. D., Liu, Y., Ouimet, C., & Wang, Z. (2002). The Effects of Social Environment on Adult Neurogenesis in the Female Prairie Vole. *J Neurobiol, 51*(2), 115-128.

Fransen, E., Topsakal, V., Hendrickx, J. J., Laer, L. V., Huyghe, J. R., Eyken, E. V., Lemkens, N., Hannula, S., Mäki-Torkko, E., & Jensen, M. (2008). Occupational Noise, Smoking, and a High Body Mass Index are Risk Factors for Age-related Hearing Impairment and Moderate Alcohol Consumption is Protective: A European Population-based Multicenter Study. *JARO - Journal of the Association for Research in Otolaryngology, 9*(3), 264-276.

Freitas, A., & Timmis, J. (2007). Revisiting the Foundations of Artificial Immune Systems for Data Mining. *IEEE Transactions on Evolutionary Computation, 11*(4), 521-540.

Fried, L. P., Borhani, N., Enright, P., Furberg, C. D., Gardin, J. M., Kronmal, R. A., Kuller, L. H., Manolio, T. A., Mittelmark, M. B., & Newman, A., et al. (1991). The cardiovascular health study: Design and rationale. *Ann Epidemiol, 1*(3), 263-276.

Fried, L., Borhani, N., Enright, P., Furberg, C., Gardin, J., Kronmal, R., Kuller, L., Manolio, T., Mittelmark, M., Newman, A., & et, A. (1991). The Cardiovascular Health Study: Design and Rationale. *Ann Epidemiol., 1*(3), 263-276.

Fried, L., Kronmal, R., Newman, A., Bild, D., Mittelmark, M., Polak, J., Robbins, J., & Gardin, J. (1998). Risk Factors for 5-year Mortality in Older Adults: The Cardiovascular Health Study. *JAMA., 279*(8), 585-592.

Gage, F. (2002). Neurogenesis in the Adult Brain. *The Journal of Neuroscience, 22*(3), 612-613.

García-Gómez, J. M., Vidal, C., Martí-Bonmatí, D. L., Galant, J., Sans, N., Robles, M., & Casacuberta, F. (2004). Benign/Malignant Classifier of Soft Tissue Tumors using MR Imaging. *Magnetic Resonance Materials in Physics, Biology and Medicine, 16*(4), 194-201.

Garcia-Laencina, P., Vidal, A., & Sancho-Gomez, J. L. (2008). *A Robust Approach for Classifying Unknown Data in Medical Diagnosis Problems.* Paper presented at the IEEE World Automation Congress (WAC) (pp. 1-6). Hawaii, HI:

Garg, A. X., Adhikari, N. K. J., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., Sam, J., & Haynes, R. B. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA, 293*(10), 1223-1238.

Geert, M., Güiza, F., Jan, R., & Maurice, B. (2009). Machine learning techniques to examine large patient databases. *Best Practice & Research Clinical Anaesthesiology, 23*(1), 127-143.

Ginsburg, G. S., & Willard, H. F. (2009). *Essentials of Genomic and Personalized Medicine.* (1 ed.). Academic Press.

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Borden, W. B., Bravata, D. M., Dai, S., Ford, E. S., & Fox, C. S., et al. (2013). Heart Disease and Stroke Statistics--2013 Update: A Report from the American Heart Association. *Circulation, 127*, 6-245.

Golledge, J., Leicht, A., Crowther, R. G., Clancy, P., Spinks, W. L., & Quigley, F. (2007). Association of Obesity and Metabolic Syndrome with the Severity and Outcome of Intermittent Claudication. *Journal of Vascular Surgery, 45*(1), 40-46.

Gonzalez-Angulo, A., Hennessy, B., & Mills, G. (2010). Future of Personalized Medicine in Oncology: A Systems Biology Approach. *J Clin Oncol., 28*(16), 2777-2783.

Gordon, E., & Koslow, S. (2010). *Integrative Neuroscience and Personalized Medicine.* (1 ed.). Oxford University Press, USA.

Gould, E., Beylin, A., Tanapat, P., Reeves, A., & Shors, T. (1999). Learning Enhances Adult Neurogenesis in the Hippocampal Formation. *Nat Neurosci, 2*, 260-265.

Gould, E., & Tanapat, P. (1999). Stress and Hippocampal Neurogenesis. *Biol Psychiatry, 46*(11), 1472-1479.

Gray, D., Bray, G., Gemayel, N., & Kaplan, K. (1989). Effect of Obesity on Bioelectrical Impedance. *American Society for Clinical Nutrition, 255–60*, 255-260.

Guerra, S., Sherrill, D. L., Bobadilla, A., Martinez, F. D., & Barbee, R. A. (2002). The Relation of Body Mass Index to Asthma, Chronic Bronchitis, and Emphysema. *Chest, 122*(4), 1256-1263.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Habot-Wilner, Z., & Belkin, M. (2005). Obesity is a Risk Factor for Eye Diseases. *Harefuah., 144*(11), 805-809.

Harrison, R. F., Marshall. Stephen J, & Kennedy, R. L. (1991). *The Early Diagnosis of Heart Attacks: A Neurocomputational Approach.* Paper presented at the International Joint Conference on Neural Networks (pp. 1-5). Seattle:

Heer, J., Card, S. K., & Landay, J. A. (2005). *Prefuse: A Toolkit for Interactive Information Visualization.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems Portland, Oregon, USA:

Hillestad, R., Bigelow, J., Bower, A., Girosi, F., Meili, R., Scoville, R., & Taylor, R. (2005). Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Aff (Millwood), 24*(5), 1103-1117.

Himes, C. L. (2000). Obesity, Disease, and Functional Limitation in Later Life. *Demography, 37*(1), 73-82.

Holland, J. (1992). Genetic Algorithms. *Sci Am,* , 66-72.

Holliday, K., McWilliams, D., Maciewicz, R., Muir, K., Zhang, W., & Doherty, M. (2011). Lifetime Body Mass Index, Other Anthropometric Measures of Obesity and Risk of Knee or Hip Osteoarthritis in the GOAL Case-Control Study. *Osteoarthritis and Cartilage, 19*(1), 37-43.

Hossain, J., FazlidaMohdSani, N., Mustapha, A., & SurianiAffendey, L. (2013). Using Feature Selection as Accuracy Benchmarking in Clinical Data Mining. *Journal of Computer Science, 9*(7), 883-888.

Hsia, T., Chiang, H., Chiang, D., Hang, L., Tsai, F., & Chen, W. (2003). Prediction of Survival in Surgical Unresectable Lung Cancer by Artificial Neural Networks including Genetic Polymorphisms and Clinical Parameters. *J Clin Lab Anal., 17*(6), 229-234.

Huang, Y., McCullagh, P., Black, N., & Harper, R. (2007). Feature Selection and Classification Model Construction on Type 2 Diabetic Patients' Data. *Artif Intell Med, 41*(3), 251-262.

Huanga, C. L., & Wangb, C. J. (2006). A GA-based Feature Selection and Parameters Optimization for Support Vector Machines. *Expert Systems with Applications, 31*(2), 231-240.

Hubert, H. B., Feinleib, M., McNamara, P. M., & Castelli, W. P. (1983). Obesity as an Independent Risk Factor for Cardiovascular Disease: A 26-year Follow-up of Participants in the Framingham Heart Study. *Circulation, 67*(5)

Hulens, M., Vansant, G., Claessens, A. L., Lysens, R., & Muls, E. (2003). Predictors of 6-minute Walk Test Results in Lean Obese and Morbidly Obese Women. *Scandinavian Journal of Medicine amp Science in Sports, 13*(2), 98-105.

Japkowicz, N. (2000). Learning from Imbalanced Data Sets: A Comparison of Various Strategies. *AAAI Workshop on Learning from Imbalanced Data Sets,*

Jemal, A., Ward, E., Hao, Y., & Thun, M. (2005). Trends in the Leading Causes of Death in the United States, 1970-2002. *JAMA, 294*(10), 1255-1259.

Jerez, J., Molina, I., García-Laencina, P., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing Data Imputation using Statistical and Machine Learning Methods in a Real Breast Cancer Problem. *Artif Intell Med., 50*(2), 105-115.

Kane, C., Bassett, W., Sadetsky, N., Silva, S., Wallace, K., Pasta, D., Cooperberg, M., Chan, J., & Carroll, P. (2005). Obesity and Prostate Cancer Clinical Risk Factors at Presentation: Data from CaPSURE. *J Urol, 173*(3), 732-736.

Kaushal, R., Shojania, K., & Bates, D. (2003). Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. *Arch Intern Med., 163*(12), 1409-1416.

Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *,*

Kelm, B. M., Mittal, S., Zheng, Y., Tsymbal, A., Bernhardt, D., Vega-Higuera, F., Zhou, S. K., Meer, P., & Comaniciu, D. (2011). Detection, Grading and Classification of Coronary Stenoses in Computed Tomography Angiography. *Medical Image Computing and Computer-Assisted Intervention (MICCAI), 6893*, 25-32.

Kempermann, G., Kuhn, H. G., & Gage, F. H. (1997). *Genetic Influence on Neurogenesis in the Dentate Gyrus of Adult Mice.* Paper presented at the Proceedings of the National Academy of Sciences (pp. 10409-10414).

Kennard, R., & Stone, L. (1969). Computer Aided Design of Experiments. *Technometrics, 11*(1), 137-148.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., & Meltzer, P. S. (2001). Classification and Diagnostic Prediction of Cancers using Gene Expression Profiling and Artificial Neural Networks. *Nat Med, 7*(6), 673-679.

Khosla, A., Cao, Y., Lin, C. C. Y., Chiu, H. K., Hu, J., & Lee, H. (2010). An Integrated Machine Learning Approach to Stroke Prediction. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge discovery and Data Mining (KDD '10),*

Kim, J., Cho, B., Im, S., Jeon, M., Kim, I., & Kim, S. (2005). *Comparative Study on Artificial Neural Network with Multiple Regressions for Continuous Estimation of Blood Pressure.* Paper presented at the Proceedings of the IEEE Engineering in Medicine and Biology 27th Annual Conference (pp. 6942-6945).

Kitney, R., & Poh, C. L. (2006). *Geometric Framework Linking Different Levels of the Biological Continuum.* Paper presented at the Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the (pp. 4068-4071). Shanghai:

Kohavi, R., & Sommerfield, D. (1995). *Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology.* Paper presented at the First International Conference on Knowledge Discovery and Data Mining

Kohonen, T. (1990). The Self-Organizing Map. *Proceedings of the IEEE, 78*(9), 1464-1480.

Kowaliw, T., & Banzhaf, W. (2012). *The Unconstrained Automated Generation of Cell Image Features for Medical Diagnosis.* Paper presented at the GECCO

Kuhn, H., Dickinson-Anson, H., & Gage, F. (1996). Neurogenesis in the Dentate Gyrus of the Adult Rat: Age-related Decrease of Neuronal Progenitor Proliferation. *J Neurosci., 16*(6), 2027-2033.

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing Performances of Logistic Regression, Classification and Regression Tree, and Neural Networks for Predicting Coronary Artery Disease. *Expert Systems with Applications, 34*(1), 366-374.

Kwan, J. (2007). Care pathways for acute stroke care and stroke rehabilitation: from theory to evidence. *J Clin Neurosci, 14*(3), 189-200.

Latifoğlu, F., Polat, K., Kara, S., & Güneş, S. (2008). Medical Diagnosis of Atherosclerosis from Carotid Artery Doppler Signals using Principal Component Analysis (PCA), k-NN based Weighting Pre-processing and Artificial Immune Recognition System (AIRS). *J Biomed Inform, 41*(1), 15-23.

Leal, J., Luengo-Fernández, R., Gray, A., Petersen, S., & Rayner, M. (2006). Economic Burden of Cardiovascular Diseases in the Enlarged European Union. *Eur Heart J, 27*(13), 1610-1619.

Lenfant, C. (2012). Prospects of Personalized Medicine in Cardiovascular Diseases. *Metabolism, 62*

Levin, S., Harley, E., Fackler, J., Lehmann, C., Custer, J., France, D., & Zeger, S. (2012). Real-time Forecasting of Pediatric Intensive Care Unit Length of Stay using Computerized Provider Orders. *Crit Care Med., 40*(11), 3058-3064.

Levy, W., Mozaffarian, D., Linker, D., Sutradhar, S., Anker, S., Cropp, A., Anand, I., Maggioni, A., Burton, P., Sullivan, M., Pitt, B., Poole-Wilson, P., Mann, D., & Packer, M. (2006). The Seattle Heart Failure Model: Prediction of Survival in Heart Failure. *Circulation, 113*, 1424-1433.

Li, D., Liu, C., & Hu, S. (2010). A Learning Method for the Class Imbalance Problem with Medical Data Sets. *Comput Biol Med, 40*(5), 509-518.

Li, Y., Liu, L., Chiu, W., & Jian, W. (2000). Neural Network Modeling for Surgical Decisions on Traumatic Brain Injury Patients. *Int J Med Inform., 57*(1), 1-9.

Li, Y., & Agarwal, P. (2009). A Pathway-based View of Human Diseases and Disease Relationships. *PloS one, 4*(2), e4346.

Lillard, A. S., & Erisir, A. (2011). Old Dogs Learning New Tricks: Neuroplasticity Beyond the Juvenile Period. *Developmental Review, 31*(4), 207-239.

Lisboa, P. (2002). A Review of Evidence of Health Benefit from Artificial Neural Networks in Medical Intervention. *Neural Netw., 15*(1), 11-39.

Lisboa, P., & Taktak, A. (2006). The Use of Artificial Neural Networks in Decision Support in Cancer: A Systematic Review. *Neural Netw., 19*(4), 408-415.

Listgarten, J., Damaraju, S., Poulin, B., Cook, L., Dufour, J., Driga, A., Mackey, J., Wishart, D., Greiner, R., & Zanke, B. (2004). Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. *Clinical Cancer Research, 10*, 2725-2737.

Liu, H., & Motoda, H. (1998). *Feature Extraction, Construction and Selection: A Data Mining Perspective.* (1 ed.). Kluwer.

Liu, Y. (2004). Active Learning with Support Vector Machine Applied to Gene Expression Data for Cancer Classification. *J. Chem. Inf. Comput. Sci., 44*(6), 1936-1941.

Lloyd-Jones, D. M., Wilson, P. W., Larson, M. G., Beiser, A., Leip, E. P., D'Agostino, R. B., & Levy, D. (2004). Framingham Risk Score and Prediction of Lifetime Risk for Coronary Heart Disease. *Am J Cardiol, 94*(1), 20-24.

Mair, J., Smidt, J., Lechleitner, P., Dienstl, F., & Puschendorf, B. (1995). A Decision Tree for the Early Diagnosis of Acute Myocardial Infarction in Nontraumatic Chest Pain Patients at Hospital Admission. *CHEST Journal, 108*(6), 1502-1509.

Marmot, M., Syme, S., Kagan, A., Kato, H., Cohen, J., & Belsky, J. (1975). Epidemiologic Studies of Coronary Heart Disease and Stroke in Japanese Men Living In Japan, Hawaii and California: Prevalence of Coronary and Hypertensive Heart Disease and Associated Risk Factors. *Am J Epidemiol, 102*(6), 514-525.

Martini, F. H., Nath, J. L., & Bartholomew, E. F. (2011). *Fundamentals of Anatomy & Physiology.* (9 ed.). Pearson.

McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull Math Biophys, 5*(4), 115-133.

McGovern, P. G., Pankow, J. S., Shahar, E., Doliszny, K. M., Folsom, A. R., Blackburn, H., & Luepker, R. V. (1996). Recent Trends in Acute Coronary Heart Disease — Mortality, Morbidity, Medical Care, and Risk Factors. *N Engl J Med, 334*, 884-890.

McKinney, B., Reif, D., Ritchie, M., & Moore, J. (2006). Machine Learning for Detecting Gene-Gene Interactions: A Review. *Appl Bioinformatics., 5*(2), 77-88.

Menown, I., Mackenzie, G., & and Adgey, A. (2000). Optimizing the Initial 12-lead Electrocardiographic Diagnosis of Acute Myocardial Infarction. *Eur Heart J, 21*(4), 275-283.

Miller, J., & Thomson, P. (2000). *Cartesian Genetic Programming.* Paper presented at the EuroGP (pp. 121-132). Springer-Verlag.

Miller, R., & Sim, I. (2004). Physicians' use of electronic medical records: barriers and solutions. *Health Aff (Millwood)., 23*(2), 116-126.

Min, H. (2010). Artificial Intelligence in Supply Chain Management: Theory and Applications. *International Journal of Logistics Research and Applications, 13*(1), 13-39.

Minaei-Bidgoli, B., Kashy, D., Kortmeyer, G., & Punch, W. (2003). *Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-based System.* Paper presented at the 33rd ASEE/IIEEE Frontiers in Education Conference (pp. 13-18).

Moon, H., Ahn, H., Kodell, R. L., Baek, S., Lin, C. J., & Chen, J. J. (2007). Ensemble Methods for Classification of Patients for Personalized Medicine with High-Dimensional Data. *Artif Intell Med, 41*(3), 197-207.

Moses, M., & Banerjee, S. (2011). *Biologically Inspired Design Principles for Scalable, Robust, Adaptive, Decentralized Search and Automated Response (RADAR).* Paper presented at the IEEE Symposium on Artificial Life (ALIFE) (pp. 30-37).

Murley, D., Rees, S., Rasmussen, B., & Andreassen, S. (2005). Decision support of inspired oxygen selection based on Bayesian learning of pulmonary gas exchange parameters. *Artif Intell Med, 34*(1), 53-63.

Mushkudiani, N., Hukkelhoven, C., Hernández, A., Murray, G., Choi, S., Maas, A., & Steyerberg, E. (2008). A Systematic Review Finds Methodological Improvements Necessary for Prognostic Models in Determining Traumatic Brain Injury Outcomes. *J Clin Epidemiol., 61*(4), 331-343.

Neill, D. B. (2013). Using Artificial Intelligence to Improve Hospital Inpatient Care. *IEEE Intelligent Systems, 28*(2), 92-95.

Neves, G., Cooke, S., & Bliss, T. (2008). Synaptic Plasticity, Memory and the Hippocampus: A Neural Network Approach to Causality. *Nat Rev Neurosci., 9*(1), 65-75.

Nevins, J., Huang, E., Dressman, H., Pittman, J., Huang, A., & West, M. (2003). Towards Integrated Clinico-Genomic Models for Personalized Medicine: Combining Gene Expression Signatures and Clinical Factors in Breast Cancer Outcomes Prediction. *Hum Mol Genet., 12*(2), 153-157.

Nguyen, A., Moore, D., McCowan, I., & Courage, M. (2007). *Multi-class Classification of Cancer Stages from Free-text Histology Reports Using Support Vector Machines.* Paper presented at the IEEE Engineering in Medicine and Biology Society (pp. 5140-5143).

Nilsson, J., Ohlsson, M., Thulin, L., Höglund, P., Nashef, S., & Brandt, J. (2006). Risk Factor Identification and Mortality Prediction in Cardiac Surgery using Artificial Neural Networks. *J Thorac Cardiovasc Surg., 132*(1), 12-19.

Nilsson, M., Perfilieva, E., Johansson, U., Orwar, O., & Eriksson, P. S. (1999). Enriched Environment Increases Neurogenesis in the Adult Rat Dentate Gyrus and Improves Spatial Memory. *J Neurobiol, 39*(4), 569-578.

O'Donnella, C. J., & Elosua, R. (2008). Cardiovascular Risk Factors. Insights from Framingham Heart Study. *Rev Esp Cardiol., 61*(3), 299-310.

Ohlsson, M. (2004). WeAidU—a decision support system for myocardial perfusion images using artificial neural networks. *Artif Intell Med, 30*(1), 49-60.

Osheroff, J., Teich, J., Middleton, B., Steen, E., Wright, A., & Detmer, D. (2007). A roadmap for national action on clinical decision support. *J Am Med Inform Assoc., 14*(2), 141-145.

Osuna, E., Freund, R., & Girosit, F. (1997). *Training Support Vector Machines: An Application to Face Detection.* Paper presented at the Proceedings of Computer Vision and Pattern Recognition

Ounpuu, S., Negassa, A., & Yusuf, S. (2001). INTER-HEART: A Global Study of Risk Factors for Acute Myocardial Infarction. *Am Heart J., 141*(5), 711-721.

Özçift, A. (2011). Random Forests Ensemble Classifier Trained with Data Resampling Strategy to Improve Cardiac Arrhythmia Diagnosis. *Comput Biol Med, 41*(5), 265-271.

Palaniappan, S., & Awang, R. (2008). *Intelligent Heart Disease Prediction System Using Data Mining Techniques.* Paper presented at the Computer Systems and Applications (pp. 108-115).

Park, H., & Lee, S. (2011). Association of Obesity with Osteoarthritis in Elderly Korean Women. *Maturitas, 70*(1), 65-68.

Patel, S. R., Blackwell, T., Redline, S., Ancoli-Israel, S., Cauley, J. A., Hillier, T. A., Lewis, C. E., Orwoll, E. S., Stefanick, M. L., Taylor, B. C., Yaffe, K., &

Men, K. L. (2008). The Association between Sleep Duration and Obesity in Older Adults. *Int J Obes, 32*(12), 1825-1834.

Patterson, R. E., Frank, L. L., Kristal, A. R., & White, E. (2004). A Comprehensive Examination of Health Conditions Associated with Obesity in Older Adults. *Am J Prev Med, 27*(5), 385-390.

Peleg, M., & Tu, S. (2006). Decision support, knowledge representation and management in medicine. *Yearb Med Inform.,* , 72-80.

Pittman, J., Huang, E., Dressman, H., Horng, C. F., Cheng, S. H., Tsou, M. H., Chen, C. M., Bild, A., Iversen, E. S., Huang, A. T., Nevins, J. R., & West, M. (2004). Integrated Modeling of Clinical and Gene Expression Information for Personalized Prediction of Disease Outcomes. *The National Academy of Sciences of the USA, 101*(22), 8431-8436.

Poh, C. L., Kitney, R., & Shrestha, R. (2007). Addressing the Future of Clinical Information Systems—Web-Based Multilayer Visualization. *Information Technology in Biomedicine, IEEE Transactions on,* (2), 127-140.

Polat, K., & Güneş, S. (2007). An Expert System Approach based on Principal Component Analysis and Adaptive Neuro-Fuzzy Inference System to Diagnosis of Diabetes Disease. *Digital Signal Processing, 17*(4), 702-710.

Polat, K., Güneş, S., & Tosun, S. (2006). Diagnosis of Heart Disease using Artificial Immune Recognition System and Fuzzy Weighted Pre-processing. *Pattern Recognition, 39*(11), 2186-2193.

Polat, K., Şahan, S., & Güneş, S. (2007). A Novel Hybrid Method based on Artificial Immune Recognition System (AIRS) with Fuzzy Weighted Pre-processing for Thyroid Disease Diagnosis. *Expert Systems with Applications, 32*(4), 1141-1147.

Praag, H. V., Kempermann, G., & Gage, F. H. (1999). Running Increases Cell Proliferation and Neurogenesis in the Adult Mouse Dentate Gyrus. *Nat Neurosci, 2*, 266-270.

Psaty, B. M., Furberg, C. D., Kuller, L. H., Cushman, M., Savage, P. J., Levine, D., O'Leary, D. H., Bryan, R. N., Anderson, M., & Lumley, T. (2001). Association Between Blood Pressure Level and the Risk of Myocardial Infarction, Stroke, and Total Mortality - The Cardiovascular Health Study. *American Medical Association, 161*(9), 1183-1192.

Quinlan, J. R. (1992). *C4.5: Programs for Machine Learning.* (1 ed.). Morgan Kaufmann.

Quinlan, J. (1986). Induction of Decision Trees. *Machine Learning, 1*(1), 81-106.

Rish, I. (2001). *An empirical study of the naive Bayes classifier.* Paper presented at the In IJCAI-01 workshop on "Empirical Methods in AI"

Robertson, T. L., Kato, H., Rhoads, G. G., Kagan, A., Marmot, M., Syme, S. L., Gordon, T., Worth, R. M., Belsky, J. L., Dock, D. S., Miyanishi, M., & M.D., S. K. (1977). Epidemiologic Studies of Coronary Heart Disease and Stroke in Japanese Men living in Japan, Hawaii and California: Incidence of Myocardial Infarction and Death from Coronary Heart Disease. *Am J Cardiol, 39*(2), 239-243.

Roganb, J., Franklina, J., Stowb, D., Millerc, J., Woodcockd, C., & Robertse, D. (2008). Mapping Land-cover Modifications Over Large Areas: A Comparison of Machine Learning Algorithms. *Remote Sensing of Environment, 112*(5), 2272-2283.

Rosengren, A., Subramanian, S., Islam, S., Chow, C., Avezum, A., Kazmi, K., Sliwa, K., Zubaid, M., Rangarajan, S., & Yusuf, S. (2009). Education and Risk for Acute Myocardial Infarction in 52 High, Middle and Low-Income Countries: INTERHEART Case-Control Study. *Heart, 95*(24), 2014-2022.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A Review of Feature Selection Techniques in Bioinformatics. *Bioinformatics, 23*(19), 2507-2517.

Samama, M. (2000). An Epidemiologic Study of Risk Factors for Deep Vein Thrombosis in Medical Outpatients: The Sirius Study. *Arch Intern Med., 160*(22), 3415-3420.

Samuel, A. (1959). Some Studies in Machine Learning using the Game of Checkers. *IBM Journal of Research and Development, 3*(3), 210-229.

Schaefer, G., & Leung, E. (2007). Neural Networks for Exudate Detection in Retinal Images. *Advances in Visual Computing, 4842*, 298-306.

Shahlaeiab, M., Madadkar-Sobhanic, A., Saghaiebd, L., & Fassihibd, A. (2012). Application of an Expert System based on Genetic Algorithm – Adaptive Neuro-Fuzzy Inference System (GA–ANFIS) in QSAR of Cathepsin K Inhibitors. *Expert Systems with Applications, 39*(6), 6182-6191.

