A Predictive Analysis of Electronic Healthcare Records for Stroke Symptoms

Pattanapong Chantamit-o-pas

Faculty of Engineering and Information Technology
University of Technology Sydney

A thesis submitted in fulfilment of the requirement for
the degree of Doctor of Philosophy

November 2019

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Pattanapong Chantamit-o-pas, declare that the thesis is submitted in fulfilment of the requirements for the award of Doctoral of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: ....

Date: ………..21st..November..2019.................

ACKNOWLEGEMENT

First of all, I would like to express my sincere gratitude to my principal supervisor, Dr. Madhu Goyal. My priceless journey started when she accepted me to be her PhD student in December 2016. She has given me not only excellent guidance in my research, but also great support and encouragement. This really helps me to finish my research and learn how to be a good researcher.

Most importantly, my journey could not have happened without the financial support given by the Royal Thai government scholarship, offered by the Ministry of Science and Technology of Thailand. It has been an excellent opportunity to have the opportunity to gain more knowledge and experiences aboard in Australia. Furthermore, without support the secondary data from the Department of Medical Services (affiliated with Ministry of Public Health of Thailand), and Office of Educational Affairs (affiliated with Royal Thai Embassy to Australia), the research would not have been accomplished. With this acknowledgement, we would like to express our sincere appreciation to them.

Additionally, I would like to thank my parents who always stand beside me and walk through all bad situations with me. I am who I am because of them, so it is undoubtedly that this success also belongs to my parents. I cannot be here without their endless care, support, and encouragement. Likewise, I would like to thank my family, Professor Tony Moon, my friends, and my colleagues who always help me to get over all the bad situations.

Finally, I dedicate this thesis to Dr. Madhu Goyal, who is a key person of this work, my parents, and my family.

LIST OF PUBLICATIONS

1. Chantamit-o-pas, P. & Goyal, M. 2018, 'A Case-Based Reasoning Framework for Prediction of Stroke', in D.K. Mishra, A.T. Azar & A. Joshi (eds), *Information and Communication Technology: Proceedings of ICICT 2016*, Springer Singapore, Singapore, pp. 219-27.
2. Chantamit-o-pas, P. & Goyal, M. 2017, 'Prediction of Stroke Using Deep Learning Model', *International Conference on Neural Information Processing*, Springer, Guangzhou, China, pp. 774-81.
3. Chantamit-o-pas, P. & Goyal, M. 2018, 'Long Short-Term Memory Recurrent Neural Network for Stroke Prediction', *International Conference on Machine Learning and Data Mining in Pattern Recognition*, Springer International Publishing, Cham, pp. 312-23.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abstract

Cerebrovascular symptoms, commonly known as stroke, can affect different parts of the human body depending on the area of brain affected. The patients who survive usually have a poor quality of life because of serious illness, long-term disability and become a burden to their families and the health care system. There is a strong demand for the management focused on prevention and early treatment of disease by analysing different factors. However, a high volume of medical data, heterogeneity, and complexity have become the biggest challenges in stroke symptoms prediction. Algorithms with very high level of accuracy are, therefore, vital for medical diagnosis. The development of such algorithms nevertheless still remains obscure despite its importance and necessity for healthcare. Electronic Healthcare Records (EHRs) describe the details about patient's physical and mental health, diagnosis, lab results, treatments or patient care plan and so forth. The huge amount of information in these records provides insights about the diagnosis and prediction of various diseases. Currently, the International Classification of Diseases, 10th Revision or ICD-10th codes is used for representing each patient record. The huge amount of information in these records provides insights about the diagnosis and prediction of various diseases. Various machine learning techniques are used for the analysis of data derived from these patient records. The predictive techniques have been widely applied in clinical decision making such as predicting occurrence of a disease or diagnosis, evaluating prognosis or outcome of diseases and assisting clinicians to recommend treatment of diseases. However, the conventional predictive models or techniques are still not effective enough in capturing the underlying knowledge because it is incapable of simulating the complexity on feature representation of the medical problem domains. This research used aggregated files of Electronic Healthcare Records

(EHRs) from Department of Medical Services, The Ministry of Public Health of Thailand between 2015 and 2016. The empirical research is intended to evaluate the ability of machine learning and deep learning to recognize patterns in multi-label classification of stroke. This research aims at the investigation of five techniques: Support Vector Machine (SVM); k-Nearest Neighbours (k-NN); Backpropagation; Recurrent Neural Network (RNN); and Long Short-Term Memory - Recurrent Neural Network (LSTM-RNN). These are powerful and widely used techniques in machine learning and bioinformatics. First, we decoded ICD-10th codes into the health records, as well as other potential risk factors within EHRs into the pattern and model for prediction. Second, we purposed a conceptual Case Based Reasoning (CBR) framework for stroke disease prediction that uses previous case-based knowledge. A conceptual case-based reasoning framework to predict from patients' health risk factors and to recognize a particular case that probably develop stroke and prepare or warn patients to handle disease burden outcome. It describes the design, implementation and evaluation of a novel system to facilitate stroke prediction, which relies on data collected from EHRs. Finally, the effectiveness of Backpropagation; RNN; and LSTM-RNN for prediction of stroke based on healthcare records is modelled. The results show several strong baselines that include accuracy, recall, and F1 measure score. Consequently, deep learning allows the disclosure of some unknown or unexpressed knowledge during prediction procedure, which is beneficial for decision-making in medical practice and provide useful suggestions and warnings to patient about unpredictable stroke.

Chapter 1.

INTRODUCTION

1.1. Introduction

Stroke is the second or third most common cause of death in most countries (Khosla et al. 2010; Langhorne, Bernhardt & Kwakkel 2011). The patients who survive usually have a poor quality of life because of serious illness, long-term disability and become a burden to their families and the health care system. There is a strong demand for the management focused on prevention and early treatment of diseases by analysing different factors. Several health conditions and lifestyle factors have been identified as risk factors for stroke. These factors can be grouped into the risk factors that cannot be changed, the risk factors that can be changed (treat or control), and other risk factors that are less well-documented. The risk factors that cannot be changed are focused on demographic data such as age, heredity (family history), race, sex (gender), and prior serious conditions such as prior stroke, Transient Ischemic Attack (TIA) or heart attack. For the risk factors that can be changed, some patients have had some risk behaviour and/or other diseases before stroke attack. The doctors try to help them control their health condition and behaviour (such as some personality and eating behaviours) to change to a better way of life that can prevent them from stroke disease, for example, hypertension, heart diseases such as myocardial infarction (MI), valvular heart diseases, atrial fibrillation, asymptomatic carotid artery disease and others diseases or behaviour such as diabetes mellitus (DM), blood lipids, and smoking. Other risk factors that less well-documented are geographic location, socioeconomic factor, alcohol abuse and drug abuse (The American Heart Association 2016). Recognition of these risk factors is important to reduce the incidence of stroke, which has been increasing (Gorelick et al. 1999). If the first stroke or a like condition called transient ischemic attack (TIA) happened, early management such as “stroke fast track”, “pre-hospital management” have been proven to reduce patients morbidity and mortality. Also, the American Heart Association/American Stroke Association (AHA/ASA) reported a reduction of stroke, coronary heart disease, and cardiovascular risk by 25% based on 10 years data and mentioned that the reason for the success was multifactorial included improved prevention and improved care within the first hours of acute stroke (Jauch et al. 2013; Stroke Risk in Atrial Fibrillation Working Group 2008; The American Heart Association 2016). After the first stroke attack or after TIA, secondary prevention for recurrence of stroke is essential to avoid a worsened outcome for the patients. The ability to predict who are at risk for early recurrence was useful for designing preventive strategies and management protocols (Leira et al. 2004). In the post-stroke phase, an ability to predict who are at risk of hospital readmission was also useful. The tertiary prevention regime could be driven by medical record data, rehabilitation data and data from nursing homes. Prediction of functional outcomes or long-term outcomes and even survival after stroke could help in planning for supporting the patients, family guide and health policy. Lastly, the recognition of risk factors prior to development of a stroke is needed for planning a primary prevention regime that would help in reduction of stroke incidence. The research carried out for this research focuses on stroke prediction for the patient who has no history of stroke or TIA but has some risk factors.

