UNIVERSITY OF
CANBERRA
Australia's Capital University

Decision Support Framework for Cardiovascular Disease Prediction Using Machine Learning

Nitten Singh Rajliwall

A thesis submitted for the requirements of the Degree
of Doctor of Philosophy
Faculty of Science and Technology

Nov 2022

जीवेषु करुणा चापि मैत्री तेषु पवधीयताम्

Be compassionate and friendly to all living beings

Declaration

I hereby declare that the work presented in this thesis has not been submitted for any other degree or professional qualification, and that it is the result of my own independent work.

Nitten Singh Rajliwall

30 / 11 / 2022

Date

Abstract

Clinical decision making is an important and frequent task, which physicians make in their daily clinical practice. Conventionally, physicians adopt a cognitive predictive modelling process (i.e., knowledge and experience learnt from experience, their research, related literature, patient cases, etc.) for anticipating or ascertaining health problems based on clinical risk factors, that deem to be the most salient. However, with the inundation of health data, from EHR system, wearable devices, and other systems for monitoring vital parameters, it has become difficult for physicians to make sense of this massive data, particularly, due to confounding and complex characteristics of chronic diseases, and there is a need for more effective clinical prediction approaches to address these challenges.

Given the paramount importance of predictive models for managing chronic disease, cardiovascular diseases in particular, this thesis proposes a novel computational predictive modelling framework, based on innovative machine learning and data science approaches that can aid in clinical decision support. The focus of the proposed predictive modelling framework is on interpretable machine learning approaches that consist of interpretable models based on shallow machine learning techniques, such as those based on linear regression and decision trees and their variants, and model-agnostic approaches based on neural networks and deep learning methods but enhanced with appropriate feature engineering and post-hoc explainability. These approaches allow disease prediction models to be deployed in complex clinical settings, including under remote, extreme, and low-resource environments, where data could be small, big, or massive and has several inadequacies in terms of data quality, noise, or missing data. The availability of interpretable models, and model-agnostic approaches enhanced with

explainable aspects are important for physicians and medical professionals, as it will increase transparency, trust and confidence in the decision support provided by computer based algorithmic models.

This thesis aims to address the research gap that exists in the current ML/AI based disease detection models, particularly, the lack of robust, objective, explainable, interpretable and trustworthy inference available from the computer based decision support tools, with a majority of the performance metrics reported from computer based tools have been limited to quantitative measures such as accuracy, precision, recall, F-measure, AUC, ROC, without any detailed qualitative metrics, that provide insight into how the computer has arrived at a decision, and ability to explain the decision making logic, eliciting trust from the stakeholders using the system. This could be due to the problem that most of the current ML/AI tools were built using mathematically rigorous constructs, designed around black box approaches, which are hard to interpret and explain, and hence the decisions provided by them appear to be coming from a black box, offering little explanation on decision arrived.

The research proposed in this thesis is aimed at the development of a breakthrough explainable predictive modelling framework, based on innovative ML/AI algorithms for building CVD disease detection models. The proposed computation framework provides an intelligent and interpretable holistic analytics platform with improved prediction accuracy, and improved interpretability and explainability. The proposed innovation and development can help drive the healthcare system to one that is more patient-centred, and trustworthy, with potential to be tailored for several diseases such as cancer, cardiovascular disease, asthma, traumatic brain injury, dementia, and diabetes. The outcomes of this research based on innovative findings can serve as an example – that the availability of better computer-based

decision support tools, with novel computational strategies, which can address a patient's unique clinical/genetic characteristics, can result in better characterization of diseases and at the same time redefine therapeutic strategies. Some of the key contributions from this research include:

- Novel disease detection models based on traditional shallow machine learning algorithms, particularly those based on decision trees and their variants. These algorithms have shown to be inherently interpretable and accurate white box models and can serve as the baseline for comparing with previous models proposed in the literature.
- Innovative disease detection models based on model agnostic algorithms, such as deep learning networks, but augmented with appropriate pre- processing and post-processing stages to provide better interpretability and explainability and eventually make them an efficient white box model.

For an objective comparison of the methods proposed in each of the above stages, several publicly available benchmark clinical datasets, including Cleveland dataset, NHANES dataset and Framingham Heart Study/CHS dataset were used for model building and experimental validation.

Although Cardiovascular disease has been selected as the use case and disease under investigation, since it has led to an alarming increase in the burden of disease, almost at the epidemic levels, and is a major health concern in today's world, the findings from this research can lead to meaningful and significant impact towards improved self-management of chronic non-communicable diseases and make a significant contribution towards better public health management.

Form C

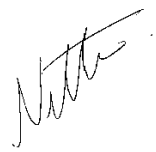
Retention and Use of Thesis by the University

I, Nitten Singh Rajliwall being a candidate for the degree of Doctor of Philosophy accept the requirement of the university relating to the retention and use of the theses deposited in the library. I agree that the original of my thesis deposited in the library should be accessible for purpose of study and research, in accordance with the normal conditions established by the Librarian for the case, loan or reproduction of the theses.

I agree to abide by any general conditions established by the University for the care loan or reproduction of theses and any special conditions of usage in relation to this thesis entitle -

Decision Support Framework for Cardiovascular Disease Prediction Using Machine Learning

By signing this document, I agree that a copy of my thesis will be stored in the university's research repository, that it will be available in full text online, and that I have obtained all the appropriate copyright permission for all copyrighted material used in my thesis.



Signature of the Candidate (Nitten Singh Rajliwall)

30 / 11/ 2022

Date

Publications associated with this research

Following is a list of my peer-reviewed conference papers published, as contributions arising from my thesis. The full papers are included in the links below:

1. Nitten S. Rajliwall, Girija Chetty, Rachel Davey, "Chronic disease risk monitoring based on an innovative predictive modelling framework" in 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, 2017, DOI: 10.1109/SSCI.2017.8285257 https://ieeexplore.ieee.org/document/8285257

2. Nitten S. Rajliwall, Girija Chetty, Rachel Davey , "Machine Learning Based Models for Cardiovascular Risk Prediction" in 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, NSW, Australia DOI: 10.1109/iCMLDE.2018.00034 https://ieeexplore.ieee.org/document/8614017

3. Nitten S. Rajliwall, Girija Chetty, Rachel Davey, "Cardiovascular Risk Prediction Based on XGBoost" in 2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), Nadi, Fiji DOI: 10.1109/APWConCSE.2018.00047 https://ieeexplore.ieee.org/document/8853798/

4. Nitten S. Rajliwall, Girija Chetty, "Deep Learning Based Decision Support Framework for Cardiovascular Disease Prediction" in 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, Australia DOI: 10.1109/CSDE53843.2021.9718459 https://ieeexplore.ieee.org/document/9718459

5. Nitten S. Rajliwall, Girija Chetty, "Cardiovascular Disease Prediction Based on Interpretable and Explainable AI" under review (submitted to Scientific Reports

ACKNOWLEDGEMENTS

I would like to thank everyone who has helped in me to complete this thesis. I am deeply grateful to the supervisor panel, mainly, Professor Girija Chetty, who in a late 2017 read my research proposal and believed enough in me to accept becoming my supervisor. She has guided me and encouraged me to carry on through these years and has contributed to this thesis with a major impact. Big thanks to A/Prof. Dat Tran and A/Prof. Wanli Ma for their support during my research journey and suggestions. Thank you as well for guiding me, often with big doses of patience, through the subtleties of academic writing.

I would like to thank all my PhD colleagues especially Kunal Rajput, with whom I have shared moments of deep anxiety but also of big excitement. Their presence was very important in a process that is often felt as tremendously solitaire.

My distinctive thanks to my wife Sunita Arora; you are complete package of family and love that gave me strength, support, and love during tough times. My sons Aarav and Kanav, you are my love, my world and reason of my smiles. Thank you.

A very special word of thanks goes for my parents, Kamlesh and Rajender Singh, who are source of power and strength in my life. I like to thank my dad for his advice, encouragement, and support through my life.

This thesis is dedicated to my parents, my wife Sunita, my sons Aarav and Kanav.

Table of contents

Contents

List of Figures

List of Tables

Acronyms

AI: Artificial Intelligence

AUC: Area Under the Curve

BMI: Body Mass Index

CAD: Coronary Artery Disease

CDC: Center for Disease Control and Prevention

CLEF: Clinical Laboratory Evaluation Forum

CNNs: Convolutional Neural Networks

COPD: Chronic obstructive pulmonary disease

CRF: Conditional Random Field

CVD : Cardiovascular Disease

CV: Cardio-vascular

MLP-NNs: Deep Neural Networks

eHealth: electronic Health

EHR: Electronic Health Record

EMR: Electronic Medical Record FFNs: Feedforward Networks

F-score: Final Score

GE: Generalized Expectation criteria

GRU: Gated Recurrent Unit

HMM: Hidden Markov Model

ICD : International Classification of Disease

ICT: Information Communication Technology

IR : Information Retrieval IE: Information Extraction

IV: Information Visualization

LSTM: Long Short-Term Memory

MALLET: Machine Learning for Language Toolkit

MaxEnt: Maximum Entropy

MICE : Multiple imputation by chained equations

ML: Machine Learning

NLPBA: Natural Language Processing in Biomedicine and its Applications

NHS: National Health System

NIH : National Institute of Health

NICTA: National Information and Communications Technology Australia

NLP: Natural Language Processing

NNR: Neural networks with regression O: Others

PHI: Protected Health Information

PID :Patient's Id

RBM: Restricted Boltzmann Machine

RegExp: Regular Expression

ReLu: Rectified Linear Unit

RegExlib: Regular Expression Library RNN: Recurrent Neural Network SGD: Stochastic Gradient Descent

TF: Term Frequency

UIMA - Unstructured Information Management Architecture UMLS: Unified Medical Language System

URI: Uniform Resource Identifier

US: United States

VM: Virtual Machine

WR: word representing

WSGI: Web Server Gateway Interface

SAF : Smart Analytical Framework

XGBoost : eXtreme Gradient Boosting

Chapter 1: Introduction

1.1 Motivation

Computer based decision support systems based on Machine Learning (ML) and Artificial intelligence (AI) have created a panacea everywhere and have become a buzzword nowadays, because of their increasing use for solving problems where an objective, data driven intelligent assessment is required. Different variants of AI based decision support technologies, such as machine learning, data mining, and pattern recognition have become pervasive in almost every type of decision support system, because of the enormous amount of data available, and the improved processing power of computers and storage devices. The progress in the state of the art in this field has enabled their diffusion and deployment in clinical decision- making workflow as well, often as a mechanism to provide second opinion to physicians and health professionals, for navigating through massive amount of data generated in the healthcare sector, through clinical notes, laboratory tests, and medical devices, wearable and portable devices used for vital parameter monitoring. Having a set of computer-based tools to make sense of this enormous data will help the burden and fatigue associated in analysing them and uncover the trends and patterns, providing the timely provided for interventions and medical care needed.

According to WHO (World Health Organisation), non-communicable diseases (NCDs), particularly cardiovascular diseases (CVDs) are a leading cause of death, with an estimated death rate of 31% caused globally. CVDs are caused by abnormal functioning of the heart and blood vessels, encompassing a range

of diseases, such as coronary heart disease, cerebrovascular disease, rheumatic heart disease, arrythmia and other conditions. It is estimated that around 18 million deaths occur worldwide due to CVDs, and it is important to detect the risk factors and vital signs of CVDs early on, and manage with appropriate interventions involving medications, diet, exercise, and behaviours.

Computer based decision support tools based on novel and sophisticated AI/ML algorithms can come to the rescue and help relieve much of the burden associated with high levels of disease management overheads throughout the life of a patient once the disease strikes. These tools can be built using the enormous data collected in the health system, at the primary, secondary, tertiary and quaternary levels, including the self-management techniques, based on using wearables and tracking devices. The levels of care refer to the complexity of medical cases, depending on the type of condition the physician treats and their specialities. Primary care involves your primary healthcare provider. You see them for things like acute illnesses, injuries, screenings, or to coordinate care among specialists. Secondary care is the care of a specialist. These specialists may include oncologists, cardiologists, and endocrinologists. And tertiary care is a higher level of specialized care within a hospital. Similarly, quaternary care is an extension of tertiary care, but it is more specialized and unusual. The decision support tools at any of these different care levels can be designed to learn patterns from the data, based on different algorithms tailored for at patient level or demographic cohort level, with appropriate parameter configuration and metrics for assessing the health status and impact of different interventions.

1.2 Background and aims

Out of several CVD variants, the coronary heart disease (CHD) is one of the leading causes of death in Australia, followed by cerebrovascular disease (which includes stroke). Obesity or overweight is one of the main reasons for this disorder and a large amount of research literature is available on the causes of disease, associated risk factors, and early interventions devised by clinical experts in managing the disease. However, with the increasing adoption of digital health technologies, there is a large amount of health-related data that gets collected in the health systems at various levels, at GP clinics, pathology labs, hospital systems, and personal health monitors. Much of the data does not get used properly, and ends up in storage and archival, as doctors, caregivers and health professionals can't navigate through this enormous data and find unusual trends, patterns and anomalies that can help inform the patients' response to a particular medical intervention. Availability of appropriate computer-based decision support tools that can analyse the massive data and provide timely decision support, can help alleviate the burden and fatigue in the health systems.

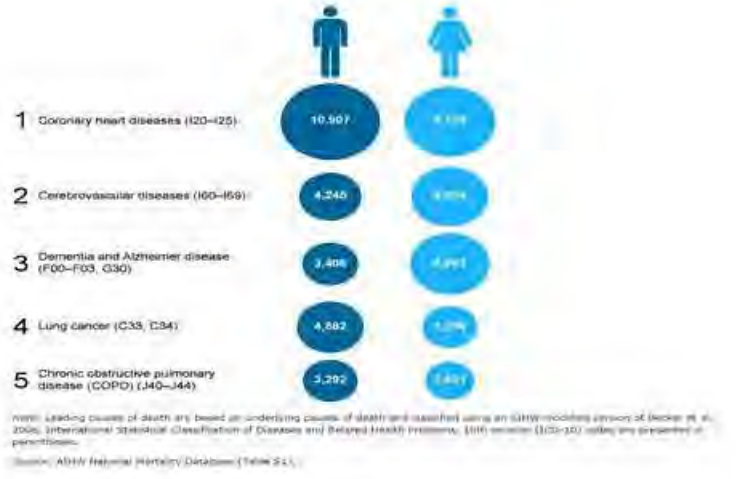


Figure 1: Leading Causes of Death in Australia [1]

As shown in Figure 1, The Australian Institute of Health and Welfare is Australia's national health and welfare statistics and information agency [1] and has conducted seminal research studies and analysis on "The relationship between overweight, obesity and cardiovascular disease", and presented the literature evaluating the association between overweight, obesity and cardiovascular disease (CVD), and whether the relationship is direct or indirect. The report also examined the relationship between CVD or its metabolic risk factors and 'excess body weight', indicating all degrees of overweight taken as one group. This is because the traditional literature does not always focus on traditional categories such as 'overweight' and 'obesity'. While this ground-breaking research study and analysis was comprehensive in its nature, it does not provide information to physicians on approaches for early detection and effective management of the disease for vulnerable individuals, as often the chronic non-communicable diseases need to be managed appropriately, often lifelong, with continuous monitoring and tracking, with different types of interventions needed, involving medication, diet, exercise, and behavioural interventions. But as outlined before, none of the data available in disparate health systems in the health care pipeline, contain detailed in-depth information on successes and failures achieved for a particular intervention strategy that the physicians and care teams can find helpful, and be able to evaluate which of the interventions have worked, and led to improvement for either an individual or a group of individuals with shared demographics [2], unless they manually spend hours at length digging into the data, and make sense of it. This also takes away, valuable time from experts who could be using it for understanding the disease better and working

more with patients directly.

In 2010, National Health and Medical Research Council (NHMRC) prepared a very detailed report about obesity and its drivers i.e. "A "state of the knowledge" assessment of comprehensive interventions that address the drivers of obesity" [3]. This report is based on literature reviews, reports and policy documents produced or published between 1980 and September 2010. A series of search strategies were defined to try to capture all recent reviews evaluating interventions addressing obesity-related behaviours as well as structural and environmental changes [3]. The above referenced studies were based on manually captured data via survey questions and available literature in these areas. While these studies indicate the alarming state of the obesity epidemic as one of the main risk factors for CVDs and other NCDs, this large set of information and insight just gets buried in the reports, ending up in archives, and does not get utilised in equipping the community and health professionals in proactive, predictive, and prescriptive management of disease. In recent years, there are several advances in computer-based analysis of complex and big data, which can provide methods and tools for capture and analysis of the health- related data and can utilise the valuable information stored in archives. The development of existing and new computer-based approaches for analysis and interpretation of health data can provide a comprehensive understanding on accurate relationship correlations between different risk factors, and hence improve our understanding of the obesity epidemic and its effect on population health, particularly concerning heart diseases. Also, sophisticated big data analytics based on machine learning and AI, can substantially improve decision making, can minimize risks, and

unearth valuable insights that would otherwise remain hidden [4].

Further, with immense advancements in wearable technologies, specifically smart watches, fitness bands and on-board biosensors on smart phones, health information technology has become ubiquitous in patients' and physicians' lives [5]. This intersection of health and technology is changing in how long-term chronic conditions can be monitored and treated at person level. Google, Apple, Fitbit, Samsung, and many more are racing to develop devices and platforms that track, aggregate, and monitor a wide range of biometrics. These vital signs can be collected at the geospatial level with on-board GPS tracking devices on smartphones, for monitoring the health at community level, for different subpopulations. The integration of these devices in our daily life is growing rapidly, and it is envisaged that such devices can seamlessly monitor and keep track of our activities, learn from them, and assist not only the physicians, but each one of us in making decisions about our health [5]. Chronic disease management, particularly the non-communicable diseases, mainly heart disease, is one of the most expensive, fastest growing, and most intractable health care problem facing healthcare providers and governments across the globe. More than 95 per cent of the world's population suffer from one or more chronic health problems, according to the Lancet Study on Global Burden of Disease Study 2013 [6], as patients live longer with a higher number of significant, expensive, debilitating health problems.

Further, this situation gets worst with inappropriate policies for different developed nations. For instance, health care costs have been on a continuous rise in many countries, including Australia. Australia is today ranked as one of the fattest nations in the developed world. The prevalence of obesity in

Australia has more than doubled in the past 20 years [2]. Obesity has become the single biggest threat to public health in Australia. It is a serious public health problem with over 61% of Australian adults and around 25% of Australian children being assessed as overweight or obese in the 2016 Australian Health Survey [2]. There is no wonder that due to this, cardiovascular disease (CVD) - heart, stroke, and vascular disease - remains the leading cause of death in Australia, the most expensive disease group in terms of direct healthcare costs and a major cause of avoidable hospital admissions and disability [3]. One way to combat this epidemic is to contain this problem, if individuals and communities take responsibility of their lives, in monitoring their vital signs regularly, follow healthy lifestyles patterns, and contribute to reducing the burden of increasing healthcare costs.

With the availability of innovative and cutting-edge computer and web-based tools, and the personal tracking devices, including smart watches, fitness bands and other wearables, it appears that this would be feasible. The tracking of vital signs by individuals on a regular basis in the community might help in enhancing the medical research into understanding causes, symptoms, and treatments of different debilitating chronic diseases, as it will allow massive amounts of data to be collected. By analysing large repositories of health-related data, it might be possible to obtain hindsight, insight and foresight behind the disease causes, and actionable interventions to reduce direct health costs, help formulate evidence-based policies, and build predictive models of disease and treatment, during early stages of disease onset.

1.3 Significance

The availability of novel computer-based decision support tools can lead to

greater good, allowing providers to look beyond the well-being of their patient populations, and be able to achieve improvement in collective health outcomes of much larger groups of individuals. This can lead to healthcare reforms and enhancement in holistic health outcomes, rather than simply providing more tests, medications, or procedures, and facilitate a value-based care. For this, providers need to be actively exploring ways of using data, analytics, and care coordination tools to manage population health and make value-based care a reality. There is still much to learn, however, and even more, the urgent task is the translation of existing knowledge into policies and effective programmes [2].

To reiterate, health care costs in Australia are increasing year by year. Australia is today ranked as one of the fattest nations in the developed world. The prevalence of obesity in Australia has more than doubled in the past 20 years [3]. Obesity has become the single biggest threat to public health in Australia. Obesity is a serious public health problem with over 61% of Australian adults and around 25% of Australian children being assessed as overweight or obese in the 2014 Australian Health Survey [1]. During one study, it was found that Aboriginal and Torres Strait Islander Australians are 1.9 times as likely as non-indigenous Australians to be obese [7]. In Australia, heart disease groups are most expensive group, putting billions of dollars' burden on the Australian budget every year. While there has been significant progress over the past five decades, CVDs- heart, stroke, and vascular disease - remain the leading cause of death in Australia, the most expensive disease group in terms of direct healthcare costs and a major cause of avoidable hospital admissions and disability. The number of people with cardiovascular

disease (CVD) is set to increase as the population grows, ages, becomes increasingly overweight and obese and some risk factors, such as poor nutrition, lack of physical activity, high blood cholesterol and high blood pressure, continue at alarmingly high rates. Here are the few facts about heart diseases in Australia [7]:

Cardiovascular disease (CVD) is the leading killer of Australians: 46,000 deaths (31.7% of all deaths) in 2010.

- CVD is the most expensive disease group: $7.9bn or 11% of direct healthcare expenditure a year.
- CVD accounts for many potentially preventable hospitalisations.
- Strokes cost Australia an estimated $2.1bn a year.
- The direct health costs of heart attacks are estimated at $1.1bn a year.
- Total economic cost of CVD is estimated to be $15.5 billion a year.
- CVD comprises 18% of the total disease burden in Australia.
- There are an estimated 3.7 million Australians with long-term CVD.
- 1.4 million Australians have a disability associated with CVD.
- Australians will suffer around 50,000 new and recurrent strokes each year.
- CVD death rate in rural/remote areas is 1.4 times higher than in major cities.
- Addressing lifestyle factors can reduce mortality risk by 66%.

Globally, about US$70 billion is spent on health research overall, with an estimated 10% spent on researching the 90% of the world's problems, thus neglecting the health needs of poor countries [2]. In Australia, less than 3.5% of total health expenditure was spent on health research and development in 2011-12 [3]. In one of the types of research reported on US health system, it was shown that by using efficient and appropriate computer-based technologies and tools, for analysing big data from health domain, has a value

of more than $300 billion every year, with two-thirds of that in the form of reductions to national health care expenditure of around 8 percent, and is depicted in Figure 2.

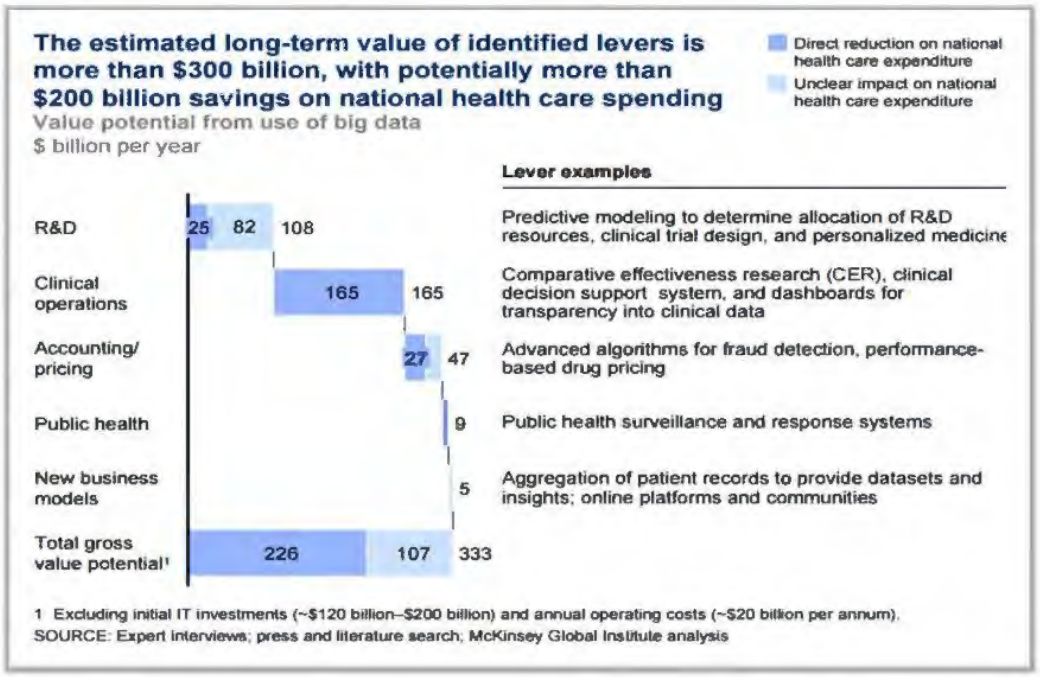


Figure 2: Long Term Value of Health Data Research in US [4]

Similar benefits can be achieved in Australian healthcare systems as well. Obesity costs Australia more than $56 billion per year [2]. In 2008, the cost of obesity in NSW alone was $19 billion, according to NSW Health [3]. Globally this phenomenon also holds. Projected health care costs have revealed that by the year 2018, obesity-related medical expenses will top $344 billion in the US, double the current expenditure level [6]. It is estimated that by 2025, close to 30 percent of the population in mature economies across the globe will be aged 60 or over, up from 20 percent in 2000 [6]. According to the World Bank, it represents anywhere from

1.6 percent of a country's GDP (South Sudan) to 19.5 percent in Liberia. In the developed world, the U.S. tops the pack at 17.9 percent [6]. The healthcare

sector faces basic challenges of operations, logistics, resource allocation, customers, and management. In addition, efficient use of resources is a perennial problem, where the significant investment is made by nations in acquiring resources, but not fully exploited and utilized. For example, a large amount of data is collected in health care systems, including electronic health care records in hospital databases, clinical notes in primary and allied health data stores. While these different care systems have been amassing data and storing them at an exponential rate, this data has not been fully utilised in understanding the disease prevalence, and its impact on health care burden, and how to mitigate this. This could be due to lack of availability of sophisticated and objective tools for quantitative and qualitative analysis and making sense of this massive data getting stored in hospital databases, and the primary, secondary, and allied health systems. For example, availability of state-of- the-art spatial health data analytics tools, can enhance the capabilities of electronic health care records in hospital databases in conjunction with GIS (Geographical Information Systems) and can help the healthcare industry, address some of the public health issues, such as the strategy, capital planning, public health administration, marketing, and operations [4]. Spatial data analysis is a set of sophisticated data analytics techniques devised to support a spatial perspective on data. To distinguish it from other traditional forms of analysis, it can be defined as a set of techniques whose results are dependent on the locations of the objects or events being analysed, requiring access to both the location and attributes of objects. Its techniques can range from simple descriptive measures of patterns of events to complex statistical tests of whether a set of events could have been generated by specific, well-defined,

processes, by using location as a logical nexus, spatial data analytics using GIS allows executives and managers to evaluate the interplay of factors that affect health care delivery and the operations of providers. Understanding the implications of such large-scale interactions can lead to better decision-making. Related to cost reductions is the task of efficient capital planning. Poor allocation of funds can hurt an institution, its operations, and its future. Geospatial analysis of demographic patterns and associated changes in types of services in demand—an aging population for example, would require more services for age-related conditions, and can will help governments to put money not just where it is needed today, but in the next five to ten years.