Shors, T. J., Miesegaes, G., Beylin, A., Zhao, M., Rydel, T., & Gould, E. (2001). Neurogenesis in the Adult is Involved in the Formation of Trace Memories. *Nature, 410*, 372-376.

Sittig, D. F., & Ash, J. S. Jules J. Berman (Ed.). (2009). *Clinical Information Systems: Overcoming Adverse Consequences.* (1 ed.). Jones and Bartlett.

Sittig, D. F., Wright, A., Osheroff, J. A., Middleton, B., Teich, J. M., Ash, J. S., Campbell, E., & Bates, D. W. (2008). Grand challenges in clinical decision support. *J Biomed Inform, 41*(2), 387-392.

Snyderman, R., & Williams, R. (2003). Prospective Medicine: The Next Health Care Transformation. *Acad Med., 78*(11), 1079-1084.

Snyderman, R., & Langheier, J. (2006). Prospective Health Care: The Second Transformation of Medicine. *Genome Biol, 7*(104)

Song, X., Mitnitski, A., Cox, J., & Rockwood, K. (2004). Comparison of Machine Learning Techniques with Classical Statistical Models in Predicting Health Outcomes. *Medinfo, 107*(Pt 1), 736-740.

Stangl, D., & Thuret, S. (2009). Impact of Diet on Adult Hippocampal Neurogenesis. *Genes Nutr, 4*(4), 271-282.

Stein, P., Beemath, A., & Olson, R. (2005). Obesity as a Risk Factor in Venous Thromboembolism. *Am J Med, 118*(9), 978-980.

Stokes, J., Kannel, W. B., Wolf, P. A., D'Agostino, R. B., & Cupples, L. A. (1989). Blood Pressure as a Risk Factor for Cardiovascular Disease. The Framingham Study - 30 Years of Follow-up. *Hypertension, 13*

Su, C. T., & Hsu, J. H. (2005). An Extended Chi2 Algorithm for Discretization of Real Value Attributes. *Knowledge and Data Engineering, IEEE Transactions on,* (3), 437-441.

Suka, M., Oeda, S., Ichimura, T., Yoshida, K., & Takezawa, J. (2008). Neural Networks Applied to Medical Data for Prediction of Patient Outcome. In O. Castillo, L. Xu & S. I. Ao (Eds.), *Lecture Notes in Electrical Engineering* pp. 309-325). Springer US.

Syme, S., Marmot, M., Kagan, A., Kato, H., & Rhoads, G. (1975). Epidemiologic Studies of Coronary Heart Disease and Stroke in Japanese Men Living in Japan, Hawaii and California: Introduction. *Am J Epidemiol, 102*(6), 477-480.

Tanapat, P., Hastings, N., Reeves, A., & Gould, E. (1999). Estrogen Stimulates a Transient Increase in the Number of New Neurons in the Dentate Gyrus of the Adult Female Rat. *J Neurosci., 19*(14), 5792-5801.

Tanapat, P., Hastings, N. B., Rydel, T. A., Galea, L. A., & Gould, E. (2001). Exposure to Fox Odor Inhibits Cell Proliferation in the Hippocampus of Adult Rats via an Adrenal Hormone-dependent Mechanism. *Journal of Comparative Neurology, 437*(4), 496-504.

Tanner, L., Schreiber, M., Low, J. G. H., Ong, A., Tolfvenstam, T., Lai, Y. L., Ng, L. C., Leo, Y. S., Puong, L. T., Vasudevan, S. G., Simmons, C. P., Hibberd, M. L., & Ooi, E. E. (2008). Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness. *PLoS Neglect Tropical Diseases, 2*(3), 1-9.

Taupin, P. (2006). *Adult Neurogenesis and Neural Stem Cells in Mammals.* Nova Publishers.

Tay, D., Poh, C., Goh, C., & Kitney, R. (2014). A Biological Continuum based Approach for Efficient Clinical Classification. *J Biomed Inform, 47*, 28-38.

Tay, D., Poh, C., & Kitney, R. (2013). *An Evolutionary Data-Conscious Artificial Immune Recognition System.* Paper presented at the Genetic and Evolutionary Computation Conference (GECCO) Amsterdam, The Netherland:

Tay, D., Poh, C. L., & Kitney, R. (2014). Artificial Neural Cell System for Classification. *Submitted to IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,*

Taylor, J., Ankerst, D., & Andridge, R. (2008). Validation of Biomarker-Based Risk Prediction Models. *Clinical Cancer Research, 14*(19), 5977-5983.

Tehrani, A. S. S., Lee, H., Mathews, S. C., Shore, A., Makary, M. A., Pronovost, P. J., & Newman-Toker, D. E. (2013). 25-Year Summary of US Malpractice Claims for Diagnostic Errors 1986–2010: An Analysis from the National Practitioner Data Bank. *BMJ Quality and Safety, 22*, 672-680.

Timmis, J., & Neal, M. (2001). A Resource Limited Artificial Immune System for Data Analysis. *Knowledge-Based Systems, 14*, 121-130.

Tison, G., Ndumele, C., Gerstenblith, G., Allison, M., Polak, J., & Szklo, M. (2011). Usefulness of Baseline Obesity to Predict Development of a High Ankle Brachial Index (from the Multi-Ethnic Study of Atherosclerosis). *Am J Cardiol, 107*(9), 1386-1391.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. (2001). Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics., 17*(6), 520-525.

Tsai, D., & Watanabe, S. (1999). A Method for Optimization of Fuzzy Reasoning by Genetic Algorithms and its Application to Discrimination of Myocardial Heart Disease. *IEEE Transactions on Nuclear Science, 46*(6), 2239-2246.

Tunstall-Pedoe, H., Kuulasmaa, K., Amouyel, P., Arveiler, D., Rajakangas, A., & Pajak, A. (1994). Myocardial Infarction and Coronary Deaths in the World Health Organization MONICA Project: Registration Procedures, Event Rates, and Case-fatality Rates in 38 Populations from 21 Countries in Four Continents. *Circulation., 90*(1), 583-612.

U.S. Department of Health & Human Services (2013). *Genome-Wide Association Studies (GWAS).* Retrieved 5/6/2013, from http://gwas.nih.gov/index.html

U.S. Department of Health and Human Services (2009). *The Health Information Technology for Economic and Clinical Health (HITECH) Act.* Retrieved 10/9/2013, from http://www.healthit.gov/policy-researchers-implementers/hitech-act-0

Uzark, K. (2003). Clinical pathways for monitoring and advancing congenital heart disease care. *Progress in Pediatric Cardiology, 18*(2), 131-139.

Vapnik, V. (1999). An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks, 10*(5), 988-999.

Vellido, A., Biganzoli, E., & Lisboa, P. (2008). *Machine Learning in Cancer Research: Implications for Personalized Medicine.* Paper presented at the European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning

Wang, S. J., Middletonc, B., Prosser, L. A., Bardon, C. G., RN, C. D. S., MBA, RNN, P. J. C., Kittler, A. F., Goldszer, R. C., Fairchild, D. G., Sussman, A. J., Kuperman, G. J., & Batesc, D. W. (2003). A cost-benefit analysis of electronic medical records in primary care. *Am J Med, 114*(5), 397-403.

Watkins, A. (2001). *AIRS: A Resource Limited Artificial Immune Classifier.* Master's Thesis, Mississippi State University, United States.

Watkins, A., & Boggess, L. (2002). *A New Classifier Based on Resource Limited Artificial Immune Systems.* Paper presented at the Proceedings of Congress on Evolutionary Computation Honolulu, USA:

Watkins, A., Timmis, J., & Boggess, L. (2004). Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm. *Genetic Programming and Evolvable Machines, 5*(3), 291-317.

Wei, W., Visweswaran, S., & Cooper, G. (2011). The Application of Naive Bayes Model Averaging to Predict Alzheimer's Disease from Genome-wide Data. *J Am Med Inform Assoc, 18*, 370-375.

Wiener, J., & Tilly, J. (2002). Population Ageing in the United States of America: Implications for Public Programmes. *Int J Epidemiol., 31*(4), 776-781.

Wiens, J., Guttag, J., & Horvitz, E. J. (2012). *Patient Risk Stratification for Hospital-Associated C. Diff as a Time-Series Classification Task.* Paper presented at the Neural Information Processing Systems (NIPS)

Wilson, W., D'Agostino, R., Levy, D., Belanger, A., Silbershatz, H., & et al. (1998). Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation, 32*(18), 560-565.

Winter, Y., Rohrmann, S., Linseisen, J., Lanczik, O., Ringleb, P., Hebebrand, J., & et al. (2008). Contribution of Obesity and Abdominal Fat Mass to Risk of Stroke and Transient Ischemic Attacks. *Stroke, 39*, 3145-3151.

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining.* Paper presented at the Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining

Wiskott, L., Rasch, M. J., & Kempermann, G. (2006). A Functional Hypothesis for Adult Hippocampal Neurogenesis: Avoidance of Catastrophic Interference in the Dentate Gyrus. *Hippocampus, 16*(3), 329-343.

Wolk, R., Berger, P., Lennon, R., Brilakis, E., & Somers, V. (2003). Body Mass Index: A Risk Factor for Unstable Angina and Myocardial Infarction in Patients with Angiographically Confirmed Coronary Artery Disease. *Circulation, 108*, 2206-2211.

World Health Organization (2008). *Disease and Injury Regional Estimates.* Retrieved 9/16/2012, from http://www.who.int/healthinfo/global_burden_disease/estimates_regional/en/index.html

World Health Organization (2013). *World Health Organization: Cardiovascular Diseases (CVDs).* Retrieved 10/8/2013, from http://www.who.int/mediacentre/factsheets/fs317/en/

Wu, W., Walczak, B., Massart, D., Heuerding, S., Erni, F., Last, I., & Prebble, K. (1996). Artificial Neural Networks in Classification of NIR Spectral Data: Design of the Training Set. *Chemometrics and Intelligent Laboratory Systems, 33*(1), 35-46.

Yan, Q., Yan, H., Han, F., Wei, X., & Zhu, T. (2009). SVM-based decision support system for clinic aided tracheal intubation predication with multiple features. *Expert Systems with Applications, 36*(3), 6588-6592.

Yang, J., & Honavar, V. (1998). Feature Subset Selection Using a Genetic Algorithm. *IEEE Intelligent Systems and their Applications, 13*(2), 44-49.

Yang, P., Zhou, Y., Chen, B., Wan, H., Jia, G., Bai, H., & Wu, X. (2009). Overweight, Obesity and Gastric Cancer Risk: Results from a Meta-Analysis of Cohort Studies. *Eur J Cancer, 45*(16), 2867-2873.

Yano, K., Reed, D. M., & Mcgee, D. L. (1984). Ten-year Incidence of Coronary Heart Disease in the Honolulu Heart Program - Relationship to Biologic and Lifestyle Characteristics. *Am J Epidemiol, 119*(5), 653-666.

Yeh, D. Y., Cheng, C. H., & Chen, Y. W. (2011). A Predictive Model for Cerebrovascular Disease using Data Mining. *Expert Systems with Applications, 38*(7), 8970-8977.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J., & Hua, L. (2012). Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *J Med Syst., 36*(4), 2431-2448.

Yu, Y., Chen, S., Wang, L. S., Chen, W. L., Guo, W. J., Yan, H., Zhang, W. H., Peng, C. H., Zhang, S. D., Li, H. W., & Chen, G. Q. (2005). Prediction of Pancreatic Cancer by Serum Biomarkers Using Surface-Enhanced Laser Desorption/Ionization-Based Decision Tree Classification. *Oncology, 68*(1)

Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., & Liu, L. (2004). Effect of Potentially Modifiable Risk Factors Associated with Myocardial Infarction in 52 Countries (the INTERHEART study): Case-Control Study. *The Lancet, 364*(9438), 937-952.

Yusuf, S., Hawken, S., Ounpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., Lisheng, L., & INTERHEART Study Investigators (2004). Effect of Potentially Modifiable Risk Factors Associated with Myocardial Infarction in 52 countries (the INTERHEART study): Case-Control Study. *Lancet, 364*(9438), 937-952.

Appendix A

This appendix serves to provide an introduction to Genetic Algorithm (GA). GA (Holland, 1992) is a type of evolutionary computing algorithm inspired by Darwinian evolutionary theory. It works in a parallel manner - where the algorithm explores the solution space in multiple directions. This differs from conventional mathematical analysis, making it a promising search heuristic method that is less likely to be trapped in a local optimal position.

GA works with a population of candidate solutions that searches for the optimal solution probabilistically. It iteratively transforms the initial set of possible solutions encoded in chromosome-like data structure (each associated with a fitness value) into a population of offspring that aims to find the global optimum within a reasonable number of iterations. Each successive offspring is generated and optimized based on the Darwinian principle of natural selection, together with operations patterned after the natural occurring genetic operations (e.g. crossover and mutation).

One application of GA is to perform feature selection, a process that identifies informative subset of predicators within a dataset (Huanga & Wangb, 2006). This process is of paramount importance in view of the exponential growth of clinical data in recent years - making analysis of large number of clinical features difficult. Extraction of the least number of highly relevant and informative predictors that can comprehensively describe the underlying pattern present in the dataset results in two significant advantages. Firstly, a boost in the accuracy can be obtained by a classifier (e.g. SVM) - as removing irrelevant and redundant features can effectively ameliorate the learning capability of the classifier. Secondly, the computational time needed for developing the classification model can be decreased because, with less features, it reduces the data complexity - allowing the classifier to learn at a faster pace.

Figure A.1 illustrates a generic GA evolutionary process that can be used for feature selection. At onset, an initial population of fixed-length bit-string chromosome is formulated. The length of the chromosomes is tantamount to the

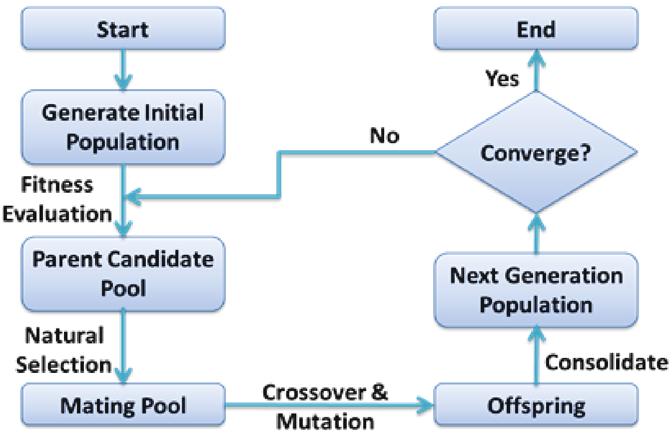


Figure A.1: Canonical Flow of Genetic Algorithm
The basic form of GA consists of operators like natural selection, crossover and mutation.

number of features in the dataset. The value of the chromosomes is initialized randomly so that it spreads across the feasible solution space. Each bit in the chromosome is assigned with a value of either '1' or '0', indicating whether that feature is selected or eliminated from consideration by the classifier, respectively. Next, the fitness of each chromosome is evaluated using the fitness function. The choice of the fitness representation is important as only effective representation and meaningful fitness evaluation will result in a successful GA application. Typically, when GA is used with a classifier to restrict the feature set, the accuracy obtained by the classifier is used as the fitness value. Once the chromosomes (or quality of the solutions) are assessed by the fitness function, it enters into the parent candidate pool.

Natural selection operation is then applied. An example is the stochastic universal sampling selection (SUS) technique, where it selects chromosomes that exhibit no bias and with minimal spread (Baker, 1984). In other words, SUS selects the chromosomes by sampling the population of candidates that are uniformly spaced. This offers an opportunity for chromosomes with low fitness value to be selected, avoiding the fittest chromosomes from saturating the candidate space.

The selected chromosomes subsequently enter into the mating pool where genetic operations, such as crossover and mutation, are applied. This results in the initial pool of candidate solution to stochastically transit to a new pool of possible solution (i.e. offspring). One type of crossover operation is uniform crossover, where it exchanges information between 2 parent strings at the bit (locus) level. This differs from one-point or two-point crossovers which exchange information at the segment level. Mutation, on the other hand, is a bit-wise operation that changes a bit in the chromosome from its original state (e.g. from ‘0’ to ‘1’, or vice versa). The purpose of mutation is to instil some form of randomness into the algorithm, thus avoiding a situation where candidate solutions get trapped in local minima.

The generated offspring consequently forms the new population for the next generation. This generation of population is then verified for convergence. If it fails the convergence criteria, the entire process of fitness evaluation, selection, crossover and mutation operations will be repeated. Otherwise, the best chromosome will be picked and returned as the result.

Appendix B

This appendix serves to provide an introduction to Support Vector Machine (SVM). SVM (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1999) is a robust supervised learning algorithm that is capable of yielding excellent generalization performance on an extensive area of problems - such as intrusion detection, face detection, biomedical research, etc (Chen et al., 2005; Osuna et al., 1997; Listgarten et al., 2004). It is derived from statistical learning theory and is capable of solving linearly and non-linearly separable problems. Fundamentally, SVM performs classification through the construction of an N-dimensional hyper-plane that optimally separates the data into two or more categories whereby the margin of separation between the different categories is maximized.

Considering a binary class classification problem with training dataset $\{(x_i, d_i)\}_{i=1}^{N}$, where x_i is the input pattern for the i^{th} example, d_i is the corresponding desired output ($d_i = +1$ or $d_i = -1$) and N is the total number of training data; SVM attempts to construct a linear separating hyper-plane $\mathbf{w}^T\mathbf{x} + b = 0$ with maximal distance between the soft margins, where **w** is an adjustable weight vector (normal to the plane) and b is the bias. The classification condition may be expressed in the following form:

$$\mathbf{w}^T\mathbf{x} + b \geq 0, \qquad \text{for } d_i = +1 \qquad (10)$$

$$\mathbf{w}^T\mathbf{x} + b < 0, \qquad \text{for } d_i = -1 \qquad (11)$$

In order to maximize the distance between the data vectors that belong to different classes, the gap between the soft margins that separates the two classes of data need to be maximized. These soft margins are defined as follow:

$$\mathbf{w}^T\mathbf{x} + b = 1, \qquad \text{for } d_i = +1 \qquad (12)$$

$$\mathbf{w}^T\mathbf{x} + b = -1, \qquad \text{for } d_i = -1 \qquad (13)$$

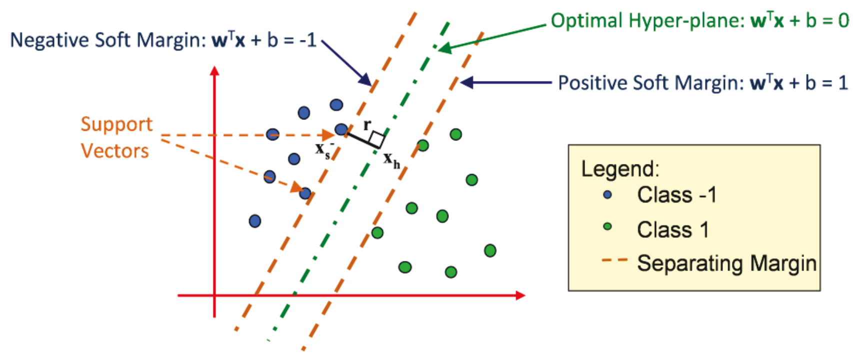


Figure B.1: Geometric Representation of Optimal Hyper-plane and the Soft Margins

With reference to Figure B.1, assume that x_h is a point on the optimal hyper-plane and x_s is a support vector (i.e. data point that satisfies either of the soft margins equations), then the distance between the soft margin and the optimal hyper-plane can be calculated and maximized. Mathematically, vector $\mathbf{X_s}$ (the vector from the origin to the point x_s) can be defined with the following expression:

$$\boldsymbol{X_s} = \boldsymbol{X_h} + r\frac{\boldsymbol{w}}{||\boldsymbol{w}||} \tag{14}$$

where $\mathbf{X_h}$ is the vector from the origin to the point x_h; $\mathbf{w}$ is the weight of the perpendicular vector from the optimal hyper-plane to the soft margin; and r is the scalar distance between the optimal hyper-plane and the soft margin. Since the optimal hyper-plane is defined as $h(x) = \mathbf{w}^T\mathbf{x} + b$, the expression for the distance between the soft margin and the optimal hyper-plane can be calculated as follow:

$$h(x) = \boldsymbol{w}^T(\boldsymbol{X_h} + r\frac{\boldsymbol{w}}{||\boldsymbol{w}||}) + b$$

$$h(x) = \boldsymbol{w}^T\boldsymbol{X_h} + b + r\frac{\boldsymbol{w}^T\boldsymbol{w}}{||\boldsymbol{w}||}$$

$$h(x) = r\frac{||\boldsymbol{w}||^2}{||\boldsymbol{w}||}, \qquad \text{since} \quad \boldsymbol{wX_h} + b = 0$$

$$h(x) = r||\boldsymbol{w}||$$

$$r = \frac{\boldsymbol{h}(\boldsymbol{x})}{||\boldsymbol{w_h}||} \tag{15}$$

Therefore, the distance between the positive (and negative) soft margin and the optimal hyper-plane is defined as:

$$r = \begin{cases} \frac{1}{||\boldsymbol{w}||}\ , & for\ d^{+} = +1 \\ \frac{-1}{||\boldsymbol{w}||}\ , & for\ d^{-} = -1 \end{cases} \tag{16}$$

Hence, the distance between the soft margins is equivalent to $\frac{2}{\|\boldsymbol{w}\|}$. The maximization of $\frac{2}{\|\boldsymbol{w}\|}$, which is tantamount to minimizing $\frac{||w||^2}{2}$, can be optimized with the employment of a constrained optimization technique such as the Lagrange theory. The Lagrangian dual problem used in SVM is expressed as:

$$L(\boldsymbol{w}, b, \lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j\, d_i d_j k(\boldsymbol{x_i}, \boldsymbol{x_j}) \tag{17}$$

subjected to the following constraints:

1. $\sum_{i=1}^{N} \lambda_i\, d_i = 0$
2. $0 \le \lambda_j \le c, for\ i = 1, 2, \dots, N$

where λ_i is the Lagrange multiplier, $k(\boldsymbol{x_i}, \boldsymbol{x_j})$ is the kernel function and c is a user-specified regularization parameter. The kernel function can be a nonlinear function which enables SVM to effective solves nonlinear classification problems like the classical XOR problem. This is because when the input space of a nonlinearly separable problem is nonlinearly casted into a higher dimensional space, it is more likely to be separated linearly than in a lower dimensional space, as suggested by Cover's separability theorem (Cover, 1965).

Several commonly used nonlinear kernel functions include polynomial - $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$; sigmoid - $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$; and radial basis function (RBF) - $K(x_i, x_j) = \exp(-\gamma\| x_i - x_j\|^2)$, where γ , r and d are kernel parameters. In this work, RBF is used as the kernel function due to its ability to solve non-linearly separable problems, low complexity involved during model selection and excellent performance. It is noteworthy that the linear kernel is a special case of RBF kernel due to the fact that with certain cost and gamma settings, RBF can achieve the same performance as linear kernel with certain cost value (Keerthi & Lin, 2003).

Appendix C

Journal of Biomedical Informatics 47 (2014) 28–38

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A biological continuum based approach for efficient clinical classification

Darwin Tay [a,b], Chueh Loo Poh [b,*], Carolyn Goh [a], Richard I. Kitney [a]

[a] Department of Bioengineering, Imperial College London, UK
[b] Division of Bioengineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 9 April 2013
Accepted 3 September 2013
Available online 12 September 2013

Keywords:
Classification
Dimensionality reduction
Etiological network
Feature selection
Genetic algorithm
Support vector machine

ABSTRACT

Clinical feature selection problem is the task of selecting and identifying a subset of informative clinical features that are useful for promoting accurate clinical diagnosis. This is a significant task of pragmatic value in the clinical settings as each clinical test is associated with a different financial cost, diagnostic value, and risk for obtaining the measurement. Moreover, with continual introduction of new clinical features, the need to repeat the feature selection task can be very time consuming. Therefore to address this issue, we propose a novel feature selection technique for diagnosis of myocardial infarction – one of the leading causes of morbidity and mortality in many high-income countries. This method adopts the conceptual framework of biological continuum, the optimization capability of genetic algorithm for performing feature selection and the classification ability of support vector machine. Together, a network of clinical risk factors, called the biological continuum based etiological network (BCEN), was constructed. Evaluation of the proposed methods was carried out using the cardiovascular heart study (CHS) dataset. Results demonstrate a significant speedup of 4.73-fold can be achieved for the development of MI classification model. The key advantage of this methodology is the provision of a reusable (feature subset) paradigm for efficient development of up-to-date and efficacious clinical classification models.

1. Introduction

The efficient development of accurate clinical classification models has been a challenge for many reasons. One problem that is commonly encountered is the 'curse of dimensionality' [1], where the linear growth of clinical features (i.e. predicators) results in an exponential growth in the search space. This inevitably hinders the development of classification models as it becomes computationally expensive to investigate a plethora of clinical features simultaneously using search heuristics that analyze features in combinations (particularly, when performing multivariate analysis based on wrapper approach). This situation is exacerbated by the fact that up-to-date and sophisticated clinical classification models need to be constantly developed in order to continually improve the quality of clinical diagnosis. Specifically, the clinical classification models need to be rebuilt whenever new clinical risk factors that could potentially ameliorate the performance of the classification model are introduced. An example of such clinical effort is the perpetual studies of different types of clinical risk factors and approaches that could improve the ability to identify events of myocardial infarction (MI) [2,3]. This is of paramount importance as MI is a leading cause of morbidity and mortality in many developed countries, such as the United States (US) and the United Kingdom (UK) [4–6]. Despite considerable advances in medicine, MI approximately occurs every 34 s in the US and about 15% who experience MI will die from it [4]. Moreover, MI is difficult to ascertain in patients presenting to the emergency department with anterior chest pain [2]. This advocates for the need of an efficient approach to develop up-to-date MI classification models for performing accurate diagnosis.