Normally, a prediction technique uses previous cases for decision-making for new cases, called “Case-based reasoning (CBR)”. CBR is a methodology for solving a problem that uses previous data or memorized problem situations called cases. The processes of CBR system proceed in four main steps such as *retrieve, reuse, revise, and retain* (Fig. 1)

(Aamodt & Plaza 1994) . The new case starts at the top of stage, where an input is entered into the system. The previous case is compared to the new case and starts a *retrieve* step. A practical CBR system is a comparison between all the cases in the system and a new case; the result will list the ranking of similar cases.

Data mining has become an essential instrument for researchers and clinical practitioners in medicine and numerous reports using these techniques are available. However, one of the biggest problems in data mining in medicine is that medical data is voluminous, heterogeneous and complex. The need for algorithms with very high accuracy is required as medical diagnosis is considered quite a significant task that needs to be carried out accurately and efficiently.

Most common techniques that are used to induce predictive models from data sets are Naïve Bayesian classifier and the decision tree. A Bayesian classifier is one of the simplest yet a fairly accurate predictive data mining method (Kononenko 1993). A clinical decision support system (Amin, Agarwal & Beg 2013b) is used for prediction and diagnosis in heart disease. This approach is able to extract hidden patterns and relationships among medical data for prediction of heart disease using major risk factors. It applies genetic algorithms and neural networks and is called 'hybrid system'. It uses a genetic algorithm feature for initialization of the neural network weights.

Deep learning algorithm is a technique that focuses on how computers learn from data. It is the intersection of statistics, computer science, and mathematics - which generates the algorithm of building patterns and models from massive data sets, as well as is applicable to billons or trillions of data records.(Deo 2015; LeCun, Bengio & Hinton 2015). A Deep learning technique employs learning from data together with multiple levels of abstraction deriving from computational models that are associated with multiple processing layers.

Several health conditions and lifestyle factors have been identified as risk factors for stroke. These factors have three groups that consist of the risk factors which cannot be changed, the risk factors which can be changed (treated or controlled), and other risk factors which are less well-documented (The American Heart Association 2016). Recognition of these risk factors is important to reduce the incidence of stroke, which has been increasing (Gorelick et al. 1999).

For stroke, the predictive techniques of stroke vary from simple to more complex models. The risk factors of stroke are complex and applicable to find different convolutions of disease and uncertainty from direct and/or indirect sources. The analysis of stroke patients who were admitted in the TOAST study was done by using stepwise regression methods (Leira et al. 2004). This research was conducted among 1,266 patients selected from database, provided that those patients must have had suffered a transient ischemic attack (TIA) or recurrent stroke within 3 months after the first stroke. Additionally, 20 clinical variables were chosen for finding performance and evaluation.

The prognostic significance of blood pressure for stroke risk was examined by using the Cox proportional hazards regression model, which was adjusted for possible confounding factors. The results from a number of measurements showed that the predictive value of home blood pressure increased progressively. The initial home blood pressure values (one measurement) showed a significantly greater relation with stroke risk than conventional blood pressure values (Ohkubo et al. 2004).

The Cox proportional hazards model and machine learning approach have been compared for stroke prediction on the Cardiovascular Health Study (CHS) dataset (Khosla et al. 2010). Specifically, they considered the common problems of prediction in medical datasets, feature selection, and data imputation. This research proposes the use of an innovative algorithm for automatic feature selection - which chooses robust features based on heuristic: conservative mean. This algorithm was applied in combination with Support Vector Machines (SVMs). The feature selection algorithm achieves a greater area under the ROC curve (AUC) in comparison with the Cox proportional hazards model and L1 regularized Cox model. The method was also applied to clinical prediction of other diseases - where missing data are common, and risk factors are not well understood.

The Bayesian Rule Lists generated stroke prediction model employing the Market Scan Medicaid Multi-State Database (MDCD) with Atrial Fibrillation (AF) symptom (Letham et al. 2015b). The database categorised 12,586 patients on the basis of AF diagnosis. The observation was divided into two phases: a 1-year observation prior to the diagnosis; and 1-year observation after the diagnosis. The result found that 1,786 patients had a stroke within a year after suffering the atrial fibrillation. With regards to evaluation, the Bayesian Rule List (BRL) point is estimated by constructing a receiver operating characteristic (ROC) curve and measuring area under the curve (AUC) for each fold.

The classification algorithms i.e. Neural Network, Naïve Bayes, and Decision Tree are used for predicting the presence of stroke with various related attributes. The principle component analysis algorithm is commonly used for reducing the dimensions. Also, it is used for determining more relevant attributes towards the prediction of stroke and predicting whether the patient is suffering from stroke or not (Sudha, Gayathri & Jaisankar 2012).

In summary, stroke is a devastating phenomenon which usually occurs when blood flow to a portion of the brain stops for an extended period of time and the brain tissues in that area is severely injured or dead. Patients who survive from stroke are likely to suffer various impairments such as cognitive, visual and motor losses depending on the location and extent of brain tissue damage. The impairment decreases the patient's ability to perform activities of daily living such as dressing, and feeding themselves. Most of stroke survivors cannot live independently or return to the workforce (Alankus 2011).

1.2. Motivation

Stroke disease cannot be predicted with certainty whether or when it will attack the patient who has related risk factors but it is possible to predict who is likely to develop stroke from some risk factors. There are many statistical models based on clinical research and medical field research data to find risk factors in order to develop preventive medical planning for the high-risk patients and many risk factors had been filtered to focus on the most relevant. However, there are some limitations and criticisms that there is more complexity for the prediction of a stroke occurring. It might not be sufficient to use just only previously identified risk factors. There were also questions for faster and scalable modelling processes from the research and how to better integrate with the stroke registry databases.

Recently, machine learning has been applied for prediction model utilizing data from the Electronic Healthcare Records (EHRs). The results of the study are able to show percentages of the new patients that would likely suffer a stroke in the future and a new modelling pattern of risk factors was proposed and founded by the system from data mining and machine learning techniques. However, the prediction from statistical models and prediction from machine learning study are as the same that all the models do not compare with previous case(s). But in medicine, the experts' judgments normally use analogies from previous cases to analyse for clinical reasoning and clinical decision making to predict, diagnose and make prognoses in terms of achieving a solution for solving problem. So, this research will focused on the Case Based Reasoning (CBR) and machine learning to predict risk of stroke diseases and assist clinicians in prediction, diagnosis, making prognosis and treatment planning for the patients from EHR data. The machine learning technique will classify patients' data from EHRs of previous patients and new patients by groups of risk factors. Then it will find the risk of stroke by comparing with similarity of previous case(s). Finally, the outcome will show the percentages of the risk of having a stroke and show a care plan from previous case(s) proposed as solution to solving problem. The result of the research would not only support medical professionals for stroke symptoms decision making, but also provide suggestion and warnings to patients before they visit a hospital or go for costly medical check-ups.