As mentioned before, healthcare providers have been continuously and quite proactively, making enormous progress in equipping the health care systems with an ability to collect, store, and archive massive amounts of data, and perform routine tasks to identify, diagnose, treat, manage common chronic conditions. Recently, a growing number of physicians and health care providers have started to appreciate the potential of electronic health records installed in most hospitals, and the capability inherent in the massive data stores created each day, by collecting from inpatient claims to lab tests to patient-generated health data (PGHD) streaming, or the smart wearables, and medical devices [4]. In Australia, this technological adaptation is a bit slow, as less than 3.5% of total health expenditure was spent on health-related research and development work in 2011-2012 [7]. As there are limited resources, research and development in healthcare domain will benefit the society and reduce the burden of disease, if appropriate computer-based technologies and decision support tools are available, targeted towards the

predictive epidemiology domain.

Further, by making care coordination and patient supervision a priority, it is possible to cut the average costs of chronic disease care by $4500 per patient over a three-year period, if providers use intelligent, computer-based population health management strategies to reduce the fragmentation of services, says a recent study published in the American Journal of Managed Care. It can also help in raising patient satisfaction and reducing preventable hospital readmissions [9].

In addition, for making sense of this big data from large number of sources, and using the insights to obtain actionable preventative strategies, there is a need for intelligent, automatic computer-based decision support tools, that use sophisticated machine learning, AI, and data analytics algorithms, and can work under challenging data capture/collection settings.

1.4 Research Gap

Though applications of AI in healthcare are evolving fast, and hyped as an essential tool in clinical practice, this has not happened as expected. This could be due to lack of trust on the decision support provided by the automatic tools, without clear reasoning and rationale, on how computer arrived at a particular decision related to patient's health. This could be due to a problem called the black-box problem. The computer-based models take inputs and decides without the user involvement in making predictions. This leads to lack of trust, confidence, and transparency, and can sometimes lead to catastrophic results when the application is of vital importance for life threatening situations in emergency hospital settings and extreme environments.

Most of the previous studies using computer-based tools in clinical decision support pipeline, usually focused on providing absolute decisions without a clear reasoning on predictions made on the disease diagnosis. Though providing a prediction can be useful, it is not sufficient if it is a blind decision with certain performance metrics, based on black-box models, without explicit analysis supported with clear interpretations and proper explanations. It is important for stakeholders and end users (experts and patients) to be fully convinced with the decisions provided by computer-based tools, so that it can be seamlessly embedded in the decision- making clinical workflow.

Hence, there is a need to develop computer-based disease prediction models based on innovative algorithm pipelines, that can not only provide accurate predictions but also be interpretable in how the model has arrived at a decision and be able to explain the rationale behind the decisions made. This is particularly important for better management of non-communicable chronic diseases, such as cardiovascular diseases, diabetes, asthma, obesity, and high blood pressure, where there are no short-term fixes, but involve lifelong disease management, and requires efficient and interpretable decision support to be provided by computer-based tools. For this, there is a need to rethink on type of machine learning algorithms to be used or developed, which not only have to be efficient and robust enough to work under challenges of big data with poor data quality often associated with multitude of data sources used in monitoring CVDs, but also need to be simple, easy to interpret, and can be traceable on the decision-making rationale the tool has used. Currently there are not many approaches that are available to deal with these challenges, and there is an urgent need to develop decision prediction models based on innovative

machine learning/AI algorithms to address these requirements. In this research, an innovative predictive modelling framework for addressing these challenges is proposed. The proposed framework allows the disease prediction models based on novel algorithm approaches for addressing several of these challenges and shortcomings associated with clinical decision support tools.

1.5 Challenges

There are several challenges in developing predictive modelling and decision support systems based on novel machine learning, AI, and data science approaches. Some of the key challenges are as summarized below:

1. Unlike manufacturing process, in which the products are standardized, patients are humans, and 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, poor prediction performance and incorrect decision support can happen.
2. Evolving medical knowledge and continual addition of new clinical information result in inclusion of large number of clinical features in the predictive models, leading to complexity in analysis and interpretation. This situation, commonly known as the '*curse of dimensionality*', 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 lies beneath 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.

4. Hence, there is a need to develop and use algorithms and methods, that can (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), (3) recognize the importance of employing learning algorithms with high generalization ability valuable for the development of patient-oriented prediction models. (4) Interpretable and explainable in providing the rationale and reasoning behind the decisions provided, so that they elicit trust, transparency and confidence in stakeholders and end-users. Achieving these outcomes would ultimately lead to enhancement in 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, a novel computational framework is proposed, for addressing some of the limitations and shortcomings of the algorithms and methods that are based on machine learning, AI, and data science approaches. The focus is on cardiovascular diseases (CVD), for the development of the proposed algorithmic framework, as it continues to be one of the leading causes of death worldwide since last couple of decades [6].

1.6 Research Questions

The research questions addressed in the thesis can be outlined as follows:

- What are limitations of current disease prediction models based on machine learning/AI algorithms, in terms of interpretability and challenging data settings?
- Is it possible to enhance disease prediction models based on novel machine learning /AI algorithms, so that they are more interpretable and explainable eliciting increased trust, transparency, and confidence?

1.7 Research Objective

The aims and objectives for the proposed research attempt to address the above- mentioned research questions are outlined below:

1. To investigate current and existing machine learning and data mining approaches for building disease prediction models, which can work well under challenging CVD data settings and are inherently interpretable.
2. To adapt and extend the disease prediction models with innovative machine learning and AI algorithms, so that they are more accurate, robust, interpretable, and explainable.

1.8 Research Methodology

A data driven quantitative approach and research methodology will be used for achieving the aims and objectives of the proposed research and will involve

following stages:

Stage 1: Investigating current disease prediction models based on ML/AI algorithms for CVD and extending/adapting the algorithms for enhancing the performance and robustness.

Stage 2: Enhancing the disease prediction models based on novel ML/AI algorithms for CVD with improved interpretability and explainability.

1.9 Innovative and Original Contributions

There are several original and innovative contributions made in this thesis and are as outlined below:

- Developed disease detection models based on shallow machine learning/AI approaches, with model building and validation on Cleveland dataset [10], for setting a baseline reference for comparison with previous work.

- Extended the disease detection models based on shallow machine learning/AI approaches, which are high performing and inherently interpretable, with model building and validation using two publicly available CVD datasets, NHANES [11] and Framingham Heart Study [12].

- Developed Disease prediction model based on innovative XGBoost algorithm which is high performing with better interpretability, with model building and validation using two publicly available CVD datasets, NHANES [11] and Framingham Heart Study [12].

3. Disease prediction model based on agnostic black box model based on deep learning equipped with innovative feature optimisation based on LASSO regression algorithms for enhancing robustness, in terms of working under challenging imbalanced big data settings. The model building and validation was done on a super dataset from multiple NHANES datasets [11] with simulated imbalanced data settings.

The results from these findings were published/under review in several international peer reviewed conferences/journals listed below.

1. Nitten S. Rajliwall, Girija Chetty, Rachel Davey, "Chronic disease risk monitoring based on an innovative predictive modelling framework" in 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, 2017, DOI: 10.1109/SSCI.2017.8285257 https://ieeexplore.ieee.org/document/8285257

2. Nitten S. Rajliwall, Girija Chetty, Rachel Davey , "Machine Learning Based Models for Cardiovascular Risk Prediction" in 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, NSW, Australia DOI: 10.1109/iCMLDE.2018.00034 https://ieeexplore.ieee.org/document/8614017

3. Nitten S. Rajliwall, Girija Chetty, Rachel Davey, "Cardiovascular Risk Prediction Based on XGBoost" in 2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), Nadi, Fiji

DOI: 10.1109/APWConCSE.2018.00047

https://ieeexplore.ieee.org/document/8853798/

4. Nitten S. Rajliwall, Girija Chetty, "Deep Learning Based Decision Support Framework for Cardiovascular Disease Prediction" in 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, Australia
   DOI: 10.1109/CSDE53843.2021.9718459
   https://ieeexplore.ieee.org/document/9718459

5. Nitten S. Rajliwall, Girija Chetty, "Cardiovascular Disease Prediction Based on Interpretable and Explainable AI" under review (submitted to Scientific Reports on 07 May 2022) under review.

1.10 Thesis structure and road map

This thesis is organised into seven chapters:

- Chapter 1: Introduction
- Chapter 2: Literature review
- Chapter 3: Disease Detection Models Based on Shallow Machine Learning
- Chapter 4: Smart Predictive Modelling Framework
- Chapter 5: Cardiovascular Risk Prediction based on XGBoost
- Chapter 6: Deep Learning Based Decision Support Framework for Cardiovascular Disease Prediction
- Chapter 7: Conclusions and Future Work

- Chapter 8: Bibliography

The introductory chapter explains the problem description, motivation and objective of this research, research questions, the contribution to the scientific knowledge, the methodology, and the structure of the thesis

Chapter two explains the background literature and related work reviewed during the research.

Chapter three describes the work done on building disease prediction models based on shallow machine learning using Cleveland dataset, which sets the baseline for comparison with prior work and examines the interpretability of the shallow machine learning models.

Chapter four extends the disease prediction models based on shallow machine learning techniques, which are inherently interpretable, using two different datasets NHANES and Framingham Heart Study /CHS dataset.

Chapter five focus on the developing a novel disease detection model, based on XGBoost approach for CVDs.

The focus of the Chapter six is on enhancing the model performance and robustness, based on deep learning algorithms, which can work well under imbalanced big data settings.

And finally, the thesis concludes with Chapter seven, with some key contributions from this research, and future directions on extending the work.

1.11 Chapter Summary

This Chapter presented an overview of the entire thesis, the motivation, and the objective, and involves defining the research problem, identifying the research gap, formulation of research questions, defining the aims and objectives, identifying approach and methodology for addressing the research questions, innovative research contributions from this work, and providing a road map for the thesis.

The next chapter provides a review of literature, and related work on methods and approaches based on different machine learning algorithms and data analytics techniques and provides a background related to the work in this thesis.

Chapter 2: Literature review

2.1 Introduction

Computer aided decision support tools for disease prediction, including CVD, can be categorised based on clinical activity or the type of data used. Some of the types include demographic data, medical notes, radiological scans, pathology slides, skin lesions, retinal images, electrocardiograms, vital signs, electronic recordings, physical examinations and clinical laboratory images. ML/AI based approaches can be used for building disease prediction models and compared with ground truth or physicians' assessment based on different metrics such as the area under the curve (AUC) obtained from the plot of true-positives versus false-positive rates, known as the Receiver Operating Characteristic (ROC) to assess their performance and suitability in clinical decision-making workflow [13].

This Chapter aims to review some of the methods and approaches proposed in the literature for disease prediction based on ML/AI approaches relevant to this thesis and is not meant to be exhaustive and extensive. Hence, the literature review covers some methods based on the data type and some related work done on detecting cardiovascular diseases.

2.2 Biomedical Image Analysis Based Methods

An extensive body of work has been done based on medical image data, often categorised as biomedical image analysis. Several methods for processing and analysing biomedical images have been proposed in health technology literature. Proliferation of methods based on this type of image data could be due to the abundance of radiological and medical imaging scans stored in large quantities in most of the hospitals' Picture Archiving and Communication Systems (PACS).

Also, due to significant advances made in image analysis and computer vision algorithms for other non-medical fields, it was easy to adapt and extend it to analyse biomedical images.

Amongst biomedical images, radiological scans based on X-rays, particularly chest X- rays, currently represent the most common type, with more than 2 billion Chest X-rays performed worldwide. Analysis of Chest X-rays allows diagnosis of different stages of lung diseases. An AL/ML approach based on 121-layer convolutional neural network (CNN) was proposed by Wang et al [14] to detect pneumonia in 112,000 frontal chest X- ray images, and even with an accuracy of around 76%, their study concluded that their model performed better than four radiologists' opinion. However, this does not ideally reflect the superior performance of computer-based method proposed, because radiologists can detect other anomalies in the X-rays, which could be significant which computer-based model was not trained to do. Another approach proposed by authors in [15] was based on the same dataset and their method was able to detect 14 different thoracic diseases, with an accuracy range of 67% for pneumonia to 91% for emphysema.

Other than radiological image scans, pathological images form another frequently used modality for building disease detection models, particularly the whole slide imaging (WSI) scans, which are digitized versions of glass slides. An AL/ML approach based on different deep learning algorithms was proposed by authors in [16] to analyse the whole slide images to detect metastases in tissue sections of lymph nodes of breast cancer images and compared with a ground truth diagnosis from 11 pathologists. By using a test protocol of pathologists' skill in assessing 129 whole slide images within a small time window(less than one minute per slide), results turned out to be in favour of their proposed ML/AI algorithms. The

performance metrics achieved for algorithms included an AUC ranging from 0.556 to 0.994, whereas the mean AUC score achieved for pathologists' skill under time constraints was 0.810. However, without imposing time constraints, the mean AUC score achieved by top five algorithms was similar to pathologists' skill, with a score of 0.966 for pathologists, as compared to 0.960 for algorithms.

Use of dermatology images is another type of image data, that had several approaches proposed by researchers based on AI/ML. One of the most significant works was based on using Google Inception CNN model on more than 100,000 skin cancer images for detecting around 2000 different skin diseases, with a matching performance achieved when compared with 21-member board certified dermatologists for skin cancer detection and classification [17]

The ophthalmology images have received increased interest in the research community, due to their importance on detecting diabetic retinopathy, affecting more than 90 million individuals worldwide, mainly causing adult blindness. One of the popular and highly efficient AI/ML based system, called IDx-DR, proposed by researchers in [18], was used for a trial conducted on 900 patients at primary care clinics, in an autodidactic mode, and then locked for testing. The system performance achieved was a sensitivity, true positive rate of 87% and specificity, true negative rate of 91%, and exceeded FDA regulations. However, FDA regulations stipulated the switching off the autodidactic mode, and stopping its auto-learning function, and was allowed to be used as just a non-AI based assistive diagnostic tool with ophthalmologists involved in the decision-making loop, thus hindering the systems' capability.

2.3 Physiological Sensor Signal Analysis Based Methods

Several methods based on physiological signals from different sensors attached externally to the skin region corresponding to different body parts constitute an important area for detecting diseases. Some of these sensors, such as the Electromyogram (EMG), electroencephalogram (EEG), electrooculogram (EOG) and Electrocardiogram (ECG) sensors have been used for building disease detection models by analysing the physiological signals from these sensors [19].

For analysing the electrical activity produced by skeletal muscles such as muscle state, activation of muscle and force generated, Electromyogram (EMG) sensors were used. By placing these sensors on various parts of the body and capturing the signals and analysing them with appropriate ML/AI methods, muscle and limb disorders could be detected. Several deep learning based physiological signal analysis methods were proposed by authors in [20]. Other similar works include methods for limb movement estimation [21], gesture recognition [22], and hand movement classification [23].

For recording electrical activity in the brain, Electroencephalography (EEG) sensors were used, and brain-computer interfaces were developed. The electrophysiological monitoring of EEG signals involves collective processing of signals corresponding to charges coming from the neurons in the brain and is often very noisy and hard to interpret. In these settings, the computer-based approaches perform well as compared to humans in interpreting these noisy signals, and several methods have been proposed in the literature to address this complex task. Notable amongst these works include sleep state identification [24], seizure detection [25] and emotion classification [26].

For measuring corneo-retinal standing potential between the front and back of human eye, electrooculography (EOG) sensors were used, and by analysing the eye movements from EOG sensor signals, ophthalmological diagnosis on eye health is obtained. However, these signals could be affected significantly by noise and difficult to interpret, and appropriate AI/ML based methods perform better than humans. Some of the methods proposed in this area include approaches for drowsiness detection [27], sleep stage classification [28] and driver fatigue detection [29].

For recording the electrical activity of the heart, Electrocardiogram (ECG) sensors are used, which involve placing electrodes on the chest and ECGs signals analysed in different intervals to detect cardiac activity. The cardiologists look for changes in the morphological pattern produced in different sections of the time series signals captured from ECG electrode sensors and assess cardiac ailments. Due to difficulty in distinguishing the ECG signals for a healthy heart activity from an abnormal heart activity from these ECG sensor signals, computer-based methods were found to be useful, and several methods have been proposed for analysing ECG recordings, for detecting coronary heart disease [30], classification or irregular heartbeats [31], and detection of congestive heart failure [32].

2.4 Methods for Analysing Electronic Health Records

Several methods for analysing electronic health records (EHRs) have also been proposed in the literature, as they constitute primary carriers of health information, and contain a wealth of patient information related to vital parameters tracking needed for diagnosis, medications, and laboratory tests (structured records) and free-text clinical notes (unstructured records). This massive data in structured and unstructured format has

been underutilised so far and digging deep into this enormous data using appropriate AI/ML based methods can provide an opportunity to address larger health issues with public health and global health perspective and develop personalised patient level health interventions and can guide new treatments and therapies. Not much research has been pursued in analysing this type of data, due to the complexities associated with the nature of data from EHRs, such as high dimensionality, sparse due to poor data quality, temporal but irregularly spaced, and biased. One of the notable approaches includes the use of four-layer convolution neural network model for feature extraction and electronic phenotyping from patient EHRs and predicting congestive heart failures and chronic obstructive pulmonary diseases, with significant improvement in performance as compared to the logistic regression model [33].

Another approach for analysing EHRs efficiently for personalised care and treatment, called DeepCare [34] uses a cascade of several neural networks, including recurrent neural network (RNN) and long short-term memory (LSTM). This dynamic neural network architecture does several tasks including reading the medical records, storing previous illness history, can infer current illnesses, and predicts medical outcomes in future. By conducting the studies on diabetes and mental health data, this model allows modelling of disease progression, recommending appropriate interventions and prediction of adverse event risk in future.

As the EHRs also contain unstructured data in terms of free-text clinical notes, in addition to structure data captured from vital parameter tracking devices, several natural language processing (NLP) approaches have been proposed for extracting useful information and medical concepts for assessing the patients' health status. One of the notable works includes automatic discovery of peripheral arterial disease (PAD) from free-text clinical narrative notes based on NLP text processing approaches [35].

Another study involving a NLP model that can learn on concepts extracted from structured ontologies and free-text notes, used a with semantic similarity measure between the medical concepts, patient records and journal abstracts, and reported promising outcomes, with results correlating well with ground truth annotations from expert human assessors, and exceeding the state-of-the-art benchmarks in medical semantic similarity [36].

This section reviewed the prior work on different AI/ML based disease detection models based on different input data sources, such as the biomedical images from radiological and pathology slide scans, sensor signals from physiological sensor arrays, and structured and unstructured data from EHRs containing structured records from vital measurements, and unstructured free text clinical notes. But the methods proposed were targeted towards different diseases, affecting different body organs. The focus of the next Section is to review methods related to cardiovascular disease, particularly those studies where the same dataset was used, for consistent treatment and assessment on prior work done.

2.5 Related Work on CVD Prediction Models

As reviewed in the previous section, AI/ML based approaches appear to be promising in early detection and tracking of several debilitating diseases based on analysing input information from different types of data, including the data from biomedical images, bio-sensor signals, and structured and unstructured EHRs. Some of these methods and data types have also addressed cardiovascular diseases. For instance, one of the important studies involved detection of chronic myocardial infarction base on MRI radiological scans [37], and several methods for analysing ECGs exist in literature, with some of them reviewed in the previous section.

For diagnosing CVDs, it is interesting to note that a wealth of information is contained in EHRs, particularly the structured longitudinal records from vital parameter monitoring, such as the raised blood pressure, glucose, lipids, weight, and obesity status. Some of the computer-based methods used for processing these records for CVD detection are as discussed next.

One of the famous datasets used for developing AI/ML based disease detection models is the Cleveland database [10]. The authors in [38] proposed a heart disease prediction model based on Cleveland database and Naive Bayes classifier and classified the prediction output into five different disease risk categories such as none, low, average, high and very high risk. The system reported a prediction accuracy of 89.58%, and twenty-five misclassified instances, but it was not clear on how much data was used for model building, validation, and testing.

Another study used Naïve Bayes classifier, but another private dataset was reported by authors in [39]. Here, the system was developed with data collected from 500 patient records and reported a prediction accuracy of 74%. However, since the data was not publicly available, nor the details on how the model building, validation and testing were provided, it cannot be compared with other methods reported using Cleveland database.

An approach based on Support Vector Machine (SVM) and Multilayer Perceptron (MLP) was proposed by authors in [40], who used the Cleveland database for developing the disease detection models, and achieved a prediction accuracy of 80.41% with SVM, and 97.5% with MLP algorithm. While SVM was trained on detecting two classes (presence or absence of disease), the MLP model was built for detecting five different categories of heart disease, like the system in [38].

The authors in [41] proposed a CVD disease detection model based on decision trees and Cleveland database, and used different variants of decision tree algorithms, such as J48, logistic model tree (LMT) and Random Forest (RF) algorithm to compare the prediction performance. The prediction model was built using 10-fold cross validation, and designed to predict five different heart disease categories, with J48 algorithm being the best performing algorithm with 56.76% accuracy. As can be seen from the performance achieved with other similar studies [38] and [39], this model does not perform well.

Another disease prediction model [42], using Cleveland database and J48 decision tree with additional feature selection step involving only thalassemia, chest pain type and several features, such as major vessels, heart rate, chest pain, gender and age showed improved performance as compared to previous methods. The method used 240 instances only and with a train-test set split of 50:50, 75.83% - 85% accuracy was achieved, by removing irrelevant features and retaining selected features in model building (thalassemia, chest pain type and some of the features like major vessels, heart rate, chest pain, gender, and age).

The authors in [43] proposed a model based on hybrid neural network (involving a combination of fuzzy neural network and Artificial Neural Network), and Cleveland dataset, and achieved a prediction accuracy of 86.8%, with fuzzy concepts implemented allowing clustering, where patterns can belong to more than one cluster different degrees of membership.

The method proposed by authors in [44], was based on principle of neural network ensemble, and achieved a classification accuracy of 89% with Cleveland dataset. Another interesting study reported in [45] was based on a newer approach of ensemble neural networks, achieving a middle ground between ease of

comprehension possible with decision trees, and generalization ability of neural networks. The proposed neural network ensemble 'NeC4.5', uses C4.5 a decision tree algorithm and employs neural networks at the for generation of newer training examples, which are then fed to the C4.5 for classification. With the model built on Cleveland heart dataset, the performance achieved was significantly improved, when newer training examples were added with a larger decision tree with more nodes. However, the model building time was significantly higher than the other models, which could be one of the drawbacks for real time prediction settings.

2.6 Recent Work on CVD Prediction Models and the Gaps

Some of the more recent works reported, particularly the state-of-the-art studies and approaches on cardiovascular disease prediction based on machine learning are reviewed in this section. Nissa, N. et al. proposed an algorithm based on random forests by analysing different algorithms comparatively on publicly available UCI machine learning repository and reported a performance of 99.35% accuracy. However, the gap was that their time complexity was not determined, details of the methodological framework and model building process was not specified [61]. Dhankhar, A. and S. Jain reported a comparative study of different algorithms based on machine learning and showed that the random forest variant they have proposed performs better than decision trees, ANN, SVM for the UCI publicly available dataset with around 90% accuracy [62]. However, the dataset used was very small, with no details on methodological framework and the feature selection aspects. Geetha S. et al proposed a framework for cardiac infection using neural system estimation with model building and evaluation on Cleveland dataset and showed that their approach based on decision trees performs best with an accuracy of 99.7%, when compared with KNN, NB, DT, SVM and AdaBoost algorithms. However, the

architecture proposed was complicated, model didn't scale up for larger dataset, and did not include the features in model building with a scope for early detection [63]. Ghosh et al. proposed an architecture based on a specific feature selection algorithm for heart disease detection and showed that the random forest variant performs the best with 99.05% accuracy, with model building and evaluation done on several public and private databases such as UCI repository Satlog, Cleveland, Switzerland, VA Long Beach and Hungary and with similar algorithms as reported in other works, including DT, KNN, GB, SVM algorithms. However, the use of several private databases does not provide enough confidence on the capabilities of superior performance for their proposed architecture [64]. Maini, E. et al developed a practical and economic prediction system for early CVD diagnosis based on a variant of random forest algorithm again with a performance of 93.6% achieved when compared to other machine learning algorithms based on KNN, LR, DT, Adaboost, RF and ANN. However, the model building and evaluation was done on a private database, with no studies shown on any publicly available database. Hence it is hard to assess the merit of the proposed approach [65]. Abdellatif A. et al proposed a method based on the Synthetic Minority Oversampling Technique (SMOTE) to handle imbalance distribution issue, typical for health data sets, and evaluated six different ML classifiers to detect the patient status, and used Hyperparameter Optimization (HPO) to find the best hyperparameter for ML classifier, based on Extra Trees (ET) together with SMOTE [66]. The authors use two public datasets to build and test the model using all features. The results show that SMOTE and ET optimized using hyperband achieved higher results than other models and outperformed the state-of-the-art works by achieving 99.2% and 98.52% in CVD detection, respectively. Also, the developed model converged to 95.73% severity classification using the Cleveland dataset. However, the Cleveland dataset

is a balanced dataset already, and investigating SMOTE on already balanced dataset seems to be of not much value.