Furthermore, investigation of the association between a range of clinical observations (e.g. medical history, chemotherapy, stage of disease, gene, etc.) and the disease at the human population level is important as it has demonstrated promising potential for improving disease classification performance [7,8]. However, when such an investigation is carried out on a larger scale, this would involve a large amount of clinical features, making analysis challenging and even computationally infeasible. Additionally, it also hinders the ability for any machine learning method to perform accurate disease classification. One approach to mitigate the aforementioned problems is through dimensionality reduction – where significant clinical risk factors are identified, reducing the total number of predicators that need to be analyzed.

In this paper, we introduce a novel clinical feature selection methodology for the development of MI classification model. This

* Corresponding author. Address: 70 Nanyang Drive, N1.3-B2-09, Singapore 637457, Singapore. Fax: +65 6791 1761.

E-mail addresses: darwintay@imperial.ac.uk (D. Tay), CLPoh@ntu.edu.sg (C.L. Poh), c.goh@imperial.ac.uk (C. Goh), r.kitney@imperial.ac.uk (R.I. Kitney).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.09.002

approach utilizes on the conceptual framework of biological continuum (BC) [9,10], the optimization capability of genetic algorithm (GA) [11] for performing feature selection and the classification ability of support vector machine (SVM) [12–14] for dichotomizing patients experiencing a phenotypic manifestation from healthy individuals. The BC is the hierarchy of the human organism comprising body, systems, viscera, tissue, cells, proteins and genes. In this study, it provided the biological paradigm necessary for segregating a range of available clinical features; offering the advantage of reducing the number of clinical features that needs to be analyzed concurrently. A GA based wrapper approach using SVM, which selects significant clinical features capable of dichotomizing patients experiencing a phenotypic manifestation from healthy individuals, was implemented. This hybrid algorithm (called GA-SVM) was used to identify important clinical features at each level of the BC and incrementally built a network of clinical risk factors, called the biological continuum based etiological network (BCEN). The primary advantage of BCEN used for the construction of up-to-date clinical classification model is that it allows new clinical features to be considered for incorporation into the classification model without the need for a total reanalysis from scratch.

The reliability of the constructed BCEN was assessed by comparing the set of identified risk factors found in the (obesity-system) sub-network, with the risk factors found in previous clinical studies. Promising results were obtained from this analysis. An MI classification model was subsequently developed based on the clinical features identified and present in the BCEN. Significant reduction in the computational time required to develop the classification model was achieved. It is noteworthy that comparable classification accuracy was obtained between the proposed method (i.e. pre-selection of clinical features using BCEN) and the baseline approach (i.e. no pre-selection was performed). The Cardiovascular Health Study (CHS) [15] dataset was analyzed in this study.

The rest of the paper is organized as follows. Section 2 provides the background information on feature selection. In section 3, the experimental methodology involved in the development of the clinical feature selection technique and the clinical classification model is presented. The experimental results are presented in Section 4 and discussed in Section 5. Finally, conclusions are drawn in Section 6.

2. Background

Conventionally, clinical predictions which provide the disease diagnosis for an individual are based on expert knowledge. However, with the exponential growth of clinical data generated in healthcare industries, this approach has become more and more difficult and costly. An approach to mitigate this challenge is to process and analyze the large amount of clinical data, extracting knowledge that enables support for cost-containment and decision making [16]. Machine learning is one method that has been proposed to address this issue. It provides the techniques necessary for the analysis of the data, discovery of hidden patterns and provides healthcare professionals with an additional source of knowledge for decision making. In the parlance of literature, machine learning is defined as a branch of artificial intelligence that postulates a set of computer-based methods for automatic analysis of information and recognition of patterns through repeated learning from the training data [17], and is a more powerful and sophisticated descendant of traditional statistical models. It is generally model-free and is capable of efficiently detecting and modeling the non-linear interactions in high dimensional datasets. Additionally, the associations or patterns detected by machine learning methods tend to be logical and can be identified by human experts if they analyze the problem carefully enough [18]. Clearly, this entails that machine learning is capable of saving both the time and effort necessary for the discovery of underlying patterns.

Clinical prediction (e.g. diagnosis of cardiovascular disease) based on machine learning approaches has gained popularity over the years [2,16,19–24] and shown to be an extremely useful tool in medical innovation [21]. It is often based on the patient's unique clinical, genetic and environmental characteristics and plays a significant role in healthcare decision making and planni ng. Since each clinical feature collected is associated with a different financial cost, diagnostic value and risk [25], it is highly desirable to reduce the number of clinical tests that need to be taken by a patient. This would inevitably reduce the financial cost, and the time incurred on both the analysts and patients. One approach commonly adopted by machine learning techniques to reduce the number of clinical features while improving the diagnostic/classification accuracy is feature selection.

Feature selection is the process of selecting a subset of relevant features for model construction and provides better insights into the target concept of a real-world problem [21]. It differs from other dimensionality reduction techniques like project and compression where their original representation of the variables is modified. Therefore, feature selection has the advantage of preserving the original semantics of the features which enables domain experts to interpret the selected features. Furthermore, it has shifted from being an illustrative example to one of real prerequisite for developing classification models [26]. This is, in part, because of the exponential increase in the dimensionality of the data (e.g. in clinical and bioinformatics domains), the fact that most classifiers were originally not designed to handle plethora of irrelevant features, and the need to generate more accurate classifiers efficiently. In general, feature selection aims to identify a parsimonious subset of useful features (from a large set of features) that (1) does not decrease the classification accuracy, (2) reduces the computational time needed to learn a sufficiently accurate classification model, (3) does not acutely changes the class distribution while adequately representative for descripting the target concept, and (4) reduces the amount of examples that need to be collected in order to develop a classification model with the desired accuracy [27,28].

Feature selection algorithms typically fall under 4 categories depending on how it is performed in relation to the classification algorithm. They include (1) selection based on expert knowledge, (2) filter approach, (3) wrapper approach, and (4) embedded approach. Each has its own competitive advantages and drawbacks. Selection based on expert knowledge (e.g. human domain expert or referencing the scientific literature) offers a set of features with high interpretability in relation to the target concept. However, its major drawbacks are that it can be time consuming and human expert is required to perform the task. An illustration of this approach is demonstrated in [25], where the number of interaction tests that need to be performed can be limited with the use of experimental knowledge of the biological network. More specifically, knowledge extracted from protein interaction databases reduces the number of interaction tests from 1.25×10^{11} to 7.1×10^{4}, allowing more efficient analysis of genome-wide studies to be carried out.

Filter methods, on the other hand, evaluate the relevance of each feature by assessing only the intrinsic characteristics of the data. Although this approach does not need a domain expert to intervene, is simple, efficient and can easily scale to very high-dimensional datasets, it does not always guarantee improved performance [29] as it ignores the inductive bias associated with the classifier [30]. Examples of filter techniques include chi-square test, *t*-test, information gain, correlation-based feature selection and Markov blanket filter.

Wrapper methods embed the inductive bias associated with the classifier within the feature selection process. In this case, subsets of features are generated and their performance is assessed by training and testing them on a specific classification algorithm. The advantages of this approach are: (1) the freedom to choose the desired classification algorithm, (2) allowing interaction between feature selection and model selection, and (3) ensuring that feature dependencies are taken into consideration (i.e. the need to add or remove more than 1 feature at the same time in order to improve the performance [25]). Consideration of feature dependencies is important, especially in the medical field, as it has become evident that multiple genes collectively contribute to the etiology and clinical manifestation of human diseases [31]. Hence, important genotypic factors might be missed if they have been examined in isolation or in a linear fashion – without allowing for potential interactions. This situation would be exacerbated when performing genome-wide association studies where hundreds of thousands of single nucleotide polymorphisms (SNPs) need to be analyzed. Wrapper approach, on the downside, becomes computationally intensive when the number of features grows exponentially. This is because every feature subsets generated need to be executed on the selected learning algorithm. Moreover, it has a higher risk of over-fitting the classifier than filter approach. Examples of this technique include sequential forward selection, sequential backward selection, simulated annealing, genetic algorithm and estimation of distribution algorithm.

Finally, embedded approach integrates the process of identifying the optimal subset of features within the learning algorithm. Based on this mechanism, it has the advantage of being more computationally efficient (compared to wrapper approach) while maintaining interaction with the classifier. Examples include decision trees and weighted naïve Bayes.

3. Methodology

3.1. Dataset

The CHS dataset, as described in [15], is an epidemiology study of the elderly (defined as adults aged 65 and older). It comprises of elderly subjects from four US communities, namely Forsyth County, North Carolina; Sacramento County, California; Washington County, Maryland; and Pittsburgh, Pennsylvania. A total of 5888 individuals from urban and rural areas form the baseline cohort of CHS. Eligible individuals were sampled from Medicare eligibility lists in each area. Eligible participants included all individuals sampled from the Health Care Financing Administration (HCFA) sampling frame – they were 65 years or older at the time of examination, non-institutionalized, expected to remain in the area for the next 3 years, able to give informed consent and do not require a proxy respondent at baseline. Individuals who were wheelchair-bound at home at baseline, receiving hospice treatment, radiation therapy or chemotherapy for cancer were excluded. The eligible individuals were examined yearly from 1989 to 1999. Extensive physical and laboratory evaluations were carried out to identify the presence and severity of cardiovascular disease (CVD) risk factors – such as hypertension; hypercholesterolemia and glucose intolerance; subclinical disease, such as carotid artery atherosclerosis; left ventricular enlargement; and transient ischemia. Criteria for identification of MI events include: observation of evolving Q-wave, cardiac pain and abnormal enzymes together with an evolving ST-T pattern or new left bundle branch block. A total of 355 clinical features related to the individual's health status were selected from the CHS dataset for this study.

The dataset was chosen because of (1) the relatively high prevalence of coronary heart disease (CHD) among the elderly, (2) worldwide demographic aging, (3) paucity of information regarding risk factors for CHD among elderly, and (4) the changing clinical characteristics of CHD with advancing age [4,15,32,33].

3.2. Biological continuum based etiological network (BCEN)

Several steps were taken to construct the BCEN for MI with the canonical flow illustrated in Fig. 1. A succinct description of the key steps taken is given below while we dedicate separate sections for the discussion of the details:

1. Sparse records were removed and missing entries in the dataset were imputed to ensure good quality data is used to model the risk factors associated with MI. This was performed with the K-nearest neighbor (KNN) algorithm [34] – it calculates the missing value by taking the K nearest training set vectors (based on Euclidean distance) into consideration.
2. Healthy individuals, forming a large proportion of the dataset in relation to the number of patient records, were sampled to avoid jeopardizing the ability of SVM to learn and generalize. This is carried out with Kohonen Self-Organizing Map (SOM) [35], where a representative subset of the majority class (i.e. healthy individuals) present in the CHS dataset was selected, a process known as under-sampling.
3. Clinical features, such as blood pressure, electrocardiography (EKG) readings, ultrasound data, hematology data, etc., were segregated along the BC – the hierarchy of the human organism. It comprises 7 levels, namely the body, system, viscera, tissue, cell, protein and gene.
4. GA-SVM, a hybrid algorithm used to identify significant clinical features, was implemented. It is used repeatedly at each level of the BC to identify significant risk factors that are related to the different phenotypic manifestations, and ultimately MI.
5. With the significant risk factors identified at the different levels of the BC, they were consolidated to construct a consensus network, known as the BCEN in this work. These risk factors, in turn, were used to perform MI classification using the GA-SVM algorithm.

3.2.1. Data Imputation

As with many datasets collected from real subjects and patients, missing data is unavoidable. This may be due to various factors, e.g. the refusal of respondents, malfunction of equipment, data not entered correctly and the death of patients [36]. Moreover, since the quality of the results is largely determined by the quality of the data used in the analysis, detailed consideration was given before using the CHS dataset. It was found that the CHS dataset contains a significant percentage of missing information. Hence, data imputation was first conducted.

Data imputation, the process of substituting missing values in a dataset with plausible values, was performed using KNN. KNN imputation was used because of its excellent performance in estimating missing values [37–40] and its ability to estimate both qualitative and quantitative attributes. This makes it highly suitable for extrapolating the missing entries in the CHS dataset.

Firstly, individuals with unknown MI status were removed from the analysis. Next, to foster more accurate data imputation, individuals and clinical features with high percentage of missing entries were removed. It is important to have low percentage of missing values because the accuracy of the imputed result would suffer if too little complete entries were available for KNN to reference when estimating the missing values [37,40,41]. Hence, individuals and clinical features with more than 20% and 4.5% missing entries, respectively, were removed. Consequently, the resultant dataset was normalized to unit variance before data imputation was performed using KNN. This is important as it

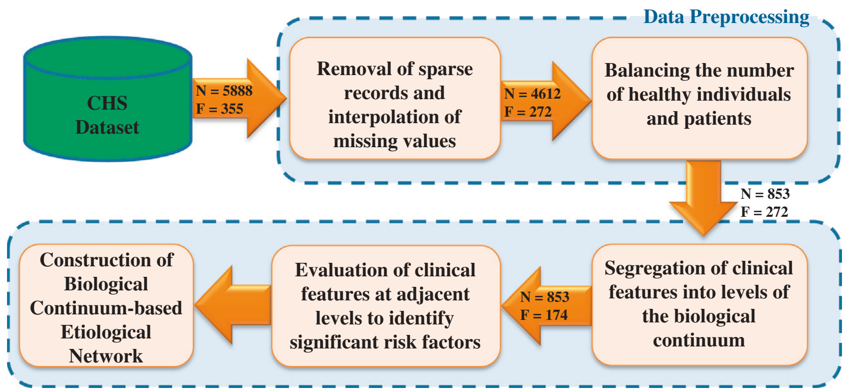


Fig. 1. Canonical flow of the methods adopted to construct BCEN. 'N' denotes the number of instances and 'F' represents the number of features present in the dataset at different stages.

ensures that variables with large scale do not dominate the (Euclidean) distance measure [42].

The optimal value of K for each clinical feature was determined by 10-fold cross-validation. After the value of K for each clinical feature had been determined, data imputation for each missing attribute was performed. The type of replacement method used depends on the type of data present in each clinical feature. For instance, if the data is categorical, a reliable choice is to use the mode of the K nearest neighbors to assign the value for the missing entries [34,38]. On the other hand, if the data is continuous, the weighted-mean of the K nearest neighbor is used instead to calculate the missing value. Weighted-mean estimation has been demonstrated in [37,43] to be robust and accurate.

3.2.2. Class imbalance data problem

The class imbalance data problem is not uncommon in medical datasets where the data is predominated by the healthy subjects (i.e. controls), with only a small number of disease-affected subjects (i.e. cases). Consequently, this limited the effectiveness ability of standard machine learning algorithms – where the algorithms tend to be overwhelmed by the major class and ignore the minor one. This, in turn, hinders performance [44,45]. This class imbalance data problem prevails in the CHS dataset as well. Therefore, data balancing was performed before deploying the data to GA-SVM.

SOM, an unsupervised (neural network) learning algorithm, was employed to under-sample the major class. This algorithm was chosen because it is capable of generating high quality samples that are representative of the original dataset [35] and it has been shown in [46] that SOM outperforms random selection. Once the imputed dataset was obtained, the SOM was trained in two phases; namely, the ordering phase and the tuning phase. Two key adaptive parameters, neighborhood size and learning rate, were used when training the SOM. Neighborhood size defines the number of neurons that surround the winning neuron (i.e. most stimulated neuron) at each epoch, while the learning rate controls the degree of change for the adapting neurons.

During the ordering phase, large initial neighborhood size (i.e. 10) and learning rates (i.e. 0.9) were used. Conversely, small neighborhood size (i.e. 1) and learning rates (i.e. 0.02) were used during the tuning phase – where the neighborhood size will shrink progressively to 1. This is to allow the SOM to adjust quickly to the input pattern during the ordering phase and to stabilize the feature map during the tuning phase [35]. The following value for the SOM parameters was determined experimentally and used in this study: number of neurons: 21 by 21; topology function: hexagon; distance function: Euclidean; epoch: 1000; ordering phase learning rate: 0.9; tuning phase learning rate: 0.02; initial neighborhood size: 10; final neighborhood size: 1. The reason for using these values is because they have shown to provide reasonable performance.

3.2.3. Segregation of clinical features

The Biological Continuum was central to the development of the BCEN. It was utilized in this case to provide the necessary biological paradigm to relate the disease mechanisms to the clinical manifestations at various levels of the biological continuum. Upon analyzing the clinical features, it was found that these features fall under 4 key levels along the BC, namely: body, system, viscera and protein level. Clinical features related to medication were removed from the study as it was difficult to adjudicate to which level of the BC they belong. Categorization of the rest of the clinical features, in relation to the levels of the BC, was undertaken using the following guidelines:

- *Body level* – Contains clinical features related to individuals' personal statistics (e.g. age, weight), lifestyle (e.g. smoking status, exercise intensity) and cardiovascular events which that individual is experiencing.
- *System level* – Consists of clinical features related to individuals' medical history (e.g. arthritis, diabetes), symptoms (e.g. hearing/vision problems) that the individual is experiencing and blood pressure measurements.
- *Visceral level* – Clinical measurements, e.g. EKG, ultrasound data and treatment specific to an organ were classified under this level.
- *Protein level* – Clinical features related to hematology were grouped under this level.

3.2.4. GA-SVM

GA-SVM, a hybrid algorithm that comprises of (1) SVM that models the statistical properties necessary to distinguish healthy individuals from patients experiencing a clinical phenotype, and (2) GA that selects the significant features that contribute to the construction of an accurate SVM model, was implemented. In this work, SVM uses radial basis function (RBF) as its kernel function and is defined as:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{1}$$

where γ is a variable used to adjust the width of the Gaussian functions of the kernel. RBF is used due to its ability to solve non-linearly separable problems, low complexity involved during model selection and excellent performance. Two parameters, namely the regularization cost and gamma (used in RBF) parameters, were tuned over the recommended range $[2^{-5}, 2^{13}]$ and $[2^{-15}, 2^{3}]$

respectively [47]. Optimization of SVM parameters were performed by evaluating a set of cost-gamma combinations defined using uniform design (UD) method [48]. UD is a technique that scatters a set of points uniformly across the cost-gamma landscape, proposed to alleviate the computational loads associated with the search for the optimal cost-gamma pair [49]. This search process begins by initializing a 30-points UD (global) search across the defined cost-gamma landscape. Next, it identifies the top 5 most accurate (global) cost-gamma pairs, where they form the centroid for 10-points UD (fine) search. If improved accuracy was achieved, the points will form the centroid for another 10-points UD search. This process repeats until no further improvement is achieved. Fig. 2 provides an illustration of this method.

Fig. 3 provides the schematic illustration of GA-SVM algorithm. The flow of the algorithm is as follow: GA first (randomly) initializes a pool of clinical feature subsets (Fig. 3 – chromosome 1 to *N*) from the CHS dataset (consisting of M clinical features). Each bit in the chromosome is assigned with a value of either '1' or '0', indicating whether that feature is selected or eliminated from consideration by the classifier, respectively. This produces a pool of chromosomes representing different input features. Consequently, each chromosome was evaluated by SVM (where optimization of SVM parameters was performed independently for each chromosome) in an attempt to determine how informative and discriminative the clinical features are in relation to the associated clinical or subclinical manifestation. This evaluation is conducted by performing a 10-fold stratified cross-validation. Subsequently, these subsets of clinical features undergo natural selection, crossover and mutation phases postulated by GA. The process repeats until GA converges or the maximum number of generations has been reached. GA is considered to have converged if the maximum fitness value (i.e. balanced accuracy – the average of sensitivity and specificity) does not improve after 20 consecutive generations. Upon termination, the subset of clinical features that yielded the highest balanced accuracy will be selected and considered as significant risk factors. A consensus network was constructed if several combinations of clinical feature subset yielded the same fitness performance. The reason for doing this is to build a parsimonious model that maximizes the likelihood of the clinical features that are most influential to the development of the phenotypic manifestation. It was derived by identifying clinical features that existed in more than 75% of the highest-performing clinical feature combinations. The parameters value used by GA are as follow: population size: 250; maximum generation: 300; natural selection: stochastic universal sampling; crossover type: uniform crossover; crossover probability: 0.8; mutation probability: 0.01. These values were chosen because they provided satisfactory result when experimented over a range of values. The algorithm was written in Matlab (MathWorks Inc., Natick, MA) and executed in parallel using a high performance computer (HPC) cluster.

3.2.5. Construction of BCEN

The underlying cause of MI is multifactorial and subtle, with nonlinear causal dynamics. Moreover, with the plethora of clinical predicators available, analysis of all of them becomes computationally impractical. In view of such challenges, GA-SVM, together with the conceptual framework of the BC, were used to construct the BCEN for MI.

Firstly, by segregating the clinical features into various levels along the BC, the number of clinical features to be analyzed is effectively reduced to the number of clinical features present at each level (i.e. dimensionality reduction). Secondly, with the employment of GA, which is capable of performing global heuristic searches both effectively and efficiently, the computational burden of discovering significant risk factors is alleviated. Finally, facilitated by SVM, which outperforms popular technique like multifactor dimensionality reduction (MDR) [50], it ensures that accurate estimation of the association between the clinical features at adjacent levels of the BC is being carried out.

At onset, clinical features grouped under the "body level" of the BC were input into GA-SVM for investigation. This step aims to identify clinical features that contribute significantly to the development of an accurate inference model for MI. Consequently, significant risk factors, defined in this work as risk factors that can potentially contribute to the manifestation of a clinical or subclinical risk, were identified - forming the top level of the BCEN. If any of these identified risk factors are continuous, it is discretized based on the extended χ^2 algorithm [51]. The reason for performing this step was to alleviate the associated computational complexity when analysis was performed with SVM.

Next, clinical features categorized under the "system level" of the BC were input into GA-SVM for investigation. This, similar to the earlier step, aims to identify clinical features that have a significantly impact to the inference of the phenotypic manifestation previously identified at the "body level". The resultant output from this step forms the "system level" of BCEN. This procedure is repeated for the rest of the levels along the BC, constructing a probabilistic tree-structured BCEN at the end of this propagation. The resultant BCEN is capable of scrutinizing how, for instance, clinical

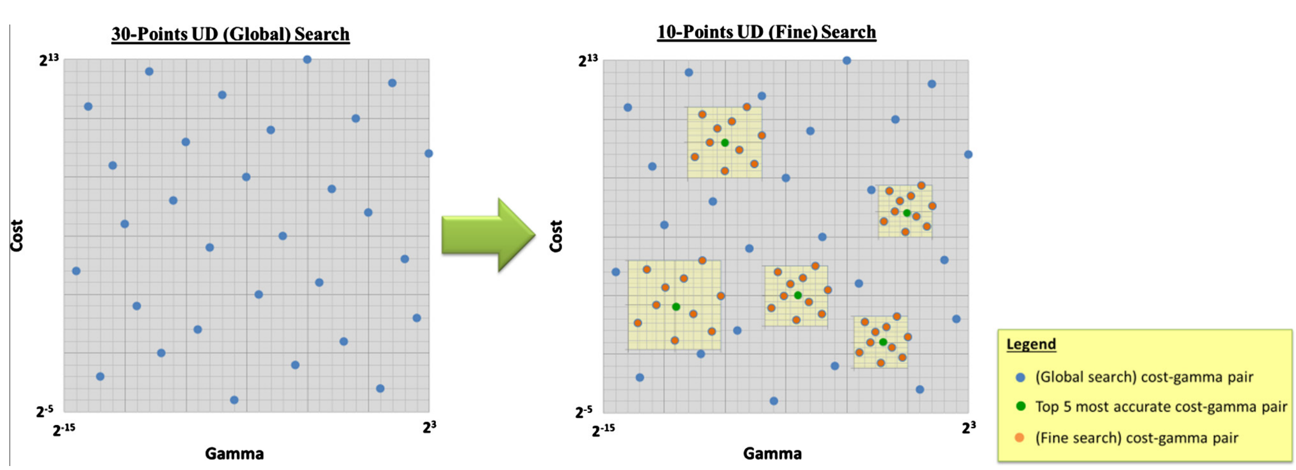


Fig. 2. Graphical illustration of SVM parameter optimization using UD technique. 30-Point UD (Global) search is first performed to determine regions with cost-gamma combinations that would produce the optimal SVM model. Subsequently, 10-point UD (fine) search is carried out to determine the optimal parameter set.

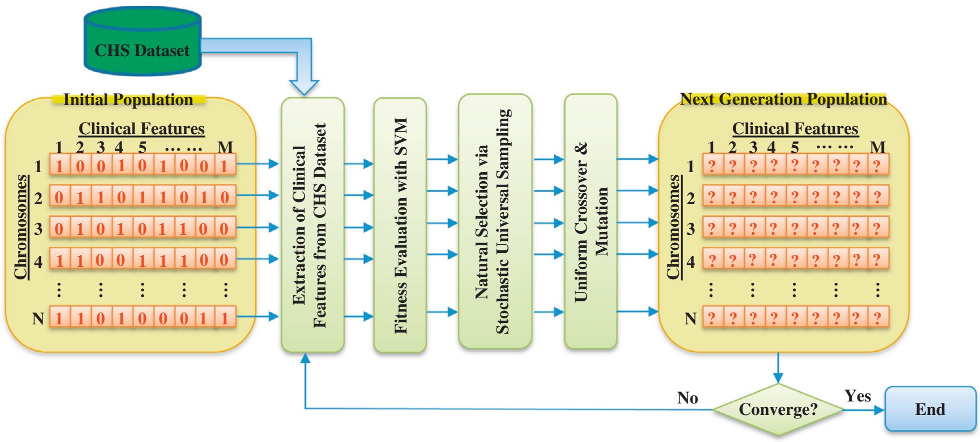


Fig. 3. A schematic illustration of clinical feature selection based on GA-SVM. A string of binary value in each chromosome (of size M) represent the present (value of 1) or absent (value of 0) of a risk factor during evaluation. This set of chromosomes (of size N) is randomly generated at onset and undergo the selection process postulated by GA to find the set of risk factors that produce the highest fitness value. Fitness evaluation is performed with SVM.

features at the visceral level are associated with those at the system level and, in turn, how these features at the system level are associated with those at the body level. This concept is graphically illustrated in Fig. 4.