1.3. Research Question

Based on the literature review, I identified the following research question as follows:

RQ1: How to select and identify risk factors for prediction of stroke?

Stroke symptom are complex and share relevant data with other diseases. Thus, the risk factors have complicated attributes and are hard to predict. Moreover, there are no warning signs to classify new patients as ones with strokes risk factors or strokes non-risk factors.

RQ2: How can previous patient records be used for stroke prediction?

Normally, the Electronic Healthcare Records (EHRs) is recorded in the hospital database when patients visit. The records contain demographic data, disease type, diagnosis codes, etc. The dataset includes historical records of stroke patients, both of patients who have potential stroke risk factors and patients who have non-potential stroke risk factors. The information of previous case(s) can be applied for prediction with stroke symptoms for new case(s).

RQ3: How to develop a prediction model to focus on stroke symptoms using Electronic Healthcare Records (EHRs)

The prediction model can be applied for prediction as well as decision-making, using a machine learning approach and Case-Based Reasoning system for making-decision and treatments. EHRs are needed to create a model for stroke symptoms and other diseases. At present, the predictive model uses many techniques such as Case-Based Reasoning system, k-Nearest Neighbor (k-NN), Support Vector

Machine (SVM), and Deep Learning technique. These techniques can be applied for creating a model for predicting stroke symptoms.

1.4. Objective and Aims

The purpose of this research is to apply clinical decision-making strategies by using automated reasoning in an attempt to predict stroke disease from the patient data in Electronic Healthcare Records (EHRs). This may potentially facilitate clinicians in predicting future stroke patients.

In order to achieve this objective, this research has two aims.

1. The first aim is to predict the likelihood of developing stroke symptoms in patients who are in risk groups by using information from Electronic Healthcare Records in terms of automated reasoning concepts. The research will identify and select risk factors for stroke symptoms using feature selection technique. These aims are achieved by developing an algorithm for feature selection with the International Classification of Diseases, and 10th Revision code (ICD) are used for electronic health records and for classifying diseases and other health problems appearing in many types of health and vital records. Irrelevant data are filtered out. The filtering process takes a significant amount of time due to the large size of the data as well as the large number of stroke symptoms.

2. This research aims to apply different machine learning and deep learning techniques for prediction. In term of prediction, we propose a prediction model for stroke symptoms using Case-based Reasoning system and machine learning comprising of relevant risk attributes. Furthermore, this research aims to compare all the machine learning techniques and present research showing the different

performance metrics. However, the result of the research would not only support medical professionals' decision making for stroke symptoms, but also provide suggestions and warnings to patients before they visit a hospital or decide to take a costly medical check-up.

1.5. Plan of the thesis

The focus of this thesis is on the predictive analysis for stroke symptoms. To accomplish this goal, the thesis is organized into eight chapters. In this section, a brief summary of each chapter is detailed as follows:

Chapter 2 Literature Review: This Chapter provides a review of existing literature in the areas related to stroke symptoms; stroke risk factors, Case-Based Reasoning system (CBR), Machine Learning, Deep Learning, and predictive models in healthcare. The research in stroke symptoms and stroke risk factors aims to understand the diseases, whereas the research in CBR is about a concept able to be applied in healthcare sectors. Deep Learning's intended aim is in applications for predicting future outcomes. The details of each section are discussed in chapter 2.

Chapter 3 The Electronic Healthcare Records (EHRs): This section describes the EHRs of cerebrovascular disease patients containing medical and other information recorded in the hospital. In most cases, Electronic Health Records are used for storing most of this medical or patient data whereas the International Classification of Diseases, 10th Revision code (ICD) is used for electronic health records and for classifying diseases and other health problems appearing in many types of health and vital records. These details of each section are discussed in chapter 3.

Chapter 4 Feature Selection Framework: This chapter proposes the feature selection frame work which eliminates anomalous data from the EHRs by applying a filtering process to remove irrelevant data. The filtering process takes a considerable amount of time due to the large size of the data as well as the large number of stroke symptoms.

Chapter 5 Case-based Reasoning Framework for stroke: This chapter proposes the case-based reasoning framework for stroke. It contains the main components of the framework and techniques applied in CBR for stroke domain. It gives a brief description of case base management issues, data collection and the prediction functional implementations.

Chapter 6 Deep Learning Framework for Stroke: Deep learning algorithm is applied on EHRs for prediction. Deep Learning (DL) is a process of training a neural network to perform the given task. Stroke is predicted by using Long Short-Term Memory - Recurrent Neural Network (LSTM-RNN); Recurrent Neural Network (RNN), and Backpropagation, which are currently the most suitable approaches. The details of each algorithm are discussed in this chapter 6.

Chapter 7 Experiments and Results: Chapter 7 presents the implementation of the experiments and the results. The experiment includes the investigation of five techniques: Support Vector Machine (SVM); k-Nearest Neighbours (k-NN); Backpropagation; Recurrent Neural Network (RNN); and Long Short-Term Memory - Recurrent Neural Network (LSTM-RNN). These are powerful and widely used techniques in machine learning and bioinformatics. The empirical research is intended to evaluate the ability of machine learning and deep learning to recognize patterns in multi-label classification of stroke. First, we proposed a conceptual Case Based Reasoning (CBR) framework for stroke disease prediction that uses previous case-based knowledge. Finally, we modelled the effectiveness of Backpropagation; RNN; and LSTM-RNN for prediction of stroke based on healthcare records. The results show several strong baselines that include accuracy, recall, and F1 measure score.

Chapter 8 Conclusion: This chapter outlines the contribution of the research, concludes the thesis, and discusses of the experimental results of the proposed CBR and Deep Learning approaches. Finally, recommendations for future work are presented.

Chapter 2.

Literature Review.

2.1. Introduction

This chapter provides the necessary background information for this research. Section 2.1 introduce the significance and characteristics of stroke symptoms. Section 2.2 describes the risk factors for stroke patients who have health conditions and the potential to suffer a stroke in the future. Section 2.3 gives an overview of the prediction model and how it relates to this research. Section 2.4 presents the prediction model in the healthcare sector that covers machine learning, data mining techniques, and the standard medical prediction score. Section 2.5 presents the feature selection approach applied in healthcare to filter the missing data in the medical database. Section 2.6 presents Case Based Reasoning system (CBR), widely used in the healthcare sector which has previously provided solutions for diagnosis and treatment of diseases based on past experiences. Section 2.7 presents the data mining and predictive analysis that is used for stroke symptoms. Finally, section 2.8 introduces the deep learning method used in this thesis for stroke prediction.