2.7 Limitations of Current CVD Prediction Models

As can be seen from the review of methods for CVD detection/prediction in the earlier section, several studies have been reported on different epidemiological databases, similar to Cleveland database, for obtaining a deeper understanding and characterisation of myocardial infarction (MI) disease, commonly known as heart attack, one of the life-threatening cardiovascular diseases. From these studies, major risk factors associated with CVD were identified and these include cholesterol, gender, age, obesity, diabetes, hypertension, alcohol, smoking, sedentary lifestyle, unhealthy diet and psychosocial factors. The risk factors outlined above can be broadly 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, amongst others. Identification of these risk factors early on, is important as they are measurable elements or characteristics that are causally correlated to an increased risk of a disease. However, caution needs to be taken when analysing risk factors as their degree of impact on individuals' health may change as one ages. Therefore, careful monitoring, analysis and management of these risk factors. could reduce the mortality rate. This can be done by continuous analysis of rich information available in the EHRs, as they get recorded and stored during every visit to GP clinics and routine health check-ups and stored in the health system regularly. However, it is interesting to note that much of the rich information available in these records do not get exploited fully for getting a deeper insight into

the health status of individuals, nor there are many computer-based decision support systems based on novel ML/AI algorithms are available for developing proactive interventions for people at high risk of developing cardiovascular diseases.

Another shortcoming of the current ML/AI based disease detection models, is the lack of robust, objective, explainable, interpretable and trustworthy inference available from the existing decision support tools, as most of the performance metrics reported have been limited to quantitative measures such as accuracy, precision, recall, F-measure, AUC, ROC without any qualitative metrics that are easy to interpret and explain, eliciting trust from the stakeholders using the system. This could be due to the problem that most of the current ML/AI tools were built using mathematically rigorous constructs, that are hard to interpret and explain, and hence the decisions provided by them appear to be coming from a black box, offering little explanation on decision arrived.

The research proposed in this thesis is aimed at the development of a breakthrough predictive modelling framework, based on innovative ML/AI algorithms for building CVD disease detection models. The proposed computation framework provides an intelligent and interpretable holistic analytical platform with improved prediction accuracy, and improved interpretability and explainability. The proposed innovation and development can help drive the healthcare system to one that is more patient-centred, and trustworthy, with potential to be tailored for several diseases such as cancer, cardiovascular disease, asthma, traumatic brain injury, dementia, and diabetes. The outcomes of this research based on innovative findings can serve as an example – that the availability of better computer-based decision support tools, with novel computational strategies, which can address a

patient's unique clinical/genetic characteristics, can result in better characterization of diseases and at the same time redefine therapeutic strategies.

The proposed predictive modelling framework in this thesis attempts to achieve these objectives, using a line of investigation with different strategies, such as:

- Developing disease detection models based on traditional shallow machine learning algorithms, particularly those based on decision trees and their variants. These algorithms have shown to be inherently interpretable and accurate white box models and can serve as the baseline for comparing with previous models proposed in the literature.
- Developing disease detection models based on model agnostic algorithms, such as deep neural networks (which are black box models), but are ideal for complex clinical data settings, with large data sizes, and poor data quality with imbalanced class distributions.
- Developing disease detection models based on model agnostic algorithms, such as deep learning networks, but augment with appropriate pre-processing and post-processing stages to provide better interpretability and explainability and eventually make them an efficient white box model.

To the best of our knowledge, no prior work on such a comprehensive predictive modelling framework, based on innovative machine learning approaches has been reported so far.

For an objective comparison of the methods proposed in each of the above stages, several publicly available benchmark clinical datasets, including Cleveland [10], NHANES [11] and Framingham Heart Study/CHS [12] datasets were used for model building and experimental validation.

2.8 Chapter Summary

This chapter described the relevant literature and prior research and has set the background and identified the line of investigation for this research. First, the previous work on diverse types of data and algorithms used for CVD in the research literature was discussed. Based on the review of the earlier studies in the related area, challenges that need to be addressed for enhancing the performance, robustness, and interpretability of algorithms were identified. The Chapter concluded with a systematic plan for pursuing research by addressing each of the challenges identified, and the proposal for an innovative predictive modelling framework for CVD diagnosis is defined.

Next few Chapters describe in detail, contributions towards different disease detection models based on novel AI/ML algorithms, addressing each challenge.

Chapter 3 focusses first, on setting a baseline reference with Cleveland dataset used in previous studies (discussed in Chapter 2/Literature Review Chapter), and traditional shallow learning algorithms and examines the accuracy and interpretability of the disease detection model.

Chapter 3: Disease Detection Models Based on Shallow Machine Learning

3.1 Introduction

This Chapter provides details on studies conducted to examine the disease detection models based on traditional shallow machine learning algorithms using the Cleveland dataset. The outcomes from the study will set the baseline for comparison with previous work done, and for assessing the performance, robustness, and interpretability of complex data settings with different databases. The rest of the Chapter is organised as follows. The next section describes the details of Cleveland database, one of the most famous datasets used in the prior work for building disease prediction models, followed by the exploratory analysis of the data in Section 3 for understanding the data. Section 4 describes the experimental work done for developing disease prediction models based on shallow machine learning algorithms and assesses its performance in terms of prediction accuracy and model interpretability. Section 5 concludes with a summary of the findings from this Chapter.

3.2 Description of Cleveland Heart Disease Dataset

The Cleveland heart disease dataset is a publicly available dataset available in UCI machine learning repository [10] and consists of 13 variables representing vital parameter measurements on 303 individuals. The 14^{th} variable is the class label, a binary label, indicating the presence or absence of heart disease. The description of the dataset is shown in Table 1. Out of 303 instances available, six instances were dropped as they

appeared to be duplicate instances. The Target variable shown in Figure 1 represents the class label, indicating the presence or absence of cardiovascular disease.

TABLE I. CLEVELAND HEART DISEASE DATASET

Variable	Description	Type
Age	Age in years	Integer
Sex	1=Male, 0=Female	Binary
Cp	cp: chest pain type 0: asymptomatic 1: atypical angina 2: non-anginal pain 3: typical angina	Categorical
Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	Continuous
Chol	Serum cholesterol in mg/dl	Continuous
Fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	Binary
Restecg	Resting electrocardiographic results 0: showing probable or definite left ventricular hypertrophy by Estes' criteria; 1: normal; 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)	Categorical
Thalach	Maximum heart rate achieved	Continuous
Exang	Exercise induced angina (1 = yes; 0 = no)	Binary
Oldpeak	ST depression induced by exercise relative to rest	Continuous
Slope	The slope of the peak exercise ST segment 0: downsloping; 1: flat; 2: upsloping	Categorical
Ca	number of major vessels (0-3) colored by fluoroscopy *(4 missing values)	Integer
Thal	Thallium stress test result 1 = fixed defect; 2 = normal; 3 = reversable defect) *(2 missing values)	Categorical
Target	Presence of heart disease: 0 = disease, 1 = no disease	Binary

3.3 Exploratory Analysis of Cleveland Dataset

The exploration analysis started with visualising the class distribution, assessing the class imbalance, and examining the correlation between variables. As can be seen from Figure 3, about 160 patients have a class label of ‘1’, indicating they are healthy, and 136 patients are associated with a class label of ‘0’, signifying they are

prone to heart disease.

The dataset is balanced, which means the model built will not be biased towards decisions favouring one disease class. Further, by looking at the correlation matrix, it can be ascertained that there is no redundancy in attributes or variables (if the variables are highly correlated, they indicate high redundancy, and need to be eliminated from the model building). The correlation matrix shown in Figure 3 does not indicate a high correlation between any of the attributes, and all the attributes may be contributing differently towards disease characterisation, hence need to be included in model development, and cannot be eliminated.

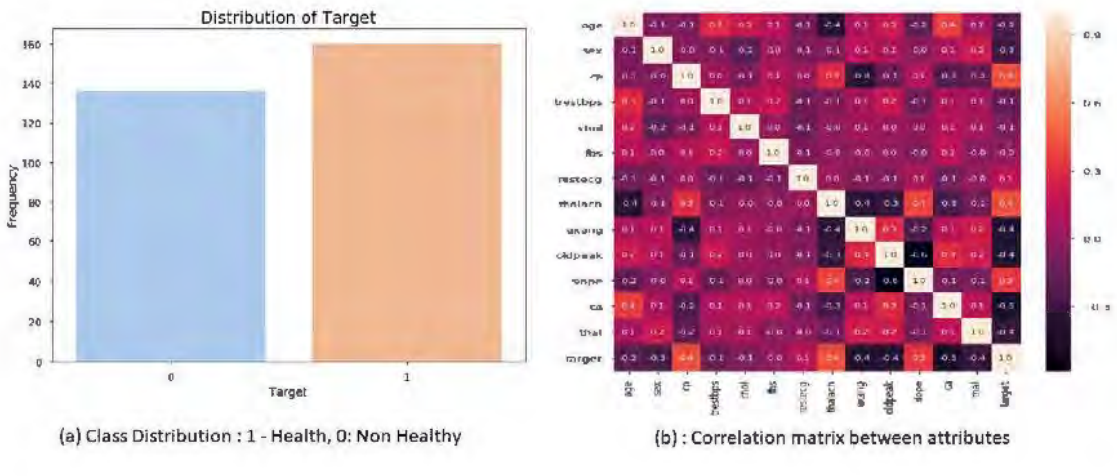


Figure 3: Cleveland Dataset: Class Distribution and Correlation Matrix

The next important aspect of exploratory analysis is to examine the relationship between the classes and different attributes, which is as shown in Figure 4. As can be seen in Figure 4, depending on the prevalence of the disease in the dataset (160 healthy and 136 unhealthy), different attributes (continuous and categorical) affect the disease class differently. For instance, around 80% unhealthy are from the male gender cohort, and 20% unhealthy are from the female cohort. Further, it can be noted that most of the unhealthy patients have asymptomatic anginal chest pain and prevalence of few other indicators, as it appears in the relationship plot

between disease class and continuous attributes. Hence, it can be concluded that the people with a diagnosis of heart disease in Cleveland dataset, are mostly older males, with high blood pressure and high cholesterol as compared to those who are not diagnosed with a heart ailment. The exploratory analysis presented in this section can give a good understanding of the data in the dataset, which is important before a model building task based on any ML/AI algorithm is done. The next Section describes the model building step based on several types of traditional shallow machine learning algorithms.

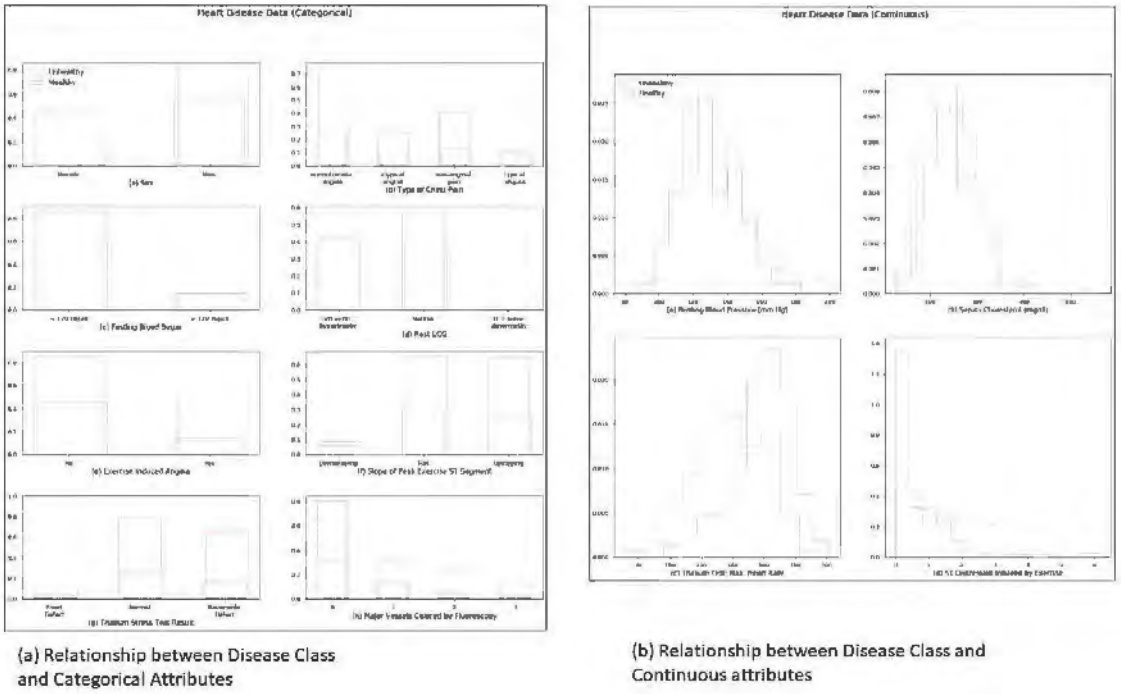


Figure 4: RELATIONSHIP BETWEEN DISEASE CLASS AND ATTRIBUTES

3.4 Model Development Based on Shallow Machine Learning

In this section, details of model development based on diverse types of traditional shallow machine learning algorithm used in prior work for analysing Cleveland dataset is presented, before discussing the performance evaluation of the models in terms of their different classification metrics.

For developing the disease detection models based on shallow machine learning algorithms, three different approaches used in previous work for Cleveland dataset analysis were used, namely the naïve Bayes (NB), the logistic regression (LR) and the decision tree (DT) algorithm.

The Naive Bayes classifier uses the Bayes' theorem of conditional probabilities, and builds the model, by calculating the probability for a class depending on the value of the feature over all the features. The naïve term is used to represent the assumption of independence of the features. For Naïve Bayes Modelling, the conditional probability of a class C_k is obtained as:

$$P(C_k|x) = \frac{1}{Z} P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \quad (1)$$

Here, Z represents a scaling parameter, which ensures the sum of all class probabilities is 1. The conditional probability of a class can be defined as the class probability multiplied by the probability of each feature given the class, normalized by scaling parameter Z. Naive Bayes is inherently an interpretable model because of the independence assumption, and it is easy to explain how each feature contributes towards a certain class prediction, providing a clear interpretation of the conditional probability. For Naïve Bayes modelling, a mixed naïve Bayes approach was used, as it allows the attributes to be considered by their distribution. This means the categorical attributes would be considered with a categorical distribution and continuous attributes from the normalised Gaussian distribution.

Logistic regression is an extension of the linear regression model and applied for classification problems. In logistic regression modelling, a logistic function tries to squeeze the output of a linear equation between 0 and 1. The logistic function is

defined as:

$$\log(\eta) = \frac{1}{1 + \exp(-\eta)} \quad (2)$$

Logistic Regression is used as a linear model for classification rather than regression, despite its name. It is also known as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier, and in this model, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function. By using different parameters for the mode, the model can be optimized and regularized. For solving the overfitting problems in machine learning model building, a regularization technique can be used, involving the selection of a linear and C parameter, indicating the reverse regularization strength. Different values were set for C and the selected one was 0.5 for this work.

The performance of naïve Bayes and logistic regression models deteriorates, if the relationship between the attributes and class is nonlinear or the independence assumption does not hold good, and the attributes are interacting with each other. An alternate modelling approach under these scenarios is to use decision trees. The decision trees use a measure such as information gain for splitting the data several times, by evaluating how much information is gained by each split. The splitting process involves different subsets to be created, called leaf nodes, and actual prediction occurs on the leaf nodes, and the average outcome of the training data subset is the prediction outcome. The structure of the tree depends on different algorithms for growing the decision tree. The relation between an outcome y and features x can be mathematically described as:

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^{M} C_m I\{x \in R_m\} \quad (3)$$

where, Rm represents the leaf node subset and $I\{x \in Rm\}$ is the identity function that returns 1 if the instance belongs to that subset or 0 otherwise. The predicted outcome will be $\hat{y}=c_l$, if an instance falls into a leaf node Rl, where c_l represents the average of all training instances in Rl. Interpreting the reasoning being decision trees is quite straightforward. Following the edges of the tree from the root node to the leaf node explains how a certain prediction decision is made. However, this is the case when the decision tree is short. If the tree is deeper, it is harder to interpret the decision rules of the tree. The deeper the tree, the harder it gets to understand the decision rules of the tree.

Figure 5 shows the confusion matrix obtained for the disease prediction model for three different types of shallow machine learning approaches considered in this study.

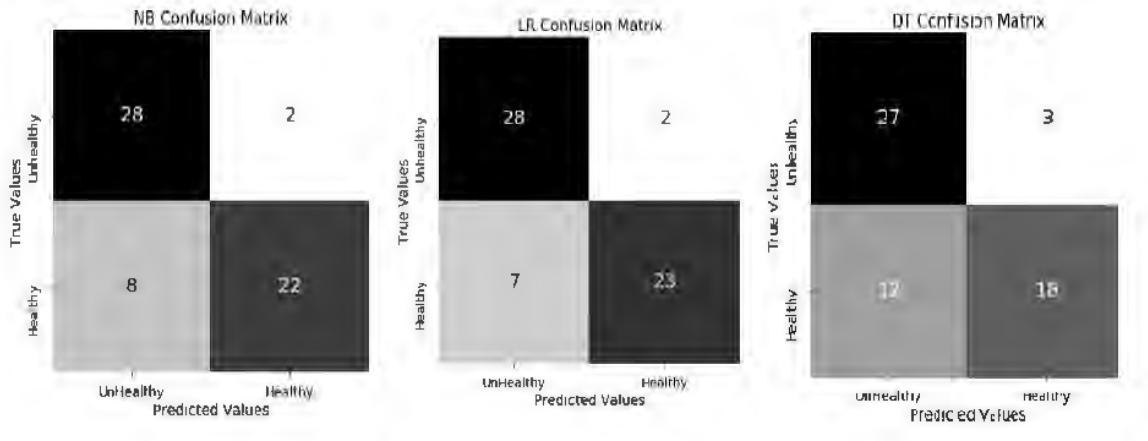


Figure 5 : Confusion Matrix for NB, LR and DT Models

As can be seen from the confusion matrix for the test set shown in Figure 5, the NB and LR models work well as compared to the DT model. Though the True positive rates for each of the models are similar, the false negative rates are poor

for these models (DT model in particular). Other performance measures such as ROC-AUC curves and precision, recall (sensitivity), and F-1 score are as in Figure 6 and as performance matrix in Table II.

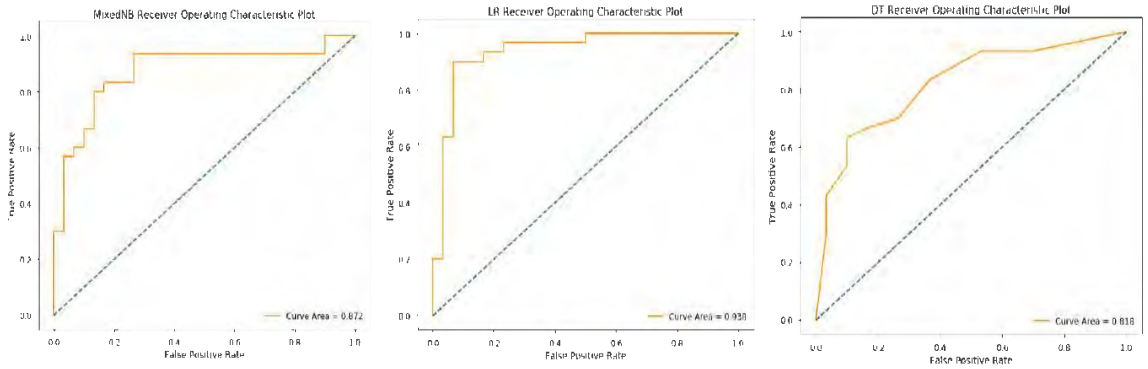


Figure 6 : ROC-AUC curves for NB, LR and DT Models

TABLE II. PERFORMANCE MATRIX FOR NB, LR AND DT MODELS

		Precision		Recall		F1-Score		
Model	Accuracy	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1	AUC
NB	0.83	0.78	0.92	0.93	0.73	0.85	0.81	0.87
LR	0.85	0.80	0.92	0.93	0.77	0.86	0.84	0.94
DT	0.75	0.69	0.86	0.90	0.60	0.78	0.71	0.81

As can be seen in Table II, for the disease model developed using Naïve Bayes approach, the prediction accuracy was 83%, and the precision obtained was between 0.78 and 0.92 for classes (0= unhealthy) and (1=healthy) respectively, and recall was 0.93 and 0.73. Further, the F-1 score, which measures how perfect the precision and recall was 0.85 and 0.81 for the two classes ((0= unhealthy) / (1=healthy)).

Logistic regression performed similarly to NB, with an accuracy of 85%, and precision of 0.8 for class 0 and 0.92 for class 1, and a recall of 0.93 for class 0 and 0.77 for class 1. Also, F1-scores for each class were 0.86 and 0.84, and an AUC of 0.94.

For decision trees, though the training set performance matrix was good with 84% prediction accuracy with 10-fold cross-validation, the test set performance dropped to 75%

decision tree suffered from the problem of overfitting and the accuracy dropped to 0.75 on the test set. It seems there is an over-fitting happening for DT model, and the other performance metrics as seen in the performance matrix in Table II, reflects similar poor performance due to over fitting.

The performance matrix shown in Table II, indicates that all three shallow learning models considered for this study performed similarly in terms of prediction accuracy, although the decision tree model performance was poor due to the overfitting problem. The logistic regression model is best out of the three, for all metrics shown in the performance matrix in Table II, particularly achieving a prediction accuracy of 85% and AUC of 94%.

3.5 Model Interpretability and Explainability

In this section, we discuss the interpretability and explainability of the best performing model - the logistic regression model, out of three shallow learning algorithms examined for building CVD disease detection models on Cleveland dataset.

For this, the foremost consideration is to examine the impact of different attributes on the model, by calculating the feature importance. This will help in understanding which feature is influential in determining prediction model performance. One of the strategies that can be used for assessing the feature importance is by calculating permutation importance. Permutation feature importance is a model inspection technique that can be used for any fitted estimator when the data is tabular. The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature

[46]. The permutation feature importance measure for the variables for the LR model built is as shown in Figure 7. Table 7 shows the description of each attribute for feature importance interpretation.

Weight	Feature
0.0767 ± 0.0806	ca
0.0333 ± 0.0365	thalach
0.0267 ± 0.0267	cp_2
0.0267 ± 0.0618	oldpeak
0.0200 ± 0.0249	cp_3
0.0100 ± 0.0163	trestbps
0.0067 ± 0.0163	fbs
0.0067 ± 0.0340	slope_2
0.0033 ± 0.0133	slope_1
0 ± 0.0000	restecg_2
-0.0000 ± 0.0211	exang
-0.0033 ± 0.0533	thal_3
-0.0067 ± 0.0163	cp_1
-0.0067 ± 0.0267	chol
-0.0067 ± 0.0267	age
-0.0100 ± 0.0400	sex
-0.0100 ± 0.0340	restecg_1
-0.0167 ± 0.0211	thal_2

Figure 7 : Feature Permutation Importance

Figure 7 shows how the accuracy LR model gets impacted by shuffling different features. Green or positive values represent the most important features and change in accuracy when the features are shuffled, randomly. The randomness of multiple shuffles gets captured after the ± symbol. For example, *ca* or a number of major vessels (0-3) coloured by fluoroscopy is the most important feature for this model and shuffling it would impact the model accuracy by around 8%. On the contrary, negative values represent an increase in prediction accuracy, as the features are

shuffled. This impact could be due to either random chance or because of the small size of the dataset.

TABLE III. ATTRIBUTE DESCRIPTION FOR FEATURE IMPORTANCE STUDY

Attribute no.	Attribute name	Attribute Description
1.	Age	Age in years
2.	Sex	gender (1 = male; 0 = female)
3.	cp	Chest pain type -- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic
4.	trestbps	resting blood pressure (in mm Hg on admission to the hospital)
5.	chol	serum cholestoral in mg/dl
6.	fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7.	restecg	resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8.	thalach	maximum heart rate achieved
9.	exang	exercise induced angina (1 = yes; 0 = no)
10.	oldpeak	ST depression induced by exercise relative to rest
11.	slope	slope: the slope of the peak exercise ST segment -- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping
12.	Ca	number of major vessels (0-3) colored by

Another measure to assess the model interpretability and explainability is called

the **LIME measure**. The LIME - Local Interpretable Model-agnostic Explanations (LIME) measure was proposed by Ribeiro et al [47] and aims to provide evidence for users to trust machine learning models. By using the LIME measure, it is possible to interpret and explain the predictions made by the model and gain trust in the decision support provided by the algorithm. The LIME measure is obtained as follows:

(i) Sample instances are close and far from the interpretable representation of the original input.

(ii) Calculate the prediction of these instances from their interpretable representation and builds a weighted linear model by minimizing the loss and complexity.