3.3. MI classification with BCEN

After the construction of BCEN for MI, the distinct risk factors present in the network were used to develop an MI classification model. The performance (both classification accuracy and computational time) yielded with this approach was compared with an MI classification model that uses all clinical features present in the CHS dataset. GA-SVM was used as the classification algorithm for both the postulated approaches; hence, any benefits or drawbacks of using this classifier would prevail in both approaches.

4. Experimental results

4.1. Data preprocessing

Records and clinical features with considerable missing entries were removed. In addition, only records with known MI status were selected. This resulted in a dataset comprising of 4612 instances and 272 clinical features, with less than 1% of missing values (with respect to the entire dataset) and 40.8% of records with complete entries. The training and query datasets thus have 1881 and 2731 instances (both with 272 features), respectively. Subsequently, the K neighbor value for each clinical feature was determined based on the normalized training dataset. This yielded an average K value of 9.80, with standard deviation of 9.38. Data imputation was next performed to impute the missing entries found in the query dataset.

The imputed dataset obtained has a high fraction of controls (i.e. without MI – 4200 instances) and a relatively small portion of cases (i.e. with MI – 412 instances). SOM was thus employed to resolve this class data imbalanced problem. Under-sampling was performed on the major class (i.e. controls), yielding 441 instances. The final dataset produced has 853 instances and 272 clinical features.

4.2. Segregation of clinical features

The construction of a BCEN involved the segregation of the clinical features (173 diagnostic measurements and 1 MI status) along the BC. These 173 clinical features (after excluding medication) satisfied the characteristics of only 4 levels of the BC; namely,

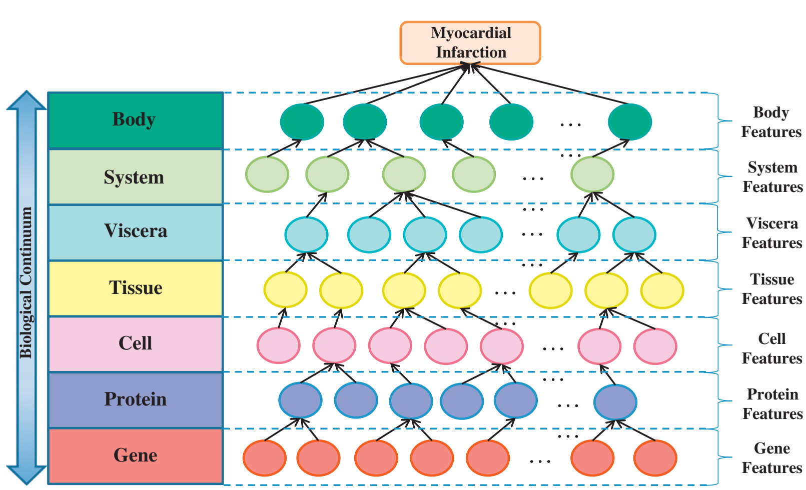


Fig. 4. Graphical illustration of BCEN. The circles represent clinical feature that belong to the respective levels of the BC. The arrows linking the clinical features indicate that a significant correlationship was found between them.

body, system, viscera and protein. Among these clinical features, 38, 74, 41 and 20 belong to the body, system, viscera and protein levels, respectively. A description of the segregated clinical features is provided online as an Appendix at http://www.bg.ic.ac.uk/jtay/web/chs_appendix.html. Readers may refer to the CHS data dictionary made available at the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) website for more information (https://biolincc.nhlbi.nih.gov/studies/chs/).

4.3. Construction of BCEN and classification of MI

Clinical features at the body level were first deployed to GA-SVM to determine the set of risk factors that were highly correlated to MI (root node). A total of 11 risk factors, namely ANGBASE (angina status at baseline), CHFBASE (congestive heart failure at baseline), STRKBASE (stroke status at baseline), CBD (self-reported stroke, transient ischemic attack (TIA) and cardiac endarterectomy), SCORE03 (social support score), AMOUNT (cigarettes smoked per day), WGTEEN (teenage weight category), OVRWT120 (obesity > 120% ideal), EDUC (education level), WAIST (waist circumference – cm) and ALCOH (number of alcoholic beverages per week) were identified at the body level (note that these modifiable risk factors are also identified in earlier reported clinical studies [52,53]).

When extending the network, only clinical feature subsets (child nodes) that yielded a balanced accuracy of at least 0.7 were considered. This threshold was imposed to reflect only child nodes that are highly correlated to their parent node. This resulted in 5 inner nodes at the body level – namely ANGBASE, CHFBASE, STRKBASE, CBD and OVRWT120. This criterion was applied to the rest of the levels of the BC.

The resultant inner nodes identified at the system level include ANBLMOD (angina modified at baseline status), CLBLMOD (claudication modified baseline status), SUPPUL16 (supine reading: 30 s heart rate), CHSTPN (chest pain) and VISPROB (vision problem). Table 1 provides the details of the best-performing clinical feature subsets that satisfy the aforementioned criteria. Note that none of the clinical features at the protein level correlated well with those at the visceral level. The authors believe that this could be due to the discontinuity in continuum along the BC (i.e. missing data at the tissue and cell levels) when estimating the association between the clinical features and phenotypic manifestation that resulted in the low performance.

The resultant BCEN consists of 111 distinct nodes (Body level: 11; System Level: 63; Viscera Level: 37) in total, accounting for 64.1% of the original number of clinical features analyzed. The complete BCEN for MI (created using prefuse toolkit [54]) is illustrated in Web Fig. 1 – available at http://www.bg.ic.ac.uk/jtay/web/chsBCENFull.html. The BCEN provides a visual and interactive etiological network for the user to visualize and comprehend the relationship among the different risk factors along the BC for MI. For our discussion here, a sub-network of the BCEN was analyzed because of its complexity and numerous interrelated risk factors present in the complete network. This sub-network is presented in Fig. 5.

Referring to Fig. 5, it can be seen that obesity (OVRWT120), a risk factor of MI, has 34 risk factors at the system level that are highly correlated with it. These risk factors are related to rheumatology, physical function, oncology, pulmonology, thromboembolism, sleep disorder, ophthalmology, otolaryngology, cognitive function and endocrinology. They account for 45.9% of the clinical features analyzed at the system level. This suggests that not all clinical features at the system level are good predictors of obesity and it could be more fruitful to focus investigations on significantly contributing clinical features.

MI classification, with GA-SVM algorithm, was next performed with the 111 clinical features that were present in BCEN. Baseline comparison was made with the original set of 173 clinical features present in the imputed CHS dataset. Results, as shown in Table 2, were obtained from averaging 3 runs of GA-SVM. For each method, the best-performing clinical feature subset for the different runs is the same. Comparable classification performance was achieved for both the methods. However, the computational time required by the proposed method (i.e. deploying only risk factors present in the BCEN to GA-SVM algorithm) to develop the MI classification model was much lower (approximately 14.7 h).

5. Discussion

To develop MI classification models efficiently in high dimensional datasets, we introduced a novel methodology for the reduction of clinical features to be analyzed without compromising the performance of the classification model. Classification (without feature selection) conducted on a large number of clinical risk factors often produced low-performing classification models, as the performance is often jeopardized by the present of irrelevant or

Table 1
Details of best-performing clinical feature subsets.

Parent node	Child nodes	# Inner nodes	# Leaf nodes	Total nodes	ACC	SN	SP	PR	FM	BA
MI Status	Clinical features at body level	5	6	11	0.828	0.786	0.866	0.846	0.815	0.826
Body level										
ANGBASE	Clinical features at system level	4	19	23	0.814	0.741	0.929	0.416	0.428	0.835
CHFBASE		2	16	18	0.958	0.559	0.855	0.596	0.575	0.707
STRKBASE		0	12	12	0.958	0.701	0.905	0.878	0.672	0.803
CBD		3	18	21	0.955	0.734	0.983	0.841	0.784	0.858
OVRWT120		2	32	34	0.737	0.737	0.738	0.704	0.720	0.737
System level										
ANBLMOD	Clinical features at viscera level	0	25	25	0.785	0.562	0.931	0.841	0.673	0.746
CLBLMOD		0	18	18	0.955	0.426	0.991	0.767	0.548	0.709
SUPPUL16		0	9	9	0.828	0.865	0.749	0.834	0.849	0.807
CHSTPN		0	16	16	0.717	0.794	0.609	0.695	0.741	0.702
VISPROB		0	21	21	0.829	0.667	0.981	0.568	0.609	0.824

Column 1 provides the best-performing clinical features at different levels of the BC: ANGBASE = angina status at baseline; CHFBASE = congestive heart failure at baseline; STRKBASE = stroke status at baseline; CBD = self-reported stroke, transient ischemic attack and cardiac endarterectomy; OVRWT120 = obesity > 120% ideal; ANBLMOD = angina modified at baseline status; CLBLMOD = claudication modified baseline status; SUPPUL16 = supine reading: 30 s heart rate; CHSTPN = chest pain; VISPROB = vision problem.
Columns 6 to 11 represent the various performance measurements: ACC = accuracy; SN = sensitivity; SP = specificity; PR = precision; FM = F-measure; BA = balanced accuracy.

Fig. 5. Sub-network of BCEN for MI. Eleven clinical features at the body level were found to be potential etiological factors of MI. Obesity, one of the risk factor of MI, consists of 34 highly correlated clinical features at the system level.

Table 2
Performance of classification with and without BCEN.

Experiment	#Features considered	#Gen	Time taken, hours (Mean ± SD)	ACC	SN	SP	PR	FM	BA
Baseline method: classification with original set of risk factors	173	73	69.6 ± 0.136	0.941	0.993	0.893	0.897	0.942	0.943
Proposed method: classification with risk factors present in BCEN	111	21	14.7 ± 0.005	0.931	0.995	0.871	0.878	0.933	0.933

These experiments were executed in parallel over an 8-core computer server. The best-performing clinical feature subset is the same for the different runs. '#Gen' denotes the number of generations taken by GA before it converges.

redundant predicators. On the other hand, the development of classification models with feature selection (e.g. the baseline method used in this work) conducted on a large number of clinical risk factors is usually computationally expensive. Therefore, pre-selection of clinical risk factors is vital to mitigate this problem contributed by the 'curse of dimensionality'. This was performed by segregating the clinical features along the various levels of the BC. The segregation process effectively reduces the data dimension, where its size is dependent on the number of clinical features categorized under each level of the BC. In this study, for example, analysis performed at the "body level" requires only 38 clinical features to be considered at a time. This, in contrast to the initial 173 clinical features, offers a reduction of 4.55-fold in the data dimension. Having to analyze a smaller number of clinical features inevitably reduces the amount of computational time required to develop the classification model. Moreover, if prior knowledge is available the data dimension can be further restricted. For instance, Emily et al. [29] utilize knowledge from protein databases to reduce the search of SNPs to gene pairs that are known to interact and reference. A similar concept can be applied to other levels of the BC to alleviate the search effort required.

Although effort is required to construct the BCEN, the resultant network has several advantages. Firstly, with the introduction of new clinical risk factors the entire BCEN need not be reconstructed. It provides a reusable framework where only the level of the BC, at which the new clinical risk factor belong to, need to be redeveloped. If the newly introduced clinical risk factor is identified as an etiological factor (i.e. risk factor contributing to the cause of the disease), then starting with that clinical risk factor as the root node, the network is extended for levels of the BC that is below that of the newly inserted etiological factor. This approach thus provides a significant reduction in the time and effort required to build up-to-date clinical classification models. Secondly, the BCEN provides an excellent paradigm for the illustration of the potential biological pathways that underpin the different phenotypic manifestations and has the significant advantage of analyzing only clinical risk factors that are biologically plausible. This not only allows the identification of significant risk factors that can be used for efficient development of accurate classification models, but, also, (1) reveals relationships that are not readily apparent from the study of individual disorders, (2) provide a global perspective of the different risk factors and etiologic pathways associated with the disease, and (3) identify new risk factors that could pave the way to the development of novel diagnostic, preventive or therapeutic strategies. Therefore, BCEN may be a simple etiological network, but it has the potential to provide significant insights into the mechanisms of a disease.

The constructed BCEN was validated by comparing the identified inter-relationship among different risk factors with those reported in previous clinical studies. All risk factors found at the body level of BCEN were also identified in previous clinical studies. Further, comparisons of a sub-network of BCEN (i.e. obesity-system sub-network) have shown that there is a large overlap (of 82.4%) between the identified relationships and those found in previous work. A possible reason for the identification of the additional inter-relationships is the employment of machine learning techniques. Since previous clinical studies tend to use linear statistical models to perform the analysis, non-trivial and non-linear relationships may go undetected. Therefore, the use of machine learning techniques in this work could potentially identify the non-trivial, non-linear and interacting etiological factors. This enables one to better understand the underlying causes of the disease, allowing more appropriate and focus interventions to be recommended to the patients. Table 3 lists the risk factors found to be highly associated with obesity and their presence in the clinical literature.

Arthritis, for instance, has been reported previously to be more prevalent among obese patients [55,56]. This is primarily due to the presence of excess biomechanical stress, inducing deleterious effect on the joints. Similarly, obese individuals have a higher risk of cancer related to endometrium, prostate, colon, esophagus and stomach [57,58]. Previously reported investigations have also shown association between obesity and bronchitis, pneumonia, emphysema, deep vein thrombosis, intermittent claudication, duration of sleep, blindness, hearing impairment, activities of daily living, pulmonary embolism, ankle-arm index, loss of balance, walking capacity, cognitive function, unstable angina, stroke, transient ischemic attack, hypertension and diabetes [59–75].

This suggests that the BCEN is feasible and effective in characterizing a disease and identifying the possible etiological factors. It is noteworthy that analysis of the obesity-system sub-network identified 6 new clinical features that were not previously identified in previous work. This could indicate that these clinical features are potential etiological factors of MI where further investigations could improve the understanding and treatment of the disease. We hypothesize that the reconstruction of the etiologic pathways is of major importance in healthcare as it would allow a more proactive approach for providing medical interventions to eradicate or delay the onset of a disease. This differs from the traditional reactive approach where individuals visit a physician only when they are sick or in pain, which sometimes results in a situation where treatment is too late to achieve complete recovery. Early medical interventions can be realized with BCEN by monitoring and controlling the risk factors (especially at the lower levels of the BC) that contribute to the development of a disease (e.g. MI).

Table 3
Obesity-system level risk factors.

Variable	Description
ARTH01[a]	Arthritis
DIAG01[a]	Ever diagnosed with cancer
BRONCH[a]	Bronchitis confirmed by doctor
PNEUMON[a]	Pneumonia detected by doctor
EMPHYSEM[a]	Emphysema detected by doctor
THROMB[a]	Deep vein thrombosis
ROSEIC[a]	Intermittent claudication by rose questionnaire
GROGGY[b]	Groggy in morning
TRSLEEP[a]	Trouble falling asleep
WKERLY[a]	Wake up far too early
RECOGN[a]	See enough to recognize person
TELE[a]	Hear enough to use phone
CONVER[a]	Hear enough to converse
ADL[a]	Activities of daily living (ADL)
IADL[a]	Instrumental ADL score
UES[b]	Upper extremity score
BLEED12[b]	Bleed or bruise easily
CLOT12[a]	Disorder related to blood clotting
LTAAI[a]	Left ankle-arm index (%)
SUPPUL16[b]	Supine reading: 30 s heart rate
BIORES21[a]	Bioelectric impedance – resistance
BAL22[b]	Dizziness, loss of balance screen
LOSBAL22[a]	Loss of balance
DIZZY22[b]	Dizzy/light-headed when stand up quickly
TIMEWLK[a]	15 feet walk time-sec
DIGCOR[a]	Digit symbol score
SCOR3510[a]	Mini-mental score (35pt)
SCORE30[a]	Mini-mental score (30pt)
ANBLMOD[a]	Angina modified baseline status
CHBLMOD[a]	CHF modified baseline status
STBLMOD[a]	Stroke modified baseline status
TIBLMOD[a]	TIA modified baseline status
HYPER[a]	Calculated hypertension status
DIABADA[a]	ADA guidelines diabetic status

[a] Risk factors found in previous work.
[b] Potential risk factors not found in previous studies (to the best of our knowledge).

The employment of BCEN to reduce the number of clinical features to be analyzed significantly alleviated the computational demands. Without acutely compromising the classification performance, a speedup of approximately 4.73-fold was achieved. This was possible due to the earlier convergence of GA, suggesting that significant risk factors are already identified and present in BCEN. This facilitates the identification of risk factors that contribute significantly to the modeling of accurate MI classification model.

This study has a few limitations. Firstly, only a single dataset (i.e. CHS dataset) was used to build the etiological network for MI. This inevitably limits the power to detect all the associated risks and conclusively state that the BCEN has described the complete etiology of MI. Additionally, it limits the ability to state that the proposed method provides efficiency for all clinical classification problems. Nonetheless, it does shed some light to a novel approach for investigating the etiology of MI and efficient clinical classification. Secondly, only a single classification algorithm (i.e. SVM) has been used to identify the association between the clinical features and for developing MI classification model. This may hinder the discovery of the underlying associations and the performance of the classification model, as no single machine learning technique or statistical model is optimal for every problem. The reason for this is because each method would have its own inductive bias [76]. Hence, it is suggested in [18] that comparison between multiple machine learning techniques, traditional statistical models and expert-based schemes should be conducted in order to assess the suitability of each method for a particular problem. Finally, the CHS dataset only contains risk factors that fall under the body, system, visceral and protein levels. This hinders the construction of a complete BCEN, limiting the ability to provide a more comprehensive illustration of the underlying etiology of a disease and the development of a more accurate classification model.

Nevertheless, the constructed BCEN is potentially capable of presenting the etiology of a disease in a biologically-structured manner that could facilitate the understanding and management of a disease. Moreover, it offers an effective and efficient approach for the development of MI classification model.

6. Conclusions

In view of the high prevalence of MI worldwide, better ability to characterize and classify the disease is both appropriate and necessary. In this paper we have presented an integrated approach to build a single probabilistic network (i.e. BCEN which identifies and relates the etiological factors associated with MI) that aims to provide an efficient approach for the development of MI classification model.

Validation of the constructed BCEN was conducted and our results indicate that the network is reliable and capable of identifying significant etiological factors. There is a large overlap between the relationships identified by our approach and those found in previous work. Out of the 34 clinical features identified at the obesity-system level, 28 (82.4%) of them were found in the previous clinical studies. However, 6 new clinical features, that had not been identified previously, were found to be associated with obesity in this study. These new clinical features could be probable risk factors for MI. They indicate the need for further clinical investigations to improve the understanding and treatment of the disease.

Based on the distinct risk factors identified and present in BCEN, a classification model for MI was developed. The classification model obtained demonstrated high balanced accuracy of 0.933. It was developed at a rate of 4.73-fold faster than its counterpart that does not adopt any pre-selection strategy. This suggests that BCEN may be a desirable approach for developing clinical classification models when a large number of clinical features need to be considered.

Although further validation of this methodology is necessary, this approach may be valuable in exploring and identifying risk factors that underpin a disease. To conclude, the BCEN is an etiological network that is simply built but profoundly useful. It has the potential to provide insights, from a novel perspective, into the characteristics of (current/new) diseases - allowing more efficient and effective understanding, analysis, management and classification to be undertaken. We look forward to a more comprehensive understanding of the disease etiology and eventually, towards personalized medicine.

Acknowledgments

Darwin Tay would like to express his sincere gratitude for his scholarship funding provided by the Nanyang Technological University-Imperial College London Joint PhD programme.

The authors would like to thank the National Heart, Lung and Blood Institute (NHLBI) for providing the CHS dataset.

Professor Richard Kitney and Dr Carolyn Goh wish to acknowledge the support of The Engineering and Physical Science Research Council (EPSRC) in this study.

Darwin Tay would also wish to thank EPSRC for partial support at Imperial College.

References

[1] Bellman R. Adaptive control processes. Princeton New Jersey: Princeton University Press; 1961. pp. 274.
[2] Baxt W, Skora J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. Lancet 1996;347(8993):12–5.
[3] Menown IBA, Mackenzie, Adgey AAJ. Optimizing the initial 12-lead electrocardiographic diagnosis of acute myocardial infarction. Eur Heart J 2000;21(4):275–83.
[4] Roger V, Go A, Lloyd-Jones D, Adams R, et al. Heart disease and stroke statistics-2011 update: a report from the American heart association. Circulation; 2011.
[5] Wilson W, D'Agostino R, Levy D, Belanger A, et al. Prediction of coronary heart disease using risk factor categories. Circulation 1998;32(18):560–5.
[6] British Heart Foundation Statistics Database. Coronary heart disease. <http://www.bhf.org.uk/publications/view-publication.aspx?ps=1001546; 2012 [accessed 08.08.12].
[7] Hsia T, Chiang H, Chiang D, Hang L, et al. Prediction of survival in surgical unresectable lung cancer by artificial neural networks including genetic polymorphisms and clinical parameters. J Clin Lab Anal 2003;17(6):229–34.
[8] Pittman J, Huang E, Dressman H, Horng CF, et al. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. Natl Acad Sci USA 2004;101(22):8431–6.
[9] Kitney R, Poh CL. Geometric framework linking different levels of the biological continuum. Engineering in Medicine and Biology Society, 2005. In: IEEE-EMBS 2005, 27th annual international conference of the; 2006.
[10] Poh CL, Kitney J, Shrestha R. Addressing the future of clinical information systems - web-based multilayer visualization. IEEE Trans Inf Technol Biomed 2007:127–40.
[11] Holland J. Genetic algorithms. Sci Am 1992:66–72.
[12] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory; 1992.
[13] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.
[14] Vapnik V. An overview of statistical learning theory. IEEE Trans Neural Networks 1999;10(5):988–99.
[15] Fried L, Borhani N, Enright P, Furberg C, et al. The cardiovascular health study: design and rationale. Ann Epidemiol 1991;1(3):263–76.
[16] Bhatla Nidhi, Jyoti Kiran. An analysis of heart disease prediction using different data mining techniques. Int J Eng Res Technol 2012;1(8):1–4.
[17] Roganb J, Franklina J, Stowb D, Millerc J, et al. Mapping land-cover modifications over large areas: a comparison of machine learning algorithms. Remote Sens Environ 2008;112(5):2272–83.
[18] Cruz JosephA, Wishart DavidS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2006;2:59–77.
[19] Eggers K, Ellenius J, Dellborg M, Groth T, et al. Artificial neural network algorithms for early diagnosis of acute myocardial infarction and prediction of infarct size in chest pain patients. Int J Cardiol 2007;114(3):366–74.
[20] Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. Comput Syst Appl 2008.
[21] Hossain Jafreen, FazlidaMohdSani Nor, Mustapha Aida, SurianiAffendey Lilly. Using feature selection as accuracy benchmarking in clinical data mining. J Comput Sci 2013;9(7):883–8.

[22] Latifoğlu F, Polat K, Kara S, Güneş S. Medical diagnosis of atherosclerosis from carotid artery doppler signals using principal component analysis (PCA), *k*-NN based weighting pre-processing and artificial immune recognition system (AIRS). J Biomed Inform 2008;41(1):15–23.

[23] Ohlsson M. WeAidU-a decision support system for myocardial perfusion images using artificial neural networks. Artif Intell Med 2004;30(1):49–60.

[24] Nilsson J, Ohlsson M, Thulin L, Höglund P, et al. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. J Thorac Cardiovasc Surg 2006;132(1):12–9.

[25] Yang J, Honavar V. Feature subset selection using a genetic algorithm. IEEE Intell Syst Appl 1998;13(2):44–9.

[26] Kohavi Ron, Sommerfield Dan. Feature subset selection using the wrapper method: overfitting and dynamic search space topology. In: First international conference on knowledge discovery and data mining; 1995.

[27] Saeys Yvan, Inza Inaki, Larranaga Pedro. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23(19):2507–17.

[28] Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997;1(3):131–56.

[29] Emily M, Mailund T, Hein J, Schauser L, et al. Using biological networks to search for interacting loci in genome-wide association studies. Eur J Hum Genet 2009;17(10):1231–40.

[30] Chu C, Hsu A, Chou K, Bandettini P, et al. Does feature selection improve classification accuracy? Impact of Sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage 2012;60(1):59–70.

[31] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[32] Wiener J, Tilly J. Population ageing in the United States of America: implications for public programmes. Int J Epidemiol 2002;31(4):776–81.

[33] Abbott R, Curb J, Rodriguez B, Masaki K, et al. Age-related changes in risk factor effects on the incidence of coronary heart disease. Ann Epidemiol 2002;12(3):173–81.

[34] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory 1967(1):21–7.

[35] Kohonen T. The self-organizing map. Proc IEEE 1990;78(9):1464–80.

[36] Batista G, Monard MC. A study of K-nearest neighbour as an imputation method. In second international conference on Hybrid Intelligent Systems, Santiago, Chile, Soft Computing Systems: design, management and applications. IOS Press 2002:251–60.

[37] Troyanskaya O, Cantor M, Sherlock G, Brown P, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17(6):520–5.

[38] Acuña E, Rodriguez C. The treatment of missing values and its effect on classifier accuracy. In: Studies in classification, data analysis, and, knowledge organization, vol. 0; 2004. pp. 639–47.

[39] Batista G, Monard M. A study of K-nearest neighbor as an imputation method. In: Second international conference on hybrid intelligent systems; 2002.

[40] Jerez J, Molina I, García-Laencina P, Alba E, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med 2010;50(2):105–15.

[41] Garcia-Laencina P, Vidal A, Sancho-Gomez JL. A robust approach for classifying unknown data in medical diagnosis problems. IEEE World Auto Congress (WAC) 2008.