2.2. What is stroke?

Basically, a stroke affects brain dysfunction and arises when a part of brain is deprived of its blood circulation. This can occur by an obstruction of blood vessel, named an "*ischemic stroke*", which accounts for the majority of strokes, or another type happens when a blood vessel in the brain is ruptured, in which case it is called a "*haemorrhagic stroke*". Ischemic strokes (80%) can either be a blood vessel which is slowly blocked as a result of cholesterol and other blood products building up inside the blood vessels, or it can come from a small clot dislodging, for example, from the heart and going up into a brain blood vessel and blocking it. Haemorrhagic strokes have a lower occurrence (20%) but have critical effects, due to uninhibited bleeding in the brain. A haemorrhagic stroke can be divided into two types: *subarachnoid haemorrhage* and *intracerebral haemorrhage.* Subarachnoid haemorrhage is the result of bleeding in the subarachnoid space which is the area between the skull and the surface of brain tissue, while Intracerebral haemorrhage refers to bleeding into the cerebral tissue (Mayo Clinic Staff Oct 2018). Irrespective of the cause of stroke, the inadequate blood supply causes a deficiency of oxygen and glucose in the location of the lesion. Typically, however, we can see that area of the brain that controls the function of the arms or speech or areas of the brain that control the function of the legs is affected. The main regions of the brain consists of cerebrum, cerebellum, and brain stem.

The cerebellum is involved with body movement and muscle coordination. In addition, the brain stem comprises motor and somatosensory nervous system. All information passing between the body and the brain passes through the brain stem. The brain stem includes multiple functions, such as breathing, blood pressure, normal heart rate, body reflexes, pain response, etc. Therefore, all functions of our body are controlled by the brain (Clifford 2010).

In summary, stroke is a devastating phenomenon of brain dysfunction resulting from an inadequate blood flow in the brain for an extended period of time and the cerebral tissues in that location are severely injured or dead. Stroke survivors are likely to suffer various impairments depending on the lesions and extension of cerebral tissue damage such as cognitive, visual and motor losses. The impairment decreases the ability of self-care in patients with stroke to coping with activities of daily living such as dressing, bathing and eating unaided. Most of patients who survive from stroke cannot live alone and rarely go back to the workforce (Alankus 2011).

2.3. Risk factors effecting stroke patients

Stroke disease has many risk factors to indicate those patients with risk of stroke or not. Even though the demographic data can show basic information, this is not enough to diagnose what disease(s) they will suffer. Normally, patients that have got some diseases will be taken for laboratory investigation and some value taken from laboratory result will be recorded. For stroke, many investigations will be used for detection, diagnosis and treatment. However, it is more complicated to predict stroke occurrence. Some factors from other diseases become stroke risk factors. That is one of the diseases somehow related to or results from other diseases. This means that there might be interrelationship between risk factors and/or diseases that a patient has which are of significance for stroke prediction.

The risk factors for stroke disease are found in various variables such as demographic data and laboratory results from other disease such as diabetes, obesity, cardiac diseases, hypertension etc. These factors can be grouped into three categories as the unchangeable risk factors, the changeable risk factors (that can be treated or controlled) and other risk factors that are less well-documented (Goldstein et al. 2006; Goldstein et al. 2001; The American Heart Association 2016)(see in Figure 2-1).

2.3.1. The unchangeable risk factors of stroke

The unchangeable risk factors are the elements of personal data such as age, heredity (family history), ethnicity (race), gender (sex), and prior stroke, TIA or heart attack. The details of information are as follows:

Age: this factor leads to stroke as a patient becomes older. The incidence of having a stroke approximately doubles when you are older than age 75. The CHA_2DS_2-VASc score, which is a predicting score of risk factors, showed that a patient has zero point under 65 years old, one point in the age range 65-75 years old, and two points at over 75 years old (Guerra et al. 2016; Morillas et al. 2015).

Heredity (family history): the stroke risk depends upon family history if any of family members, such as parents, grandparents, sisters or brothers has ever had a stroke (Francis, Raghunathan & Khanna 2007; Genetic Alliance 2009).

Ethnicity (race): nationality and stroke disease are related. The evidence showed that Africans have a far greater incidence of stroke due to having high blood pressure, diabetes, and obesity than other nationalities (The American Heart Association 2016).

Gender (sex): the statistics indicated that females have a higher stroke incidence than males and stroke also caused death more often in females than in males (Lisabeth & Bushnell 2012; The American Heart Association 2016).

Prior stroke, TIA or heart attack: A person who experienced a stroke at least once has a significantly higher recurrence of stroke than those who never had one. Transient ischemic attacks (TIAs) are "warning strokes" that produce temporary symptoms of stroke but no lasting injury in cerebral area. TIAs are strong risk factors of stroke. The chance of inducing stroke in a person who had previous TIAs is nearly 10 times of those of the same age and sex who never has had one. TIAs recognition and treatment can reduce the possibility of a repeat stroke. TIA should be treated as an urgent condition and followed up with a qualified health professional. If patients have experienced a heart attack, they were at greater chance of having a stroke as well (Clifford 2010; The American Heart Association 2016).

2.3.2. The changeable risk factors of stroke

This section describes the details of the risk factors that can be changed. Some patients have had some personal behaviour (as physical activity and eating habits) and/or other past illnesses before a stroke attack. These factors may be related to a health condition and precipitate a stroke attack. Trying to control the health condition, modification of life style or behaviour and undergoing diseases treatment may improve the quality of life that can reduce the incidence rate of stroke. The information is described as follows:

High blood pressure (Hypertension): A diagnosis of hypertension is recorded when patients have systolic blood pressure of 140 mmHg or higher and/or diastolic blood pressure above 90 mmHg (Hitman et al. 2007). The studies from 32 countries showed that Hypertension stands as the main cause of stroke and it is a serious risk factor that can be controlled for the disease (O'Donnell et al. 2016). Therefore, the effective treatment of hypertension is an essential key for a decrease in the mortality rates of stroke (The American Heart Association 2016).

Cigarette smoking: For decades, many research studies found that cigarette smoking is an important predictor for stroke (O'Donnell et al. 2016). The nicotine and carbon monoxide of tobacco smoke damage the cardiovascular system and cause atherosclerosis (plaque build-ups in artery walls) in brain. These will be increased the incidence of stroke. Moreover, the risk factor of stroke is greatly increased when patients have used birth control pills combined with cigarette smoking (Morales-Vidal & Biller Dec 16, 2003).

Diabetes mellitus: Diabetes, either type one or two is an important risk maker for stroke due to the role of blood sugar in the pathogenesis of cerebral atherosclerosis (Chait & Bornfeldt 2009). Diabetic patients also have hypertension, dyslipidaemia and are overweight. These are related with accelerated stroke incidence (Md Mahfuj Ul et al. 2017). The most common type of stroke in diabetes is Ischaemic stroke compared to intracerebral haemorrhage stroke (Tun et al. 2017). Although diabetes is treatable, the existence of this disease still increases the rate of stroke incidence (The American Heart Association 2016).

Carotid or other artery diseases: Blood to the brain is supplied by the carotid arteries in the neck. Fatty deposit from atherosclerosis can narrow a carotid artery, called carotid artery stenosis, and also create a thrombus (clot) in the blood vessel. This can block the blood flow into brain and cause a stroke (Flaherty et al. 2012).

Peripheral artery disease or PAD: This disease is the atherosclerosis of blood vessels serving limbs and abdominal organs. Patients with peripheral artery disease have a greater possibility of carotid artery stenosis, which increases their chance of stroke (Banerjee, Fowkes & Rothwell 2010; Rahman et al. 2017).

Atrial fibrillation: this disease causes the risk for stroke because the irregular heartbeats can lead to the formation of a blood clot. The clot can move and block blood circulation in the brain and a stroke will occur (Lip, Frison, et al. 2010; Lip, Nieuwlaat, et al. 2010; The American Heart Association 2016). Atrial fibrillation is a strong predictor of stroke (Kamel et al. 2016).