The samples are weighted based on the proximity from their original point and the weights decrease as the distances increase. Figure 8 presents the intuition, with a toy example on the details of how LIME measure can be calculated [47]. The black-box model's complex decision function *f (*unknown to LIME*)* is represented by the blue/pink background, which cannot be approximated well be a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by the size). The dashed line is the learned explanation that is locally (but not globally) trustworthy or faithful.

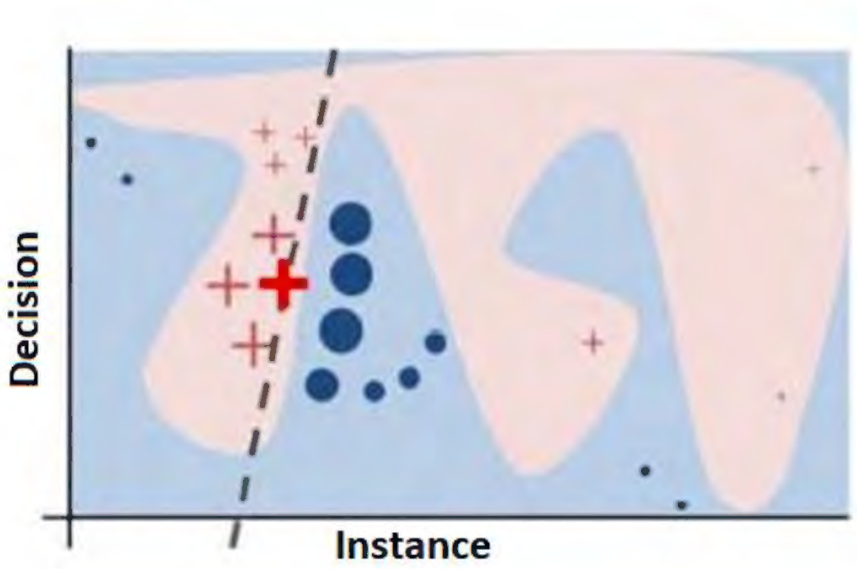


Figure 8: Interpretability/Explainability with LIME [47]

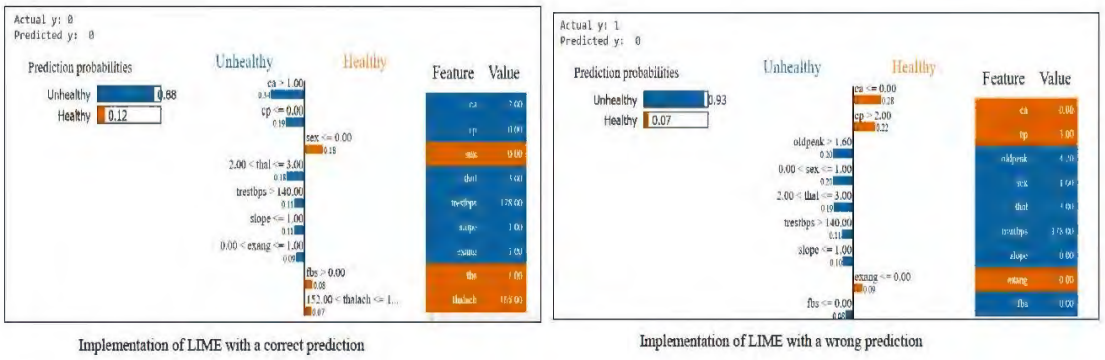


Figure 9: LIME Visualisation for DT model

Figure 9 shows how decision tree model interpretability and explanations can be visualised with LIME approach and depict how the model weighted its prediction heavily on feature values of *ca=2, cp=0, and thal=3* for a correct prediction. In contrast, for incorrect prediction or misclassification, the model explains the reason behind incorrect prediction by highlighting the feature values *for oldpeak, sex, thal, and trestbps* features, as compared to the weights of *ca* and *cp*. Visualisation of the model with LIME can give the user, doctor or medical professional whether it is indeed making the decision correctly and whether it can be trusted.

3.6 Chapter Summary

In this Chapter preliminary studies to examine the disease detection models based on shallow machine learning were conducted. Three different types of shallow machine learning algorithms, the naïve Bayes (NB), logistic regression (LR) and decision tree (DT) were used for developing disease detection models using Cleveland database. The models built were assessed in terms of different performance metrics used for classification models, and their interpretability/explainability was examined with two different measures, the feature permutation importance and LIME map. This preliminary study helped in setting a baseline reference for performance comparison in terms of prediction accuracy metrics as well as interpretability/explainability metrics.

The study has shown that the shallow machine learning models based on decision trees are better in terms of interpretability and explainability, though they may not have excellent prediction performance. Another important observation from this study, once the model performance is acceptable in terms of detection and prediction accuracies, it can be augmented and supplemented with appropriate processing stages to provide better interpretations and explanations and can be made trustworthy for inclusion in clinical decision-making workflow. While a small database (Cleveland database) was used in this Chapter for building disease detection models, the focus of the next Chapter is to extend this study to larger data sets, and aim to enhance the prediction performance, which is more challenging, as the data size becomes larger.

Chapter 4: Smart Predictive Modelling Framework

4.1 Introduction

This Chapter provides details of a generic and unified computational framework proposed called "*Smart Predictive Modelling Framework*", for building white box models for disease prediction based on shallow machine learning algorithms discussed in the previous chapter, and some new variants based on decision trees.

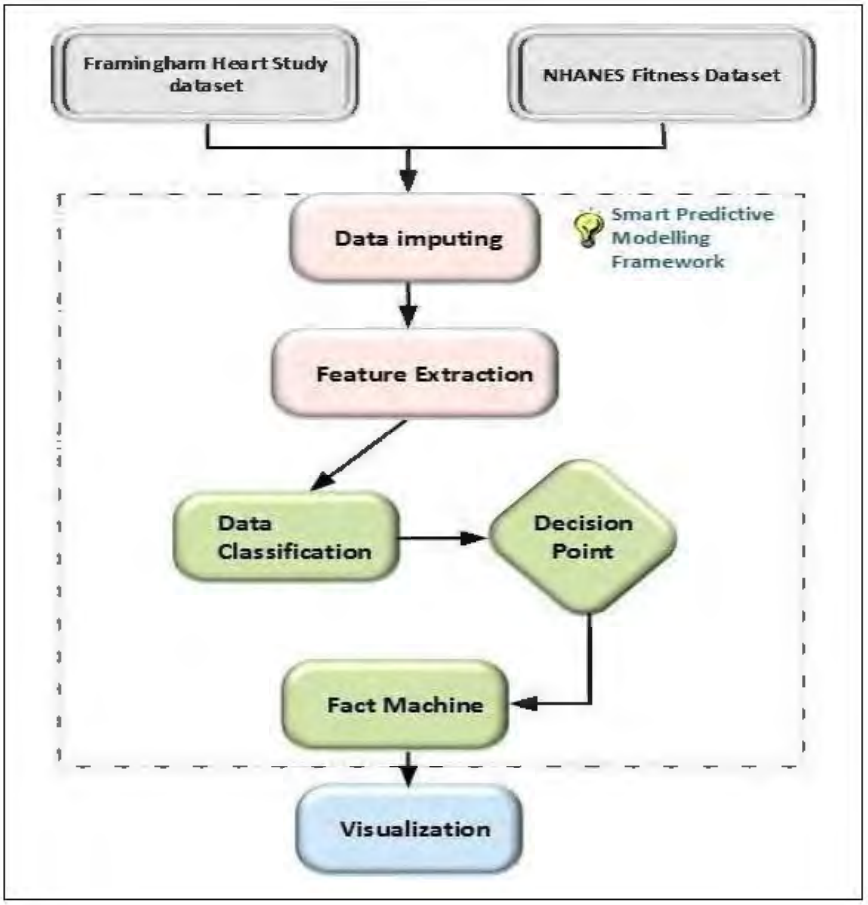


Figure 10 : Smart Predictive Modelling Framework

The details of the datasets used in this Chapter are discussed first in the next Section, before delving into the methods used.

4.2 Description of Datasets

Two different real-world and publicly available datasets were used for building the disease prediction models:

(1) The "*NHANES Physical Activity and Cardiovascular Fitness Data*" from National

Centre for Health Statistics in Centre for Disease Control [11], and

(2) The "*Framingham Heart Study dataset*" [10].

NHANES datasets contain physical activity and cardiovascular fitness data collected through the physical activity questionnaire, the cardiovascular fitness examination, and the physical activity monitor examination. These components were combined for analysis. Also, different demographic cohorts and population categories were considered separately, with analysis of each component done individually to assess health outcomes. The inclusion of objective measures from the cardiovascular fitness and physical activity monitor examinations provides a unique approach to characterize physical activity performance and estimating cardiovascular fitness [13]. The dataset we used in this chapter was a combination of two files "Cardiovascular Fitness_CVX_B- Model" and "Physical Activity_PAQ_B" from NHANES dataset. Both these files contain 65 and 21 variables respectively which indicate the cardiovascular status. Different data cleaning and pre-processing techniques were applied to merge these files to make a resultant dataset, suitable for building the disease detection models based on different shallow AL/ML algorithms.

The second dataset i.e., "Framingham Heart Study dataset" was sourced from a publicly available section of Framingham Heart institute dataset. The Framingham Heart Study is a long-term prospective study of the aetiology of cardiovascular disease among a population of free-living subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology, in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. This dataset

contains 22 different parameters which affect the cardiovascular health. The publicly available section of the dataset that was used for this study is a subset of the data collected as part of the Framingham study and includes laboratory, clinic, questionnaire, and adjudicated event data on 4,434 participants. Participant's clinical data was collected during three examination periods, approximately 6 years apart.

Each of these datasets was subdivided into multiple subsets based on different cohorts, such as education level, gender and age to achieve the in-depth analysis. As the dimensionality of features was very high (22-65 features), as compared to the data available, leading to a sparse data situation, and adversely affecting the performance of the prediction engine, we used a feature selection technique to select the most significant features, based on an information theory-based ranking of features, as an additional pre-processing step. In this feature selection technique, the features or attributes were ranked using information gain as the criterion, and other insignificant features were discarded. This has worked surprisingly well as compared to other attribute selection methods, given that in this application context, one is dealing with a very high dimensional database, and where there is a need to use around half the features of the database to achieve the same level of performance and accuracy.

4.2.1 Data acquisition and pre-processing

The following steps were taken for data preparation before model creation.

- Outlier detection and elimination were done with the help of the Grubb's test. This method is beneficial as it considers complete dataset values for outlier detection and elimination.
- Missing data were replaced with the modes and the mean.

- Scaling of data was performed by using mean and standard deviation
- F-Score feature selection method was used to select the most important and related features from the datasets. F-Score is a simple method that distinguishes the two classes with real values.

4.3 Experimental Results Using Supervised Learning

Different sets of experiments were performed to validate the analytics and modelling algorithms for the prediction engine for the proposed unified predictive modelling framework. To ensure the validity of results, different publicly available benchmark databases were used, and a standard protocol was developed. The performance of supervised machine learning models is described in the next few sections. Also, instead of sequence-based learning or tracking-based learning models such as HMM (Hidden Markov Models) and Condition Random Fields (CRF), which are traditional models for complex data sets, light weight interpretable machine learning models based on decision trees and its variants were used.

4.3.1 Experimental Results for National Health and Nutrition Examination Survey (NHANES) dataset

NHANES dataset focused on oversampling many groups within the U.S. population aged 2 months and over. Each record in the dataset consists of several attributes, such as high blood pressure, Body mass index, Pain in the right arm, Pain in the left chest and activity hours etc. A 65-features vector and the risk of having a cardiovascular disease label, as an identifier (disease class label) was extracted for building the models for each subject in

the training phase. We used a 5-fold cross validation for dividing the training and test subset. Table III shows the performance matrix for each of the shallow learning algorithms, in terms of prediction accuracy and the time it takes for learning/model building. For instance, for the simple decision tree classifier (J48) the accuracy for disease prediction is ~97.6% and model building time is around 56.74 seconds.

TABLE III. NHANES PERFORMANCE COMPARISON W.R.T. MODEL BUILDING TIME

Algorithms	***Time Taken (Sec.)***	***Accuracy %***
DECISION TREE (J48)	56.74	97.6
BAGGING	3.5	96.5
KNN	2.7	80.8
LOG. REG. (GLM)	1.71	96.4
NAIVE BAYES	7.2	95.7
RANDOM FOREST	214	98.5
SVM	1.8	95.4

Further, the performance of the kernel support vector machine (KSVM) based model was better, in terms of both prediction accuracy, and model building time, with 95% accuracy and 1.80 seconds model building time. However, the random forests, one of the popular ensembles learning approaches, although better in terms of prediction accuracy, with 98.5% accuracy, has a large model building time, the maximum in the list is 214 seconds. The best performer is the logistic regression classifier (GLM); with a prediction accuracy of 96.4% and model building time is 1.71 seconds. In comparison, surprisingly, K-nearest Neighbours classifier performs poorly with ~81% accuracy; however, its model building time is the third best in the list. Three classifiers namely, ensemble (Bagging), Logistic regression (GLM) and Support vector machine performed

better with a model creation time of less than 4 seconds and an accuracy percentage more than 95%.

4.3.2 Results for Framingham heart study dataset

The Framingham Heart Study (FHS) is a long-term prospective study of the aetiology of cardiovascular disease among a population of free-living subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology. It was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects [14]. The core research of the Heart Study has focused on many aspects of cardiovascular and cerebrovascular disease. Access to this data publicly, for conducting a wide range of investigations , has enabled research discoveries related to other diseases such as stroke, osteoporosis, arthritis, obesity, eye, and hearing disorders, and cancer. It contains 22 features vector and has a cardiovascular disease label, and an identifier (disease class label) for each subject. Experimental results with FHS dataset are shown in Table IV.

As is evident from Table IV, the Logistic regression classifier accomplishes good accuracy for this dataset as well i.e. with 90% prediction accuracy, and model building time of

16.17 seconds. Further, the decision tree achieved the same level of accuracy as Logistic regression, whereas RPART algorithm perform better than both algorithms in terms of model creation time, but its accuracy fell marginally behind with ~89%.Support Vector Machine classifier showed the most impressive accuracy of 90.2%, but it is overshadowed by the long model building time i.e. 468 seconds. Also, random forests performed worst in terms of model creation time which is 2180 seconds with ~90% accuracy. Further, K-Nearest neighbour also

performed well with ~91% accuracy and consumed only ~89 seconds for data model creation. Three classifiers namely, ensemble (RPART), Logistic regression and Naïve Bayes performed better with a model creation time of less than 30 seconds and an accuracy percentage of more than 89%.

TABLE IV. FHS PERFORMANCE COMPARISON W.R.T. TIME TAKEN

Algorithms	***Time Taken (Sec.)***	***Accuracy %***
KNN	88.8	90.1
RANDOM FOREST	2180	90.1
SVM	468	90.2
NAIVE BAYES	28.9	89.9
DECISION TREE (J48)	77.4	90
LOG. REG. (GLM)	16.17	90
ENSEMBLE(RPART)	15.1	89.3

For understanding how the prediction models perform for different cohorts or subsets of the dataset, we examined model performance, with different cohorts of data (Filtered data models) based on age, gender and education level cohort segmentation. The performance improvement achieved for the supervised learning approach based on shallow machine learning models for this age filtered data group is shown in Table V, with 3 cohorts: Senior (above 59 years), Medium (between 49- 59 years) and Young (below 49 years).

TABLE V. FHS PERFORMANCE COMPARISON W.R.T. MOBEL BUILDING TIME (AGE SEGMENTATION)

	Algorithms	***Time Taken (Sec.)***	***Accuracy %***	
Young	KNN	24.21	90.8	
	RANDOM FOREST	607.2	90.8	
	SVM	144	91.6	
	NAIVE BAYES	9.78	90.7	
	DECISION TREE (J48)	41.22	90.9	
	LOG. REG. (GLM)	3.78	91	
	ENSEMBLE(RPART)	3.3	89.9	
Medium	KNN	26.5	92.08	
	RANDOM FOREST	606.6	92.11	
	SVM	54.23	92.03	
	NAIVE BAYES	8.8	92.05	
	DECISION TREE (J48)	47.7	92.08	
	LOG. REG. (GLM)	3.2	91.8	
	ENSEMBLE(RPART)	3.1	91.4	
Senior	KNN	27.5	87.5	
	RANDOM FOREST	10.48	88.6	
	SVM	59.11	89	
	NAIVE BAYES	8.8	88	
	DECISION TREE (J48)	41.13	87.9	
	LOG. REG. (GLM)	2.9	87.5	
	ENSEMBLE(RPART)	3.35	86.5	

Table VI shows prediction performance with gender filtered models unfiltered data models.

	Algorithms	***Time Taken (Sec.)***	***Accuracy %***
Female	KNN	72	94.9
	RANDOM FOREST	1113	94.7
	SVM	78	94.8
	NAIVE BAYES	12.3	94.5
	DECISION TREE (J48)	52.8	94.7
	LOG. REG. (GLM)	4.1	94.2
	ENSEMBLE(RPART)	3.9	94.01
Male	KNN	29.55	89.9
	RANDOM FOREST	672	89.5
	SVM	126	88
	NAIVE BAYES	12.11	88.1
	DECISION TREE (J48)	51.8	87.1
	LOG. REG. (GLM)	3.4	86.9
	ENSEMBLE(RPART)	3.3	85.3

TABLE VII. FHS PERFORMANCE IMPROVEMENT (AGE FILTER/ SEGMENTATION) OVER UNSEGMENTED AGE GROUP

Algorithms	***Accuracy improvement %***
KNN	0.03
RANDOM FOREST	0.45
SVM	0.85
NAIVE BAYES	0.41
DECISION TREE (J48)	0.33
LOG. REG. (GLM)	0.11
ENSEMBLE(RPART)	0.04

TABLE VIII. FHS PERFORMANCE IMPROVEMENT (GENDER FILTER/ SEGMENTATION) OVER UNSEGMENTED GROUP

Algorithms	***Accuracy improvement %***
KNN	2.55
RANDOM FOREST	2.22
SVM	1.33
NAIVE BAYES	1.56
DECISION TREE (J48)	1
LOG. REG. (GLM)	0.61
ENSEMBLE(RPART)	0.4

The age level segmentation involved cohorts corresponding to Senior (above 59 years), Medium (between 49- 59) and Young (below 49) as shown in Table V. As can be seen from Table IV and Table V, there is marginal improvement in the overall % accuracy for each model except ensemble (RPART).

Table VI shows the performance achieved based on gender level segmentation, and Table VII and Table VIII shows the performance improvement achieved for each segmented cohorts, in comparison with unsegmented cohorts. As can be seen from results in Table IV to VIII, all the models based on shallow machine learning algorithms showed positive improvement and in some cases accuracy improvement was significantly better, such as the k-nearest neighbour and random forest algorithms where improvement was achieved more than 2%. Other models were also created based on cohort segmentation with several other variables like education level. Within the education level field, four categories were used, namely the education level "Below School", "High School Student", "College" and "Vocational Degree". Table IX shows only two of the categories' results.

TABLE IX FHS PERFORMANCE IMPROVEMENT (GENDER FILTER/ SEGMENTATION) OVER UNSEGMENTED GROUP

	Algorithms	*Time Taken (Sec.)*	*Accuracy %*
Below School	KNN	19.7	91.9
	RANDOM FOREST	575	90.9
	SVM	103	91.3
	NAIVE BAYES	5.68	90.2
	DECISION TREE (J48)	29.9	90.1
	LOG. REG. (GLM)	3.2	91
	ENSEMBLE(RPART)	2.9	89.5
College	KNN	20.1	91.6
	RANDOM FOREST	418	90.9
	SVM	97.2	90.8
	NAIVE BAYES	4.32	91.1
	DECISION TREE (J48)	27.1	90.1
	LOG. REG. (GLM)	2.3	90.7
	ENSEMBLE(RPART)	2.1	89.6

As is apparent from Table V to Table IX, for these filtered / segmented, age, gender and education specific cohort models, Logistic regression and K-nearest neighbour performed well with around 91% accuracy and a decent model building timeframe, and turn out to be low resource models, consuming a low computational footprint. Also, it can be noted, that the improvement in accuracy was marginal, around ~1% in most of the cases for these cohort-specific models as compared to original unsegmented models. This shows that for challenging data settings, with appropriate cohort segmentation based on different demographic filters, it is possible to build disease prediction models based on simple and traditional shallow learning algorithms, which are inherently more interpretable/explainable (with feature importance and LIME metrics) as discussed in Chapter 3) and work well with smaller data sizes. In other words, these models based on decision trees and their variants turn out to be low-resource consuming, lightweight prediction algorithms, and perform well even with any data size (high and low), and performance deterioration for reduced sized data subsets is not very pronounced, even after filtering the cohorts based on age, gender, and education levels. In addition, these models are inherently more interpretable and explainable, eliciting trust in the decision made and can be more acceptable in embedding in the clinical workflow.

4.4 Experiments with unsupervised models

As it is difficult to obtain ground truth labels annotated by experts for large datasets, it will be worthwhile to explore, if the prediction engine, can work with unsupervised learning algorithms. To examine the performance of unsupervised

learning algorithms, we used three different unsupervised learning algorithms as shown in Table X. The benefit of unsupervised learning could be significant, as it requires no manual work for expert annotations and labels, for building the prediction models, as compared to the need for expert annotations required for supervised learning approaches. But the outcomes of experiments for both datasets were not quite promising, and we found that unsupervised learning-based models showed poor accuracy. As can be seen, in the Table X, the prediction accuracy for different unsupervised models is between ~56 % to ~88% whereas the supervised learning-based models showed much better prediction performance. Perhaps there is a need to develop better unsupervised learning models.

TABLE X. FHS PERFORMANCE WITH UNSUPERVISED SEGMENTATION MODELS

Algorithms	***Accuracy %***
Hierarchical Structure Model	88.6
Kmean Clustering Model	69.37
PAM Model	56.01

Figure 11 : Performance of Unsupervised Learning Predictive Models

4.5 Chapter Summary

In this chapter, a generic computational framework was developed for investigating different shallow learning models based on supervised and unsupervised machine learning and cohort segmentation and filtering based on age, gender, and education level. For the first dataset i.e. NHANES, optimal attribute selection techniques based on information theory-based ranking resulted in better performance. For Framingham heart study dataset, a combination of supervised and cohort segmentation using filtered variables based on different demographic attributes such as age, gender and education level was better. One of the important findings from this Chapter was, that the proposed

Smart Predictive Modelling Framework, and the core prediction engine, based on the concepts of segmenting big data sets into small subsets of data with appropriate demographics filters, and with models built with established traditional shallow learning algorithms, it is possible to achieve graceful performance. This can help in building fast AI/ML based prediction models, which are inherently more interpretable and explainable, and could be more trustworthy.

In the next Chapter, the proposed Smart Predictive Modelling Framework is extended for building white box disease detection models based on a novel algorithm based on XGBoost technique, considered to be an efficient, scalable, distributed gradient- boosted decision tree (GBDT) machine learning algorithm set, and provides a parallel tree boosting capabilities, suitable for regression, classification, and ranking problems.

Chapter 5: Cardiovascular Risk Prediction based on XGBoost

5.1 Introduction

This Chapter presents the details of extending the core prediction engine for the proposed smart predictive modelling framework, based on a powerful algorithm - the XGBoost technique. The details of the technique and its experimental evaluation is presented in the rest of the sections of this Chapter.

5.2 XGBoost Algorithm

Boosting refers to a class of learning algorithms, involving model fitting by combining many simpler models [48]. The beauty of this powerful algorithm lies in its scalability, which drives fast learning through parallel and distributed computing and offers efficient memory usage. These simpler models are typically referred to as base models and are learnt using a weak learner. Also, these simpler models tend to have limited predictive ability, but when selected carefully using a boosting algorithm, they form a relatively more accurate model. It is sometimes referred to as an ensemble model, as it can be viewed as an ensemble of weak models. XGBoost uses the craft of penalization of individual trees. The trees are consequently allowed to have a varying number of terminal nodes. XGBoost can shrink the leaf using penalization. The benefit of this approach is that the leaf weights are not all shrunk by the same factor, but leaf weights estimated using less evidence in the data will be shrunk more heavily. Also, the bias- variance trade-off happens during model fitting. XGBoost is thus a scalable machine learning algorithm for decision tree boosting that was originally proposed by [49]. This algorithm was used in more than 50% of winning solutions in several machine learning competitions during 2015. The superior performance of XGBoost in

supervised machine learning is the reason why it was chosen to train the classifier for this work. Gradient boosting is the original model of XGBoost, combining weak base learning models into a stronger learner in an iterative fashion.