[42] Minaei-Bidgoli B, Kashy D, Kortmeyer G, Punch W. Predicting student performance. An application of data mining methods with an educational web-based system. Frontiers in Education; 2003. FIE 2003. 33rd Annual, 2003.

[43] Dudani S. The distance-weighted k-nearest-neighbor rule. IEEE Trans Syst, Man Cyber 1976;6(4):325–7.

[44] Japkowicz N. Learning from imbalanced data sets: a comparison of various strategies. AAAI workshop on learning from imbalanced data sets; 2000.

[45] Li D, Liu C, Hu S. A learning method for the class imbalance problem with medical data sets. Comput Biol Med 2010;40(5):509–18.

[46] Wua W, Walczaka1 B, Massarta D, Heuerdingb S, et al. Artificial neural networks in classification of NIR spectral data: design of the training set. Chem Intell Labor Syst 1996;33(1):35–46.

[47] Chang CC, Lin C. LIBSVM: a library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.

[48] Fang KT, Lin DKJ, Winker P, Zhang Y. Uniform design: theory and application. Technometrics 2000;42(3).

[49] Chow Rick, Zhong Wei, Blackmon Michael, Stolz Richard, Dowell Marsha. An efficient SVM-GA feature selection model for large healthcare databases. Genetic and evolutionary computation conference (GECCO); 2008.

[50] Chen SH, Sun J, Dimitrov L, Turner AR, et al. A support vector machine approach for detecting gene-gene interaction. Genet Epidemiol 2008;32(2):152–67.

[51] Su CT, Hsu JH. An extended χ^2 algorithm for discretization of real value attributes. Knowl Data Eng, IEEE Trans 2005(3):437–41.

[52] Yusuf S, Hawken S, Ounpuu S, Dans T, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet 2004;364(9438):937–52.

[53] Rosengren A, Subramanian S, Islam S, Chow C, et al. Education and risk for acute myocardial infarction in 52 high, middle and low-income countries: INTERHEART case-control study. Heart 2009;95(24):2014–22.

[54] Heer J, Card SK, Landay JA. Prefuse: a toolkit for interactive information visualization. In: Proceedings of the SIGCHI conference on Human factors in, computing systems; 2005.

[55] Holliday K, McWilliams D, Maciewicz R, Muir K, et al. Lifetime body mass index, other anthropometric measures of obesity and risk of knee or hip osteoarthritis in the GOAL case-control study. Osteoarthr Cartilage 2011;19(1):37–43.

[56] Park H, Lee S. Association of obesity with osteoarthritis in elderly korean women. Maturitas 2011;70(1):65–8.

[57] Kane C, Bassett W, Sadetsky N, Silva S, et al. Obesity and prostate cancer clinical risk factors at presentation: data from CaPSURE. J Urol 2005;173(3):732–6.

[58] Yang P, Zhou Y, Chen B, Wan H, et al. Overweight, obesity and gastric cancer risk: results from a meta-analysis of cohort studies. Eur J Cancer 2009;45(16):2867–73.

[59] Guerra S, Sherrill DL, Bobadilla A, Martinez FD, et al. The relation of body mass index to asthma, chronic bronchitis, and emphysema. Chest 2002;122(4):1256–63.

[60] Corrales-Medina V, Valayam J, Serpa J, Rueda A, et al. The obesity paradox in community-acquired bacterial pneumonia. Int J Infect Diseas 2011;15(1).

[61] Samama M. An epidemiologic study of risk factors for deep vein thrombosis in medical outpatients: the sirius study. Arch Intern Med 2000;160(22):3415–20.

[62] Golledge J, Leicht A, Crowther RG, Clancy P, et al. Association of obesity and metabolic syndrome with the severity and outcome of intermittent claudication. J Vasc Surgery 2007;45(1):40–6.

[63] Patel SR, Blackwell T, Redline S, Ancoli-Israel S, et al. The association between sleep duration and obesity in older adults. Int J Obes 2008;32(12):1825–34.

[64] Patterson RE, Frank LL, Kristal AR, White E. A comprehensive examination of health conditions associated with obesity in older adults. Am J Prev Med 2004;27(5):385–90.

[65] Habot-Wilner Z, Belkin M. Obesity is a risk factor for eye diseases. Harefuah 2005;144(11):805–9.

[66] Fransen E, Topsakal V, Hendrickx JJ, Laer LV, et al. Occupational noise, smoking, and a high body mass index are risk factors for age-related hearing impairment and moderate alcohol consumption is protective: a European population-based multicenter study. JARO – J Assoc Res Otolaryngol 2008;9(3):264–76.

[67] Himes CL. Obesity, disease, and functional limitation in later life. Demography 2000;37(1):73–82.

[68] Stein PD, Beemath A, Olson RE. Obesity as a risk factor in venous thromboembolism. Am J Med 2005;118(9):978–80.

[69] Tison GH, Ndumele CE, Gerstenblith G, Allison MA, et al. Usefulness of baseline obesity to predict development of a high ankle brachial index (from the multi-ethnic study of atherosclerosis). Am J Cardiol 2011;107(9):1386–91.

[70] Gray D, Bray G, Gemayel N, Kaplan K. Effect of obesity on bioelectrical impedance. Amer Soc Clin Nutr 1989;255–260:255–60.

[71] Corbeil P, Simoneau M, Rancourt D, Tremblay A, et al. Increased risk for falling associated with obesity: mathematical modeling of postural control. IEEE Trans Neur Syst Rehab Eng 2001(2):126–36.

[72] Hulens M, Vansant G, Claessens AL, Lysens R, et al. Predictors of 6-minute walk test results in lean obese and morbidly obese women. Scandinavian J Med Sci Sports 2003;13(2):98–105.

[73] Elias MF, Elias PK, Sullivan LM, Wolf PA, et al. Lower cognitive function in the presence of obesity and hypertension: the framingham heart study. Int J Obes 2003;27(2):260–8.

[74] Wolk R, Berger P, Lennon R, Brilakis E, et al. Body mass index: a risk factor for unstable angina and myocardial infarction in patients with angiographically confirmed coronary artery disease. Circulation 2003;108:2206–11.

[75] Winter Y, Rohrmann S, Linseisen J, Lanczik O, et al. Contribution of obesity and abdominal fat mass to risk of stroke and transient ischemic attacks. Stroke 2008;39:3145–51.

[76] Freitas A, Timmis J. Revisiting the foundations of artificial immune systems for data mining. IEEE Trans Evol Comput 2007;11(4):521–40.

Appendix D

An Evolutionary Data-Conscious Artificial Immune Recognition System

Darwin Tay[1,2], Chueh Loo Poh[2,*], Richard I. Kitney[1]

[1] Department of Bioengineering, Imperial College London, UK

[2] Division of Bioengineering, Nanyang Technological University, Singapore

Email: darwintay@imperial.ac.uk, clpoh@ntu.edu.sg, r.kitney@imperial.ac.uk

ABSTRACT

Artificial Immune Recognition System (AIRS) algorithm offers a promising methodology for data classification. It is an immune-inspired supervised learning algorithm that works efficiently and has shown comparable performance with respect to other classifier algorithms. For this reason, it has received escalating interests in recent years. However, the full potential of the algorithm was yet unleashed.

We proposed a novel algorithm called the evolutionary data-conscious AIRS (EDC-AIRS) algorithm that accentuates and capitalizes on 3 additional immune mechanisms observed from the natural immune system. These mechanisms are associated to the phenomena exhibited by the antibodies in response to the concentration, location and type of foreign antigens. Bio-mimicking these observations empower EDC-AIRS algorithm with the ability to robustly adapt to the different density, distribution and characteristics exhibited by each data class. This provides competitive advantages for the algorithm to better characterize and learn the underlying pattern of the data. Experiments on four widely used benchmarking datasets demonstrated promising results – outperforming several state-of-the-art classification algorithms evaluated. This signifies the importance of integrating these immune mechanisms as part of the learning process.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning – *Concept learning, Induction, Knowledge acquisition, Parameter learning.*

Keywords

Artificial immune recognition system, Classification algorithm, Evolutionary computation.

*Corresponding author

1. INTRODUCTION

The human immune system is a highly sophisticated, distributed, complex and powerful natural defense mechanism that comprises of several functional mechanisms, positioned in strategic locations, conferring resistance against viruses and foreign pathogens. It has the ability to learn the characteristics of the foreign antigens and contrive a defense strategy to detect and neutralize them. Specifically, the immune system possesses properties such as the capability of recognition, memory acquisition, diversity and self-regulation, making it highly suitable for learning patterns that underlie a data. On this note, it has inspired the development of the artificial immune system capable of solving many problems related to computer science and engineering (e.g. computer security, anomaly detection, optimization, machine learning, etc.) [1, 2]. One such algorithm that has received escalating interests is the Artificial Immune Recognition System version 2 (AIRS2) [3].

Although AIRS2 algorithm has shown to be an effective classification algorithm, some useful immune mechanisms are yet to be exploited by the algorithm. For instance, artificial recognition balls (ARBs) are used in AIRS2 algorithm to denote a representative subset of B-Cells. They would compete for survival based on the idea of resource limited system [4]. However, the creation and elimination of the ARBs do not correspond to the density of the data in which they cover (i.e. a larger number of ARBs do not survive in regions that are more densely populated with data). This contradicts with the natural immune system where macrophages would flood the extracellular space of the infected regions (attempting to eliminate the harmful agents) and B-Cells would proliferate and secrete antibodies profoundly in response to pathogenic agents. In other words, a larger concentration of defense agents would be present in regions that has received intense invasion from harmful antigens. Another area that the original AIRS2 algorithm did not explore and exploit is the distributed diversity exhibited by the lymph nodes found in the natural immune system. The AIRS2 algorithm uses a common parameter set to model the distribution of different data classes. This is undesirable in cases where the distribution of different data classes differ by a considerable degree. Observation of the strategic positioning of the lymph nodes in human bodies (which promotes better immune defense) advocates for the need of a more specific and distinct parameter set (e.g. affinity threshold scalar, density and total resources parameters) to model each data class (i.e. instances that belong to a specific class). Finally, it is important to generate B-Cells that can affiliate/bind well with the antigens. This is realized biologically through the production of highly specific surface receptors on the B-Cells which facilitates the detection and eradication of the foreign antigens. To mimic this concept computationally, feature selection can be performed where highly informative features that can describe the underlying association were identified and used for classification.

This paper presents a novel algorithm called the evolutionary data-conscious AIRS (EDC-AIRS) algorithm, which extends the existing AIRS2 algorithm by contextualizing the immune response to the concentration, distribution and characteristics of the antigens and is no longer a global centralized response.

GECCO'13, July 6–10, 2013, Amsterdam, The Netherlands.

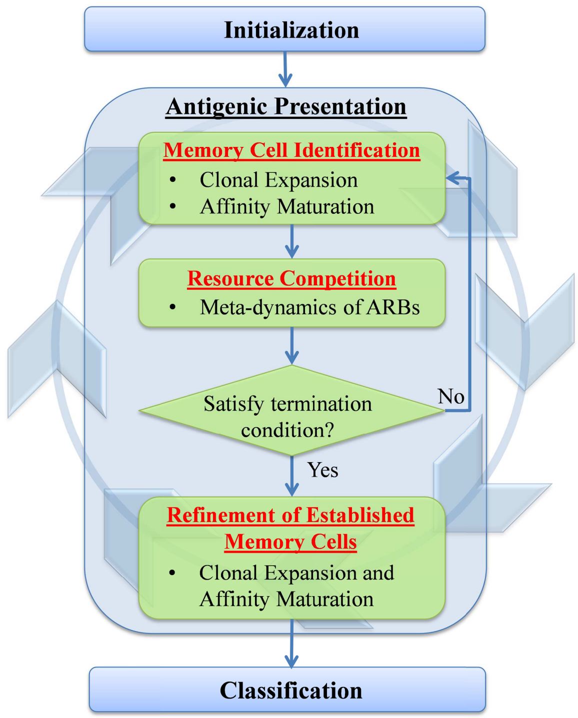


Figure 1: Canonical Flow of the AIRS2 Algorithm
The AIRS2 algorithm consists of 4 stages during the learning phase, namely the initialization, memory cell identification, resource competition and refinement of established memory cells stages. The last 3 stages will repeat for every data instance presented. After learning the underlying pattern of the data, classification is performed on the unseen data using K-nearest neighbor.

When evaluated using 4 widely used benchmarking datasets, our method has exhibited improved learning ability and classification accuracy.

The rest of the paper is organized as follows. Section 2 provides a brief introduction to the AIRS2 algorithm. A description of the proposed EDC-AIRS algorithm is presented in Section 3. The experimental results are offered in Section 4 and discussed in Section 5. Finally, conclusion is drawn in Section 6.

2. ARTIFICIAL IMMUNE RECOGNITION SYSTEM

The natural immune system is a highly rapid and efficient biological self-defense mechanism that protects a given host against infections, for example, from foreign antigens or pathogens. The immune system functions by detecting a wide variety of agents and distinguish the foreign antigens (e.g. viruses) from the organism's own healthy cells or molecules (also known as self-antigens). The immune system consists of a number of components. Two examples are macrophages and lymphocytes (e.g. B-cell ad T-Cell) which are responsible for the recognition and elimination of the determined infectious agents. The lymphocytes have highly specific surface antigenic receptors to a given antigenic determinant, in which they would only proliferate in response to a specific infection. Therefore, the type of antibodies present in an individual could reflect the infections to which they are infected with.

The antibody's polypeptide chains composed of a highly variable amino-terminal region (V-region) and a carboxy-terminal region (C-region) that can be of a few types. The V-region is responsible for the antigenic detection while the C-region is responsible for a variety of effector functions. The polypeptide chain of an antibody is formed through the genetic recombination and somatic hyper-mutation of multiple gene segments scattered along the chromosome of the genome. Such formation mechanism used to generate antibodies introduces diversity into the underlying immune defense [2], ameliorating the ability of the antibodies to recognize/bind to the antigens.

Inspired by the robustness exhibited by the natural immune system, the AIRS2 algorithm [3] - a novel one-shot incremental supervised learning algorithm was developed and applied to solve classification problems. It has several attractive characteristics such as the ability to (1) adaptively develop an appropriate architecture during the learning process, (2) achieve competitive accuracy compared to other classification algorithms, (3) develop a generalized model by generating a representative set of memory cells and (4) achieve accuracy comparable to those obtained with the optimal parameter set when experimented over a wide range of parameter values [5].

The AIRS2 algorithm consists of 4 stages during the process of learning the underlying patterns of the data. They are the initialization, memory cell identification, resource competition and refinement of established memory cells stages. The canonical flow of the AIRS2 algorithm is presented in Figure 1.

The initialization stage is responsible for normalization of the data, parameter discovery and seeding of memory cells. The data items found in the dataset is first normalized so that the Euclidean distance between the feature vectors of any 2 data items is in the range [0, 1]. Affinity threshold, the average Euclidean distance between each data item in the training dataset, is then calculated. This value controls the quality of the memory cells maintained and utilized for classification. The mathematical expression for computing the affinity threshold is as follow:

$$\text{affinity threshold} = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\text{affinity}(ag_i, ag_j)}{\frac{n(n-1)}{2}} \quad (1)$$

where n is the number of training instances (antigens), ag_i and ag_j are the ith and jth training antigens in the training data, and affinity(ag_i,ag_j) returns the Euclidean distance between the two antigens. The initial memory cell pool (MCP), a collection of classifier cells that will be used for classification at the end of the training lifecycle, is then seeded by randomly selecting data item(s) from the training dataset.

A process known as the antigenic presentation is then undertaken where each training instance is subsequently presented to the AIRS2 algorithm. For each training instance presented, it first undergoes the memory cell identification stage where its affinity with the memory cells in MCP (that reside in the same class) is computed. The most stimulated memory cell (also known as matched memory cell) is then selected and cloned in proportion to its stimulation value (i.e. clonal expansion phase). This value is calculated based on the following equation:

$$\text{stimulation}(x, y) = 1 - \text{affinity}(x, y) \quad (2)$$

where x is the presented training instance and y is the memory cell. These cloned memory cells forms the artificial recognition ball (ARB) pool where an ARB [4] is a single representation for a number of similar memory cells. This allows a reduction in duplication and manages the survival of classifier cells within the population. The cloned ARBs are then mutated at a rate inversely proportional to the antigenic affinity, introducing diversity into the system (i.e. affinity maturation phase).

```
CandStim ← stimulation(ag, mc_candidate)
MatchSim ← stimulation(ag, mc_match)
CellAff ← affinity(mc_candidate, mc_match)
if (CandStim > MatchSim)
    if (CellAff < AT * ATS)
        MC ← MC - mc_match
    end
    MC ← MC ∪ mc_candidate
end
```

Figure 2: Pseudo-code for Memory Cell Introduction used in AIRS2 Algorithm – adopted from [3]

CandStim (and MatchSim) denotes the stimulation level between the presented antigen and the candidate (and matched) memory cell. CellAff refers to the affinity between the candidate and matched memory cell. MC represents the memory cell pool.

The range of the maturated value assigned to a selected attribute is centered at the attribute's initial value and spanned over the difference between 1 and the ARB's stimulation value. In other words, mutated ARB offspring of highly stimulated cells are only allowed to explore and mutate to a value near its initial state while less stimulated ARB offspring are allowed to mutate over a lager range.

Next, the ARBs will compete for survival based on the concept of resource allocation mechanism [4], where the ARBs are allocated a number of resources proportional to their normalized stimulation values. The resulting ARBs with insufficient resources are subsequently pruned (i.e. meta-dynamic phase). The average simulation level for the ARBs is then computed based on the following equation:

$$\text{avg\_stimulation}_i = \frac{\sum_{j=1}^{|AB_i|} ab_j.stimulation}{|AB_i|}, \; ab_j \in AB_i \qquad (3)$$

where AB refers to the ARB pool, ab ε AB; $|AB_i|$ is the number of ARBs in class i. The average stimulation is then compared with the user-defined stimulation threshold. If it is greater than the user-defined threshold, the training cycle stops for that training instance. Otherwise, the training cycle repeats.

Once the termination condition is satisfied, the most stimulated ARB is selected as the candidate memory (CM) cell. If this CM cell's stimulation level is higher than all the memory cells in the established memory (EM) set (i.e. collection of ARBs that have survived the resource competition stage), then it is added into the EM set. Otherwise, this CM cell is discarded. Finally, replacement of the EM cells is carried out first by computing the memory cell replacement cutoff value as defined as:

$$\text{Cutoff} = \text{AT} * \text{ATS} \qquad (4)$$

where AT refers to affinity threshold and ATS denotes affinity threshold scalar. If the affinity between this CM cell and the best affiliated memory cell found previously (i.e. EM cell) is below the cutoff value, the EM cell will be removed and replaced with the CM cell. Consequently, the next training instance is deployed to the AIRS2 algorithm until all the training instances are presented. This process ultimately identifies a set of representative memory cells that provides a generalized representation of the pattern that underlies the data, which will then be used for classification. The classification algorithm employed is K-nearest neighbour (KNN) where the classification outcome for each unseen data instance is determined by taking the majority vote of the k most stimulated EM cells. For a more detailed description of the algorithm, readers can refer to [3, 6].

```
CellAff ← affinity(mc_candidate, mc_match)
Density_count ← 0
foreach (ag_i in AG)
do
    AntigenAff ← affinity(ag_i, mc_candidate)
    if (AntigenAff < AT*Radius_density)
        Density_count ← Density_count + 1
    end
done
Density_ratio ← Density_count / Density_max
if (CellAff < (1- Density_ratio) * AT * ATS)
    MC ← MC - mc_match
end
MC ← MC ∪ mc_candidate
```

Figure 3: Pseudo-code for Memory Cell Introduction used in EDC-AIRS Algorithm

$Density_{count}$ represents the number of antigens that is proximal to the candidate memory cell. $Density_{max}$ denotes the maximum number of antigen present in the training data.

3. Material & Methods

3.1 Evolutionary Data-Conscious AIRS (EDC-AIRS) Algorithm

This study formulates a novel immune-inspired (EDC-AIRS) algorithm that employs several natural immune mechanisms. In particular, how antibodies evolve and adapt to the different concentration, location and type of foreign antigens are being mimicked in addition to those proposed by the AIRS2 algorithm. This, when implemented as a high fidelity computational technique, empowers the algorithm with the ability to independently adapt to the distinct (1) density, (2) distribution and (3) characteristics of each data class.

Firstly, the ability to adapt to the different (local) density present in the data was addressed by the observation of the rapid growth of macrophages and B-Cells in response to the invasion of foreign antigens (particularly, at the regions of infection). More specifically, a relative proportion of antibodies to antigens were necessary to neutralize the harmful agents. This mechanism was incorporated in the EDC-AIRS algorithm by allowing a relatively larger number of ARBs to survive in regions that are more densely populated with training data. Implementation was carried out by removing and modifying some of the criteria present in the original AIRS2 algorithm. In particular, the way the memory cells are introduced into the system is modified. The original pseudo-code for memory cell introduction [3] used in AIRS2 algorithm is shown in Figure 2. The criterion that requires the candidate memory cell ($mc_{candidate}$) to be more stimulated (by the training antigen, ag) than the matched memory cell (mc_{match}) before it was added to the memory cell pool was first removed. The reason for doing so is to encourage new ARBs that are highly stimulated (ensured by the high stimulation threshold adopted) to survive within the system. Secondly, computation of the density ($Density_{count}$) proximal to $mc_{candidate}$, based on the initial set of training antigens (AG), was implemented in the algorithm. The degree of proximity was determined by a user-defined parameter, $Radius_{density}$. Additionally, the maximum density ($Density_{max}$) present in AG was also computed based on the same approach. However, another user-defined parameter ($Radius_{max}$) was used to determine the size of the region to be considered.

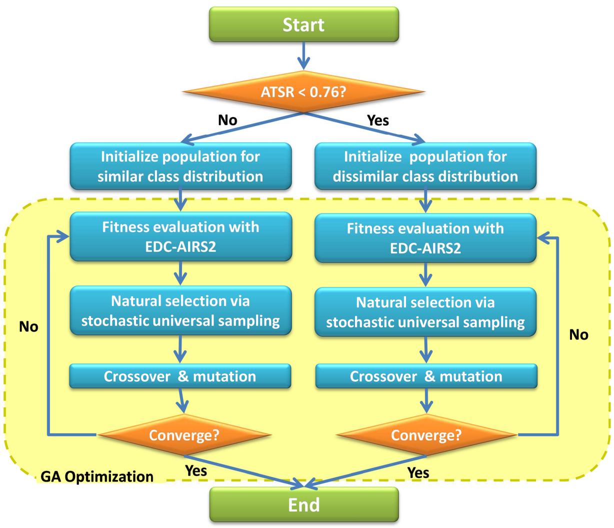


Figure 4: Proposed Methodology for Optimization of Parameter Set for Binary Class Classification Problems
The ATSR of the dataset was first computed. If it was smaller than the empirically derived threshold (0.76), an independent set of parameters for each data class was created. Otherwise, a common parameter set was used for all data classes. These parameters were then optimized (with respect to classification accuracy) using GA.

Finally, a weighting coefficient ($\text{Density}_{\text{ratio}}$) was derived based on the following equation:

$$\text{Density}_{\text{ratio}} = \frac{\text{Density}_{\text{count}}}{\text{Density}_{\text{max}}} \qquad (5)$$

This weighting coefficient was subsequently multiplied with the product of affinity threshold (AT) and the affinity threshold scalar (ATS) to determine whether the new memory cell should be introduced into the memory cell pool. The revised pseudo-code for memory cell introduction used in EDC-AIRS algorithm is depicted in Figure 3.

The strategy that was delved into next is associated with the distribution characteristic of different data classes. This is vital according to the mechanism observed in the natural immune system, where lymph nodes are located in strategic positions – producing antibodies that could detect and eradicate the foreign antigens more efficiently. The spatial independency of the lymph nodes [7] and the circulatory networks in the immune system is of significant importance as it enables decentralized immune defense while protecting the human body in a global fashion. Therefore, if the distribution of different data classes differs by too much, this could indicate that the location at which the antibodies are produced (i.e. the position of the lymph nodes) would need to be adjusted so that the antibodies produced could detect and eradicate the antigens found in each data class in a more efficient manner. On the contrary, if the distribution of different data classes is near symmetry, this could indicate that no additional lymph nodes are required for more efficient neutralization of antigens; mitigating the required search effort as a result. This was empowered within the EDC-AIRS algorithm by having an independent set of parameters for evolving the memory cells if the distribution similarity between the data classes was below an empirically derived threshold, known as the Affinity Threshold Similarity Ratio (ATSR). Otherwise, a common set of parameters was used for all data classes. The ATSR was calculated by first computing the affinity threshold (i.e. the average affinity value over all training data) associated with each data class.

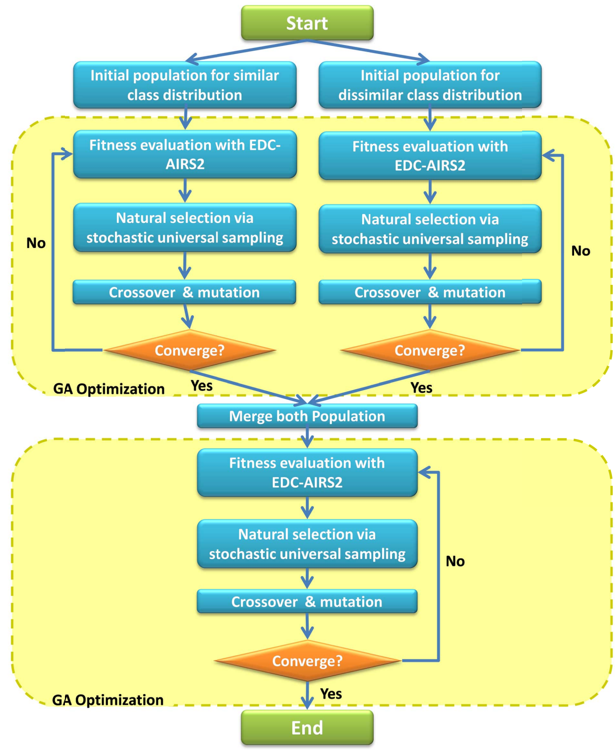


Figure 5: Proposed Methodology for Optimization of Parameter Set for Multiclass Classification Problems
Both similar and dissimilar populations were initialized and optimized concurrently. Upon convergence, these populations were merged and re-optimized by GA.