Other heart disease: people who have past illness of heart failure or coronary disease are associated with a greater risk of stroke. Congestive cardiomyopathy (a cardiomegaly), valvular heart diseases and some classifications of congenital heart defects also increase the risk of stroke (Amin, Agarwal & Beg 2013b).

Sickle cell disease (also called **sickle cell anaemia**): patients with sickle cell anaemia have a genetic disorder and present an abnormal type of red blood cell (called "S haemoglobin"). Normally, red blood cells carry oxygen to supply human body but S haemoglobin are inadequate in carrying oxygen. The blood arteries in the brain are blocked as cells tend to be stick to the blood vessel walls and then cause a stroke. Predominant populations of this disease are African-American and Hispanic. The incidence of having a stroke in people with sickle cell disease by 20 years of age is approximately 11% (Kassim et al. 2015; Nichols 2018).

High blood cholesterol: High plasma level of cholesterol was significantly associated with increased stroke rate. A low level of HDL ("good") cholesterol seems to increase the incidence in males, but more studies need to be conducted to determine its effect in women (Go et al. 2013).

Poor diet: Cholesterol levels in blood can be increased by consuming foods high in trans and saturated fat. Diets high in sodium (100 mmol/day) are a cause of higher blood pressure and the higher sodium consumption is related to a chance of stroke (Medeiros et al. 2012). Diets with too many calories can increase body weight. Similarly, people who consume at least five servings of fruits and vegetables per day may cut the incidence of stroke. Hu et al. (2014) conducted a meta-analysis to examine the effects of fruit and vegetable intake on stroke rates. They found that the high consumption of fruits and vegetables reduced the incidence of stroke by 14-23%.

Physical inactivity and obesity: previous studies showed that regular physical activity was significantly related to reduce the chance of a stroke (Howard & McDonnell 2015; O'Donnell et al. 2016). Physical activity can decrease blood pressure, reduce blood cholesterol, control blood glucose, and maintain weight (Howard & McDonnell 2015). Therefore, an exercise of at least 30 minutes per day may help controlling of diseases (e.g. hypertension, dyslipidaemia, diabetes, coronary disease, stroke, etc.) and hence reduce stroke risk (The American Heart Association 2016).

2.3.3. Other risk factors that are less well-documented

This section describes the details of other risk factors. The financial, environmental, location, and habitation situation all have a relationship with stroke risk. These depend on life style (Hanchaiphiboolkul et al. 2014). The following four factors can be described:

Geographic location: The hometown of a person can be used to predict if the person has a risk of stroke or not (O'Donnell et al. 2016; O'Donnell et al. 2010; Owolabi et al. 2018). For example, people in south eastern United States tend to have more strokes than those in other parts (Genetic Alliance 2009). These are recognized as the "stroke belt" states (Benamer & Grosset 2009; The American Heart Association 2016).

Socioeconomic factors: people who have a low socioeconomic status had a significant positive association with a greater incidence of stroke and mortality rates of stroke compared to those who have a high socioeconomic status (Addo et al. 2012; Langhorne et al. 2018; Owolabi et al. 2018).

Alcohol abuse: Alcohol consumption can be a cause of different medical complications, including stroke. It is suggested that a man should drink no more than two glasses per day and no more than one glass per day for a woman who is not pregnant. The American Heart Association (2016) found that people who consumed 30 drinks per month showed a significant rise in the rates of stroke, especially the intracerebral haemorrhage stroke.

Drug abuse: Commonly drug addiction, such as heroin, amphetamines and cocaine, have be known to create a stroke risk. Young adults have a significantly clinical higher risk of all stroke types resulting by drug abuse. Fonseca & Ferro (2013; Mullen (1996; Westover, McBride & Haley (2007) found that the incidence of haemorrhagic stroke was significantly increased by amphetamine addiction, while cocaine addiction was a strong indicator for increasing ischemic and haemorrhagic stroke rates.

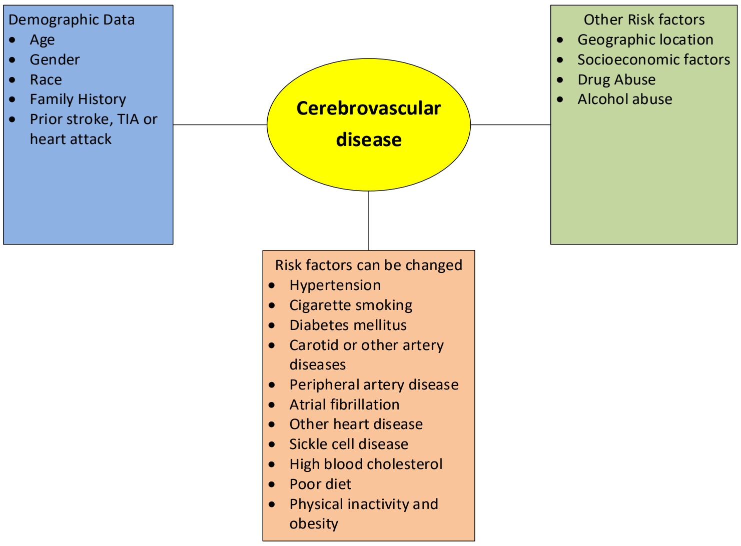


Figure 2-1 : Risk factors effective stroke symptoms

2.4. Why is prediction required?

Stroke is a disease that has no warning signs to patients, yet is the second or third most common cause of death in most countries (Khosla et al. 2010; Langhorne, Bernhardt & Kwakkel 2011). Information Technology can support and help personnel in the healthcare sector by using prediction techniques to show the likelihood of developing a stroke before the attack occurs. The prediction results may help for the reduction of stroke incidence for future patients. The prediction analysis model is a calculation performed on complex data of various data types, with the results of some significance for diagnosis and forecast even through stroke disease is complex to forecast. The prediction from a previous stroke patient may be able to support medical staff in good decision-making for treatment and also help them find the real factors of a disease. The prediction may also provide suggestions and warnings to a new patient before they visit a hospital or go for costly medical check-ups. The recognition of risk factors prior to the development of a stroke can help physicians in the planning of a primary prevention regime that would help in reduction of stroke incidence.

2.5. Predictive models in healthcare

2.5.1. Predictive Data mining

Most researchers and clinical practitioners extensively use data mining techniques for prediction in medicine. It is necessary to comprehend the mechanism of these approaches and the application of the settled and accepted procedures in order to deploy and disseminate the results. There is an area of research that deals with the combination of molecular and clinical data occurring within genomic medicine. This research has both obtained a favourable motivation and marked a new collection of complex issues. (Bellazzi & Zupan 2008).

Srinivas, Rani & Govrdhan (2010) stated that the abundance of data in healthcare business has not been well utilized due to the absence of practical analysis tools to detect covered association and inclination of the data. However there are many applications of Knowledge discovery and data mining in the business and scientific sphere. Useful information can possibly be learned from application of these data mining techniques in the healthcare system. Their study investigates the possibility of using classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to the huge amount of healthcare data. The healthcare industry gathers vast amounts of healthcare data, much of which has not been utilised to reveal concealed information. One Dependency Augmented Naïve Bayes classifier (ODANB) and naive credal classifier 2 (NCC2) are applied in the data pre-processing and effective decision-making parts.