The superior performance of XGBoost in supervised machine learning is the reason why it was chosen to train the classifier in this work. Gradient boosting is the original model of XGBoost, combining weak base learning models into a stronger learner in an iterative fashion. As shown in Figure 11 below, at each iteration of gradient boosting, the residual will be used to correct the previous predictor so that the specified loss function can be improved. As an enhancement, regularization is added to the loss function to establish the objective function in XGBoost measuring the model performance, which is given as

$$J(\Theta) = L(\Theta) + \Omega(\Theta) \qquad (1)$$

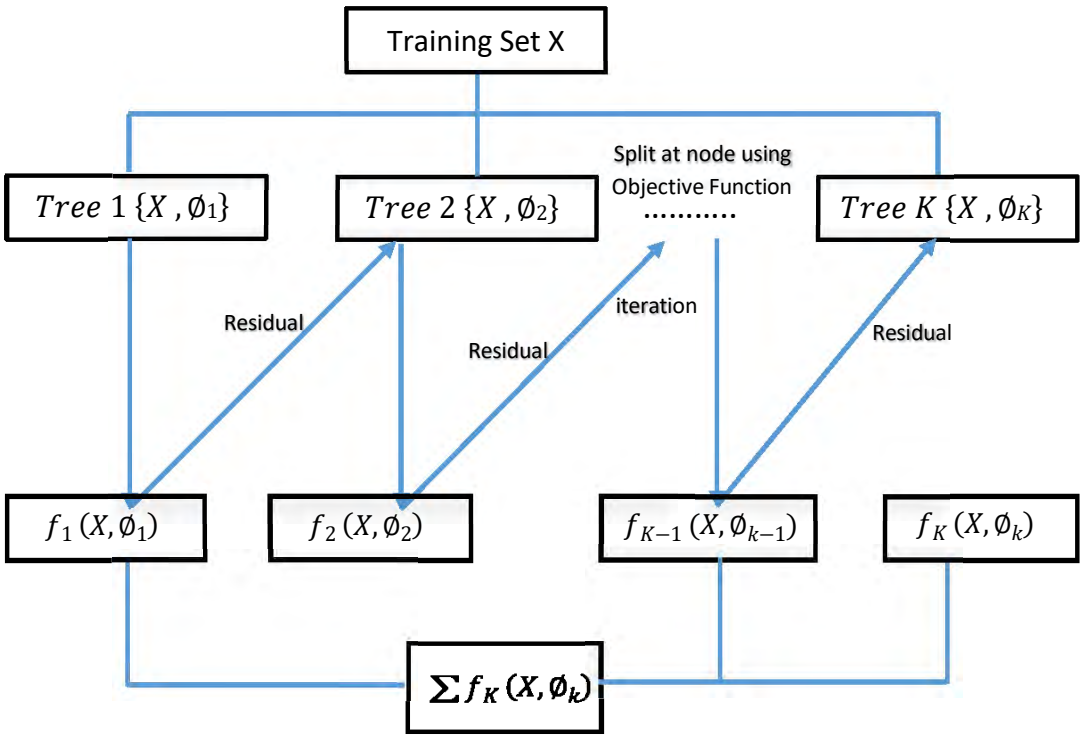


Figure 12 : Flow chart of extreme gradient boosting

The parameters trained from given data are denoted as Θ. L is the training loss function, such as square loss or logistic loss, which measures how well the model fits on training data. Ω is the regularization term, such as L1 norm or L2 norm, which measures the complexity of the model. Simpler models tend to have a better performance against overfitting. Since the base model is a decision tree, the output of model $y_i^{\wedge}$ is voted or averaged by a collection F of k trees:

$$\widehat{y_i} = \sum_{i=1}^{k} f_k\,(x_i), f_k \in F \tag{2}$$

Objective function at the t time iteration can be defined as:

$$J^{(t)} = \sum_{i=1}^{n} \overset{k=1}{L(y_i, \widehat{y_i})} + \sum_{k=1}^{t} \Omega\,(f_k) \tag{3}$$

Where the n is the number of predictions. Here the $\widehat{y_i^{(t)}}$ can be given as

$$\widehat{y_i^{(t)}} = \sum_{k=1}^{t} f_k\,(x_i) = \widehat{y_i^{(t-1)}} + f_t(x_i) \tag{4}$$

where,

γ is the complexity of each leaf.

T is the number of leaves in a decision tree.

υ is a parameter to scale the

penalty. w is the vector of

scores on leaves

Then, the second-order Taylor expansion, instead of first-order in general gradient boosting, is taken to the loss function in XGBoost. Assumed that the loss function is mean square error (MSE), the objective function can be finally derived as

$$J^{(t)} \approx \sum_{i=1}^{n}\left[g_i \bar{w}_{q(x_i)} + \frac{1}{2}\left(h_i \bar{w}^2_{q(x_i)}\right)\right] + \bar{\gamma T} + \frac{1}{2}\upsilon\sum_{j=1}^{t} w_j^2 \tag{5}$$

with the constants removed. Here the q(·) is a function that assigns data points to the corresponding leaf. gi and hi is the first and second derivative of MSE loss function. The loss function is determined by the sum of the loss value for each data sample. Because each data sample corresponds to only one leaf node, the loss function can also be expressed by the sum of loss value for each leaf node. Thus,

$$J^{(t)} \approx \sum_{j=1}^{T}\left[\left(\sum_{i\epsilon I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\epsilon I_j} h_i + \upsilon\right)\omega_j^2\right] + \gamma T \tag{6}$$

where I_j represents all the data samples in leaf node *j*. Hence, the optimization of the objective function can be transformed into a problem of finding the minimum

of a quadratic function. In other words, after a certain node is split in the decision tree, the change in model performance can be evaluated based on the objective function. If the decision tree model performance is improved after this node split, this change will be adopted, otherwise, the split will be stopped. Besides, because of the regularization when optimizing the objective function, a predictive classifier can be trained against

overfitting. In this work, the machine learning models were trained on python 3.6.7 with several scientific computing libraries, such as NumPy, TensorFlow and pandas, which provides efficient data structures and pre-processing methods.

5.2.1 Experimental Results for (NHANES) dataset

As mentioned in the previous section the XGBoost algorithm [49], uses a more regularised model formalisation to manage the over-fitting problem, which usually happens in classic gradient boosting regression trees (GBRT) and allows the performance and execution time to be improved. The weight for XGBoost is the sum of gradients scaled by the sum of Hessians. The comparison with traditional gradient boosting machine (GBM) can be outlined as follows:

$$w_{jm} = \begin{Bmatrix} -\frac{G_{jm}}{n_{jm}} & GBM \\ -\frac{G_{jm}}{H_{jm}} & XGBoost \end{Bmatrix} \tag{7}$$

As mentioned in the previous Chapter, the publicly available NHANES dataset was developed by focusing on oversampling many groups within the U.S. population aged 2 months and over. The same set of records comprising several attributes, such as high blood pressure, Body mass index, Pain in the right arm,

Pain in the left chest and activity hours etc., were used for the experimental evaluation of XGBoost algorithm. A feature vector of 65 features and a class label representing the risk of having cardiovascular disease (as an identifier (disease class label)) was used for building the disease prediction models.

TABLE XI. XGBOOST MODEL PERFORMANCE FOR NHANES DATASET

Algorithms	***Time Taken (Sec.)***	***Accuracy %***
DECISION TREE (J48)	56.74	97.6
BAGGING	3.5	96.5
KNN	2.7	80.8
LOG. REG. (GLM)	1.71	96.4
NAIVE BAYES	7.2	95.7
RANDOM FOREST	214	98.5
SVM	1.8	95.4
XGBoost	**1.6**	**97.7**

As can be seen in Table XI, the XGBoost algorithm performed well for NHANES dataset (97.6% prediction accuracy). However, it was marginally better than decision tree-based

model. By extending the experimental evaluation of the XGBoost based prediction engine for another publicly available dataset - the Framingham heart study dataset for validating the performance of the prediction engine, it is possible to have an insight into improvements possible.

As mentioned in the previous chapter, Framingham heart study (FHS) dataset contains 22 features vector and its cardiovascular disease label, an identifier (disease class label) of the subject. Experimental results with FHS dataset are shown in below Table XII. As it is evident from Table XII, the XGBoost algorithms results in similar prediction accuracy (aroung~90% prediction accuracy) and a

model run time of 15 seconds which is the fastest among all. Table XII shows the performance achieved with XGBoost algorithm for FHS dataset.

TABLE XII. XGBOOST MODEL PERFORMANCE FOR FHS DATASET

Algorithms	***Time Taken (Sec.)***	***Accuracy %***
KNN	88.8	90.1
RANDOM FOREST	2180	90.1
SVM	468	90.2
NAIVE BAYES	28.9	89.9
DECISION TREE (J48)	77.4	90
LOG. REG. (GLM)	16.17	90
ENSEMBLE(RPART)	15.1	89.3
XGBOOST	**15.0**	**89.9**

Further, several experiments were conducted based on filtered data models for different demographic cohorts, at the age, gender, and education level.

As can be seen in Table XII to XV, XGBoost algorithm has performed better overall and for different filtered data models for different demographic cohorts (age, gender and education level), with an improvement of accuracy of around ~ +/-1%, and promising improvement in model building time for filtered data models, with smaller data sizes. This makes the XGBoost models suitable for low resource sparse data settings for real time deployments due to their faster model building times. Further, being a variant of the decision tree, the XGBoost algorithm is inherently more interpretable and explainable and would be better accepted in clinical decision-making workflow.

TABLE XIII. XGBOOST MODEL PERFORMANCE FOR FHS DATASET (AGE FILTER/ SEGMENTATION)

	Algorithms	*Time Taken (Sec.)*	*Accuracy %*
Young	KNN	24.21	90.8
	RANDOM FOREST	607.2	90.8
	SVM	144	91.6
	NAIVE BAYES	9.78	90.7
	DECISION TREE (J48)	41.22	90.9
	LOG. REG. (GLM)	3.78	91
	ENSEMBLE(RPART)	3.3	89.9
	XGBoost	**1.4**	**92.2**
Medium	KNN	26.5	92.08
	RANDOM FOREST	606.6	92.11
	SVM	54.23	92.03
	NAIVE BAYES	8.8	92.05
	DECISION TREE (J48)	47.7	92.08
	LOG. REG. (GLM)	3.2	91.8
	ENSEMBLE(RPART)	3.1	91.4
	XGBoost	**2**	**91**
Senior	KNN	27.5	87.5
	RANDOM FOREST	10.48	88.6
	SVM	59.11	89
	NAIVE BAYES	8.8	88
	DECISION TREE (J48)	41.13	87.9
	LOG. REG. (GLM)	2.9	87.5
	ENSEMBLE(RPART)	3.35	86.5
	XGBoost	**1.3**	**88**

TABLE XIV. XGBOOST MODEL PERFORMANCE FOR FHS DATASET (GENDER FILTER/SEGMENTATION)

	Algorithms	*Time Taken (Sec.)*	*Accuracy %*
Female	KNN	72	94.9
	RANDOM FOREST	1113	94.7
	SVM	78	94.8
	NAIVE BAYES	12.3	94.5
	DECISION TREE (J48)	52.8	94.7
	LOG. REG. (GLM)	4.1	94.2
	ENSEMBLE(RPART)	3.9	94.01
	XGBoost	**2.2**	**94.8**
Male	KNN	29.55	89.9
	RANDOM FOREST	672	89.5
	SVM	126	88
	NAIVE BAYES	12.11	88.1
	DECISION TREE (J48)	51.8	87.1
	LOG. REG. (GLM)	3.4	86.9
	ENSEMBLE(RPART)	3.3	85.3
	XGBoost	**1.7**	**84.9**

TABLE XV. XGBOOST MODEL PERFORMANCE FOR FHS DATASET (EDUCATION LEVEL FILTER/SEGMENTATION)

	Algorithms	*Time Taken (Sec.)*	*Accuracy %*
Below School	KNN	19.7	91.9
	RANDOM FOREST	575	90.9
	SVM	103	91.3
	NAIVE BAYES	5.68	90.2
	DECISION TREE (J48)	29.9	90.1
	LOG. REG. (GLM)	3.2	91
	ENSEMBLE(RPART)	2.9	89.5
	XGBoost	1.8	89
College	KNN	20.1	91.6
	RANDOM FOREST	418	90.9
	SVM	97.2	90.8
	NAIVE BAYES	4.32	91.1
	DECISION TREE (J48)	27.1	90.1
	LOG. REG. (GLM)	2.3	90.7
	ENSEMBLE(RPART)	2.1	89.6
	XGBoost	**0.5**	**89.98**

5.3 Chapter Summary

In this Chapter, disease prediction models based on XGBoost algorithm was investigated for the proposed smart predictive modelling framework, and experimental evaluation on two different public health datasets (NHANES and FHS) showed its' merit as a white box model, along with improved prediction performance, low model building time, and with better interpretability/explainability, being a variant of decision tree algorithm.

In the Next Chapter, a new disease prediction model based on a deep machine learning algorithm that is robust to complex data settings involving imbalance in class distribution is presented for the prediction engine, and the performance on the same two benchmark publicly available datasets, NHANES and FHS datasets were examined.

Chapter 6: Deep Learning Based Decision Support Framework for Cardiovascular Disease Prediction

6.1 Introduction

In this Chapter, a new disease prediction model based on a deep machine learning algorithm is investigated to address the challenges of complex datasets with the imbalanced class distribution. While the models based on deep learning and advanced machine learning can lead to black-box models, they can enable robust models to be developed with improved prediction performance to be achieved under complex data settings.

Some of the challenges of complex health data include dependency on large number of risk factors, imbalance in the distribution of positive and negative samples, and poor performance of existing shallow machine learning based algorithms for disease prediction models developed using complex big data, making the deployment of accurate computer-based decision support tools in clinical settings and workflow difficult. The development of robust AI and machine learning based algorithms for building disease prediction models, to work under several constraints, and still can uncover the complex relationships between multiple risk factors and biomarkers, and approaches for building disease prediction models that can handle disorganised data settings, such as an imbalance in class distribution, insufficient annotations, and dependence on set of risk factors/biomarkers is the need of the hour. The next Section presents the relevant background necessary for building robust disease prediction models.

6.2 Background

Some of the prior related work on developing robust disease prediction models include, neural network algorithms based on multilayer perceptron (MLP) and

radial basis function (RBF) networks, and use of datasets with very small data sizes (around 1245 samples/subjects) for building the models. The conclusions from these research studies were that the MLPs were more efficient as compared to any other machine learning models, with an AUC_ROC performance metric of 0.78 achieved. Another set of approaches was based on optimal feature selection techniques, such as the hybrid forward selection technique, involving smaller subsets of data, along with smaller set of features, and a weak classifier, for detecting the presence of the CVD disease. Several other studies based on different types of neural network models have been reported in the literature for building robust disease prediction models, and include the methods based on ensemble models, neural networks, fuzzy logic models and deep learning approaches for detection and prediction of CVD status [50] [51]. However, for most of the previous studies, the patient cohort or sample size was relatively small, with limited risk factors dependencies, and inherent class imbalance, that exists in these data sets under real world clinical contexts, leading to poor performance of minority classes (often positive disease class). An ideal requirement for addressing this situation is the development of novel approaches and strategies for building robust disease detection models based on novel deep learning approaches, and validating the performance under complex data settings, particularly imbalanced data, a typical situation in clinical environments. Some of the strategies include, use of appropriate data augmentation or feature optimization and regression approaches, to address the shortcomings of disorganised, low quality large datasets. Several data augmentation and feature optimization approaches have been proposed in the literature for studies involving unstructured data, such as imaging datasets, particularly for segmentation tasks, where the data is homogenous,

but these approaches do not translate well to augmentation and/or feature optimisation strategies for structured clinical datasets, which consist of a combination of categorical and continuous variables.

In this Chapter, a novel deep learning algorithm based on stacked Dense-CNN cascade architecture, along with new model development and validation protocol is proposed for building CVD disease prediction model. The algorithm has shown better performance in addressing the complex relationship between multiple risk factors, and significant class imbalance in class distribution. The experimental evaluation of the proposed stacked Dense-CNN cascade neural architecture was done on a large imbalanced super-dataset - synthesized from a fusion of several subsets of NHANES datasets. The super-dataset comprised information corresponding to clinical, laboratory and examination data for patients obtained from multi-year studies of publicly available NHANES data. This innovative neural model uses a novel feature optimization and pre- processing stage, based on LASSO regression and regularization as the first pre- processing stage, for feature voting and elimination, and a subsequent model building stage involving a deep Dense-CNN cascade, with a novel data partitioning protocol for train and test phases, similar to the simulated-annealing type training and model building schedule. The performance of the disease prediction model based on this novel deep learning architecture, shows a significant improvement in robustness and prediction performance when compared to several other traditional shallow machine learning approaches, and can handle significant imbalance in the class distribution, allowing better minority class detection.

Rest of the chapter is organized as follows: Next Section presents the high-level overview of the proposed deep learning architecture, based on stacked Dense-CNN

cascade model, along with the feature learning and optimization pre-processing stage based on LASSO regression and regularization. Subsequent sections present the experimental setup for building the disease prediction model based on this algorithm and present the novel train-test data partitioning protocol for model training and testing. The experimental evaluation of the proposed disease prediction model, and results obtained is discussed in the last section.

6.3 Proposed Multi-stage Deep Learning Architecture

In this section, the details of the proposed multi-stage deep learning architecture, comprising the stacked Dense-CNN deep learning stage and pre-processing stage involving LASSO regression and regularization for optimal feature selection is presented. First, optimal feature learning stage is described, followed by the stacked Dense-CNN stage of the proposed architecture.

6.3.1 Optimal Feature Learning Stage

The front end for the optimal feature learning stage includes the LASSO (Lease Absolute Shrinkage and Selection Operator) regression and regularization stage. LASSO is a feature selection and optimization approach that can enhance the interpretability and accuracy of the machine learning models and involve a set of regression and regularization processing steps. The LASSO operator helps in selecting optimal variables/features and elimination of redundant parameters, by shrinking the data values towards a central point. LASSO performs well for model development with data comprising high multicollinearity, due to its ability to add a penalty, in terms of the absolute value of the magnitude of coefficients, leading to some coefficients reducing to zero and get eventually eliminated from the model. This leads to an optimal model, with fewer co-efficients, due to the redundant

feature elimination strategy. LASSO algorithms can be interpreted as the solutions to the quadratic optimization problem, with the goal minimization defined as sum of squares minimization with the constraint:

$$\sum_{i=N}^{n}\left(y_i - \sum_{j} x_{ij}\,\gamma_j\right)^2 + \lambda\sum_{j=1}^{p}\lfloor\gamma_j\rfloor \tag{8}$$

where, the γ factor here represents the importance assigned to a feature and models the underlying data variation contributions. With a value of zero for γ, it is considered unimportant. A regression model which is easier to interpret is obtained by having certain γ values of the model reduced to zero. Also, the strength of the regularization penalty is influenced by another factor, the tuning parameter λ, which represents the amount of shrinkage in parameters. No parameters get eliminated when λ is zero, and more coefficients get reduced to zero, and subsequently eliminated, with an increase in λ values. The model intercept, however, remains unchanged.

However, applying LASSO on the entire data for an unbalanced dataset can lead to misleading results, as it may trigger the incorrect selection of important variables. Hence, we perform LASSO processing on a subsample of the dataset. Since the dataset used in this work has mixed data types, with a subset of categorical variables, a variant of LASSO called the "Group LASSO" was implemented, though it is being referred to as LASSO in this Chapter for simplicity. In Group LASSO, the effect of imbalance is mitigated by adopting a strategy involving a random subsampling and iterative implementation of LASSO

multiple times. With majority voting performed on the set of γ factors, the features with nonzero γ factors are identified in the major number of iterations. For example, if we implement LASSO N times on N randomly subsampled datasets, with each random subsample set comprising balanced class distribution (equal number of examples with positive and negative CVD status), the selection of that feature for further analysis is done by manual thresholding as follows:

$$\chi(\gamma) = \begin{cases} 1, & if\ \gamma \neq 0 \\ 0, & otherwise \end{cases}$$

$$\chi(\gamma_{1,c})\,\chi(\gamma_{2,c}) \cdots \cdots \chi(\gamma_{N,c}) \geq \frac{N}{\alpha} \Longrightarrow \boldsymbol{c}\ is\ selected \tag{9}$$

6.3.2 Stacked Dense-CNN Cascade Stage

The aim of the stacked Dense-CNN cascade stage is two-class/binary classification and the prediction of the existence of CVD in patients. There are several research works reported in the literature on the use of neural networks for the task of binary classification as a supervised learning task [52] [53]. Also, for problems involving unstructured data (radiological image data for example), deep learning networks with a large number of application-specific hidden layers have shown to be promising, with significant performance improvements achieved in areas such as speech processing, medical image analysis, and computer vision [54]. Several deep learning architectures were proposed, for training with big datasets, involving successive training and transformation of input data over successive hidden layers, error estimation and backpropagation and iterative fine tuning of layer weights using gradient descent algorithm. There were also several methods proposed in improving the gradient descent algorithm, involving optimization of nonlinearity in the layers,

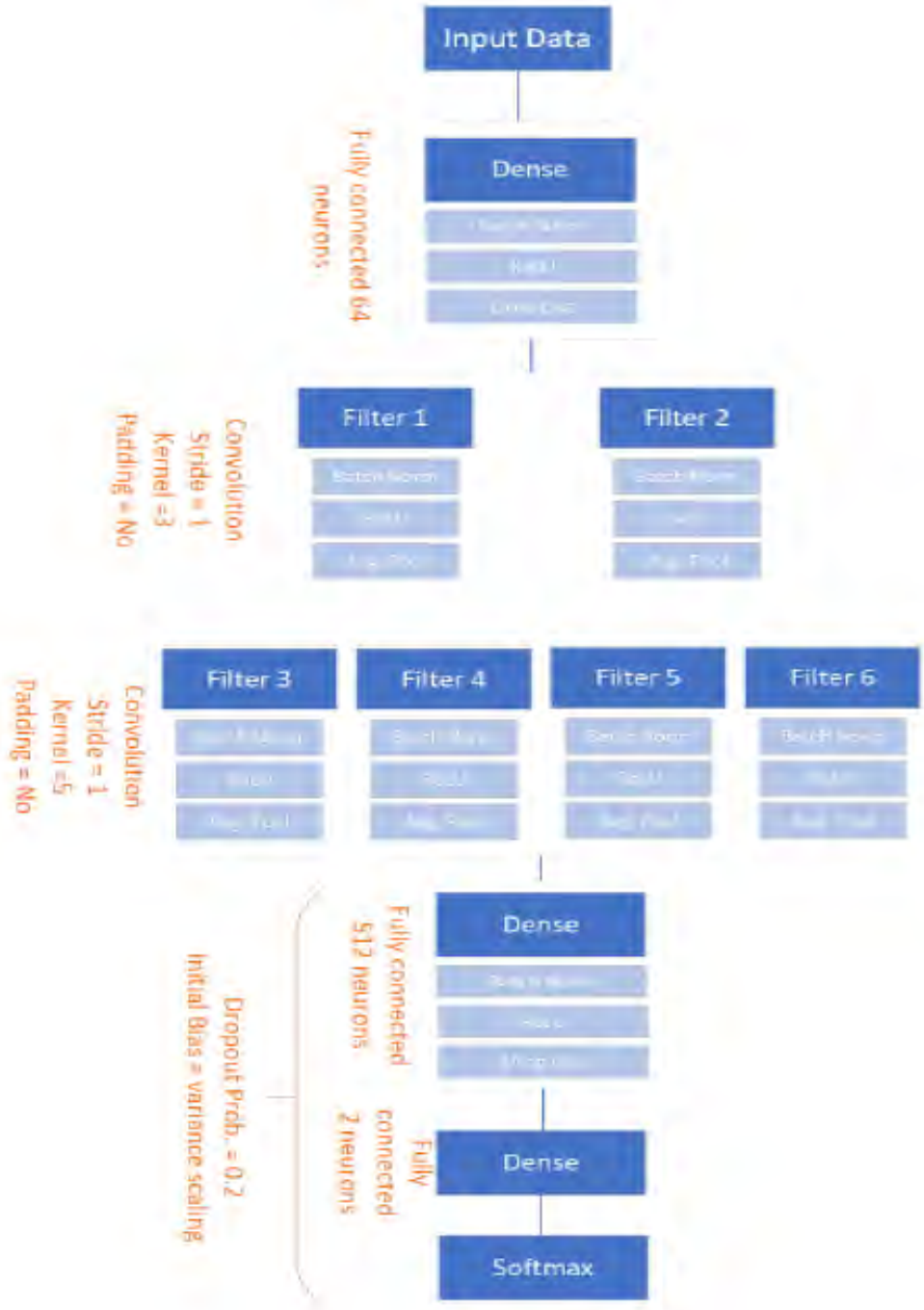
reduction of overfitting, schedule of training, and visualization of hidden layers, and other enhancements. Despite several enhancements proposed, the underlying principle of deep learning networks continues to be ill understood, seems like a black-box, and their performance could be vulnerable to adverse examples, overfitting and underfitting, especially in cases where the data is scarce or of poor quality. One of the popular approaches proposed in several works, particularly for medical image analysis problems is to address this challenge by data augmentation strategy, where artificially generated examples were used to increase the size of small datasets. This strategy could be biologically questionable for non-image type datasets with continuous and categorical datasets. For example, in the case of CVD datasets containing continuous and categorical variables, the blood platelet count is one of the features for prediction, and by using data augmentation here, the augmented samples may not quite correspond to the realistic readings for a patient, as the underlying principles of statistical synthesis used in data augmentation technique could be fundamentally different to the biological source of variations in the blood platelet count. Further, the imbalance in the augmented data distribution, and ineffective training architectures, could lead to unrealistic performance to be achieved, which is hard to elicit trust in clinical settings. While this practice may be popular in other computer vision and medical image analysis application settings, such as semantic labelling, disease tissue segmentation, and deep fake image synthesis use cases [34], in structured dataset settings

comprising categorical and continuous attributes, as in CVD datasets, the consequences of erroneous predictions, misclassifications and unrealistic classification performances could be adverse. For instance, incorrect prediction of a normal healthy patient as a CVD patient, could result in unnecessary medications

being prescribed, and could cause adverse drug reactions and side effects, and incorrect prediction of a CVD patient as a normal healthy patient could lead to irreversible damage and death, that could have been avoided if detected in time.

Therefore, the novel deep learning architecture proposed in this chapter addresses this shortcoming in prior reported work, by focusing on improving the performance of both classes, the majority class (normal healthy patients with no CVD), and the minority class (patients with CVD), not by relying on data augmentation technique, but by using a novel LASSO based feature learning strategy in the first stage, and a stacked Dense-CNN cascade model as the subsequent 2nd stage, as shown in Figure 12. The stacked Dense-CNN architecture, at 'Input' stage, consists of a one-dimensional feature vector, representing all the features obtained from LASSO regression and Majority Voting pre-processing stage. The next layer, the first fully connected 'Dense' layer, immediately after the 'Input', fuses/combines all the features, and a weighted combination of all these features is used as the output, comprising an identical mix of different feature types. The two convolution layers stacked in between dense layers try to learn different representations of the input, through a bank of different sized filters. The two subsequent fully connected 'Dense' layers after the CNN stack tend to aggregate the feature representations and extract higher level information and culminate in a 'Softmax' layer. This stacked architecture has a benefit, as the layers before the Softmax layer (last two Dense layers) can help implement a transfer learning strategy, by retraining them, for new data settings. The hyper parameters needed for training, including the parameters of convolution filters, dropout probability, ReLU activation function, number of neurons, and average pooling types are tuned iteratively as part of a tailored training schedule, as described next.