After which, the minimum affinity threshold found among the different data classes was divided by the maximum affinity threshold found. The mathematical expression is given below:

$$ATSR = \frac{\boldsymbol{min_{i=1,\ldots,C}(affinity_{threshold_i})}}{\boldsymbol{max_{j=1,\ldots,C}(affinity_{threshold_j})}} \qquad (6)$$

where C is the total number of classes. This provides a yardstick to determine how similar the distributions of different data classes were. The parameters that orchestrate the evolution, survival and existence of the memory cells include the total resources, affinity threshold scalar (ATS), $\text{Radius}_{\text{density}}$ and $\text{Radius}_{\text{max}}$ parameters. Therefore, these parameters were duplicated and optimized independently for each data class if the ATSR computed for the dataset was below the pre-defined threshold value. All parameters were optimized using Genetic Algorithm (GA) [8], a search heuristic that imitates the process of natural evolution. The optimization algorithm was developed using MATLAB GA toolbox [9] and was executed in parallel over a high performance computer (HPC) cluster.

Figure 4 and 5 illustrate the canonical flow of the proposed methodology used for binary and multiclass classification problems respectively. For binary classification problem, we assume either a similar or dissimilar distribution based on the calculated ATSR and optimize that parameter set only (i.e. distinct parameter set for each data class if the computed ATSR is below the pre-defined threshold and a common parameter set if ATSR is above the pre-defined threshold). In contrast, both similar and dissimilar distributions were assumed for multiclass classification problems. In other words, both set of parameters were optimized concurrently by GA for multiclass datasets. Upon convergence of both runs (i.e. no improvement after 10 generations or the maximum number of generations has been reached), both the populations were merged and re-optimized by GA once again.

Finally, the ability to adapt to the characteristics of the data was performed by mimicking the genetic recombination and somatic hyper-mutation of gene segments scattered along the chromosome of the genome when forming a natural antibody. This process produces highly specific surface receptors of B-Cell necessary to recognize and bind to a certain type of antigen (that possess distinct structure). From a computational perspective, this was achieved through feature selection where a subset of informative features, that could capture the true patterns underlying the particular dataset, was selected for the learning process. GA was selected to perform this feature selection task as it has the potential to generate the optimal feature subset [10]. The GA parameters were determined experimentally and kept constant between benchmarks. The setup details of GA are as follow: population size: 100; maximum generation: 100; natural selection: stochastic universal sampling; crossover type: discrete recombination; crossover probability: 0.8; mutation rate: 1/P, where P is the number of parameters. The value of the EDC-AIRS parameters that was either assigned (i.e. given as a constant value) or tuned by GA (i.e. given as a range of value) are as follow: seed: 1; clonal rate: 10; hyper-mutation rate: 2; stimulation threshold: 0.9; initial memory pool size: [0, 200]; K-nearest neighbor value: [1, 15]; affinity threshold scalar: [0, 1]; total resource: [150, 300]; $Radius_{density}$ = [0, 3]; $Radius_{max}$ = [0, 3].

The performance of EDC-AIRS algorithm was evaluated with 4 benchmarking datasets, namely the Fisher's Iris, Ionosphere, Pima Indians Diabetes and Sonar Datasets. Hold-out validation was performed on the Ionosphere dataset while cross-validation was performed on the remaining 3 datasets. More specifically, the first 200 data items of the Ionosphere dataset was selected as the training data and was tested on the remaining 151 data items. As for the Iris, Pima Indians Diabetes and Sonar datasets, 5, 10 and 13-fold cross-validation was carried out respectively. The reason for choosing such validation strategy was to remain comparable to other experiments reported in the literature. Further details about the validation procedures applied on these benchmarking datasets can be found in [11].

3.2 Dataset

Four benchmarking datasets obtained from [12] were used to evaluate the performance of the novel EDC-AIRS algorithm. A brief description of these datasets is as follow:

1. Fisher's Iris Dataset – Consists of 4 features that describe the length and width of the sepal and petal. Three classes exist which represent the type of the iris plant (i.e. Iris Sentosa, Iris Vericolour and Iris Virginica). It has a sample size of 150 with 50 instances per class. The Iris Sentosa class is linearly separable from the other 2 classes while the Iris Vericolour and Iris Virginica classes are not linearly separable from each other.
2. Ionosphere Dataset – A binary class classification problem that contains 351 instances and 34 features. The 2 classes represent "good" or "bad" radar returns. The "good" radar returns refer to those that show some types of structure in the ionosphere while "bad" radar returns have their signals passed through the ionosphere.
3. Pima Indians Diabetes Dataset – Patients in this dataset are all females who are at least 21 years of age and are of Pima Indian heritage. It is a binary class classification problem that aims to distinguish between patients tested positive for diabetes and those who are not. It contains 768 instances and 8 features.
4. Sonar Dataset – The objective of this experiment is to determine whether an object is a mine (metal) or rock by bouncing sonar signal off the object at various angles and conditions. It contains 208 instances and 60 features.

In order to investigate on how different data class distribution affects the performance of the classification algorithm, several additional benchmarking datasets were acquired from [12]. Both similar and dissimilar distributions among the data classes were assumed for these datasets. Experiments were then conducted using these 10 datasets, with different degree of data class distribution (as determined by the computed ATSR value), to determine the impact of data class distribution on the algorithm's classification performance.

Table 1: Empirical Experiments with ATSR based on Datasets with Different Data Class Distribution

Measurement	Ionosphere	Iris	Wine	ks_yr50611	MAGIC	Pima Indians Diabetes	Hill-Valley	Bupa-Liver Disorder	Sonar	Statlog Heart
#Instances	200	100	178	270	19020	768	606	345	208	270
#Attributes	34	4	13	253	10	8	100	6	60	13
#Classes	2	3	3	2	2	2	2	2	2	2
#Class1 Instances	99	50	59	135	12332	268	305	145	97	120
#Class2 Instances	101	50	71	135	6688	500	301	200	111	150
#Class3 Instances	-	50	48	-	-	-	-	-	-	-
Validation Type	Holdout	5-CV	LOO	10-CV	5-CV	10-CV	Holdout	10-CV	13-CV	10-CV
Class 1 AT	0.437	0.106	0.162	0.308	0.160	0.217	0.121	0.157	0.271	0.427
Class 2 AT	0.266	0.129	0.223	0.408	0.209	0.183	0.107	0.167	0.283	0.408
Class 3 AT	-	0.152	0.185	-	-	-	-	-	-	-
Overall AT	0.371	0.288	0.266	0.366	0.187	0.202	0.114	0.164	0.283	0.448
ATSR	**0.609**	**0.698**	**0.727**	**0.756**	**0.764**	**0.842**	**0.885**	**0.937**	**0.957**	**0.957**
Acc. for Similar Distribution	96.7%	99.0%	98.9%	65.9%	**83.1%**	**77.3%**	**56.3%**	**69.9%**	**88.5%**	**84.8%**
Acc. for Dissimilar Distribution	**97.4%**	**99.6%**	**99.6%**	**67.0%**	82.8%	77.1%	55.7%	69.6%	87.0%	83.7%

Accuracy (Acc.) was used to evaluate how datasets with varying degree of data class distribution affects the performance of the algorithm. The dataset 'ks_yr50611', which uses the CHS dataset, predicts the occurrence of MI (from year 6 to 11) based on a balanced case-control sample obtained in year 5.
AT means affinity threshold, CV denotes cross-validation and LOO refers to leave-one-out cross-validation.

Table 2: Classification Performance of the Benchmarking Datasets with Different Issues Addressed

Experiment	Description	Iris	Ionosphere	Pima Indians Diabetes	Sonar
1	AIRS2	96.0%	95.6%	74.2%	84.9%
2	GA-AIRS2	98.7%	97.4%	**77.3%**	86.5%
3	Density	98.7%	96.7%	**77.3%**	88.5%
4	Density & Distribution	**99.6%**	97.4%	**77.3%**	88.5%
5	Density, Distribution and Characteristics (EDC-AIRS)	**99.6%**	**98.0%**	**77.3%**	**90.9%**

Using GA-AIRS2 as the base algorithm, the techniques described in experiments 3, 4 and 5 are implemented respectively.

A succinct description of these datasets is as follow:

1. Wine Dataset – Contains results obtained from the chemical analysis of 3 different cultivars grown in the same region in Italy. It is a tri-nary classification problem that consists of 178 instances and 13 features.
2. Magic Dataset – This dataset, obtained from the Major Atmospheric Gamma Imaging Cherenkov (MAGIC) Telescope project, is a Monte Carlo generated data that aims to simulate the registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope. It is a binary class classification problem which contains 19020 instances and 10 features.
3. Hill-Valley Dataset – This dataset consists of 606 instances, 100 features and 2 classes. Each instance represents 100 data points. When plotted (in the given order) on a 2-dimensional graph, the resultant plot would represent either a hill (a "bump" in the terrain) or a valley (a "dip" in the terrain).
4. Bupa Liver Disorder Dataset – This dataset contains examination results (e.g. quantity of alcoholic beverages consumed per day and blood tests) of males which are used to investigate liver disorders. It has a total of 345 instances and 6 features.
5. Statlog Heart Dataset – Investigation of the presence or absence of heart disease in an individual is carried out based on various medical diagnoses. This result is dictated in this dataset, which contains 270 instances and 13 features.
6. Cardiovascular Health Study (CHS) Dataset – This dataset, as described in [13], is an epidemiology study of risk factors for cardiovascular diseases in elderly aged 65 and above.

Table 3: Performance Comparison of Different Classification Algorithms – Modified from [3]

	Iris		Ionosphere		Pima Indians Diabetes		Sonar	
Rank	**Algorithm**	**Acc**	**Algorithm**	**Acc**	**Algorithm**	**Acc**	**Algorithm**	**Acc**
1	Grobian (rough)	100%	3-NN + Simplex	98.7%	Logdisc	77.7%	TAP MFT Bayesian	92.3%
2	**EDC-AIRS**	**99.6%**	**EDC-AIRS**	**98.0%**	IncNet	77.6%	**EDC-AIRS**	**90.9%**
					DIPOL92	77.6%		
3	SSV	98.0%	3-NN	96.7%	**EDC-AIRS**	**77.3%**	Nave MFT Bayesian	90.4%
	C-MLP2LN	98.0%	IB3	96.7%	Linear Disc. Analysis	77.5 – 77.2%	SVM	90.4%
	PVM 2 rules	98.0%					Best 2-layer MLP + BP, 12 hidden	90.4%
4	PVM 1 rule	97.3%	MLP + BP	96.0%	SMART	76.8%	AIRS2	84.9%
					GTO DT (5xCV)	76.8%		
5	AIRS	96.7%	AIRS2	95.6%	ASI	76.6%	MLP+BP, 12 hidden	84.7%
	FuNe-I	96.7%						
	NEFCLASS	96.7%						
6	AIRS2	96.0%	AIRS	94.9%	Fischer Disc. Analysis	76.5%	MLP+BP, 24 hidden	84.5%
	CART	96.0%	C4.5	94.9%				
7	FUNN	95.7%	RIAC	94.6%	MLP+BP	76.4%	1-NN, Manhanttan	84.2%
8			SVM	93.2%	LVQ	75.8%	AIRS	84.0%
					LFC	75.8%		
9			FSM + rotation	92.8%	RBF	75.7%	FSM	83.6%
10			1-NN	92.1%	kNN, k=22, Manh	75.5%		
					MML	75.5%		
					NB	75.5 – 73.8%		
…					… …			
n					AIRS2	74.2%		
n+1					AIRS	74.1%		

'Acc' denotes the classification accuracy.

The cohort consists of elderly subjects from four U.S. communities, namely Forsyth County, North Carolina; Sacramento County, California; Washington County, Maryland; and Pittsburgh, Pennsylvania. Data collected in year 5 of the CHS study was utilized. The balanced case-control sample size consists of 270 instances and 253 features. It is a binary class classification problem (i.e. with or without myocardial infarction).

4. EXPERIMENTAL RESULTS

EDC-AIRS algorithm was developed by extending AIRS2 algorithm. Three areas of optimization were carried out, each addressing an aspect of the phenomenon observed in the natural immune system (i.e. the concentration, distribution and characteristics of the antigens). In order to better generate a set of representative memory cells, it is necessary to empirically determine the ATSR threshold first. To perform this investigation, 10 datasets with different degree of data class distribution were evaluated. The ATSR value of these datasets ranges from 0.609 to 0.957, where a lower value indicates that the distribution of the data classes differs by a larger degree. Both classification with a common set of parameter (assuming similar distribution among data classes) and a distinct set of parameters for each data class (assuming dissimilar distribution among data classes) were performed. Based on the results shown in Table 1, it is indicative that with an ATSR value of 0.756 and below, a distinct parameter set for each data class is capable of achieving a higher accuracy. Therefore, an ATSR threshold of 0.76 was used for the rest of the experiments.

The performance of the proposed EDC-AIRS algorithm was evaluated using 4 benchmarking datasets. The algorithm was evaluated 3 times with consistent classification result obtained each time (i.e. standard deviation of 0). The classification accuracy for the incremental implementation of the 3 aforementioned mechanisms is given in Table 2. Baseline comparison was made with GA-AIRS2 algorithm - an AIRS2 algorithm with its parameters tuned via GA. It is noteworthy that GA-AIRS2 performs better than AIRS2 for all 4 benchmarking datasets.

With the implementation to address the density issue (Table 2 – experiment 3), ameliorated performance was observed for the Sonar dataset. However, the performance on the Ionosphere dataset exacerbates while the performance for the rest of the datasets remains comparable. With the additional implementation to amortize the impact of different distribution exhibited by each data class (Table 2 – experiment 4), the deterioration in performance observed previously on the Ionosphere dataset vanished. Moreover, the accuracy obtained for the Iris dataset improved while the accuracy for both Pima Indians Diabetes and Sonar datasets remain the same. Finally, when the characteristic of the dataset was delved into (Table 2 – experiment 5), further improvement in accuracy for Ionosphere and Sonar datasets was obtained. Accuracy for Iris and Pima Indians Diabetes datasets remains unchanged, probably due to the limited features available for selection (i.e. 4 and 8 features respectively).

A comparison of EDC-AIRS algorithm with other well-known classifiers (as presented in [3]) is provided in Table 3. The EDC-AIRS algorithm has shown promising results, clinching a place in the top 3 positions for all the datasets evaluated.

5. Discussion

We have developed an immune-inspired supervised classification algorithm called EDC-AIRS that have shown improved learning and classification capability. The success of the algorithm is primarily due to the recognition of the importance of additional immune metaphors, namely the ability to adapt to the different concentration, distribution and characteristics of the antigens. However, the EDC-AIRS algorithm did not achieve ameliorated performance for all classification problems investigated in this study (e.g. Pima Indian Diabetes dataset). This is not surprising as every learning algorithm has an inductive bias that would work reasonably well for some, but not all, datasets or application domains [1]. This phenomenon has been described as the selective superiority problem [14].

The AIRS2 parameters reported in [3] has been tuned manually. This apparently hinders the true potential of the AIRS2 algorithm. As demonstrated, the employment of GA to optimize the AIRS2 parameters (i.e. GA-AIRS2) improved the classification accuracy (ranging from 1.6% to 3.1% improvement) for all the 4 benchmarking datasets investigated. Clearly, this indicates that optimization of parameters with an evolutionary computing algorithm (e.g. GA) that is capable of dynamically searching through the defined search space is invaluable in discovering the optimal parameter setting. This is especially so when dealing with datasets from various application domains where the patterns that underlie these data would be very different, causing exhaustive manual tuning of the parameters to flounder as it would be very time consuming to carry out this task.

The EDC-AIRS algorithm, when juxtaposed with the AIRS2 algorithm, has several distinctive strengths when learning the underlying patterns within the data. Firstly, by adopting a mechanism to handle the different data density exhibited at different regions, it is capable of producing representative memory cells that could better characterize and capture the real data pattern. As a result, it is at an advantage when applied on datasets (such as Sonar dataset) that have data density which tends to fluctuate at different regions. Secondly, the EDC-AIRS algorithm is more capable at dealing with difference in distribution among data classes, generating representative memory cells for each data class. The ability to do so is important because it is unlikely for different data classes to have the same distribution and even more unlikely for a classifier to recognize and robustly adapt to such deviation without explicitly allowing for it. Efforts were therefore taken in this work to calculate the ATSR value and to determine whether to optimize a common or distinct parameter set. Ten datasets from diverse domains with different characteristics were used to evaluate the importance of implementing this technique. Results shown that for datasets with ATSR value lesser than 0.76 (e.g. Iris and Ionosphere datasets), it is more desirable to have a distinct parameter set for each data class. The need to compute the ATSR value and differentiate them into similar or dissimilar distribution is not an essential step but is advantageous to do so. This is because it is theoretically possible for GA to tune the parameter set meant for dissimilar distribution to one suitable for similar distribution. However, it is computationally intensive to do so. Therefore, by performing this simple step of differentiation, it can help to alleviate the complexity involved when tuning the parameters with GA. This complexity is introduced by the (linear) increase in the number of parameters that needs to be tuned, which in turn contributed to an exponential increase in the search space. This makes the task of discovering the optimal value for the parameters very challenging. This problem is commonly referred to as the 'curse of dimensionality' [15].

Finally, the EDC-AIRS algorithm is capable of selecting features that are highly informative and relevant. This avoids some of the difficulties when dealing with datasets (e.g. Ionosphere and Sonar datasets) that have irrelevant or redundant features which often jeopardize the algorithm's ability to learn and generalize. Moreover, it has the crucial advantage of identifying important features that best associate with an outcome, building a parsimonious classification model as a result. This property is highly desirable in accordance to the law of parsimony (Occam's razor principle [16]) where a simpler model with minimal complexity is preferred.

When EDC-AIRS algorithm was benchmarked with 4 datasets, promising results were obtained consistently. It outperforms AIRS2 algorithm in all the 4 cases. The increase in classification accuracy is 3.6%, 2.4%, 3.1% and 6% for Iris, Ionosphere, Pima Indians Diabetes and Sonar dataset respectively. This suggests that EDC-AIRS algorithm is a robust learner that is capable of adapting to different profound data patterns and structures.

6. CONCLUSION

Further inspired by the characteristics of the natural immune system, we have developed an adaptive and robust supervised classification algorithm called the EDC-AIRS algorithm. The performance of the proposed algorithm was evaluated with 4 benchmarking datasets. When ranked with other classifiers, the classification performance of EDC-AIRS algorithm is in the top 3 positions for all the datasets evaluated. Ameliorated performance achieved by the algorithm signifies the importance of empowering an algorithm with the ability to independently adapt to the distinct density, distribution and characteristics of each data class. However, this approach does not guarantee improved performance for all classification problems in face of the selective superiority problem.

7. ACKNOWLEDGMENTS

Darwin Tay would like to express his sincere gratitude for his scholarship funding provided by the Nanyang Technological University-Imperial College London Joint PhD programme.

The authors wish to acknowledge the support of The Engineering and Physical Science Research Council (EPSRC) in this study.

Dr. Chueh Loo Poh and Darwin Tay would also like to thank the Ministry of Education (Singapore) for the support in this work.

8. REFERENCES

[1] A. Freitas, and J. Timmis, "Revisiting the Foundations of Artificial Immune Systems for Data Mining", *IEEE Transactions on Evolutionary Computation,* vol. 11, no. 4, pp. 521-540, 2007.

[2] L. N. D. Castro, and J. Timmis, *Artificial Immune Systems: A New Computational Intelligence Approach.* Springer, 2002, pp. 398.

[3] A. Watkins, J. Timmis, and L. Boggess, "Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm", *Genetic Programming and Evolvable Machines,* vol. 5, no. 3, pp. 291-317, 2004.

[4] J. Timmisa1, and M. Nealb, "A Resource Limited Artificial Immune System for Data Analysis", *Knowledge-Based Systems,* vol. 14, pp. 121-130, 2001.

[5] A. Watkins and L. Boggess, "A New Classifier Based on Resource Limited Artificial Immune Systems"*, Proceedings of Congress on Evolutionary Computation, 2002,* Honolulu, USA: 2002.

[6] J. Brownlee, "Artificial Immune Recognition System (AIRS). A Review and Analysis", Technical Report No. 1-02, Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology (SUT), Victoria, Australia: 2005

[7] M. Moses and S. Banerjee, "Biologically Inspired Design Principles for Scalable, Robust, Adaptive, Decentralized Search and Automated Response (RADAR)"*, IEEE Symposium on Artificial Life (ALIFE),* 2011.

[8] J. Holland, "Genetic Algorithms", *Sci Am,* pp. 66-72, 1992.

[9] A.J. Chipperfield and P.J. Fleming, "The MATLAB Genetic Algorithm Toolbox"*, IEE Colloquium on Applied Control Techniques Using MATLAB,* 1995.

[10] C. L. Huanga, and C. J. Wangb, "A GA-based Feature Selection and Parameters Optimization for Support Vector Machines", *Expert Systems with Applications,* vol. 31, no. 2, pp. 231-240, 2006.

[11] A. Watkins, "AIRS: A Resource Limited Artificial Immune Classifier," Master's Thesis, Mississippi State University, United States, 2001.

[12] C.L. Blake and C.J. Merz, "UCI Repository of Machine Learning Databases," Available at: http://www.ics.uci.edu/~mlearn/MLRepository.html. Accessed 8/11/2012.

[13] L. Fried, N. Borhani, P. Enright, and C. Furberg *et al.*, "The Cardiovascular Health Study: Design and Rationale", *Ann Epidemiol.,* vol. 1, no. 3, pp. 263-276, 1991.

[14] C. Brodley, "Addressing the Selective Superiority Problem: Automatic Algorithm/Model Class Selection"*, In Proc. 10th Machine Learning Conf.,* 1993.

[15] R. Bellman, *Adaptive Control Processes.* Princeton New Jersey: Princeton University Press, 1961, pp. 274.

[16] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Occam's Razor", *Information Processing Letters,* vol. 24, no. 6, pp. 377-380, 1987.

Appendix E

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2330898, IEEE Journal of Biomedical and Health Informatics

The Effect of Sample Age and Prediction Resolution on Myocardial Infarction Risk Prediction

Darwin Tay[1,2], Chueh Loo Poh[2,*], Eric Van Reeth[2], Richard I. Kitney[1]

[1] Department of Bioengineering, Imperial College London, UK

[2] Division of Bioengineering, Nanyang Technological University, Singapore

***Abstract* — Myocardial infarction (MI) is one of the leading causes of death in many developed countries. Hence, early detection of MI events is critical for effective preventative therapies, potentially reducing avoidable mortality. One approach for early disease prediction is the use of risk prediction models developed using machine learning techniques. One important component of these models is to provide clinicians with the flexibility to customize (e.g. the prediction range) and use the risk prediction model that they deemed most beneficial for their patients. Therefore, in this paper, we develop MI prediction models and investigate the effect of sample age and prediction resolution on the performance of MI risk prediction models. The cardiovascular health study (CHS) dataset was used in this study. Results indicate that the prediction model developed using SVM algorithm is capable of achieving high sensitivity, specificity and balanced accuracy of 95.3%, 84.8% and 90.1% respectively over a time span of 6 years. Both sample age and prediction resolution were found not to have a significant impact on the performance of MI risk prediction models developed using subjects aged 65 and above. This implies that risk prediction models developed using different sample age and prediction resolution is a feasible approach. These models can be integrated into a computer aided screening tool which clinicians can use to interpret and predict the MI risk status of the individual patients after performing the necessary clinical assessments (e.g. cognitive function, physical function, electrocardiography, general changes to health/lifestyle, and medications) required by the models. This could offer a means for clinicians to screen the patients at risk of having MI in the near future and prescribe early medical intervention to reduce the risk.**

***Index Terms* — Classification, clinical decision support system, clinical risk prediction, medical screening, myocardial infarction.**

* Corresponding author

Manuscript received Apr 09, 2014.

Darwin Tay is with the Department of Bioengineering, Imperial College London, U.K., and also with the Division of Bioengineering, Nanyang Technological University, Singapore (e-mail: darwintay@imperial.ac.uk).

Chueh Loo Poh is with the Division of Bioengineering, Nanyang Technological University, Singapore. (e-mail: clpoh@ntu.edu.sg).

Eric Van Reeth is with the Division of Bioengineering, Nanyang Technological University, Singapore. (e-mail: eric.vanreeth@ntu.edu.sg).

Richard I. Kitney is with the Department of Bioengineering, Imperial College London, U.K. (e-mail: r.kitney@imperial.ac.uk).

I. Introduction

THE best practice to avoid human mortality caused by life threatening diseases like myocardial infarction (MI) is to detect them early and prevent their onset. One approach is to devise computational methods that capitalize on clinical biomarkers to better screen the patients for their potential risk of experiencing (future) MI. Broadly, clinical screening/risk prediction tools are very important as it could potentially lead to the following benefits at the individual patient-level: for example, (1) when patients become knowledgeable of their health risk and with good physician-patient therapeutic relationship, they would be more willing to make changes to their lifestyle and adhere to treatment regimens [1], (2) allows clinicians to promptly recommend effective therapeutic or preventive measures (e.g. lifestyle changes, treatment of subclinical manifestation, etc.) to their patients [2], and (3) if such screening tools were to be integrated into electronic health record system and executed automatically to analyze individuals' health risk, the number of unscreened patients who are at risk of a disease could be reduced dramatically [3]. The key ramification of wide adoption of clinical screening tools is the possibility of significantly reducing the number of avoidable mortality. However, the development of versatile, reliable and accurate computer aided MI screening tools which the clinicians can use in the clinics/hospitals to instantly predict patients' risk remains a challenge.