For the business and scientific domain, knowledge discovery (KDD) and data mining tools have been applied to find new knowledge. A similar application of data mining techniques can possibly discover new knowledge in the healthcare system. Most of the research in this area used classification based data mining techniques such as Decision tree, Rule based, Artificial Neural Network, and Naïve Bayes to extract information from a large volume of medical data. The healthcare sector, collects huge amounts of medical data which are not “mined” to explore hidden information. Thus they developed the One Dependency Augmented Naïve Bayes classifier (ODANB) and naïve credal classifier 2 (NCC2) for data pre-processing and decision making. This represents a stretch of Naïve Bayes to inaccurate probabilities that are intended to deliver powerful classifications even when dealing with inadequate data sets. Revelation of undisclosed patterns and relationships repeatedly become of no use. From medical profiles such as age, sex, blood pressure and blood sugar, it is possible to anticipate the likelihood of patients having a heart disease. It allows important knowledge, e.g. patterns, relationships between medical factors connecting to heart disease, to be formed.

Soni et al. (2011) proposed that application of data mining is accomplished within the e-business, marketing and retail sectors. Its success has expanded into other industries and sectors. The healthcare system is a new sector that is beginning to apply data mining techniques to discover a wealth of medical data. Current techniques of data mining were surveyed in medical research in order to predict coronary diseases. Three models are compared; Decision Tree; k-NN; Neural Networks; and Classification based on clustering, using the same dataset. They concluded that Decision Tree exceeds others in precision, and in some cases Bayesian classification is comparably accurate as the Decision Tree but other predictive methods such as k-NN, Neural Networks, and Classification based on clustering do not do well. Furthermore, after using genetic algorithm to cut down the data size to get the best subset of attributes enough for heart disease prediction, the accuracy of the Decision Tree and Bayesian Classification further improves.

2.5.2. Decision Support systems

Amin, Agarwal & Beg (2013a) proposed clinical decision support systems that can accurately predict and diagnose various diseases. Due to their ability in finding undisclosed patterns and relationships in medical data, these techniques have been decidedly competent in designing clinical support systems. One of the most critical applications of such systems is the diagnosis of coronary disease as it is one of the dominant causes of mortality in the world. Nearly all systems that predict coronary diseases employ clinical datasets that contain parameters and inputs from complex tests handled in labs. There is no system that can predict coronary diseases based on risk factors such as gender, family history, hypertension, high cholesterol, diabetes, alcohol intake, tobacco smoking, obesity or physical inactivity, etc. There are visible risk factors which are commonly shared by heart disease patients. This can be effectively used for diagnosis. A system based on such risk factors would not only benefit medical professionals but it would also provide patients with the correct information about the possible presence of heart disease even before the patient decides to visit a hospital or take an expensive medical check-ups. Therefore, they propose a technique that can predict coronary disease using the prime risk factors. The technique is associated with two of the most successful data mining tools, neural networks and genetic algorithms. The hybrid system implemented makes use of the global optimization advantage of genetic algorithm for initialization of the neural network weights. Compared to the backpropagation approach, its learning is faster, more stable and accurate. The system was implemented by using Matlab and can predict the risk of coronary disease with an accuracy of 89%.

Predictive analytics gives considerable operational level benefits for healthcare providers. Enterprises count on predictive analytics to improve their perception of the effectiveness of clinical treatments. The ability to promptly and precisely diagnose and personalise patient treatment when they are admitted for the first time promotes patient assurance in the healthcare system, fosters the patient-doctor relationship and cuts back the incidence of pricey readmissions. Palem (April 2013) in the practice of predictive analytics in healthcare summarised how predictive analytics was useful to different areas of the medical industry such as assisting in healthcare research and discovering new medicines in the life-sciences, support in clinical judgement and diagnosis for healthcare professionals, improving healthcare cost-effectiveness, identifying and protecting insurance fraud, monitoring in real-time health problems of individuals, pinpointing disease outbreaks in public health, and assisting in critical care intervention for individual patients. He also indicated the use of reporting-based tools and applications which help to perceive what occurred in the past and also can be utilised to analyse and categorise historical structured data. The healthcare industry is challenging predictive analytics techniques to forecast for future actions and model scenarios, improving advanced competence such as enterprise analytics, evidence-based medicine and clinical decision support systems. Some of the user-case scenarios that he lists are as follows:

Critical care intervention: this intervention gives signals to physicians when a patient's reaches some critical values after infection, as a result of adverse drug events, or affected by other complications. Physicians can monitor from a distance and solve patient problems in real time.

Diagnostic assistance: physicians can use this method to quickly determine a precise diagnosis using advanced voice and natural language processing approaches. This assistance method also helps physicians to reduce costs for diagnostic tests.

Clinical decision support: the smart decision support systems are used by clinicians for deciding best practices and guiding possible activities for treatments. This system analyses all patient data from medical records, previous cases, clinical trials and reference materials in real-time.

Disease management: the central immunity records and epidemic control used reports of clinical evidence and treatment procedures from different healthcare providers to monitor public health safety, emerging diseases and possible disease outbreaks in real-time.

Optimized healthcare costs: the model is utilised to improve the quality of life, healthcare capability, and insurance cost-effectiveness by promoting individual plans for protection and treatments. The model is used to analyse the risk factors of the patient from their individual circumstances, such as age, genetic disease, life-style, past medical records and suspicious diseases.

Fraud detection and prevention: Healthcare organizations and insurance providers are focusing on reactive measures for fraud detection and prevention. This approach monitors the main risk indicators, over a period of time or on an ad hoc testing basis to assist the insurance providers in deciding the transactions to be examined. If the tested transaction signals fraud, repeated or continuous analysis will be conducted. Unlike retrospective analyses, uninterrupted transaction auditing allows an organization to determine the possibly of fraudulent transactions in real-time or near real-time basis, such as daily or weekly. Organizations are more and more applying continuous monitoring efforts to concentrate on narrow bands of transactions or parts that pose particularly high risks.

Personal healthcare: It has been universally accepted that pervasive and context-aware applications are encouraging plans to boost the quality of life for patients experiencing chronic illness and for their family members, as well as limiting the expenses of long-term care. For example, Telemedicine is used to promote safety and adherence to prescribed medication in daily living. Wearable body monitoring sensors are popular devices increasingly being utilised to evaluate personal health.

Readmission prevention: By collecting patient-specific historical data, managing personalised care plans, comparing the efficacy of medication against industry-wide norms, and using resources on the most practical treatments re-admission rates can be significantly cut back. Approaching and managing patient discharge and follow-up care in an efficient manner significantly reduces the expense of re-admission by reducing repeated admissions.

Crockett (2013) disclosed that forecasting hospital readmissions is a highly investigated topic. In 2013, there were 36 peer-reviewed journal articles published on the subject as well as three more review articles. Emphasising this fast rising enthusiasm are recent papers aiming at simplified readmission scoring for elderly patients, the relationship between readmission and mortality rates and an organized examination of tools for predicting severe adverse events. Prediction discussion related to specific areas such as heart failures or within the paediatric population are very active as well. He noted that the trend serve the dual purpose of enhancing patient care while reducing financial and reimbursement expenditures for hospitals.