Figure 13 : BLOCK SCHEMATIC OF STACKED CNN MODEL



6.3.3 Training Schedule Strategy

The details of heterogenous feature vector extraction from NHANES super dataset, containing continuous and categorical features, and description of a training schedule used for the proposed stacked Dense-CNN model are presented here. For the sake of simplicity, we assume that the patient category with CVD disease is "1", and the subject without CVD is "0". By majority voting, we used around 50 CVD phenotypes for this study.

Let N be the number of training examples, and the dimensionality of input layer be *Nx50*.

Different types of variables are homogenized before nonlinear transformation by a dense or fully connected layer with 64 neurons acting collectively as a linear combiner of around 50 biases and variables. The non-linear transformation is performed by the rectified linear unit (ReLU) and dropout with 20% probability is performed for overfitting reduction. A stack of convolution layers was used after the fully connected layer, with the first convolutional layer, containing two filters with a kernel size of width 3 and stride

1. There was no external zero padding for this layer.

Also, after rigorous experimentation, it was decided to use average pooling, as it performed well under all constraints as compared to the max pooling. A $\mathscr{R}_N$ *X* 64 *X* 1 dimensional tensor was obtained in the first convolution layer from the output of the fully connected block with $\mathscr{R}_N$ *X* 64 dimensions.

In the first convolutional layer, there are two filters with a core width of 3, and a stride of 1. This layer does not provide external zero padding. At the grouping layer, we conducted rigorous experiments on different grouping strategies and found that under

all constraints, the average grouping performed slightly better than the maximum grouping. A tensor of dimension $\mathscr{R}_N$ X 64 X 1 is obtained by the first convolutional layer by converting the output of the fully connected block $\in$ $\mathscr{R}_N$ X 64. Then, we obtain an output tenso of $\mathscr{R}_N$ X 31 X 2 dimensions, by a series of processing steps comprising batch normalization, nonlinear transformation, and average grouping.

The filter of the last non-padded convolutional layer in the stacked CNN cascade, uses a kernel size of width 5 and stride 1, passing the output tensor from $\mathscr{R}_N$x13x4 to the next dense/fully connected layer after the average grouping layer. With a final SoftMax layer, the categorical output is obtained with categorical cross-entropy loss as the loss function. A random bias is initialized in each layer based on a truncated normal distribution with a variance of $1/\sqrt{n}$, with n as the number of "fan-in" connections to the layer. By using the Adam optimizer with a learning rate of 0.005, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and zero attenuation, the proposed stacked CNN cascade architecture consists of 32,642 trainable parameters and 1,164 non-trainable parameters. To achieve consistent accuracy across the entire class, extensive experimentation was done to tune the hyperparameters of the model, including the class weight, a number of epochs, sub- sampling of input data, the number of filters in each convolution layer, the number of neurons in each dense layer (excluding the last layer) during the training phase.

To address the class imbalance, the weighted class ratio – defined as the ratio of CVD to non-CVD class size was used as a penalty factor during training, A weighted class ratio of 5:1 for example indicates that any misclassified CVD training samples will be penalized 5 times more than the miss-classified non-CVD samples after each epoch, during error computation before backpropagation at

the output. This helped in reducing the overfitting in addition to the dropout strategy used for preventing overfitting.

The technique involved a novel training schedule, with a weight ratio of 1: N for sufficiently large number of epochs for initial model training, and gradual reduction in epochs and an increase in weight ratio. Here, the actual class weight ratio as ρ_0: 1 was assumed, where factor ρ_0 represents the learning schedule is depicted in Figure 13 here.

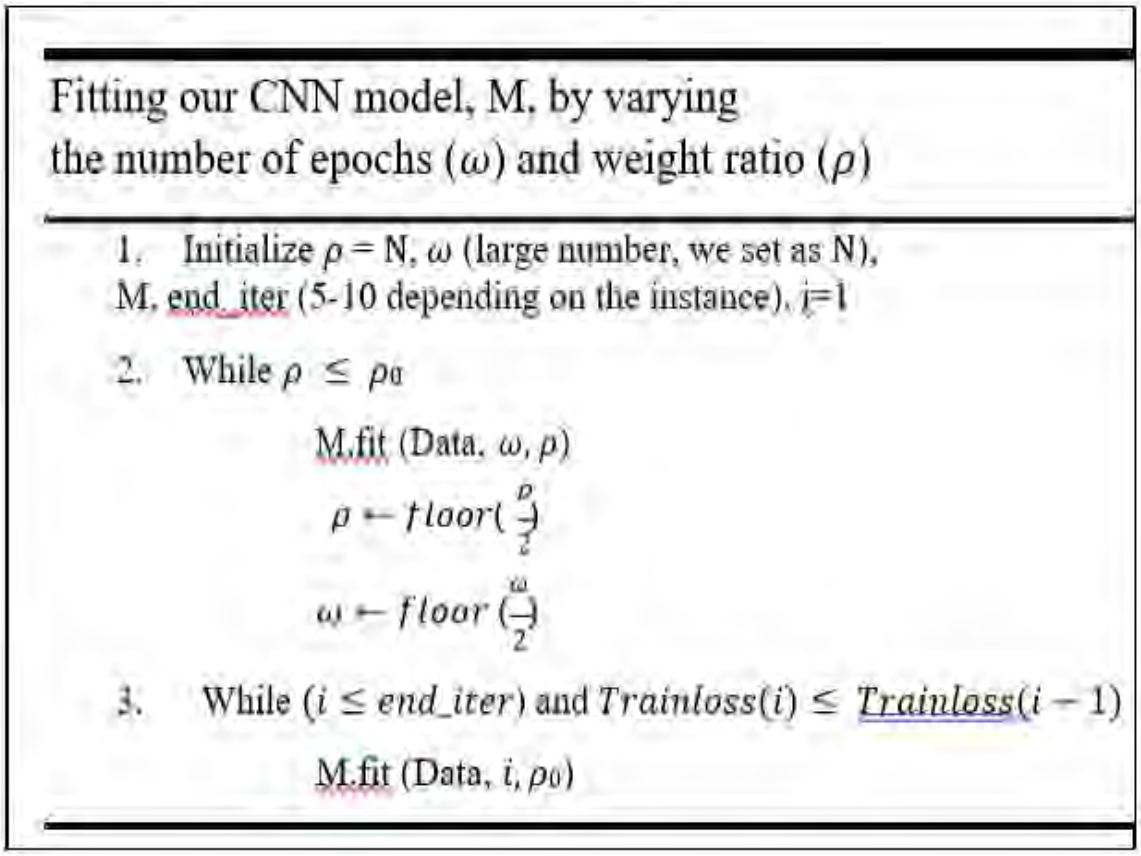


Figure 14 : The learning schedule for addressing the class imbalance

6.4 Experimental set up

In this section, the details of the experimental set up used for evaluating the performance of the proposed stacked deep learning architecture are presented. We first

discuss the details of the super dataset preparation, followed by the details of novel experimental protocol used.

6.4.1 Super dataset preparation

The super dataset preparation was done by fusing the subsets of the NHANES from 1999-2000 to 2015-2016. The merged super dataset details are as shown in Figure 14 below, which includes the demographic, examination, laboratory, and questionnaire data of around 37,079 individuals, along with their CVD disease status. The super dataset was highly imbalanced with 1300 individuals with positive CVD status, and 35,779 patients with negative CVD (no heart disease).

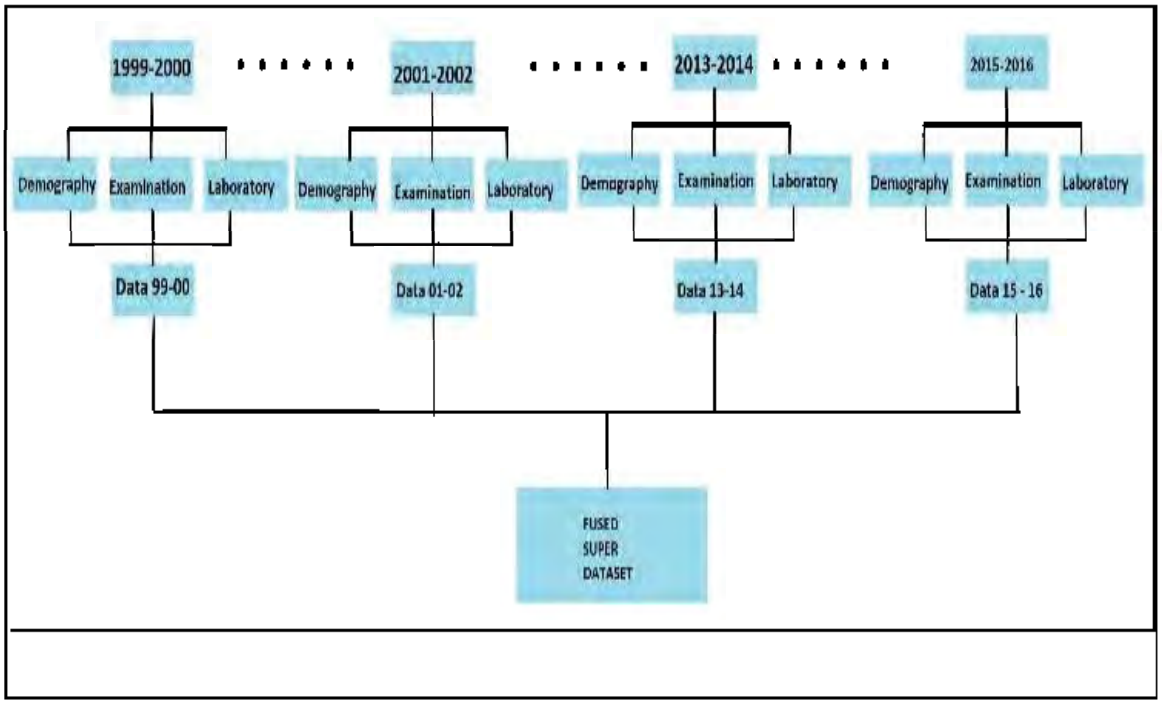


Figure 15 : Super dataset creation by fusing the subsets of NHANES data

TABLE XVI. RISK FACTOR DESCRIPTION AND ITS DEPENDENCIES.

Feature	Details	Nomenclature	Representation
Gender	Subject Gender	1/2	**Male/Female**
Activity Level (Vigorous)	One week/30 days activity level (Vigorous)	1/2/3	**Yes/No/ Inability**
Activity Level (Moderate)	One week/30 days activity level (Moderate)	1/2/3	**Yes/No/ Inability**
Diabetic status (at least 1 year)	Positive diagnosis status of Diabetes	1/2/3	**Yes/No/ Borderline**
Diabetes – genetic pre-disposition	Family members (blood related**)** diagnosed with diabetes	1/2	**Yes /No**
Stroke– genetic pre-disposition	Family members (blood related**)** diagnosed with hypertension/stroke before they turned 50 years old)	1/2	**Yes /No**
Reported CVD disease	Positive CVD disease reported by the subject	**1/2**	**Yes /No**

As can be seen in Figure 4, while the super dataset synthesized provided a large dataset to be used for building deep learning models (37,079 data samples), the class distribution is significantly imbalanced (35779:1300). The dependent and independent variables in the dataset used for the development of deep learning model is shown in Table XVI. Some of the independent variables include demographic details for the individuals at the time of screening, including the age and gender, and the variables from the examination data, such as the weight, height, blood pressure and body mass index (BMI) considered as key risk factors were included in the experimental setup, as it is important to examine their effect on the cardiovascular disease prediction. In addition, we included the laboratory

and survey/questionnaire data in the super dataset from the NHANES subsets. Table XVI also shows the relationship between the variables in the dataset used for the proposed stacked Dense-CNN cascade model development.

Different sets of variables were considered for examining their impact on the CVD status prediction, including certain demographic variables such as the gender and the age of the subjects that were surveyed during screening, as well as the subjects' blood pressure, BMI, height and weight from the data. This resulted in a total of 30 continuous and 6 categorical independent variables for predicting the likelihood of cardiovascular disease (CVD). The CVD disease status is denoted as the binary categorical (YES/NO) dependent variable.

Initially, the full list of variables in the super dataset were included, namely, blood related diabetes, blood related stroke, health insurance, diabetes, moderate and vigorous work, HDL (high density lipoprotein), glycohemoglobin, triglycerides, total cholesterol, protein, uric acid, phosphorus, bilirubin, lactate dehydrogenase (LDH), iron, glucose, GGT (gamma-glutamyl transferase), cholesterol, creatinine, alanine aminotransferase(ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), albumin, the width of red blood cells, haematocrit, neutrophils, platelets mean volume, platelet count, hemoglobin mean concentration, mean volume of cell, hemoglobin, red blood cells, basophils, eosinophils, monocyte, lymphocyte, white blood cells, BMI, height, weight, diastolic, systolic, pulse rate (60-sec), family income/poverty ratio, annual family income, age and gender.

Some of the linearly dependent variables and uncorrelated variables were removed from the dataset for subsequent processing and analysis,

such as the neutrophils(segmented), hematocrit, hemoglobin mean concentration, mean cell volume, total cholesterol, eosinophils, monocyte, lymphocyte, health insurance, pulse rate (60-sec), family income/poverty ratio, height, and annual family income.

6.4.2 Comparative Study of Shallow Learning Algorithms

For setting a baseline reference for comparing performance improvements achieved, first set of experiments involved examining behavior of NHANES super dataset for shallow machine learning algorithms. As outlined in the previous chapters of the thesis, several shallow machine learning algorithms were reported in the research literature for predicting the CVD events and their severity. Some of the shallow learning algorithms considered for this experimental validation task included methods based on logistic regression, support vector machines, random forests, bagging and boosting, and multilayer perceptron neural networks.

- Logistic regression uses a standard logistic function (a sigmoidal curve) and estimates the probabilities for the prediction of outcome of one or more variables, for modelling the relationship between the independent variables and the categorical dependent variable. The logistic function is given by, $f(x) = {}^{1}/_{1+e^{-x}}$ and the model predicts a binomial outcome.
- Support vector machine (SVM) is another established shallow machine learning algorithm used for binary classification algorithm and is based on generating a (*N-1*) dimensional hyperplane for separating feature representations as two separate and unique classes in a plane with N

dimensions. Further, a hyperplane in high dimensional is constructed for two-class classification.

- The shallow machine learning approaches based on ensemble models include Random forests based on bagging and boosting strategies, where overfitting in model building is avoided by averaging of deep-growing decision trees obtained by training on different subsets of data.

- While the random forests algorithm based on *bagging* or bootstrap aggregating strategy, involves selection of a random subset of features at each split-with each model built independently, in parallel, a sequential ensemble technique is used in the *boosting* approach, and each model is built on previous model's misclassification, and rectification of each misclassification.

- For random forests algorithm based on boosting methods, the model building involves initialization of weights on training samples and the use of single feature for training the classifier for *n* iterations and evaluating the training error. The updating of weights is done by choosing the classifier with the lowest error. A linear combination of *n* classifiers is used to obtain the final classifier.

  The random forest classifier based boosting approach is obtained in the form, $\sum_{t=1}^{T} f\ t(x)$.

- The selection of weak learner is done after each iteration *t,* and a coefficient α_t is assigned so that the minimization of error *t-stage boost classifier's in* the training stage is done.

- The multilayer perceptron (MLP) algorithm is based on a feedforward artificial neural network (ANN), and involves one or more hidden layers, in addition to the input layer, and the output layer, and uses backpropagation algorithm for building the model from the training subset of data. Further, for mapping the weighted inputs, a nonlinear activation function is used for each hidden layer's neuronal output. The ReLU (rectified linear unit) and hyperbolic tangent are the two established activation functions for MLP algorithm, represented by

  $f(x) = x^+$ (RELU) and $y(x_i) = tanh\ (x_i)$ (hyperbolic tangent).

While most of the shallow machine learning algorithms require significant feature engineering and manual expert guidance to build an efficient model, and for building MLP and other neural network models, large datasets for model building are required, in addition to feature engineering. There were several recent approaches for addressing the data shortage issue in building neural network models and involved obtaining large volumes of data synthetically for building the model, using a method called data augmentation. Some of the data augmentation approaches include, random oversampling (ROS), synthetic minority over-sampling technique (SMOTE), and adaptive synthetic sampling, in which augmentation of the minority class data was done by either replicating or synthesizing new data. While these approaches could work well for synthesizing data with moderate imbalance, and unstructured datasets such as image data, the data for severe imbalance situations (particularly for detecting rare diseases, where there are very few minority class/positive disease status samples), these augmentation approaches can lead to implausible data to be synthesized, as well as unrealistic and questionable

information synthesis. For this reason, we performed an exploratory analysis in the work reported here, for the super NHANES dataset obtained, by examining each of the data augmentations approaches mentioned above, along with the corresponding visualization of the augmented data via the t-SNE algorithm (a non-convex algorithm, generating an embedding based on initial low-dimensional embedding). Finally, after exploratory analysis, and after examining several data under sampling strategies, including the edited nearest neighbor (EDN), instance hardness threshold (IHT) and different versions of near miss algorithms (NM-v1, v2 and v3), it was decided to narrow down on random under sampling strategy for selecting a subset of data for training the stacked Dense-CNN cascade.

6.5 Experimental Results

Different sets of experiments were conducted to examine the performance of two stages of the proposed stacked Dense-CNN model including the LASSO feature pre-processing step as discussed next.

6.5.1 Performance of LASSO Regression and Regularization

For selecting the most significant features for model building from NHANES super-dataset, the correlations between the 30 continuous features for LASSO regression analysis were examined. A high degree of correction (77%) between ALT (serum alanine aminotransferase) and AST (aspartate aminotransferase) features was found. Also, AST levels can be significantly high for CVD patients, and hence a major risk factor for predicting CVD disease and need to be included as a biomedical marker and significant feature for predicting the severity of CVD events. Next, the correlation between BMI (body mass index) and weight was

found to be 89%, and the correlation between RBCs (red blood cells) and hemoglobin was found to be 74%, and hence these had to be included in the feature set as well. Further, there could be a significant association between hemoglobin, RBCs, and clinically recognized CVD, and this warrants the inclusion of these features in model building. Further, a high correlation (79%) between glycohemoglobin and glucose was noted, with significant association between high blood glucose levels and increased CVD risk. Also, a correlation of 46% was found between protein and albumin, though not significant, hence included in model building, with the dependency of lower levels of serum albumin on the higher incidence of CVD disease and mortality, and higher levels of protein significantly increasing the CVD risks. Next, the correlation between AST and LDH (Serum Lactate dehydrogenase) was found to be low (41%), as noted by findings on lower risk of CVD with an increase value of LDH. The association of several of these risk factors on the higher incidence of CVD events, was validated in the literature as well, and hence were included in LASSO regression processing, and a ranked feature list of significance and importance was obtained for inclusion in the model building stage. By using a subset of data with around 100 instances from randomly sampled data, from a data set of 1300 positive (CVD) and 1300 negative (no-CVD) samples, and by setting class-weight as '6' in equation 2, it was found that the predictor features 'ALT', 'GLUCOSE', 'HAEMOGLOBIN', and 'BMI' did not contribute significantly to CVD incidence, although, some of the earlier studies reported otherwise. Also, by majority voting, a strong correlation was noted between vigorous and moderate work, blood related stroke, presence of diabetes, gender, glycohemoglobin, HDL, triglycerides, uric acid, LDH, width of RBC (Red Blood Cell) distributions, platelet count, WBCs (White Blood Cells) and age.

6.5.2 Performance of Proposed Stacked dense-CNN Cascade

To achieve an optimal number of features for training the proposed stacked dense CNN cascade mode, the range for feature voting threshold was kept between 2 to 8. It was possible to obtain 84.17% prediction accuracy, the best accuracy for the training phase, or model building stage. Further, as shown in Figure 15, a training loss of 0,489 was achieved with a threshold value of 6, and a test phase accuracy of 82.32%. The three continuous predictor variables (ALT, BMI and GLUCOSE) and one of the categorical variables were reduced to zero, although reported to be highly correlated by authors in

[43] – [49]. The stacked dense CNN cascade model was built with training on several sets of subsampled datasets, with a threshold value of 6.

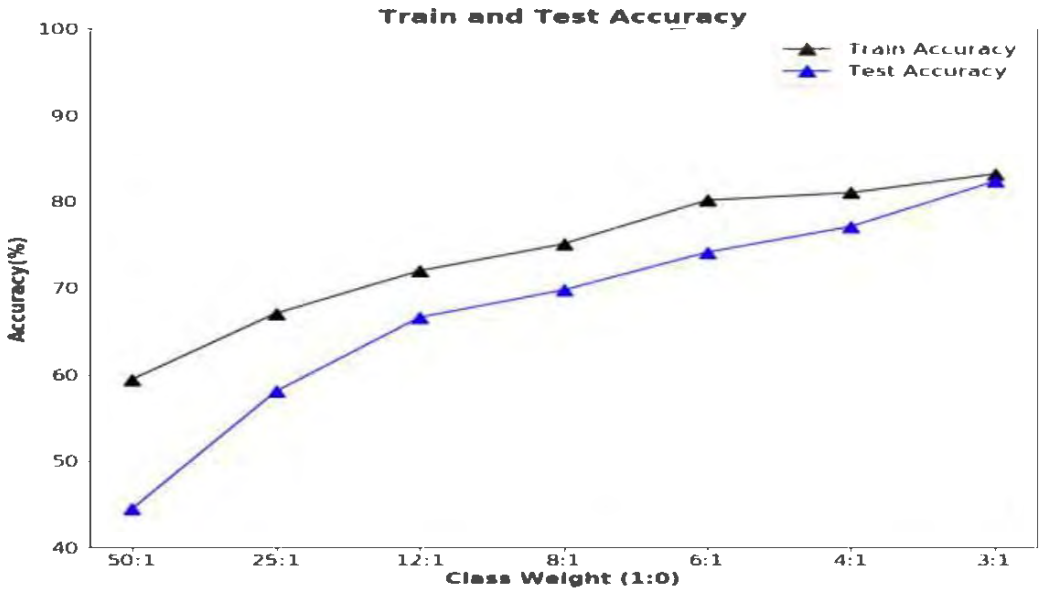


Figure 16 :PERFORMANCE OF STACKED DEEP LEARNING CASCADE

Also, as can be seen in Table XVII, the subsampling was performed with varying ratios in between 1300:13000 (CVD:Non-CVD) range, and increased to 1300:4000. The corresponding test accuracy obtained was 82.32%. The classification accuracy of CVD

disease status is impacted by the majority voting threshold. This could be due to the fact

that the selection of this threshold value determines the variables to be channelled to the proposed stacked dense CNN architecture. For smaller threshold values, the set of variables used for model building is large. The threshold value was set between 16.67 (100/6) -20(100/5), and 100 instances of LASSO appeared suitable for obtaining balanced class distribution (CVD and Non-CVD). The threshold value was set between 16.67% (~100/6) to 20 (~100/5), and with 100 LASSO instances, and balanced class distribution was obtained.

TABLE XVII. TRAINING SCHEDULE : CLASS WEIGHT RATIO VS. SAMPLING FOR OPTIMAL THRESHOLD

Class Weight	Sampling	TPR	TNR	Train Accuracy (%)	Test Accuracy (%)	Training Loss
10:1	1300:13000	0.690	0.812	78.00	81.09	0.684
8:1	1300:10000	0.706	0.827	77.28	82.67	0.676
6:1	1300:8000	0.741	0.799	80.00	79.90	0.595
5:1	1300:6000	0.735	0.80	80.48	79.96	0.548
3:1	1300:4000	0.778	0.823	83.51	82.32	0.489

It was possible to achieve maximum accuracy of 83.51%, with a minimum training loss of 0.489. And this was obtained with the misclassification penalty of 3:1 (CVD: Non-CVD) and sampling ratio of 1300:4000 (CVD: Non-CVD). The choice of subjects with CVD/Non- CVD status in each sub-sampled set included a random choice of non-CVD subjects, but all CVD subjects were taken into consideration. Also, as can be seen in Table XVII, during the training phase, the class weight, or misclassification penalty between CVD: Non-CVD classes were varied from 10:1 to10:3. This allowed a maximum accuracy of 83.51% and minimum training loss of 0.489, with a CVD/Non-CVD sampling ratio of 1300:4000 during training. The trained model was tested on an independent test set comprising a cohort of 31,779 (85.7% of the whole dataset, remaining samples), resulting in a test

accuracy of 82.32%, as can be noted in Table XVII. The final training of the network was done with an optimal sampling ratio of (1300:4000) between CVD/Non-CVD, as can be seen from Table XVIII.