The conventional approaches for assessing the risk of individuals experiencing MI include risk scoring system and survival curves [4-6]. These, however, have limitations like the inability to substantially identify minority of individuals with subsequent risk of experiencing MI [7]. Moreover, clinical biomarkers and symptoms seldom follow a linear relationship and the expected outcome at the individual patient-level does not always abide by the rules of epidemiology [8]. As a result, conventional risk scoring systems – which model relationships in a linear manner - often flounder in view of these challenges [9, 10].

In recent years, there is an exponential increase in the amount of clinical and molecular data collected from routine medical examination. To overcome the challenges associated with human scale of thinking and analysis, data mining techniques – which have been postulated as a "central feature" for future healthcare system [11] – became a popular method for extracting insights from this data deluge. Advantages of using data mining techniques include the capability of dealing with plethora of information, solving non-trivial problems, producing data-driven prediction models, and handling non-linear relationships among biomarkers. Examples of data mining techniques used to estimate disease risk include work from: (1) Wiens et al. [12] who employed support vector machine (SVM) to identify patients who are at high risk of experiencing hospital acquired Clostridium difficile (C. diff); and (2) Khan et al. [13] who used artificial neural network (ANN) for discriminating small, round blue-cell tumors (SRBCTs).

One important component of risk prediction tools is to provide clinicians with the flexible to customize (e.g. change the range and how far into the future the prediction would be) and use a risk prediction model that they deemed most beneficial for their patients. To this end, we explore the possibility of customizing MI risk prediction models to better meet the patients' needs and clinicians' expectation. Particularly, the effect of sample age and prediction resolution – 2 aspects that are not commonly examined in the literature – on the performance of MI risk prediction models constructed using Support Vector Machine (SVM) [14-16] and Evolutionary Data-Conscious Artificial Immune Recognition System (EDC-AIRS) [17] algorithms were investigated. Here, sample age refers to the average age of individuals found in the baseline (i.e. input) dataset used to construct the clinical risk prediction model while prediction resolution refers to the prediction scale (i.e. number of years into the future where prediction of MI occurrence begins) and interval (i.e. time duration, in years, that marks the start and end of MI outcomes to be considered) employed by the clinical risk prediction model.

In view of the rapid aging population worldwide and the relatively high prevalence of MI among the elderly, participants amassed from the Cardiovascular Health Study (CHS) [18] – consisting of subjects aged 65 and above - were analyzed. Further, with the wide range of clinical measurements and risk factors accrued during the CHS observational study, it makes the CHS dataset a valuable source of information for this work.

The rest of the paper is organized as follows. Section II provides details of CHS dataset, and delineates the methodology involved in developing the predictive models. Section III provides the experimental results achieved by the risk prediction models developed using different combinations of sample age and prediction resolution. Key results are discussed in Section IV and conclusions are drawn in Section V.

II. Materials and Methods

In Section IIA, details of CHS dataset are provided. This dataset, however, consists of a significant percentage of missing data and a highly skewed data distribution (commonly known as the class imbalanced data problem). Hence, for effective analysis, data imputation and class data balancing are performed and described in Section IIB and IIC respectively. Section IID explains how the various MI risk prediction models based on different combinations of baseline data and prediction resolution were developed and validated.

A. Cardiovascular Health Study (CHS) Dataset

The CHS dataset, as described in [18], is an epidemiology study of risk factors for cardiovascular diseases in elderly aged 65 and above. It contains 2 cohorts recruited at different phases. The first cohort consists of 5201 subjects from four U.S. communities, namely Forsyth County, North Carolina; Sacramento County, California; Washington County, Maryland; and Pittsburgh, Pennsylvania. An additional 687 African Americans were subsequently recruited forming the second cohort. Eligible individuals were sampled from Medicare eligibility lists in each area. Eligible participants include all individuals sampled from the Health Care Financing Administration (HCFA) sampling frame - they were 65 years or older at the time of examination, non-institutionalized, expected to remain in the area for the next 3 years, and able to give informed consent and did not require a proxy respondent at baseline. Individuals who were wheelchair-bound at home at baseline or receiving hospice treatment, radiation therapy or chemotherapy for cancer were excluded. Eligible individuals were examined yearly from 1989 to 1999. Extensive physical and laboratory evaluations were carried out to identify the presence and severity of CVD risk factors - such as hypertension; hypercholesterolemia and glucose intolerance; subclinical disease, such as carotid artery atherosclerosis; left ventricular enlargement; and transient ischemia. Criteria for identification of MI events include: observation of evolving Q-wave, cardiac pain and abnormal enzymes together with an evolving ST-T pattern or new left bundle branch block. The reason for choosing the CHS dataset was because of (1) the relatively high prevalence of CHD among the elderly, (2) worldwide demographic aging, (3) paucity of information regarding risk factors for CHD among elderly, and (4) the changing clinical characteristics of CHD with advancing age [18-21].

B. Data Imputation

Data imputation is the process of substituting missing entries in a dataset with plausible values and aims to improve the quality of the data. It was performed using weighted K-nearest neighbor (KNN) because of its excellent performance in estimating missing values [22, 23]. Moreover, it has the capability to estimate both qualitative and quantitative attributes. Hence, it is highly suitable for interpolating the missing values in the CHS dataset.

Individuals with unknown MI status and clinical features that were uninformative (i.e. features with consistent value throughout) were first removed from the analysis. Individuals and clinical features with high percentage of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2330898, IEEE Journal of Biomedical and Health Informatics

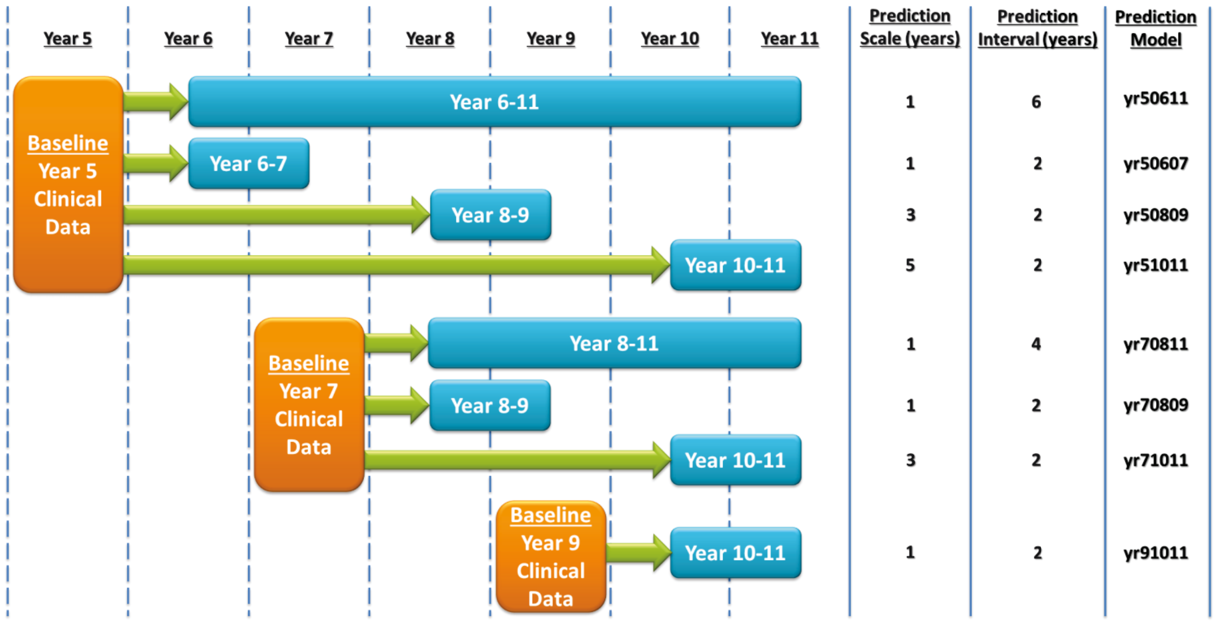


Figure 1: MI Risk Prediction Models of Various Prediction Scales and Intervals
MI risk prediction at various time scales and intervals using the CHS dataset was performed. Prediction scale refers to the number of years into the future where prediction of MI occurrence begins while prediction interval refers to the time duration (in years) that marks the start and end of MI outcomes to be considered.

missing entries were also removed. This is to ensure that there is an adequate supply of complete entries for weighted KNN to reference when estimating the missing values, which in turn promotes a more accurate data imputation process [22-24]. The resulting dataset was next normalized to unit variance to ensure that the attributes with large scale do not dominate the (Euclidean) distance measure [25]. Subsequently, the optimal value of K for each clinical feature was determined by 10-fold cross-validation and used for the data imputation process. The type of replacement method used by weighted KNN depends on the data type. For instance, if categorical (continuous) data were encountered, the weighted-mode (weighted-mean) of the K nearest neighbors was used to assign the value for the missing entries. The use of weighted KNN estimation has been demonstrated in [22, 26] to be robust and accurate.

C. Class Imbalance Data Problem

In order to create an unbiased dataset for SVM and EDC-AIRS algorithms to learn from, under-sampling of the majority class is necessary. The Kennard-Stone (KS) algorithm [27] was employed to perform this task because of its excellent performance - as demonstrated in a comparative study [28]. This algorithm sequentially selects representative data that are uniformly scattered across the data domain space. This is carried out by first selecting the data object that is closest to the mean of the dataset and is included as the first data candidate. Subsequently, the data object that is most distant from the first one (based on Euclidean distance) is included as the second data candidate. The next data object is chosen by identifying the one farthest away from the previously selected data candidates. This process repeats until the desired number of candidates has been identified [28, 29].

In this study, the KS algorithm was used to under-sample the majority class found in the imputed CHS dataset. The number of candidates to select is equivalent to the number of samples in the minority class. In other words, after this process, the number of controls and cases would be identical.

D. MI Risk Prediction Models

Two algorithms (SVM and EDC-AIRS) were employed to develop MI risk prediction models. SVM algorithm is a robust supervised learning algorithm that is capable of yielding excellent generalization performance on an extensive area of problems [30-32]. It is derived from

		EDC-AIRS	
		Misclassification	Correct Classification
SVM	Misclassification	a	b
	Correct Classification	c	d

Figure 2: Contingency Table for McNemar's Test
'a' indicates the number of data items misclassified by both SVM and EDC-AIRS algorithms; 'b' represents the number of data items misclassified by SVM algorithm but correctly classified by EDC-AIRS algorithm; 'c' denotes the number of data items misclassified by EDC-AIRS algorithm but correctly classified by SVM algorithm; 'd' dictates the number of data items correctly classified by both SVM and EDC-AIRS algorithms.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2330898, IEEE Journal of Biomedical and Health Informatics

Table 1: Details of the Imputed CHS Dataset

Prediction Model	Sample Size* (cases/controls)	#Features	Age (Mean±SD)
yr50611	3102 (6.2%/93.8%)	237	75.7 ± 5.34
yr50607	3102 (2.4%/97.6%)	237	75.7 ± 5.34
yr50809	3034 (2.1%/97.9%)	237	75.7 ± 5.34
yr51011	2978 (2.1%/97.9%)	237	75.7 ± 5.36
yr70811	2407 (2.1%/97.9%)	233	77.2 ± 5.40
yr70809	2407 (2.1%/97.9%)	233	77.2 ± 5.40
yr71011	2362 (2.0%/98.0%)	233	77.2 ± 5.40
yr91011	1909 (1.9%/98.1%)	242	78.8 ± 5.09

*This sample size refers to the number of individuals that remain in the CHS dataset after removal of records with significant missing entries.
'yrXYYZZ' denotes that the prediction model uses clinical measurements observed in year X to make prediction of whether one would experience MI from year YY to ZZ.

Table 2: Details of Datasets Used to Build the Prediction Models

Prediction Model	#Training Instances	#Validation Instances	McNemar's Test# (p-value) SVM vs EDC-AIRS
yr50611	270	114	**<0.01**
yr50607	104	42	**<0.01**
yr50809	92	38	**<0.01**
yr51011	88	36	**<0.01**
yr70811	136	58	0.31
yr70809	70	30	**0.04**
yr71011	66	28	**<0.01**
yr91011	52	20	0.07

All training and validation datasets contain equal number of cases and controls.
#The p-value of McNemar's test is presented examining whether the performance of the SVM algorithm is statistically different from EDC-AIRS algorithm.

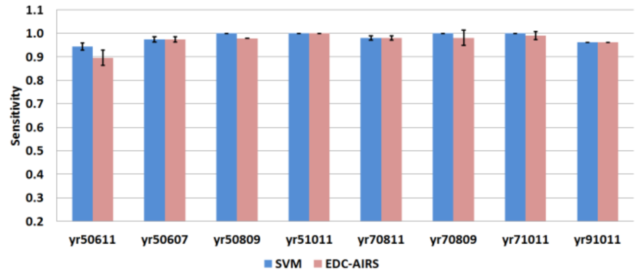
statistical learning theory and is capable of solving linearly and non-linearly separable problems. Fundamentally, SVM performs classification through the construction of an N-dimensional hyper-plane that optimally separates the data into two or more categories whereby the margin of separation between the different categories is maximized.

EDC-AIRS algorithm [17] is a supervised classification algorithm inspired by the principles and processes associated with the human immune system. It performs classification by first constructing a pool of memory cells (i.e. candidate solutions in the form of data vectors) that are representative of the training data through repetitive optimization of the (values of the) memory cells. Optimization was carried out by robustly adapting the memory cells to the different density, distribution and characteristics exhibited by each data class in the training data. Finally, with the utilization of the generated memory cells pool, KNN is used to classify unseen data observations. This algorithm, when tested on several widely benchmarked datasets, has demonstrated highly competitive classification performance [17]. To adopt a ceteris paribus experimental design, the parameters for both algorithms were first tuned using Genetic Algorithm (GA) and subsequently, feature selection was conducted (using GA) to identify predictive biomarkers. The GA parameters were determined experimentally to work well with this clinical prediction problem and kept constant for all experiments. The setup details of GA are as follow: population size: 100; maximum generation: 100; natural selection: stochastic universal sampling; crossover type: discrete recombination; crossover probability: 0.8; mutation rate: 1/P, where P is the number of parameters/features. The parameter details for SVM are: kernel function: radial basis function (RBF); cost: $[2^{-5}, 2^{13}]$; gamma: $[2^{-15}, 2^{3}]$; and for EDC-AIRS are: seed: 1; clonal rate: 10; hyper-mutation rate: 2; stimulation threshold: 0.9; initial memory pool size: [0, 200]; KNN value: [1, 15]; affinity threshold scalar: [0, 1]; total resource: [150, 300]; $Radius_{density}$ = [0, 3]; $Radius_{max}$ = [0, 3].

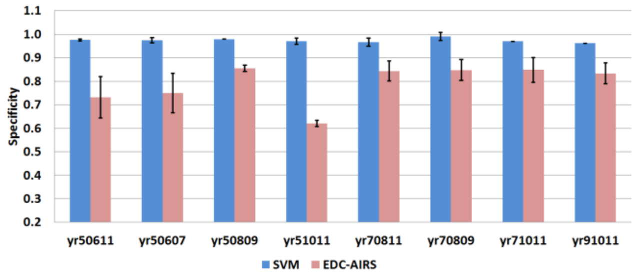
Clinical data - recorded during the 5^{th} to 11^{th} year in which the CHS clinical study was undertaken - were utilized. The reason for using clinical data recorded from year 5 onwards was because clinical examinations taken by the two different cohorts recruited at different phases synchronized from that year onward. The reason for ending the prediction at year 11 is because from year 12 onwards, participants were only monitored annually via phone calls and no clinical examinations were conducted.

To test the hypothesis, prediction models - using different baseline datasets (with different sample age) - capable of predicting the risk of experiencing MI at various prediction scales and intervals were developed. As illustrated in Figure 1, 8 different prediction models were designed to investigate how time factor in relation to the onset of MI would affect the performance of the prediction model. Three different baseline datasets were used. These datasets contain clinical examination results recorded in year 5, year 7 and year 9 of the CHS study. Each of these datasets was used to predict future. Three different prediction scales (1, 3 and 5 years) and 3 different prediction intervals (2, 4 and 6 years) were investigated. Specifically, healthy individuals present in year 5 of the CHS dataset were used as the baseline data to predict whether one would experience MI from year 6 to 11 (prediction scale: 1 year; prediction interval: 6 years), year 6 to 7 (prediction scale: 1 year; prediction interval: 2 years), year 8 to 9 (prediction scale: 3 years; prediction interval: 2 years) and year 10 to 11 (prediction scale: 5 years; prediction interval: 2 years). Similarly, clinical examination results of healthy participants in year 7 was initialized as the baseline data, where prediction of whether one would suffer from MI whether an individual would experience MI in the near from year 8 to 11, year 8 to 9 and year 10 to 11 were conducted. Likewise, clinical data recorded in year 9 was utilized to perform prediction of MI occurrence from year 10 to 11.

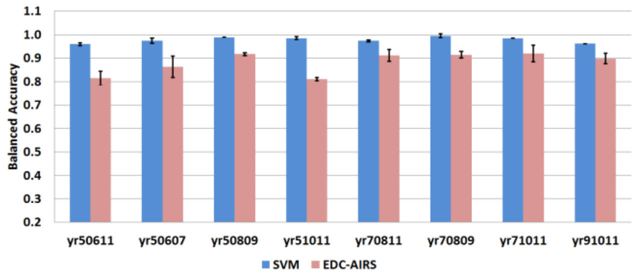
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2330898, IEEE Journal of Biomedical and Health Informatics



(a) Sensitivity Performance Metric



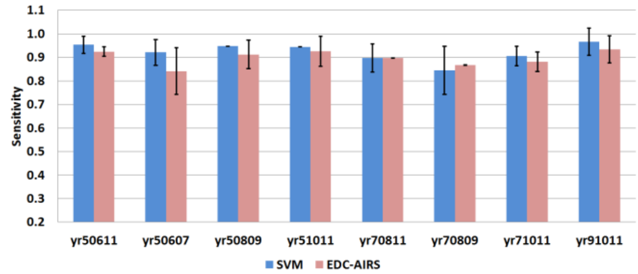
(b) Specificity Performance Metric



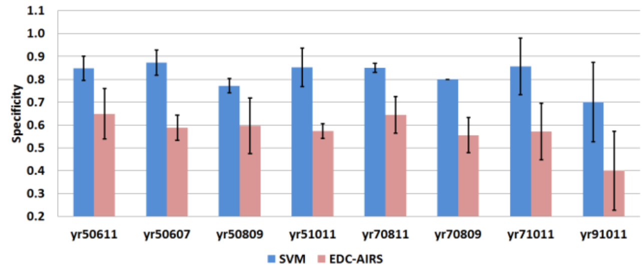
(c) Balanced Accuracy Performance Metric

Figure 3: Classification Performance of SVM and EDC-AIRS Algorithms (Cross-Validated)

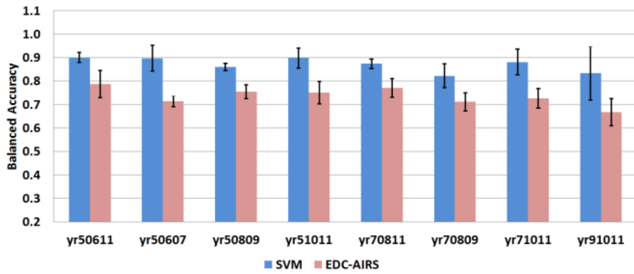
These performance measurements were obtained by performing 10-fold cross validation for each prediction model.



(a) Sensitivity Performance Metric



(b) Specificity Performance Metric



(c) Balanced Accuracy Performance Metric

Figure 4: Classification Performance of SVM and EDC-AIRS Algorithms (Tested with Validation Dataset)

These performance measurements were obtained by evaluating each developed prediction model with their respective validation dataset.

Each baseline dataset was randomly split into two subsets having balanced class distribution. The first subset contains 70% of the initial data. Using this subset, the prediction model was trained and tuned based on 10-fold cross-validation. The second subset, which contains the remaining 30% of the data, was used to validate the developed model. This splitting process was repeated 3 times and independently used to develop and test the respective prediction model. It is highly encouraged to do so to avoid the developed model from capturing not only the true associations, but, also, idiosyncratic features of the training data, which often produces an overly optimistic model [33]. Three commonly used performance measurements were employed to evaluate the prediction models developed - namely sensitivity, specificity, and balanced accuracy (i.e. average between sensitivity and specificity).

Finally, to determine whether the prediction models developed using SVM and EDC-AIRS algorithms are statistically different from each other, McNemar's test was conducted. This statistical test was chosen as it has been demonstrated to have low type 1 error [34]. For each prediction model, this test was carried out by first recording the prediction outcomes obtained (by each algorithm) when tested using each validation dataset. The results obtained from each algorithm were then used to construct the contingency table shown in Figure 2. Referring to the figure, if the sum of 'b' and 'c' is greater than 25, chi-square test with 1 degree of freedom is used for performing McNemar's test. Otherwise, to provide a better estimation of the small sample (i.e. $b + c \leq 25$), binomial distribution is used for (exact) McNemar's test. The prediction model is considered to be statistically different from the ground truth if the p-value computed using McNemar's test is smaller than 0.05.

III. Experimental Results

A. Data Preprocessing

Table 1 provides the details of the resulting CHS datasets after the removal of records and clinical features with significant missing entries.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2330898, IEEE Journal of Biomedical and Health Informatics

Table 2 offers the details of the datasets used to develop and test the MI prediction models after data imputation and class data balancing were performed.

B. MI Risk Prediction Models

Prediction models - using baseline dataset with different sample age - at various time scales and intervals were developed using the training datasets. Cross-validation was carried out to evaluate the performance of each prediction model. For all prediction models developed, results (as shown in Figure 3) indicate consistently high predictive performance was achieved by both SVM and EDC-AIRS algorithms. For example, a balanced accuracy of at least 0.95 and 0.81 was achieved by SVM and EDC-AIRS algorithms respectively.

To assess whether the prediction models developed generalize well, validation was performed using the validation datasets. Results, as presented in Figure 4, demonstrate that a balanced accuracy of at least 0.81 and 0.71 was achieved by SVM and EDC-AIRS algorithms respectively.

McNemar's test was conducted to determine whether the performance of SVM and EDC-AIRS algorithms are statistically different from each other. Results (as shown in Table 2) indicate that for most of the prediction models (except prediction models 'yr70811' and 'yr91011'), the performance of SVM and EDC-AIRS algorithms are statistically different.

IV. Discussion

MI risk prediction models developed using baseline datasets with different sample age, and based on different prediction resolution combinations were analyzed. Cross-validation was utilized during the training phase as an approach to evaluate and develop potent MI risk prediction models. The resultant prediction models developed by both algorithms achieved a relatively high sensitivity, specificity and balanced accuracy (for SVM algorithm, the respective performance achieved is at least 0.94, 0.96 and 0.95; while for EDC-AIRS algorithm, the respective performance achieved is at least 0.89, 0.62 and 0.81). An investigation of whether the prediction models developed were over-trained was conducted by validating each developed model with an unseen dataset (i.e. not used to develop the prediction model). The aim of this step was to assess the generalizability of the developed models. Results indicate that SVM algorithm (and EDC-AIRS algorithm) – across all prediction models tested - achieved a sensitivity, specificity and balanced accuracy of at least 0.84, 0.70 and 0.82 (and 0.84, 0.40 and 0.67) respectively. Furthermore, it can be observed that in general there is a drop in the validation sensitivity (SVM: 0.060±0.054; EDC-AIRS: 0.073±0.052), specificity (SVM: 0.154±0.058; EDC-AIRS: 0.219±0.124) and balanced accuracy (SVM: 0.107±0.036; EDC-AIRS: 0.146±0.070) among all the prediction models developed. It is noteworthy that the drop in performance is less severe for

Table 3: Statistical Evaluation of Prediction Resolution

Prediction Models Compared	ANOVA Test# (p-value) SVM	ANOVA Test# (p-value) EDC-AIRS
Prediction Scale		
yr50607; yr50809; yr51011	0.47	0.71
yr70809; yr71011	0.25	0.93
Prediction Interval		
yr50611; yr50607	0.92	0.12
yr70811; yr70809	0.88	0.14

#The p-value of ANOVA test is presented examining the significance of prediction scale and interval for both SVM and EDC-AIRS algorithms.

SVM algorithm (when compared to EDC-AIRS algorithm). This shows that SVM algorithm tends to perform better on noisy data even after data imputation was conducted. This observation is supported by the results obtained from the performance of McNemar's test. From this statistical evaluation, it was demonstrated that SVM algorithm outperforms EDC-AIRS algorithm for 6 out of 8 prediction models tested.

Prediction models developed (with SVM algorithm) using baseline dataset from year 5 (and year 7), and tested using their respective validation datasets have shown comparable sensitivity, specificity and balanced accuracy. Analysis of variance (ANOVA) test was conducted on the respective group of prediction models (i.e. developed using either year 5 or 7 as baseline dataset) that has a prediction interval of 2 years. Results demonstrate that they are statistically comparable - with p-value of 0.47 for prediction models using baseline dataset from year 5 (and 0.25 for prediction models using baseline dataset from year 7). This signifies that predication scale does not have a significant impact on the performance of (SVM-based) prediction models developed and tested using subjects aged 65 and above. Similar analysis was performed on prediction models developed based on different prediction interval. Results indicate that these models are statistically comparable – with p-value of 0.92 and 0.88 for prediction models developed using baseline dataset from year 5 and 7 respectively. This means that prediction interval does not have a significant impact on the performance of prediction models developed using SVM algorithm.