2.5.3. Standard medical prediction score

The original prediction score in medical practice uses various models. The CHADS2 and CHA2DS2-VASc are standards that are widely used in practice for prediction with stroke disease, atrial fibrillation, and heart disease. To compute CHADS2 score, we assign one "point" each for the presence of congestive heart failure (C), hypertension (H), age 75 years or older (A) and diabetes mellitus (D), and assign 2 points for history of stroke, transient ischemic attack or thromboembolism (S2). The CHADS2 score takes just 5 factors, while the revised CHA2DS2-VASc score also covers three more risk factors: vascular disease (V), age 65 to 74 years old (A) and female gender (Sc). Larger scores are associated with higher risk (Goldstein et al. 2006; Guerra et al. 2016; Morillas et al. 2015; Poçi et al. 2012). Whether patients get a score at higher point or not, the model does not identify patients with higher risk of stroke disease (Poçi et al. 2012). Such identification depends on decision-making by the medical staff. Some researchers used machine learning and statistics to compare the performance in predicting stroke (Gage et al. 2001). Letham et al. (2015a) used machine learning to compare the stroke prediction performance of Bayesian Rule Lists to CHADS2 and CHA2DS2-VASc. Mcheick et al. (2016) used Bayesian Belief Networks (BBN) to support medical teams for prediction that focused on TIA.

2.6. Feature Selection in Healthcare

In this section, my review of the available literature focuses on feature selection techniques used in the healthcare sector. It runs in chronological order and tries to capture the evolution over time of feature selection techniques used in clinical and Electronic Healthcare Records.

A large volume of medical data is generated and collected by hospitals. Feature selection techniques are applied in many research area of application with huge datasets with tens or hundreds of thousands of variables available (Chow et al. 2008; Guyon & Elisseeff 2003). These are selected from many types of datasets such as text processing, gene expression, image, etc. These techniques are focused on improving the prediction performance, speed and cost-effectiveness predictor, and providing a better understanding of the underling processes in datasets.

The knowledge for decision-making uses patterns and searching for relationships from data mining techniques that are useful for the prediction process. In healthcare applications, classification analysis is widely adopted to validate medical diagnosis, enhancing the condition of patient care, etc. Normally, electronic healthcare records have a good number of dimensions. If an attribute in the dataset is unrelated to a feature, classification analysis can possibly give an erroneous result. Thus data preparation is essential for data mining and machine learning to enhance the accuracy of predictions (Huang et al. 2007; Khemphila & Boonjing 2011; Weng et al. 2017). Fuzzy and neural modelling was applied searching for new knowledge from a septic shock patient database. The details of previous solutions are explained, specifically in term of sensitivity, which is beneficial for making a care plan for current patients (Fialho et al. 2010).

In feature selection, an accepted method used for high-dimensional records combines the reduction of dimensionality, the removal of irrelevant and redundant features, a cut back of the number of datasets needed for training and testing in machine learning techniques, improvement of the accuracy of the predictive algorithm, and enhancement of the constructed model's comprehensibility. Most research work in machine learning is focused on improving the predictive accuracy of the classification analysis that is applied in feature selection (Anbarasi, Anupriya & Iyengar 2010; Huang et al. 2007). This technique is important with medical data mining for diagnosis of the disease that could assist in developing a care plan for patients. The datasets following from the feature selection provides a new dataset that is smaller and more suitable for clinical measures, adding accuracy and lowering false negative values. For instance, Balakrishnan et al. (2008) proposed the feature selection approach for identifying an optimum feature for diabetes in the Pima Indian population. They used the Naïve Bayes classifier for enhancing the classification accuracy. The result is an improvement in the effectiveness of the SVM Training with Backward Search approach which leads to an enhancement on feature selection and increased classification accuracy. Moreover, this technique is applied in the G-BLUP model and the Bayesian (Bayes C) prediction method and predicts high density lipoprotein cholesterol (HDL), height, and body mass index (BMI) in the male genome. The overall performance of this technique achieved the highest accuracy in prediction (Bermingham et al. 2015).

An automatic process for extracting knowledge or previously undiscovered, valid, and actionable patterns from a large database for use in decision support has been defined as "Data Mining". In healthcare, classification analysis is a data mining technique that is widely used in medical applications for diagnostic decisions, enhancing the condition of patient care, etc. For more accuracy and understandable outcomes, two commonly used feature selection methods are employed that incorporate the use of automatic feature selection mechanisms (i.e., data-driven) or expert judgment (i.e., knowledge-driven) especially when the training dataset has some irrelevant features (i.e., attributes). These two prevailing feature selection methods can lead to a dissimilar classification inefficiency due to their unique biases arising from the differences in their underlying processes. In their study, they assessed the classification effectiveness resulting from the two feature selection techniques for a risk prediction using a cardiovascular disease dataset. Their evaluation outcomes propose that the feature subsets chosen be domain experts improve the sensitivity of a classifier, while the feature subsets chosen by an automatic feature selection mechanism boost the predictive power of a classifier on the majority class (i.e., the specificity in this study) (Tsang-Hsiang, Chih-Ping & Tseng 2006). In the United States, Canada and other countries, laws and regulations are enforced to protect the exchange of data in personal healthcare information (PHI) by e-health. A Personal information needs to be modified and categorized before being released.

Normally, attributes in the dataset will be categorized in the database. Jafer, Matwin & Sokolova (2014) proposed the TOP_DIFF algorithm to handle the privacy protection PHI, and the dataset obtained from this algorithm is an anonymous dataset.

Electronic health records (EHRs) provides healthcare professionals with a way to access more available healthcare data thereby increasing the chance, using advanced data analysis, to make more informed decisions to improve the quality of care. However, data analysis tasks from EHRs face a big challenge due to the imbalanced characteristics of medical data and its inherent heterogeneous nature. A paper addressing the challenges of imbalanced medical data using histopathological images to do morphometric analysis is one of many rapidly emerging valuable tools for the neuropathological diagnosis problem of a brain tumor called "Oligodendroglioma". This tumor has an excellent response to treatment if the tumor subtype is correctly diagnosed. The paper by Huda et al. (2016) intends to find a fast, economical and objective diagnosis of a genetic variant of the tumor employing a data mining technique using automatic image analysis and histological processing and diagnosis. The 1p-/19q- variant of the tumor has been discovered to have huge chemo- sensitivity and using its morphological attributes combining feature selection and ensemble-base classification may provide a more accurate diagnosis. In their work, 63 samples of the tumor were collected in the way that ensured the dataset would be statistically balanced. They used a technique that incorporates a global optimization-based hybrid wrapper-filter feature selection and ensemble classification. They found that this method gave a better result than the standard prevailing techniques. Chow et al. (2008) used hierarchical clustering to reduce large healthcare datasets and a multi-resolution parameter search for efficiency in SVM model selection, SVM training time, and for removing the redundancy in chromosomal dat. This approach reduced the runtime and improved the performance of the classification techniques.