TABLE XVIII. OPTIMAL THRESHOLD TRAINING SCHEDULE FOR DIFFERENT CLASS WEIGHT RATIO

Class Weight	TPR	TNR	Train Accuracy (%)	Test Accuracy (%)	Training Loss
50:1	0.980	0.383	59.43	44.48	1.591
25:1	0.923	0.568	67.06	58.04	1.236
12:1	0.860	0.664	71.98	66.60	0.965
8:1	0.836	0.686	75.08	69.76	0.765
6:1	0.817	0.740	80.13	74.12	0.620
4:1	0.788	0.770	81.00	77.10	0.550
3:1	0.773	0.818	83.17	82.32	0.489

Table XVIII depicts the performance of the stacked CNN model for different class weights between CVD and Non-CVD classes, with an improvement in training phase performance from 59.43% to 83.17% and reduced train-test mismatch, as the NonCVD: CVD class weight ratio changes from 50:1 to 3:1, indicating a better generalization error. Several hyper-parameters were fine-tuned during model building to obtain consistent class-wise accuracy, resulting in the best model fit to be obtained. ADAM Optimization was used as the activation function, with a learning rate variation from

0.006 to 60 epochs and the learning rate scheduler turned off during the training/model building phase. With the class of 3:1 (the penalty of misclassification of class 1 was set as 3 times higher than class 0 misclassification), the best test phase accuracy of 82.32% was obtained. Table XIX shows the confusion matrix of the proposed stacked CNN model, with classification parameters in the confusion matrix represented as TP, TN, FP, and FN. Here, TP represents the true positive classification cases, with true

predictions for class 1 (CVD class predicted as positive/true if the ground truth was positive /true), and TN represents true negative classification cases (CVD class predicted as negative/false when the ground truth was negative/false). The FN and FP represent False Negative and False Positive classification cases, where CVD class is predicted as negative/false when the ground truth was positive (FN), and the CVD class predicted as positive/true, when the ground truth was negative (FP). Table XX shows several of these performance metrics for different class weights.

TABLE XIX. CONFUSION MATRIX FOR THE STACKED DENSE-CNN CASCADE MODEL

Total Cohort		**True Condition**	
		Positive CVD event (Disease group)	Negative CVD event (Control group)
Predicted Condition	Presence of CVD	True Positive (TP) = 161	False Negative(FN) = 47
	Absence of CVD	False Positive (TP) = 5743	True Negative (TN) = 25838

Several other performance metrics were computed to evaluate the proposed Stacked Dense-CNN Cascade algorithm performance. This included the computation of sensitivity/recall rate, which for the positive class (class 1: presence of CVD) was 77% and for negative class (class 0: absence of CVD) it was 81. This is in fact an improved performance as compared to previous studies, particularly for a significantly imbalanced super dataset with 31,779 subjects and a novel class-weighted subsampled data partitioning strategy used between positive and negative classes, for building the model, under several class imbalance. Further by using a higher test size for each data partition (85% of

the total dataset as the test data subset, for each subsampled data partition (unlike the traditional 70:30/80:20 split between train/test subsets), it was possible to achieve a high positive/minority class detection performance despite the significant imbalance, and this could be due to an appropriate class-weighting technique, train-test partitioning approach and LASSO regularization. The promising outcomes from this indicate the generalization of this novel computational framework for other studies in healthcare with similar sparse data and imbalanced data situations.

The next performance metric assessed for the proposed stacked Dense-CNN Cascade model evaluation was computing the ROC curve. The area under the curve (AUC) measure for the ROC curve provides the probability that the binary classifier based on a model will rank a randomly chosen positive case (class 1: CVD) higher than a randomly chosen negative case (Class 0: Non-CVD). The ROC-AUC metric can thus be used as a tool for selecting the better optimal models and rejecting the suboptimal ones, independent of the class distribution. This can be done by the parametric representation of TPRs (true positive rates), and FPRs (false positive rates), at different threshold values. For our model, it was possible to achieve an AUC of 0.77 (77%) for the proposed stacked CNN model, improved performance as compared to previous work reported for CVD prediction. Also, a balanced accuracy of 79.5% ((TPR+TNR)/2) was achieved, since the balanced accuracy is often used as a more indicative metric for highly imbalanced data sets than the normal accuracy metric and is determined as an average TPR and TNR. In addition, the fall out rate or Type-I error for the model was 18.2% (obtained as ~5743/31571), and model Type-II error or miss rate was 22.6% (~47/210), showing 30% probability increase, post the positive CVD

status diagnosis, and the negative likelihood ratio was 0.27, denoting around 30% decreased probability post diagnosis, in predicting the absence of CVD status.

6.5.3 Performance Comparison with Shallow Machine Learning Models

In addition to the proposed stacked Dense-CNN Cascade model, for baseline performance comparison, several traditional models based on shadow machine learning were also built and tested on the test data subset. For all shallow machine learning models, empirically determined optimized parameters were used and assessed in terms of different performance metrics, including the train and test set accuracies, AUC and TPR/Recall and TNR/sensitivity rates. Highest test set accuracy was obtained for Logistic regression and Ensemble/Adaboost models, though they suffered from low recall values or true positive rates for detecting CVD status. The recall values or true positive rates for SVM and random forest models were comparable to that of the proposed stacked CNN model, with models predictive negative (class 0- non-CVD status) with better accuracy, as can be seen in Table XX.

TABLE XX. PERFORMANCE OF SHALLOW LEARNING MODELS WITH THE PROPOSED STACKED DENSE-CNN CASCADE MODEL

	Recall (%)	Specificity (%)	Test Accuracy (%)	AUC
Logistic Regression	51.44	91.15	90.89	71.29
SVM	77.40	77.87	77.87	77.64
Random Forest	76.44	76.06	76.06	76.25
AdaBoost	52.88	90.36	90.12	71.63
MLP	66.34	78.88	78.80	72.61
Stacked Dense-CNN Cascade model	77.3	81.8	81.78	76.78

However, the balanced accuracy of the proposed stacked CNN model was much superior (79.5%) when compared with the balanced accuracies of SVM and

random forest classifier models. Also, another variant of MLP (An optimized two-layer multilayer perceptron), without a CNN stage was examined, and led to a low TPR/recall value of 66.34% when tested on the test set cohort. As can be seen in Table V, the logistic, Adaboost and MLP classifiers performed poorly as compared to random forest and SVM classifiers, however our stacked Dense-CNN Cascade model performed better than any of these shallow learning models. The proposed stacked CNN model performs well in all the metrics. This comparative study confirms that the proposed stacked Dense-CNN Cascade model has significantly improved performance as compared to shallow machine learning models in the prediction of both positive and negative cardiovascular disease classes.

6.5.4 Performance Comparison of LASSO regularized CNN with standard CNN model

LASSO regularized CNN involves setting the α parameter appropriately for the selection of significant variables. By setting $\alpha = \infty$, the N/∞ becomes 0, and this takes into consideration all variables, leading to standard CNN model, or ∞-LASSO-CNN model. And by using a 3:1 class ratio for samples in the subsampled dataset, the naïve-CNN results in the test set accuracy of 79.42%, and is around 2% less than the average test accuracy achieved with a α value of 6, or 6-LASSO-CNN model. Although, 2 % improvement appears to be marginal, the increase in number of samples accurately labelled in the test set increases significantly due to the large size of test subset ($635 \approx 31779 * 0.2$)

6.5.5 Performance Comparison of LASSO regularised CNN with Traditional Oversampling methods

In this section, we compare the performance achieved with the proposed LASSO

regularized CNN model with the traditional oversampling techniques, such as Random Over Sampling (ROS), (Synthetic Minority Over-sampling) SMOTE technique, and Adaptive Synthesis (ADASYN) method, for imbalanced class distributions in the dataset. When the size of the test subset used in traditional oversampling methods is smaller, the oversampling of the minority class is done to be of the same dimensionality as that of the majority class. In ROS and SMOTE approach, each class of training data subset contained 32013 samples and contained 3709 samples in test data, with a class distribution of 3558 negative non-CVD samples, and 151 positive (CVD) samples in the test set. When using ADASYN method for oversampling, the training data size comprised 64134 samples (32121 positive/CVD samples and 32013 negatives / Non- CVD samples), and the size of test set was 3709 samples (3558/151 Non-CVD/CVD ratio). For all methods, the ratio of class-specific samples (non-CVD: CVD) in the test data was approximately 24:1. For all three algorithms, the test accuracy for the majority class (NonCVD/Negative) is significantly better, as compared to minority (CVD/positive) class, explaining poor capability of oversampling methods for dealing class imbalance.

6.5.6 Performance Comparison of LASSO regularised CNN with Under sampling methods

For comparing the performance of LASSO regularized CNN with traditional undersampling methods used for addressing the imbalance in class-distribution, such as Edited Distance (ED), Instance Hardness Threshold (IHT), and Near Miss (NM -v1, v2 and v3), the majority class (Non-CVD / negative class label) size in the subsets of data used was reduced against the minority class (CVD / positive class label). In the case of under sampled datasets, the size of the test dataset was

kept large, and the size of the majority class (Non-CVD) is reduced against the minority class (CVD). Any left-over samples that were not included in the train data subset after undersampling process were added to the test data subset. This has resulted in a test subset size of 6366 samples (152/6234 – CVD/Non-CVD ratio) for EDN method, and 24995 samples (152/24843 – CVD/Non-CVD) for IHT method. And for all versions of NM method, the training subset contained 4523/1357 Non-CVD/CVD samples, maintaining a ratio of 3.5:1, leading to an almost balanced data setting during model building/training phase, and similar to experimental settings for the LASSO regularized CNN model setup.

6.5.7 Robustness to Data Imbalance

As the central aim of this work is to develop a novel computational framework for imbalanced data situations (found in clinical data settings with rare disease datasets having few positive samples), we examined the robustness of the proposed Stacked Dense-CNN Cascade model to data imbalance. This will allow the model to be generalized to any dataset with similar constraints, such as: (1) Severe imbalance in data due to the acquisition restrictions or nature of how it originated, (2) Ineffectiveness of traditional data augmentation techniques based on under- sampling and over-sampling techniques, in handing the statistical distributions, and

(3) high cost of misclassification, such as the risk of not diagnosing correctly, the serious illnesses or the severity of the disease. As can be seen in Table XXI, the proposed model performs exceptionally well in handling data imbalance, particularly, the contributions from the cascaded CNN stack. Traditionally, such improved performance and robustness is found in two-dimensional and

multi-dimensional datasets, such as imaging datasets, where due to spatial correlation among pixels, standard data augmentation techniques work well and offer resilience to a n imbalance in class distribution. However, for structured datasets, containing a combination of categorical and continuous variables, this is not the case, and the proposed Stacked- Dense-CNN Cascade model offers much hope. Table VI shows the performance achieved with different CNN and dense layers in our proposed model in dealing with class imbalance. Here, *I* represent the input, *O* is the output. *C2* is a convolution layer with 2 filters, *C4* is a convolution layer with four filters, and *C8* is a convolution layer with 8 filters. *D64, D128, D512* are dense layer dimensions.

TABLE XXI. EVALUATING ROBUSTNESS TO DATA IMBALANCE

	Non-CVD acc (%)	CVD acc (%)	Overall acc (%)	Acc difference (%)	No of parameters
Model - I (I-D512-D128-O)	84.74	57.6	84.63	27.14	88706
Model- II (I-D64-D128-D256-O)	81.63	75.9	81.60	5.73	45442
Model-III (I-D128-**C4-C8**-O)	78	75.96	77.97	2.04	6306
Model-IV (I-D128-**C2-C4**-C8-O)	79.4	77	79.38	2.4	3104
Model-V (I-D128-C4-O)	77.42	75.35	77.4	.07	6426
Model-VI (I-D64-C2-D128-C4-O)	81.53	76.9	81.43	4.63	11,678
Model-VII (I-D64-**C2-C4**-D512-O)	83.17	77.88	82.3	5.29	32066

From Table XXI, it is worth noting that MLP or Fully Connected Dense

Network, when properly trained can perform well if the accuracy of the majority class is taken into consideration. However, the minority class performance suffers, due to the network architecture chosen. Model I has 27% difference in class wise accuracies, and the Model II with an extract fully connected dense layer introduced the difference between the class wise accuracies reduces remarkably to 5.73%, however, still requires a large set of trainable parameters (45,442 parameters), which means a large amount of input data requirement. By introducing a CNN layer between dense layers, the robustness to class imbalances improves, as the difference between majority and minority class accuracies becomes lesser, with 2.04% for model III, 2.4% for model IV and 0.07% for Model V. However, this improvement in robustness is achieved at the cost of overall class wise accuracies, which less than 80% accuracy for each class.

The two stacked Dense-CNN Cascade models (Model VI and Model VII), allow both robustness and performance improvement to be achieved, with Model VII performing the best. Also, it is interesting to note that the introducing CNN layer and Cascade of CNN layers in the Dense network requires lesser trainable parameters, with model III to model VII requiring lesser parameters as compared to model I and II. The appropriate combination of dense layers and CNN layers allow both performance and robustness to be achieved simultaneously, with CNN layers enhancing the robustness, and minimizing the difference between class-wise accuracies, and dense layers trying to improve the overall accuracy for the test set. Another noteworthy observation is the stability in accuracy achieved by each of the models. During the model building/training phase, the train/test accuracies are more sensitive to training epochs

(monotically increasing) when class-weight ratio is decreased from 13 to 40 gradually. This is for majority class, while the minority class performance deteriorates as the number of epochs increases. This deterioration is reduced after the inclusion of multiple CNN layers. In short, the accuracies that our proposed stacked dense-CNNcascade model yields, is far more stable for a fairly large number of epochs.

6.6 Summary

In this Chapter, a robust disease detection model for CVD prediction based on stacked Dense-CNN cascade algorithm was proposed, which performs well for sparse and highly imbalanced clinical data settings, containing both continuous and categorical variables.

Next chapter concludes the thesis with some contributions made in this work, and future directions for extending this research.

Chapter 7: Conclusions and Further Work

7.1 Conclusions and Discussion

In this thesis, a novel computational framework for CVD disease prediction was proposed, based on machine learning and AI. Several innovative contributions were made in this thesis towards the development of CVD disease prediction based on AI/ML algorithms, that address a particular challenge, in terms of either prediction performance, robustness to complex data settings, such as an imbalance in the class distribution or the interpretability and explainability of machine learning models.

The first contribution was examining simple shallow machine learning algorithms that are inherently interpretable and explainable, and investigating the performance of the disease prediction models for interpretability and explainability based on these shallow learning algorithms. The prediction performance and explainability metric was assessed with Cleveland dataset, a popular dataset though small can provide clear interpretability, and serve as a baseline reference for comparing the performance of other advanced and sophisticated methods. Three different types of shallow machine learning algorithms, the naïve Bayes (NB), the logistic regression (LR) and the decision tree (DT) algorithms were used for developing disease detection models using Cleveland database. The models built were assessed in terms different performance metrics used for classification capabilities, and their interpretability/explainability was examined with two different measures, the feature permutation importance and LIME map. This preliminary study helped in setting a baseline reference for performance comparison in terms of prediction accuracy metrics as well as interpretability/explainability metrics. The

development of this baseline reference system, along with interpretability and explainability analysis showed that machine learning based on decision trees and their variants are better in terms of interpretability and explainability, out of all shallow machine learning algorithms, though they may not have excellent prediction performance. Another important observation from this study, was once the model performance is acceptable in terms of detection and prediction accuracies, it can be augmented and supplemented with appropriate post-processing stages to provide better interpretations and explanations and can be trustworthy for inclusion in clinical workflow. While a small database (Cleveland database) was used in this Chapter for building disease detection models, the next contribution of the research was to extend this study to larger and complex data sets, to enhance the prediction performance, as the shallow machine learning algorithms find it challenging, as the data size becomes larger.

The next contribution was the development of a generic computational framework, called the "Smart Predictive Modelling Framework" based on both supervised and unsupervised machine learning algorithms, to work with partitioned data, involving cohort segmentation and filtering based on age, gender, and education level for complex and big datasets, such as the NHANES and the Framingham Heart Study datasets. For the first dataset i.e., the NHANES dataset, optimal attribute selection techniques based on information theory-based ranking resulted in better performance. For Framingham heart study dataset, a combination of supervised and cohort segmentation using filtered variables based on different demographic attributes such as age, gender and education level was better. One of the important findings from this study was, that the proposed Smart Predictive Modelling Framework, and the core prediction engine, based on the concepts of

segmenting big data sets into small subsets of data with appropriate demographics filters, and with models built with established traditional shallow learning algorithms, such as the decision tree algorithms and its variants, can achieve graceful performance. This can help in building fast AI/ML based prediction models, which are inherently more interpretable and explainable, and could be more trustworthy.

The next contribution involved the proposal for extending Smart Predictive Modelling for building disease detection models based on a novel algorithm based on XGBoost technique, considered to be an efficient, scalable, distributed gradient-boosted decision tree (GBDT) machine learning algorithm set, and provides a parallel tree boosting capabilities, suitable for regression, classification, and ranking problems. Further, being a variant based on decision tree algorithm, it has better interpretability and explainability, and more suitable for clinical workflow.

For the final contribution, a new disease prediction model was proposed based on advanced deep machine learning algorithm suite, that has better robustness to complex data settings involving imbalance in class distribution, a key issue in clinical dataset, where there are less positive data samples than negative data samples. The prediction engine based on this algorithm suite was developed and validated with two benchmark publicly available datasets, NHANES and FHS datasets. This robust disease detection model for CVD prediction based on a sophisticated machine learning architecture called as the stacked Dense-CNN cascade algorithm suite performs well for a sparse and highly imbalanced clinical data settings, containing both continuous and categorical variables. Further the inclusion of some novel pre-processing and post-processing stages, allows better visualisation and analysis of outcomes, allowing better interpretations and

explanations coming out of highly impenetrable black box model based on stacked deep learning architecture.

The next section presents future direction for research on extending the proposed line of investigation further, in terms of building disease detection models based on highly complex data settings involving big datasets from wearables such as the activity and fitness monitoring devices.

7.2 Future Directions

In this section, future directions, in terms of some of the work in progress for extending the proposed predictive modelling framework, which can allow newer disease detection models to be developed for cardiovascular diseases, particularly under highly complex big data settings, such as the dynamic data captured from fitness and activity monitoring devices, and development of models with improved prediction performance, with enhanced interpretability and explainability, and higher robustness to perform well under complex data settings, such as imbalanced class distribution, multiclass detection, missing data, poor quality data, and different types of structured and unstructured data settings, including the data from wearables, imaging devices and social media feeds, is discussed. These enhancements can address these emerging requirements well, and can allow early detection and continuous monitoring of risk factors associated with cardiovascular disease and help reduce the non-community disease burden.

7.2.1 Disease Detection Models based on Wearable Technologies

An effective strategy to mitigate the burden of cardiovascular disease is to monitor patients' vital parameters during daily activities with wearable technologies. Of late, technological advances have contributed significantly to evolution of

wearables technologies by reducing the size of the devices, and improving the accuracy of sensing vital parameters using devices with relatively low energy consumption that can manage security and privacy of the patient's medical information, are adaptable to any data storage medium, and are reasonable in cost as compared to the traditional schemes, requiring patients to visit GP health clinics or hospitals, for ECGs and other non-invasive vital parameter testing, thus contributing a serious option in early detection, continuous monitoring and treatment of CVDs.

In this section, a brief snapshot of future directions of using commercial and non-commercial wearable devices used to monitor vital parameters associated with CVD is presented, including the further plans for extending the proposed smart predictive modelling framework with better explainability and interpretability proposed in this thesis. One of the key motivations on improving the state of the art in CVD management, is to facilitate better self-management of disease at personal and individual level, instead of reliance on over-burden health care system, and how explainable and interpretable AI and ML based disease detection models based on commercial and non-commercial wearables such as the smart wristbands, patches, and smartwatches contribute. By using these wearables, it is possible to generally monitor variables such as heart rate, blood oxygen saturation, and electrocardiogram data. Non-commercial wearables focus on monitoring electrocardiogram and photoplethysmography data, and they mostly include accelerometers and smartwatches for detecting atrial fibrillation and heart failure.

7.2.2 Explainable Smart Predictive Modelling Framework

For enhancing the proposed smart predictive modelling framework for building disease detection models in this thesis, we developed a smart-watch wearable dataset, by collecting the health and behaviour data from Australian Capital

Territory, Australia residents under the project "ACTive Community Program". As ACTive Community program is volunteer based program there was insufficient data, and the quality of the captured data was not optimal for building sophisticated disease detection models. Particularly, the data was highly noisy, along with the large number of missing data. Hence, it was decided to refine the data collection protocol again, and pursue with data collection activity for further research, but these initial efforts led to an opportunity to enhance the proposed disease detection modelling framework to address the poor-quality data settings, with a novel approach developed for the data imputation described in the next section. In extending Smart Analytical Framework for developing disease detection models using smart wearables, innovative ensemble modelling techniques will be used to impute missing values in the ACTive Community dataset. The proposed imputation algorithm will be a combination of the Multiple Imputation Chained Equation Approach, Gradient boosting algorithm approach (XGBoost) and neural network ensemble, with detailed description included in Appendix I.

Some of the emerging disease detection models based on wearables and smart monitoring devices are based on Internet of Things, also called IoT devices. The current trend in IoT based wearables will progressively grow and may define the future shape of healthcare, along with concurrent technological developments in associated computational hardware and software. In addition, trends in the Internet of Things (IoT) combined with powerful machine learning and AI algorithms similar to the ones proposed in this thesis can lead to enhanced mobile health or mHealth models suitable for deployment in remote, emergency and rural settings, without elaborate hospital infrastructure.

The different sensors that make up mHealth wearables can collect daily routine

data that interact with technological cloud-based platforms. In this sense, the IoT is a promising alternative for managing data provided by wearable devices. The IoT paradigm can generate medical information and critical event alerts that can be shared with health specialists and used to interact with social networks. A major challenge in the IoT paradigm, however, remains to

find the best practices for handling confidential patient information. In this sense, multiple data privacy service providers are already working on it. Finally, CVD monitoring technologies can also be integrated in smartphones through device cameras or accelerometers. The main issue in this case is that not all smartphones have the same level of precision in terms of recording biomedical variables. Since this is a device function provided at the design level of the mobile device, it certainly goes hand in hand with the cost of the device.

Disease detection models based on privacy preserving and secure machine learning techniques and deep learning architectures, to overcome current technological barriers can improve the privacy and reliability of disease monitoring systems as well, including performance of wearable IoT devices, ambulatory systems, and the accuracy of measurements of vital parameters associated with CVD. Hence, as the privacy, reliability and accessibility of IoT and wearable based disease detection models increase, their acceptance will increase among consumers, leading to continuous day-to-day health monitoring at personalised, individual level, in collaboration with chronic disease care team, and subsequent improved public health outcomes.

Chapter 8: Bibliography

[1] *How much does Australia spend on health care - Australian Institute of Health and Welfare, Australia's Health series no.14, Cat no. AUS178, Canberra, AIHW,* 2014.

[2] *Australian Medical Society Position Statement on Obesity 2016 (https://www.ama.com.au/media/obesity-australias-biggest-public-health- challenge),* 2016.

[3] *National Health and Medical Research Council (NHMRC) (2010). A "state of the knowledge" assessment of of comprehensive interventions that address the drivers of obesty: A Rapid Assessme. University of Syndey*, 2010.

[4] S. Dash, S. K. Shakyawar and M. Sharma, "Big data in healthcare: management, analysis and future prospect," *Journal of Big Data,* vol. 6, no. 54, pp. https://doi.org/10.1186/s40537-019-0217-0, 2019.

[5] Y.-L. Zheng, X.-R. Ding and C.-C. Yan Poon et al, "Unobtrusive sensing and wearable devices for health informatics," *IEEE Transactions on Biomedical Engineering,* vol. 5, no. 61, pp. 1538-54. doi: 10.1109/TBME.2014.2309951, 2014.

[6] T. Vos , R. Barber, B. Bell and A. Bertozzi-Villa et al, "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study," The Lancet, 386 (9995), 2015.

[7] *Australia's health 2016: in brief presents highlights from the AIHW's 15th biennial report on the nation's health,* 2016.

[8] V. Raghupati and W. Raghupati, "Big data analytics in healthcare: promise and potential.," . *Health information science and systems,* vol. 2, no. 3, pp. https://doi.org/10.1186/2047-2501-2-3, 2014.

[9] . J. Bresnick, “How Healthcare Big Data Analytics Is Tackling Chronic Disease,” HealthITAnalytics, June 2015. [Online]. Available: https://healthitanalytics.com/news/how-healthcare-big-data-analytics-is-tackling-chronic-disease. [Accessed 28 May 2022].

[10] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. , “International application of a new probability algorithm for the diagnosis of coronary artery disease.,” *American Journal of Cardiology, ,* vol. 64, no. https://archive.ics.uci.edu/ml/datasets/heart+disease, pp. 304--310, 1989.

[11] Centers for Disease Control and Prevention, “National Health and Nutrition Examination Survey,” [Online]. Available: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm. [Accessed 28 May 2022].

[12] NIH , “National Longitudinal Mortality Study (NLMS),” National Heart Lung and Blood Institute, [Online]. Available: https://biolincc.nhlbi.nih.gov/studies/nlms/. [Accessed 28 May 2022].

[13] E. Topoi, “High-performance medicine: the convergence of human and artificial intelligence,” *Nature Medicine,* vol. 25, no. 1, pp. 30–41. doi: 10.1007/978-3-319- 31750-2_3, 2019.

[14] X. Wang, Y. Peng, L. Lu, Z. Lu and M. Bagheri, “ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, doi: 10.1109/cvpr.2017.369, 2017.

[15] Z. Li, C. Han, Y. Xue, W. Wei and L. Li, “Thoracic Disease Identification and Localization with Limited Supervision.,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/cvpr.2018.00865, 2018.

[16] Ehteshami Bejnordi, B., Veta, M., Johannes van Die, B. Ehteshami , M. Veta and v. D. Johannes et al, “Diagnostic assessment of deep learning algorithms for

detection of lymph node metastases in women with breast cancer," *JAMA,* vol. 318, no. 22, pp. 2199-2210. doi:10.1001/jama.2017.14585, 2017.

[17] A. Esteva, B. Kupre, R. Novoa, J. Ko and .. Swet, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature,* vol. 543, no. 7639, pp. 15-118. doi:10.1038/nature21056, 2017.

[18] M. Abràmoff, . P. Lavin and M. Birch et al, "Pivotal trial of an autonomous AI- based diagnostic system for detection of diabetic retinopathy in primary care offices," *npj Digital Med* , vol. 1, no. 39, pp. doi:10.1038/s41746-018-0040-6, 2018.