As for prediction models developed using EDC-AIRS algorithm, similar analysis was conducted. For prediction models developed using baseline dataset from year 5 (and year 7) that are based on 2-year prediction interval, and tested using their respective validation datasets, ANOVA test was conducted. Results indicate that the prediction models in their respective group are statistically comparable – having a p-value of 0.71 (for prediction model using year 5 baseline dataset) and 0.93 (for prediction model using year 7 baseline dataset). This indicates that predication scale does not have a significant impact on prediction models developed using EDC-AIRS algorithm as well. Likewise, prediction models developed based on different prediction interval were analyzed. Results show that these models are statistically comparable – having a p-value of 0.12 and 0.14

for prediction models developed using baseline dataset from year 5 and 7 respectively. This suggests that prediction interval does not have a significant impact on the performance of prediction models developed using EDC-AIRS algorithm as well. In view of these observations, we aim to investigate the effects of prediction resolution on subjects in younger age groups as part of our future work. A summary of the p-values discussed is provided in Table 3.

Analysis of prediction models that aim to predict the likelihood of MI occurrence in individuals' subsequent 2 years (i.e. 'yr50607', 'yr70809' and 'yr91011') indicate comparable performance – with p-value of 0.50 and 1.00 for SVM and EDC-AIRS algorithms respectively. Comparison of age among individuals belonging to different baseline datasets indicates that they are statistically different (p-value < 0.01). This portends that sample age does not have a significant impact on the performance of prediction models.

Among all the prediction models developed, key biomarkers identified to be statistically significant by both SVM and EDC-AIRS algorithms are related to cognitive function, physical function, depression/life events, electrocardiography, general changes to health/lifestyle, and medications. These biomarkers, in general, are also identified as clinically significant in the literature [35-38]. This suggests that statistically significant biomarkers can also be clinically significant - providing a promising avenue for identifying the potential cardiovascular risk factors to be evaluated in clinical trials.

One benefit of performing risk prediction using different prediction resolution and sample age is that it allows more refined and progressive risk prediction to be conducted (without compromising accuracy). This provides the advantage of estimating the seriousness of a disease one is experiencing; enabling clinicians to offer a more personalized management and/or therapeutic strategy to the patient.

The limitation of this investigation includes the use of a single dataset to evaluate the effects of sample age and prediction resolution in relation to the performance of MI risk prediction. This limits the power to conclusively state how each factor influences the performance of the prediction model. Nevertheless, it does provide some insights on whether sample age and prediction resolution have an impact on the performance of clinical risk prediction model. In view of the observations from this study and the importance of screening since young, we aim to investigate the effect of prediction resolution and sample age on younger subjects as part of our future work.

V. Conclusions

Early detection of individuals with high risk of experiencing MI is very important clinically, but has proved to be elusive. To this end, we investigated the effect of sample age and prediction resolution in relation to the development of accurate clinical risk prediction model. Our experiments indicate that both sample age and prediction resolution do not have a significant impact on prediction models developed using subjects aged 65 and above.

Overall, high validation sensitivity, specificity and balanced accuracy were achieved by SVM algorithm. This opens the opportunity for constructing predictive models capable of detecting MI early, allowing clinicians to take preventative measures promptly, improving the quality of individuals' life, and reducing avoidable mortality.

In view of these results, we suggest the use of different prediction resolution to provide a more detailed health screening of elderly subjects so that more appropriate preventative measurements - in relation to the individual's risk level - can be taken.

Acknowledgment

Darwin Tay would like to express his sincere gratitude for his scholarship funding provided by the Nanyang Technological University-Imperial College London Joint PhD programme.

The authors wish to acknowledge the support of The Engineering and Physical Science Research Council (EPSRC) in this study and thank the National Heart, Lung and Blood Institute (NHLBI) for providing the CHS dataset.

Dr. Chueh Loo Poh, Eric Van Reeth and Darwin Tay would also like to thank the Ministry of Education (Singapore) for the support in this work.

References

[1] L. R. Martin, S. L. Williams, K. B. Haskard, and M. R. DiMatteo, "The Challenge of Patient Adherence", *Therapeutics and Clinical Risk Management,* vol. 1, no. 3, pp. 189-199, 2005.

[2] M. A. Whooley, "To Screen or Not to Screen?: Depression in Patients With Cardiovascular Disease", *J Am Coll Cardiol,* vol. 54, no. 10, pp. 891-893, 2009.

[3] R. J. Hye, A. E. Smith, G. H. Wong, and S. S. Vansomphone *et al.*, "Leveraging the Electronic Medical Record to Implement an Abdominal Aortic Aneurysm Screening Program", *Journal of Vascular Surgery,* vol. 59, no. 6, pp. 1535-1543, 2014.

[4] T. Clayton, J. Lubsen, S. Pocock, and Z. Vokó *et al.*, "Risk Score for Predicting Death, Myocardial Infarction, and Stroke in Patients with Stable Angina, based on a Large Randomised Trial Cohort of Patients", *BMJ,* vol. 331, pp. 869-873, 2005.

[5] D. M. Lloyd-Jones, P. W. Wilson, M. G. Larson, and A. Beiser *et al.*, "Framingham Risk Score and Prediction of Lifetime Risk for Coronary Heart Disease", *Am J Cardiol,* vol. 94, no. 1, pp. 20-24, 2004.

[6] W. Levy, D. Mozaffarian, D. Linker, and S. Sutradhar *et al.*, "The Seattle Heart Failure Model: Prediction of Survival in Heart Failure", *Circulation,* vol. 113, pp. 1424-1433, 2006.

[7] S. Alty, N. Angarita-Jaimes, S. Millasseau, and P. Chowienczyk, "Predicting Arterial Stiffness from the Digital

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2330898, IEEE Journal of Biomedical and Health Informatics

Volume Pulse Waveform", *IEEE Trans Biomed Eng.,* vol. 54, no. 12, pp. 2268-2275, 2007.
[8] S. Chattopadhyay, "Mining the Risk of Heart Attack: A Comprehensive Study", *International Journal of Biomedical Engineering and Technology,* vol. 11, no. 4, 2013.
[9] X. Song, A. Mitnitski, J. Cox, and K. Rockwood, "Comparison of Machine Learning Techniques with Classical Statistical Models in Predicting Health Outcomes", *Medinfo,* vol. 107, no. Pt 1, pp. 736-740, 2004.
[10] J. Kim, B. Cho, S. Im, M. Jeon, I. Kim and S. Kim, "Comparative Study on Artificial Neural Network with Multiple Regressions for Continuous Estimation of Blood Pressure", *Proceedings of the IEEE Engineering in Medicine and Biology 27th Annual Conference,* 2005.
[11] R. Snyderman, and J. Langheier, "Prospective Health Care: The Second Transformation of Medicine", *Genome Biol,* vol. 7, no. 104, 2006.
[12] J. Wiens, J. Guttag and E. J. Horvitz, "Patient Risk Stratification for Hospital-Associated C. Diff as a Time-Series Classification Task", *Neural Information Processing Systems (NIPS),* 2012.
[13] J. Khan, J. S. Wei, M. Ringnér, and L. H. Saal *et al.*, "Classification and Diagnostic Prediction of Cancers using Gene Expression Profiling and Artificial Neural Networks", *Nat Med,* vol. 7, no. 6, pp. 673-679, 2001.
[14] B. E. Boser, I. M. Guyon and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers", *Proceedings of the fifth annual workshop on Computational learning theory,* 1992.
[15] C. Cortes, and V. Vapnik, "Support-Vector Networks", *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.
[16] V. Vapnik, "An Overview of Statistical Learning Theory", *IEEE Transactions on Neural Networks,* vol. 10, no. 5, pp. 988-999, 1999.
[17] D. Tay, C. Poh and R. Kitney, "An Evolutionary Data-Conscious Artificial Immune Recognition System", *Genetic and Evolutionary Computation Conference (GECCO),* 2013.
[18] L. Fried, N. Borhani, P. Enright, and C. Furberg *et al.*, "The Cardiovascular Health Study: Design and Rationale", *Ann Epidemiol.,* vol. 1, no. 3, pp. 263-276, 1991.
[19] J. Wiener, and J. Tilly, "Population Ageing in the United States of America: Implications for Public Programmes", *Int J Epidemiol.,* vol. 31, no. 4, pp. 776-781, 2002.
[20] A. S. Go, D. Mozaffarian, V. L. Roger, and E. J. Benjamin *et al.*, "Heart Disease and Stroke Statistics--2013 Update: A Report from the American Heart Association", *Circulation,* vol. 127, pp. 6-245, 2013.
[21] R. Abbott, J. Curb, B. Rodriguez, and K. Masaki *et al.*, "Age-related Changes in Risk Factor Effects on the Incidence of Coronary Heart Disease", *Ann Epidemiol,* vol. 12, no. 3, pp. 173-181, 2002.
[22] O. Troyanskaya, M. Cantor, G. Sherlock, and P. Brown *et al.*, "Missing Value Estimation Methods for DNA Microarrays", *Bioinformatics.,* vol. 17, no. 6, pp. 520-525, 2001.
[23] J. Jerez, I. Molina, P. García-Laencina, and E. Alba *et al.*, "Missing Data Imputation using Statistical and Machine Learning Methods in a Real Breast Cancer Problem", *Artif Intell Med.,* vol. 50, no. 2, pp. 105-115, 2010.
[24] P. Garcia-Laencina, A. Vidal and J. L. Sancho-Gomez, "A Robust Approach for Classifying Unknown Data in Medical Diagnosis Problems", *IEEE World Automation Congress (WAC),* 2008.
[25] B. Minaei-Bidgoli, D. Kashy, G. Kortmeyer and W. Punch, "Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-based System", *33rd ASEE/IIEEE Frontiers in Education Conference,* 2003.
[26] S. Dudani, "The Distance-Weighted k-Nearest-Neighbor Rule", *Systems, Man and Cybernetics, IEEE Transactions on,* vol. 6, no. 4, pp. 325-327, 1976.
[27] R. Kennard, and L. Stone, "Computer Aided Design of Experiments", *Technometrics,* vol. 11, no. 1, pp. 137-148, 1969.
[28] W. Wu, B. Walczak, D. Massart, and S. Heuerding *et al.*, "Artificial Neural Networks in Classification of NIR Spectral Data: Design of the Training Set", *Chemometrics and Intelligent Laboratory Systems,* vol. 33, no. 1, pp. 35-46, 1996.
[29] M. Shahlaeiab, A. Madadkar-Sobhanic, L. Saghaiebd, and A. Fassihibd, "Application of an Expert System based on Genetic Algorithm – Adaptive Neuro-Fuzzy Inference System (GA–ANFIS) in QSAR of Cathepsin K Inhibitors", *Expert Systems with Applications,* vol. 39, no. 6, pp. 6182-6191, 2012.
[30] W. H. Chen, S. H. Hsu, and H. P. Shen, "Application of SVM and ANN for Intrusion Detection", *Computers & Operations Research,* vol. 32, no. 10, pp. 2617-2634, 2005.
[31] E. Osuna, R. Freund and F. Girosit, "Training Support Vector Machines: An Application to Face Detection", *Proceedings of Computer Vision and Pattern Recognition,* 1997.
[32] J. Listgarten, S. Damaraju, B. Poulin, and L. Cook *et al.*, "Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms", *Clinical Cancer Research,* vol. 10, pp. 2725-2737, 2004.
[33] J. Taylor, D. Ankerst, and R. Andridge, "Validation of Biomarker-Based Risk Prediction Models", *Clinical Cancer Research,* vol. 14, no. 19, pp. 5977-5983, 2008.
[34] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms", *Neural Comput,* vol. 10, no. 7, pp. 1895-1924, 1998.
[35] M. Breteler, J. Claus, D. Grobbee, and A. Hofman, "Cardiovascular Disease and Distribution of Cognitive Function in Elderly People: The Rotterdam Study", *BMJ,* vol. 308, no. 6944, pp. 1604-1608, 1994.
[36] G. Erikssen, K. Liestøl, J. Bjørnholt, and E. Thaulow *et al.*, "Changes in Physical Fitness and Changes in Mortality", *The Lancet,* vol. 352, no. 9130, pp. 759-762, 1998.
[37] D. L. Musselman, D. L. Evans, and C. B. Nemeroff, "The Relationship of Depression to Cardiovascular Disease:

Epidemiology, Biology, and Treatment", *JAMA,* vol. 55, pp. 580-592, 1998.
[38] Kannel William B., T. Gordon, Castelli William P., and Margolis James R., "Electrocardiographic Left Ventricular Hypertrophy and Risk of Coronary Heart Disease: The Framingham Study", *Ann Intern Med,* vol. 72, no. 6, pp. 813-822, 1970.

Darwin Tay received his BEng (Hons) degree in Computer Science from Nanyang Technological University, Singapore. He is currently pursuing his doctoral degree in the department of Bioengineering, under the Imperial College London - Nanyang Technological University Joint Ph.D. program. His research interests include medical computing, nature-inspired algorithms and machine learning among others.

Chueh Loo Poh earned his B.Eng. in Electrical and Electronic Engineering from Nanyang Technological University (NTU), Singapore and Ph.D. in Bioengineering from Imperial College London. His research interests include image processing techniques, bio-inspired machine learning methods, security methods related to biomedical imaging in a web-based environment, and synthetic biology. He is currently an assistant professor in the School of Chemical and Biomedical Engineering at Nanyang Technological University.

Eric Van Reeth was born in 1985. He received an engineer degree in Electrical and Electronic Engineering with a major in image processing in 2007, and a Ph.D. degree in collaboration with STMicroelectronics in 2011 from Grenoble INP, France. In 2011, he joined Nanyang Technological University (NTU), Singapore, as a postdoctoral fellow. His current research focuses on resolution enhancement of MRI data.

Professor Richard Kitney was born in the UK, in 1948. He received his PhD in Biomedical Engineering from Imperial College and holds the Chair of Biomedical Systems Engineering at Imperial College. Kitney was Founding Head of the Department of Bioengineering; is Chairman of the Institute of Systems and Synthetic Biology and Co-director of the new EPSRC Centre for Synthetic Biology and Innovation. His research interests over the last 25 years have focused on modelling biological systems, biomedical information systems and, more recently, synthetic biology. He is a Fellow of The Royal Academy of Engineering; an Academician of the International Academy of Biomedical Engineering; a Fellow of the American Academy of Biomedical Engineering and an Honorary Fellow of both The Royal College of Physicians and The Royal College of Surgeons (UK).

ELSEVIER LICENSE
TERMS AND CONDITIONS

Feb 13, 2014

This is a License Agreement between Darwin Tay ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Darwin Tay
Customer address	- -
License number	3327120065324
License date	Feb 13, 2014
Licensed content publisher	Elsevier
Licensed content publication	Journal of Biomedical Informatics
Licensed content title	A biological continuum based approach for efficient clinical classification
Licensed content author	Darwin Tay,Chueh Loo Poh,Carolyn Goh,Richard I. Kitney
Licensed content date	12 September 2013
Licensed content volume number	
Licensed content issue number	
Number of pages	1
Start Page	0
End Page	0
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Title of your thesis/dissertation	Decision Support Continuum Paradigm for Cardiovascular Disease: Towards Personalized Predictive Models
Expected completion date	Feb 2014
Estimated size (number of	217

pages)	
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 GBP
VAT/Local Sales Tax	0.00 GBP / 0.00 GBP
Total	0.00 GBP

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. Byclicking "accept" in connection with completing this licensingtransaction, you agree that the following terms and conditions apply to thistransaction (along with the Billing and Payment terms and conditionsestablished by Copyright Clearance Center, Inc. ("CCC"), at the timethat you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementionedmaterial subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example,figures) has appeared in our publication with credit or acknowledgement toanother source, permission must also be sought from that source. If suchpermission is not obtained then that material may not be included in yourpublication/copies. Suitable acknowledgement to the source must be made, eitheras a footnote or in a reference list at the end of your publication, asfollows:

“Reprinted from Publication title, Vol /edition number, Author(s), Title ofarticle / title of chapter, Pages No., Copyright (Year), with permission fromElsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER].” Also Lancet special credit -“Reprinted from The Lancet, Vol. number, Author(s), Title of article, PagesNo., Copyright (Year), with permission from Elsevier.”

4. Reproduction of this material is confined to the purpose and/or media forwhich permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures andillustrations may be altered/adapted minimally to serve your work. Any otherabbreviations, additions, deletions and/or any other alterations shall be madeonly with prior written authorization of Elsevier Ltd. (Please contact Elsevierat permissions@elsevier.com)

6. If the permission fee for the requested use of our material is waived inthis instance, please be advised that your future requests for Elseviermaterials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specificallygranted in the combination of (i) the license details provided by you andaccepted in the course of this licensing transaction, (ii) these terms andconditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rightslicensed immediately upon issuance of the license at the end of the licensingprocess for the transaction, provided that you have disclosed complete andaccurate details of your proposed use, no license is finally effective unlessand until full payment is received from you (either by publisher or by CCC) asprovided in CCC's Billing and Payment terms and conditions. If fullpayment is not received on a timely basis,

then any license preliminarilygranted shall be deemed automatically revoked and shall be void as if nevergranted. Further, in the event that you breach any of these terms andconditions or any of CCC's Billing and Payment terms and conditions, thelicense is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of thematerials beyond the scope of an unrevoked license, may constitute copyrightinfringement and publisher reserves the right to take any and all action toprotect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respectto the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher andCCC, and their respective officers, directors, employees and agents, from andagainst any and all claims arising out of your use of the licensed materialother than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not besublicensed, assigned, or transferred by you to any other person withoutpublisher's written permission.

12. No Amendment Except in Writing: This license may not be amended exceptin a writing signed by both parties (or, in the case of publisher, by CCC onpublisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any termscontained in any purchase order, acknowledgment, check endorsement or otherwriting prepared by you, which terms are inconsistent with these terms andconditions or CCC's Billing and Payment terms and conditions. These termsand conditions, together with CCC's Billing and Payment terms and conditions(which are incorporated herein), comprise the entire agreement between you andpublisher (and CCC) concerning this licensing transaction. In the eventof any conflict between your obligations established by these terms andconditions and those established by CCC's Billing and Payment terms andconditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny thepermissions described in this License at their sole discretion, for any reasonor no reason, with a full refund payable to you. Notice of such denialwill be made using the contact information provided by you. Failure toreceive such notice will not alter or invalidate the denial. In no eventwill Elsevier or Copyright Clearance Center be responsible or liable for anycosts, expenses or damage incurred by you as a result of a denial of yourpermission request, other than a refund of the amount(s) paid by you toElsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. Translation: This permission is granted fornon-exclusive world English rights only unless yourlicense was granted for translation rights. If you licensed translation rightsyou may only translate this content into the languages you requested. Aprofessional translator must perform all translations and reproduce the contentword for word preserving the integrity of the article. If this license is tore-use 1 or 2 figures then permission is granted for non-exclusive world rightsin all languages.

16. Posting licensed content on any Website: The followingterms and conditions apply as follows: Licensing material from an Elsevierjournal: All content posted to the web site must maintain the copyright informationline on the bottom of each image; A hyper-text must be included to the Homepageof the journal from which you are licensing at

http://www.sciencedirect.com/science/journal/xxxxx orthe Elsevier homepage for books at http://www.elsevier.com; CentralStorage: This license does not include permission for a scanned version of thematerial to be stored in a central repository such as that provided byHeron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be includedto the Elsevier homepage at http://www.elsevier.com . Allcontent posted to the web site must maintain the copyright information line onthe bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clausesare applicable: The web site must be password-protected and made available onlyto bona fide students registered on a relevant course. This permission isgranted for 1 year only. You may obtain a new license for future websiteposting.

For journal authors: the following clauses are applicable inaddition to the above: Permission granted is limited to the author acceptedmanuscript version* of your paper.

*Accepted Author Manuscript (AAM) Definition: An accepted authormanuscript (AAM) is the author's version of the manuscript of an article thathas been accepted for publication and which may include any author-incorporatedchanges suggested through the processes of submission processing, peer review,and editor-author communications. AAMs do not include other publishervalue-added contributions such as copy-editing, formatting, technicalenhancements and (if relevant) pagination.

You are not allowed to download and post the published journal article (whetherPDF or HTML, proof or final version), nor may you scan the printed edition tocreate an electronic version. A hyper-text must be included to the Homepage ofthe journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx. As part of our normal production process, you will receive an e-mail noticewhen your article appears on Elsevier's online service ScienceDirect(www.sciencedirect.com). That e-mail will include the article's Digital ObjectIdentifier (DOI). This number provides the electronic link to the publishedarticle and should be included in the posting of your personal version. We askthat you wait until you receive this e-mail and have the DOI to do any posting.

Posting to a repository: Authorsmay post their AAM immediately to their employer's institutional repository forinternal use only and may make their manuscript publically available after thejournal-specific embargo period has ended.

Please also refer to Elsevier'sArticle Posting Policy for further information.

18. For book authors the following clauses are applicable in addition tothe above: Authors are permitted to place a briefsummary of their work online only.. You are notallowed to download and post the published electronic version of your chapter,nor may you scan the printed edition to create an electronic version. Posting to a repository: Authors arepermitted to post a summary of their chapter only in their institution'srepository.

20. Thesis/Dissertation: If your license is for use in athesis/dissertation your thesis may be submitted to your institution in eitherprint or electronic form. Should your thesis be published commercially, pleasereapply for permission. These requirements include permission for the Libraryand Archives of Canada to supply single copies, on demand, of the completethesis and include permission for UMI to supply single copies, on demand, ofthe complete thesis. Should your thesis be published commercially, pleasereapply for permission.

Elsevier Open Access Terms andConditions

Elsevier publishes Open Access articles in both its Open Access journals andvia its Open Access articles option in subscription journals.

Authors publishing in an Open Access journal or who choose to make theirarticle Open Access in an Elsevier subscription journal select one of thefollowing Creative Commons user licenses, which define how a reader may reusetheir work: Creative Commons Attribution License (CC BY), Creative CommonsAttribution – Non Commercial - ShareAlike (CC BY NCSA) and Creative Commons Attribution – Non Commercial – No Derivatives (CC BYNC ND)

Terms & Conditions applicable toall Elsevier Open Access articles:

Any reuse of the article must not represent the author as endorsing theadaptation of the article nor should the article be modified in such a way asto damage the author's honour or reputation.

The author(s) must be appropriately credited.

If any part of the material to be used (for example, figures) has appearedin our publication with credit or acknowledgement to another source it is theresponsibility of the user to ensure their reuse complies with the terms andconditions determined by the rights holder.

Additional Terms & Conditionsapplicable to each Creative Commons user license:

CC BY: You may distribute andcopy the article, create extracts, abstracts, and other revised versions,adaptations or derivative works of or from an article (such as a translation),to include in a collective work (such as an anthology), to text or data minethe article, including for commercial purposes without permission from Elsevier

CC BY NC SA: For non-commercialpurposes you may distribute and copy the article, create extracts, abstractsand other revised versions, adaptations or derivative works of or from anarticle (such as a translation), to include in a collective work (such as ananthology), to text and data mine the article and license new adaptations or creationsunder identical terms without permission from Elsevier

CC BY NC ND: For non-commercialpurposes you may distribute and copy the article and include it in a collectivework (such as an anthology), provided you do not alter or modify the article,without permission from Elsevier

Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC NDlicense requires permission from Elsevier and will be subject to a fee.

Commercial reuse includes:

- Promotional purposes (advertising or marketing)
- Commercial exploitation (e.g. a product forsale or loan)

· Systematic distribution (for a fee or free ofcharge)

Please refer to Elsevier'sOpen Access Policy for further information.

21. Other Conditions:

v1.7

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK501225836. Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

ASSOCIATION FOR COMPUTING MACHINERY, INC. LICENSE TERMS AND CONDITIONS

Feb 13, 2014

This is a License Agreement between Darwin Tay ("You") and Association for Computing Machinery, Inc. ("Association for Computing Machinery, Inc.") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Association for Computing Machinery, Inc., and the payment terms and conditions.

License Number	3327120800644
License date	Feb 13, 2014
Licensed content publisher	Association for Computing Machinery, Inc.
Licensed content publication	Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference
Licensed content title	An evolutionary data-conscious artificial immune recognition system
Licensed content author	Darwin Tay, et al
Licensed content date	Jul 6, 2013
Type of Use	Thesis/Dissertation
Requestor type	Author of this ACM article
Is reuse in the author's own new work?	Yes
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Decision Support Continuum Paradigm for Cardiovascular Disease: Towards Personalized Predictive Models
Expected completion date	Feb 2014
Estimated size (pages)	217
Billing Type	Credit Card
Credit card info	Visa ending in -
Credit card expiration	-
Total	4.81 GBP
Terms and Conditions	

Rightslink Terms and Conditions for ACM Material

1. The publisher of this copyrighted material is Association for Computing Machinery, Inc. (ACM). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at).

if never granted.

10. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

11. ACM makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.

12. You hereby indemnify and agree to hold harmless ACM and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

13. This license is personal to the requestor and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

14. This license may not be amended except in a writing signed by both parties (or, in the case of ACM, by CCC on its behalf).

15. ACM hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and ACM (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

16. This license transaction shall be governed by and construed in accordance with the laws of New York State. You hereby agree to submit to the jurisdiction of the federal and state courts located in New York for purposes of resolving any disputes that may arise in connection with this licensing transaction.

17. There are additional terms and conditions, established by Copyright Clearance Center, Inc. ("CCC") as the administrator of this licensing service that relate to billing and payment for licenses provided through this service. Those terms and conditions apply to each transaction as if they were restated here. As a user of this service, you agreed to those terms and conditions at the time that you established your account, and you may see them again at any time at http://myaccount.copyright.com

18. Thesis/Dissertation: This type of use requires only the minimum administrative fee. It is not a fee for permission. Further reuse of ACM content, by ProQuest/UMI or other document delivery providers, or in republication requires a separate permission license and fee. Commercial resellers of your dissertation containing this article must acquire a separate license.

Special Terms:

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or

money order referencing your account number and this invoice number RLNK501225851. Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

Requesting permission to reuse content from an IEEE publication

Title: The Effect of Sample Age and Prediction Resolution on Myocardial Infarction Risk Prediction

Author: Tay, D.; Poh, C.; Van Reeth, E.; Kitney, R.

Publication: Biomedical and Health Informatics, IEEE Journal of

Publisher: IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK | CLOSE WINDOW