[19] L. Bote-Curiel, S. Muñoz-Romero, A. Gerrero-Curieses and J. L. Rojo-Álvarez, "Deep Learning and Big Data in Healthcare: A Double Review for Critical Beginners," *Applied Sciences,* vol. 9, no. 11, p. 2331. doi: 10.3390/app9112331, 2019.

[20] O. Faust, Y. Hagiwara, T. J. Hong and O. S. Lih, "Deep learning for healthcare applications based on physiological signals: A review," *Computer Methods and Programs in Biomedicine,,* vol. 161, pp. 1-13, 2013. doi: 10.1016/j.cmpb.2018.04.005.

[21] P. Xia, . J. Hu and Y. Peng, "EMG-Based Estimation of Limb Movement Using Deep Learning With Recurrent Convolutional Neural Networks," *Artificial Organs,* vol. 42, no. 5, p. doi: 10.1111/aor.13004, 2017.

[22] W. Geng, Y. Du, W. Jin, W. Wei, Y. Hu and J. Li, "Gesture recognition by instantaneous surface EMG images.," *Scientific Reports,* vol. 6, no. 1, p. doi: 10.1038/srep36571 , 2016.

[23] M. Atzori, , ,. M. Cognolato and . H. Müller,, "Deep Learning with Convolutional Neural Networks Applied to Electromyography Data: A Resource for the Classification of Movements for Prosthetic Hands," *Frontiers in Neurorobotics,* vol. 10, p. 10 . doi: 10.3389/fnbot.2016.00009, 2016.

[24] L. Fraiwan and K. Lweesy, "Neonatal sleep state identification using deep learning autoencoders," in *IEEE 13th International Colloquium on Signal Processing & Its Applications (CSPA).*, doi: 10.1109/cspa.2017.8064956 , 2017.

[25] . U. Acharya, S. L. Oh, Y. Hagiwara, and J. H. Tan, , "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Computers in Biology and Medicine*, vol. 100, no. doi: 10.1016/j.compbiomed.2017.09.017 , p. 270–278, 2018.

[26] ,. W.-L. Zheng, ,. J.-Y. Zhu, . Y. .Peng, and B.-L. Lu, "EEG-based emotion classification using deep belief networks.," in *2014 IEEE International Conference on Multimedia and Expo (ICME).*, doi: 10.1109/icme.2014.6890166 , 2014.

[27] Zhu, X., Zheng, W.-L., Lu, B.-L., Chen, X., Chen, S., & Wang, C., "EOG-based drowsiness detection using convolutional neural networks.," in *International Joint Conference on Neural Networks (IJCNN).*, doi: 10.1109/ijcnn.2014.6889642, 2014.

[28] Xia, B., Li, Q., Jia, J., Wang, J., Chaudhary, U., Ramos-Murguialday, A., & Birbaumer, N., "Electrooculogram based sleep stage classification using deep belief network.," in *Joint Conference on Neural Networks (IJCNN).*, doi: 10.1109/ijcnn.2015.7280775 , 2015.

[29] Du, L-H; Liu, W; Zheng, W.L.; Lu, B.L, "Detecting driving fatigue with multimodal deep learning," in *8th International IEEE/EMBS Conference on Neural Engineering (NER)*, doi: 10.1109/ner.2017.8008295, 2017.

[30] Acharya, U. R., Fujita, H., Oh, S. L., Raghavendra, U., Tan, J. H., Adam, M., … Hagiwara, Y., "Automated identification of shockable and non-shockable life-threatening ventricular arrhythmias using convolutional neural network.," *Future Generation Computer Systems*, vol. 79, no. doi: 10.1016/j.future.2017.08.039, p. 952–959, 2018.

[31] Majumdar, A., & Ward, R., "Robust greedy deep dictionary learning for ECG arrhythmia classification.," in *International Joint Conference on Neural Networks (IJCNN).*, doi: 10.1109/ijcnn.2017.7966413 , 2017.

[32] Zheng, Y., Liu, Q., Chen, E., Ge, Y., & Zhao, J. L., "Time Series Classification Using Multi-Channels Deep Convolutional Neural Networks.," in *Time Series Classification Using Multi-Channels Deep Convolutional Neural Networks.*, doi: 10.1007/978-3-319-08010-9_33 , 2014.

[33] Cheng, Y., Wang, F., Zhang, P., & Hu, J., *Risk Prediction with Electronic Health Records: A Deep Learning Approach.*, SDM. DOI:10.1137/1.9781611974348.49, 2015.

[34] Pham, T., Tran, T., Phung, D., & Venkatesh, S., "DeepCare: : A Deep Dynamic Memory Model for Predictive Medicine.," in *Advances in Knowledge Discovery and Data Mining Lecture Notes in Computer Science*, doi: 10.1007/978-3-319- 31750-2_, 2016.

[35] Afzal, N., Sohn, S., Abram, S., Scott, C. G., Chaudhry, R., Liu, H., … Arruda-Olson,
A. M., "Mining peripheral arterial disease cases from narrative clinical notes using natural language processing.," *Journal of Vascular Surgery,*, vol. 65, no. 6, pp. 1753–1761, doi: 10.1016/j.jvs.2016.11.031 , 2017.

[36] Vine, L. D., Zuccon, G., Koopman, B., Sitbon, L., & Bruza, P., "Medical Semantic Similarity with a Neural Language Model," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM 14*, doi: 10.1145/2661829.2661974 , 2014.

[37] Zhang, N., Yang, G., Gao, Z., Xu, C., Zhang, Y., Shi, R., … Firmin, D., "Deep Learning
for Diagnosis of Chronic Myocardial Infarction on Nonenhanced Cardiac Cine MRI," *Radiology,* vol. 291, no. 3, p. 606–617. doi: 10.1148/radiol.2019182304 , 2019.

[38] Medhekar, D.S., Bote, M.P., Deshmukh, S.D.: , "Heart disease prediction

system using naive bayes," *Int. J. Enhanced Res. Sci. Technol. Eng.* , vol. 2, no. 3, 2013.

[39] Vembandasamy, K., Sasipriya, R. and Deepa, E. , "Heart Diseases Detection Using Naive Bayes Algorithm.," *IJISET-International Journal of Innovative Science, Engineering & Technology,* , vol. 2, pp. 441-444 , 2015.

[40] Gudadhe, M., Wankhade, K., & Dongre, S. (2010), "Decision support system for heart disease based on support vector machine and Artificial Neural Network.," in *Conference on Computer and Communication Technology (ICCCT).* , doi: 10.1109/iccct.2010.5640377 , 2010.

[41] Jaymin Patel, Prof.TejalUpadhyay and Dr. Samir Patel, "Heart Disease Prediction using Machine Learning and Data Mining Techniques," *IJCSC,* vol. 7, no. 1, pp. 129- 137, 2016.

[42] Sabarinathan V. and Sugumaran V. , "Diagnosis of heart disease using decision tree," *International Journal of Research in Computer Applications & Information Technology* , vol. 2, pp. 74-79, 2014.

[43] Kahramanli, H., & Allahverdi, N., "Design of a hybrid system for the diabetes and heart diseases.," *Expert Systems with Applications,,* vol. 35, no. 1-2, p. 82–89. doi: 10.1016/j.eswa.2007.06.004, 2008.

[44] Das, R., Turkoglu, I., & Sengur, A., "Effective diagnosis of heart disease through neural networks ensembles.," *Expert Systems with Applicatio,* vol. 36, no. 4, p. 7675–7680. doi: 10.1016/j.eswa.2008.09.013 , 2009.

[45] Z. &. J. Y. Zhou, "NeC4.5: Neural ensemble based C4.5.," *IEEE Transactions on Knowledge and Data Engineering,,* vol. 16, no. 6, pp. 770-773. doi:10.1109/TKDE.2004, 2004.

[46] L. Breiman, "Random Forests," in *Machine Learning*, 2001, pp. 45(1)5-32.

[47] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* , 2016.

[48] Robert E. Schapire, “Explaining AdaBoost,” [Online]. Available: http://rob.schapire.net/papers/explaining-adaboost.pdf. [Accessed 30 May 2022].

[49] Chen, T. and Guestrin, C.,, “Xgboost: A scalable tree boosting system,” in *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, https://arxiv.org/abs/1603.02754, 2016.

[50] Culotta, A., McCallum, A., & Betz, J, “Integrating probabilistic extraction models and data mining to discover relations and patterns in text,” in *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings of the Main Conference,*, doi:10.3115/1220835.1220873 , 2006.

[51] Dong, X., Halevy, A., & Madhavan, J., “Reference reconciliation in complex information spaces,” in *Proceedings of the ACM SIGMOD International*
Conference on Management of Data,, doi:10.1145/1066157.1066168 , 2005.

[52] Dorr, D. A., Phillips, W. F., Phansalkar, S., Sims, S. A., & Hurdle, J. F., “Assessing the difficulty and time cost of de-identification in clinical narratives,” *Methods of Information in Medicine,*, vol. 45, no. 3, pp. 246-252. doi:10.1055/s-0038- 1634080 , 2006.

[53] Douglass, M., Clifford, G. D., Reisner, A., Moody, G. B., & Mark, R. G., “Computer- assisted de-identification of free text in the MIMIC II database,” Paper presented at the Computers in Cardiology, 2004. [Online]. Available: www.scopus.com . [Accessed 30 May 2022].

[54] Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S., “Privacy-preserving data publishing: A survey of recent developments.,” *ACM Computing Surveys,* vol. 42, no. 4, p. doi:10.1145/1749603.1749605, 2010.

[55] D. V. Carvalho, M. E. Pereira, and J. S. Cardoso, , “Machine learning interpretability: A survey on methods and metrics,” *Electronics,* vol. 8, no. 8, p. p. 832, 2019.

[56] R. Elshawi, M. H. Al-Mallah, and S. Sakr, , “On the interpretability of machine learning-based model for predicting hypertension,” *BMC Medical Information Decision Making,* vol. 19, no. 1, p. 146, 2019.

[57] S. Sakr, R. Elshawi, A. Ahmed, W. T. Qureshi, C. Brawner, S. Keteyian,, *PLoS ONE,* vol. 13, no. 4, pp. Apr. 2018, Art. no. e0195344, 2018.

[58] L. Munkhdalai, K. H. Ryu, O.-E. Namsrai, and N. Theera-Umpon,, “A partially interpretable adaptive softmax regression for credit scoring,” *Applied Sciences,* vol. 11, no. 7, p. 3227, 2021.

[59] H.-C. Thorsen-Meyer, A. B. Nielsen, A. P. Nielsen, B. S. Kaas-Hansen,, “Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records,” *Lancet Digit. Health,* vol. 2, no. 4, pp. pp. e179-e191, 2020.

[60] S. El-Sappagh, J. M. Alonso, S. M. R. Islam, A. M. Sultan, and K. S. Kwak,, “A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease,” *Scientific Reports,* vol. 11, no. 1, p. Art. no. 2660, Dec. 2021.

[61] Nissa N, Jamwal S, Mohammad S. Heart Disease Prediction using Machine Learning Techniques. Wesleyan Journal of Research 2021;13(67).

[62] Annu Dhankhar, S. J. Prediction of Disease Using Machine Learning Algorithms. Smart and Sustainable Intelligent Systems. P. C. a. T. C. Namita Gupta, Wiley-Scrivener Publishing LLC. 2021: 1: 115-126.

[63] P. D. C. Geetha S, Kalaivani V, Haritha CJ, Preetha G. Prediction Techniques of Heart Disease and Diabetes Disease using Machine Learning Turkish Journal of Computer and Mathematics Education 2021;12(10):3316-25.

[64] Ghosh P, Azam S, Jonkman M, Karim A, Shamrat FJM, Ignatious E, Shultana S, Beeravolu AR, De Boer F. Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques. IEEE Access 2021;9:19304-26.

[65] Maini E, Venkateswarlu B, Maini B, Marwaha D. Machine learning-based heart disease prediction system for Indian population: An exploratory study done in South India. Medical Journal Armed Forces India 2021;77(3):302-11.

[66] A. Abdellatif, H. Abdellatef, J. Kanesan, C. -O. Chow, J. H. Chuah and H. M. Gheni, "An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods," in IEEE Access, vol. 10, pp. 79974-79985, 2022, doi:10.1109/ACCESS.2022.3191669.

Appendix I: ACTive Community Dataset

We have collected the health and behaviour data from Australian Capital Territory residents under the program "ACTive Community Program". As ACTive Community program is volunteer based program hence the quality of the captured data is not good for the research. There is high noise in the data along with many missing data samples. Hence, we dropped this dataset for research purposes, but this provided the opportunity to work on the novel approach for the data imputation.

- ACTive Community program dataset details

The ACTIVE COMMUNITY platform uses a live collection of data from wearables (Fitbit, Garmin) and provides a dashboard visualization of chronic disease risk prediction within the ACT areas on National map owned by Data 61 (previously NICTA- CISRO) in Australia. Active Community project is first ACT-wide study in Australia, involving those who use Fitbit Activity Trackers to monitor their vital signs. The University of Canberra Health Research Institute (UC-HRI) and Data61 (previously NICTA- CISRO) have developed phase 1 of the project to aggregate data on physical activity levels and health profiles to better inform city planning, health and exercise policies and initiatives.

As mentioned before, ACTive Community program is volunteer based program hence the quality of the captured data is not good for the research. To overcome the sparse nature of the data, the imputation technique is require to apply on the data set. The risk of bias due to missing data depends on the reasons why data are missing. Reasons for missing data are commonly classified as: missing completely at random (MCAR), missing at random

(MAR), and missing not at random (MNAR).

Under the ACTive community program, 43 different filed are getting collected. Few of the important fields which will help in research are:

- Date
- Dob
- Gender
- Suburb
- Walkability_score
- Steps
- Heartrateresting
- Sleeps
- Weight
- BMI
- Caloriesin
- Carbsin
- WalkingTime
- Fatin
- Fiberin
- Proteinin
- Sodiumin
- Waterin
- Caloriesburnt
- Distance
- Heartratemax
- Heartrateavg
- Deepsleepseconds
- Bodyfatpercent
- Activity_score
- Diabetes_score

To overcome the sparse nature of the data, imputation technique needs to apply on the data set. The following novel approach was looked at during the research process but resource requirements was shelved. The next section described the process at a high level.

- **Smart Analytical Framework-Imputation Technique**

In Smart Analytical Framework, innovative ensemble technique can be used to impute missing values in the ACTive Community dataset. The proposed imputation algorithm will be a combination of the Multiple Imputation

Chained Equation Approach, Gradient boosting algorithm approach (XGBoost) and neural network ensemble.

- **Multiple imputation by chained equations**

Multiple imputations is a general approach to the problem of missing data that is available in several commonly used statistical packages. It aims to allow for the uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining results obtained from each of them (Jonathan A C Sterne ; John B Carlin; Patrick Royston; Angela M Wood).

The first stage is to create multiple copies of the dataset, with the missing values replaced by imputed values. These are sampled from their predictive distribution based on the observed data—thus multiple imputations is based on a Bayesian approach. The imputation procedure must fully account for all uncertainty in predicting the missing values by injecting appropriate variability into the multiple imputed values; we can never know the true values of the missing data.

The second stage is to use standard statistical methods to fit the model of interest to each of the imputed datasets. Estimated associations in each of the imputed datasets will differ because of the variation introduced in the imputation of the missing values, and they are only useful when averaged together to give overall estimated associations. Standard errors are calculated using Rubin's rules (*Rubin D. Wiley*) which take account of the variability in results between the imputed datasets, reflecting the uncertainty associated with the missing values. Valid inferences are obtained because we are averaging over the distribution of the missing data given the observed data.

Multiple imputations has the potential to improve the validity of medical research. However, the multiple imputation procedure requires the user to model the distribution of each variable with missing values, in terms of the observed data. The validity of results from multiple imputation depends on such modelling being done carefully and appropriately.

- **MICE Algorithm steps**

The chained equation process can be broken down into four general steps:

1. A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."

2. The "place holder" mean imputations for one variable ("var") are set back to missing.

3. The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model, which may or may not consist of all of the variables in the dataset. In other words, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model. These regression models operate under the same assumptions that one would make when performing (e.g.,) linear, logistic, or Poison regression models outside of the context of imputing missing data.

4. The missing values for "var" are then replaced with predictions (imputations) from the regression model. When "var" is subsequently used as an independent variable in the regression models for other variables, both the observed and these imputed values will be used.

5. Steps 2-4 are then repeated for each variable that has missing data. The

cycling through each of the variables constitutes one iteration or "cycle." At the end of one cycle, all of the missing values have been replaced with predictions from regressions that reflect the relationships observed in the data.

6. Steps 2 through 4 are repeated for a number of cycles, with the imputations being updated at each cycle. The number of cycles to be performed can be specified by the researcher.

7. At the end of these cycles the final imputations are retained, resulting in one imputed dataset. Generally, ten cycles are performed (Raghunathan et al., 2002); however, research is needed to identify the optimal number of cycles when imputing data under different conditions. The idea is that by the end of the cycles the distribution of the parameters governing the imputations (e.g., the coefficients in the regression models) should have converged in the sense of becoming stable. This will, for example, avoid dependence on the order in which the variables are imputed. In practice, researchers can check the convergence by, for example, comparing the regression models at subsequent cycles, as discussed in He et al. (2009).

- **Gradient Boost (XGBoost)**

XGBoost is used for supervised learning problems, where we use the training data (with multiple features) xi to predict a target variable yi.

The whole idea of Gradient Boosting is to minimize the function. For a general treatment, we refer to this function as the loss function represented by L. We need to make sure that Gradient Boosting loss functions must be differentiable. An example is the squared error between the actual and predicted value i.e. L= $(y_i - h(x_i))^2$. We want to minimize $f(x) = \sum_{i=1}^{N} L(y_i, h(x_i))$ i.e. the

loss over all points (x_i,y_i). Here h(x) is the classifier/regressor (the predictor) and N is the total number of points. Further steps of the algorithm are:

- Initialize $h^{(0)}(x)=c$, a constant, such that c minimizes f(x) i.e. pick c

  $$\sum_{i=1}^{N} L(y_i, c)$$

  that minimizes

- At the ith iteration, for j = 1, 2, ...,N compute $r_{ji} = -\frac{\partial L(y_j, h^{(i-1)}(x_j))}{\partial h^{(i-1)}(x_j)}$

- The previous step gives us a value r_{ji} for each point j. With the set of tuples (x_j,r_{ji}), use these points as training data to construct a regression tree that can predict r_{ji} given x_j. This tree approximates the gradient. This takes place of the $\frac{\partial f(x^{(i-1)})}{\partial x_j^{(i-1)}}$ expression, with this one tree sort of embodying the gradient for all points x_j. refer this tree as $T_g^{(i)}$ (g for gradient, i is for the iteration). As before we want this gradient-tree to play a role in the update equation, but we are still left with the task of finding η.

- Assume that the tree $T_g^{(i)}$ has K leaves. We know that the leaves of a tree fragment the feature space into disjoint exhaustive regions. Lets refer to these regions as R_k, for k=1,2,...,K. If we send each point x_j down the tree $T_g^{(i)}$, it will end up at some region R_k. We now want to associate a constant η_k for each such region R_k such that the loss in a region, defined

as: $\sum_{x_j \in R_k} L(y_j, h^{(i-1)}(x_j) + \eta_k)$ is minimized. These are solved as k (simple) independent minimization problems for the k regions.

- Finally, we come to the update step:

  $h^{(i)}(x) = h^{(i-1)}(x) + \sum_k \eta_k I(x \in R_k)$ Here $I(x \in R_k)$ isan indicator function that has a value of 11 when x falls in the region R_k, 0 otherwise

  - Keep going back to step 2 till you have iterated the number of times - say M.
  - Finally return $h^{(M)}(x)$ as your predictor.

In essence a function $h^{(M)}(x)$ based on the values (x_j,y_j), that minimizes prediction errors f(x). The minimization is done in multiple steps: at every step we add a tree (Steps 4 and 5) that emulates adding a gradient-based correction. Using trees ensures that generalization of the gradient expression happens - because we need the gradient for an unseen/test point at each iteration as part of the calculation of $h^{(M)}(x)$. (Tianqi Chen – University of Washington).

- **Neural Network**

A neural network is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron. The processing ability of the network is stored in the interunit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns. As mentioned before that proposed imputation algorithm will be combination of the MICE algorithm, gradient boosting (XGBoosting) and Neural Network ensemble.

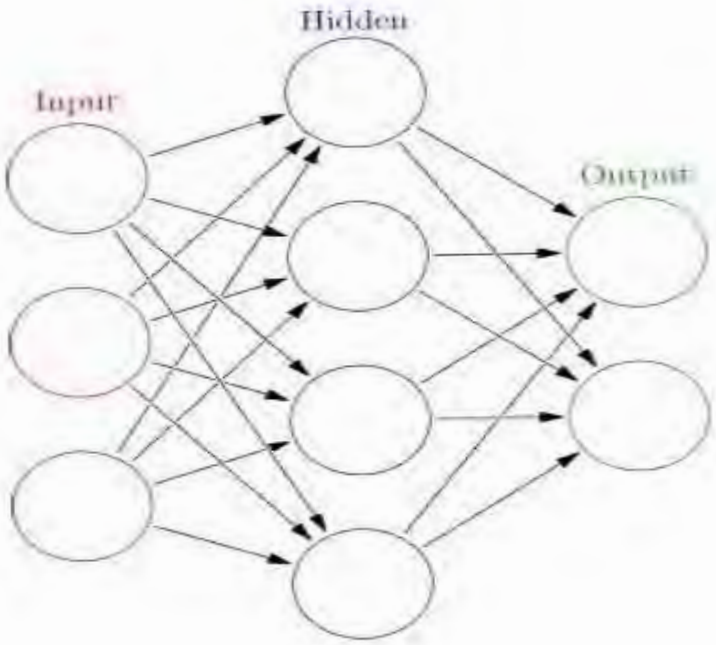


Figure 17 : Block Diagram of simple neural network is as shown blow

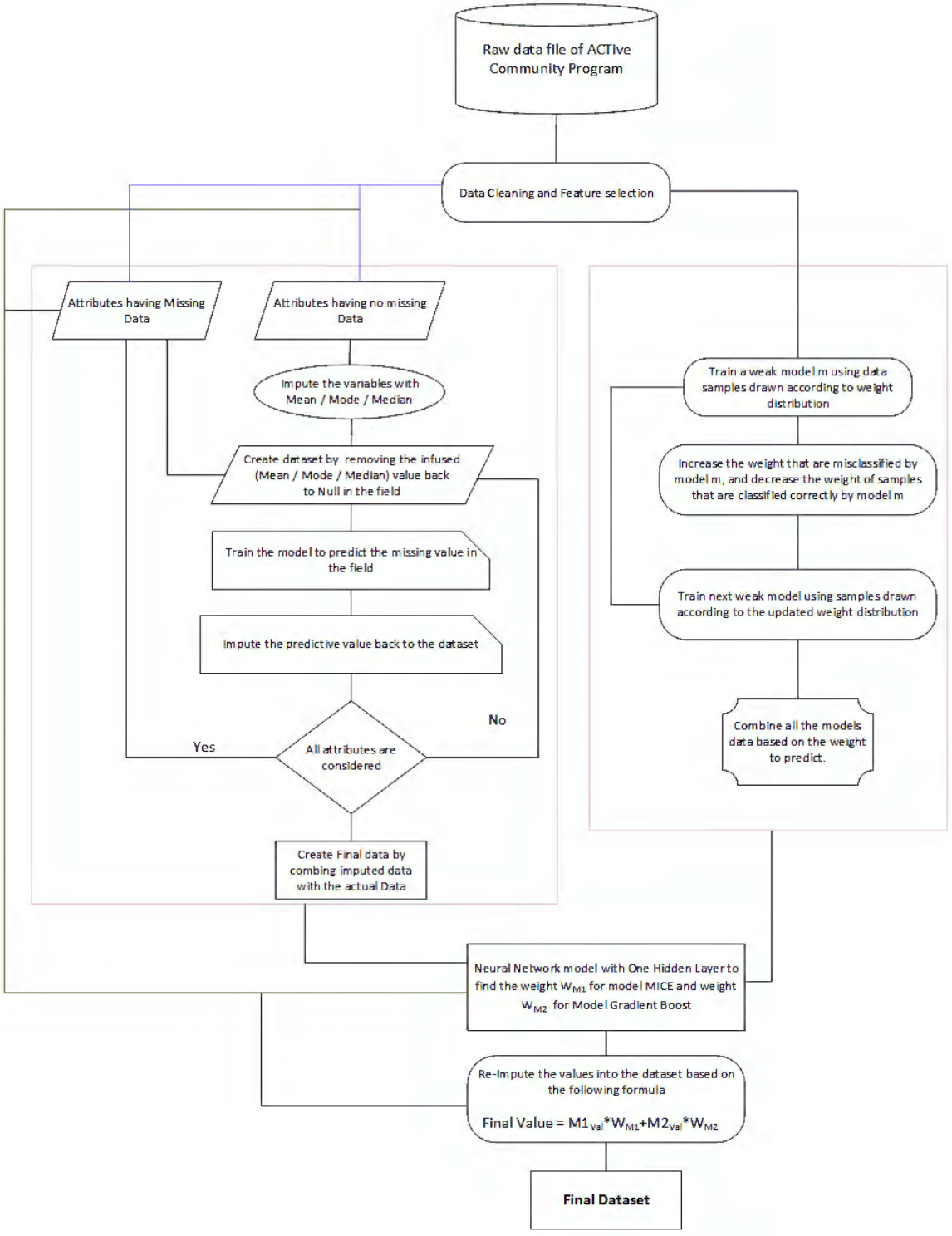


Figure 18: Proposed process flowchart