Clinical diagnosis is performed chiefly relying on a doctor's competence and experience. Some attributes are missing or have null values in the medical history records and affect the correct diagnosis of a disease. A wrong diagnosis leads to an incorrect treatment in some cases and also means the patient needs a number of tests for correct diagnosis. For some researchers, solving this problem applies a genetic algorithm, mutual information estimator and predicts using a random forest approach (Doquire & Verleysen 2012; Khalilia, Chakraborty & Popescu 2011). For example, Anbarasi, Anupriya & Iyengar (2010) proposed the prediction of heart diseases in cases where some attributes are not shown in the patient's records. In order to cope with the reduced number of heart attributes they used a genetic algorithm with the datasets which may predict a more accurate result. They also applied classification with clustering and Decision Tree for prediction. If a dataset is highly imbalanced, a random forest classifiers approach may help using repeated random sampling and dividing the training data into multiple sub-samples. The Genetic algorithm is applied to identify the aspects that benefit more for the diagnosis of heart ailments that reduces the number of tests required to be taken by a patient. Thirteen attributes were cut down to 6 attributes by means of a genetic search. As a result, three classifiers, namely Naïve Bayes, Classification by clustering and Decision Tree, were applied to anticipate the diagnosis of patients with a comparable accuracy as was achieved prior to the cutback in the number of attributes. Their examinations show that the Decision Tree data mining technique gives a better result than those of other two data mining techniques following the combination of feature subset selection with relatively high model construction time. Naïve Bayes performs consistently before and after the cut back of attributes with the invariable model construction time. When clustering is applied, classification performs less well compared to other two approaches.

Furthermore, Khalilia, Chakraborty & Popescu (2011) made use of Healthcare Cost and Utilization Project (HCUP) dataset for figuring disease chance of individuals according to their past medical diagnosis. The technique used can possibly be combined in a number of applications such as risk management, tailored health communication and decision support systems in healthcare. They used the National Inpatient Sample (NIS) data, which is openly available through Healthcare Cost and Utilization Project (HCUP), to develop random forest classifiers for disease prediction. As the HCUP data is exceedingly imbalanced, they employed an ensemble learning technique that is based on repeated random sub-sampling. This technique breaks down the training data into many sub-samples, while ensuring that each sub-sample is entirely balanced. They analysed the performance of support vector machine (SVM), bagging, boosting and RF to forecast the chance of eight chronic diseases. Overall, the RF ensemble learning approach surpassed SVM, bagging and boosting in terms of the area under the receiver operating characteristic (ROC) curve (AUC). In addition, RF has the edge of calculating the influence of each factor in the classification operation. The combination of repeated arbitrary sub-sampling with RF allowed them to overcome the class imbalance issue and attain assuring results. Using the national HCUP data set, they made a prediction of eight disease categories with an average AUC of 88.79%.

Modern healthcare is being improved by the increasing number of Electronic Medical Records (EMR), which have been accepted for their great value in constructing clinical prediction models because patients' diseases and hospital interventions are picked up through a set of diagnoses and procedures codes. These codes are typically defined in a tree form (e.g. ICD-10 tree) and the codes within a tree branch typically are deeply correlated. Comparing with conventional features selection methods (e.g. Information

Gain, T-test, etc.) that consider each factor individually and generally result in a lengthy feature list, these codes can be used as features to construct a prediction model and an applicable feature selection may help notify clinicians about critical risk factors for a disease. Recently, Lasso and related l1-penalty based feature selection methods have become accepted because of their joint feature selection property. However, the Lasso method is known to have issues of arbitrarily selecting one feature from many correlated features. This burdens the clinicians to arrive at a stable feature set, which is essential for clinical decision-making process. In their paper, they deal with this problem by applying a newly proposed Tree-Lasso model. Since the stability behaviour of Tree-Lasso is not broadly perceived, they studied the stability behaviour of Tree-Lasso and compared it with other feature selection methods. Using a synthetic and two real-world datasets (Cancer and Acute Myocardial Infarction), the result suggests that Tree-Lasso based feature selection is undoubtedly more reliable than Lasso and comparable to other techniques e.g. Information Gain, ReliefF and T-test. This result has implications in identifying stable risk factors for a number of healthcare issues and, as a result, can possibly aid clinical decision making for accurate medical prognosis (Kamkar et al. 2015).

In summary, feature selection is widely used in many applications such as knowledge extraction, modern healthcare, and clinical application, etc. (see Table 2-2) There are many potential benefits of the feature selection technique: improved data understanding, reduced storage requirements and training, finding relation of attributes in large dataset, and better utilization time. Moreover, a statistics application program has analysed data according to this technique before using a predictor such as Survival method (Lumley et al. 2002; Morillas et al. 2015), COX regression (Chambless et al. 2004; Hitman et al. 2007; Ibrahim-Verbaas et al. 2014; Lip, Frison, et al. 2010; Manuel et al. 2015; Olesen et al. 2011), Logistic regression (Baird et al. 2001; Guerra et al. 2016; Hanchaiphiboolkul et al. 2014; König et al. 2008; Lip, Nieuwlaat, et al. 2010; Weimar et al. 2002), Stepwise regression (Leira et al. 2004), and Classification (Letham et al. 2015b). We summarised all statistical methods used in previous healthcare research. Feature selection techniques can support and help the prediction process when processing a large healthcare dataset to show the likelihood of developing a predictor model. The new dataset reduces the prediction processing time and the attributes resulted from this feature selection will be ranked in order of the risk factors for stroke symptoms. This technique was analysed and found that it is difficult to forecast the result when the calculation is performed on complex data of various data types. Using the feature selection technique significantly improves the efficiency of the diagnosis and the forecast.

2.6.1. Variable ranking in feature selection

Many attributes in the dataset are related. Variable ranking is a principal or auxiliary section mechanism to find related attributes. The important properties of the resulting dataset are scalability, simplicity, and good empirical results. Most of the research uses variable ranking as a baseline method (Anbarasi, Anupriya & Iyengar 2010; Balakrishnan et al. 2008; Chow et al. 2008; Fialho et al. 2010; Khemphila & Boonjing 2011; Tazin, Sabab & Chowdhury 2016). This technique is not necessarily used to build predictors. A ranking criterion is used to find symptoms that discriminate between healthy and diseased patients as a symptom may code for “stroke (I64)”, or risk factors that are related to symptoms. Validating symptoms related to disease is a very difficult problem and is outside the scope of machine learning. as a consequence, we focused on build predictors.

2.6.2. Principle of the Method and Notations

Consider a set *m* examples {$\boldsymbol{x_i}$,$\boldsymbol{y_i}$} (*i*=1,….,*m*) having n input variable $x_{i,j}$ (j = 1,…., n) and one output variable y_i. Variable ranking applies a scoring function *S(i)* obtained from the value $\boldsymbol{x_{i,j}}$ and $\boldsymbol{y_i}$, *i*=1,….,*m*. By convention, we expect that a high score is indicative of an important variable and that we sort variables in decreasing order of S(i). To use variable ranking to construct predictors, nested subsets consolidating progressively more and more variable of decreasing relevance are characterised.

2.6.3. Correlation Criteria

At the beginning, consider the prediction of a consecutive outcome ***y***. The Pearson correlation coefficient is designated as:

$$\mathcal{R}(i) = \frac{cov(X_i, Y)}{\sqrt{var(X_i)var(Y)}} \qquad (1)$$

Where ***cov*** represents the covariance and ***var*** represents the variance. The estimate of ***R(i)*** is given by:

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i} - \bar{x_i})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i} - \bar{x_i})^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}} \qquad (2)$$

Where the bar notation stands for an average over the index ***k***. This coefficient is the cosine between vector $\boldsymbol{x_i}$ and ***y*** as well, after they have been centred (their mean subtracted). There is no need to assume that the input values are realisations of a random variable. The ***R(i)*** is derived from $\boldsymbol{\mathcal{R}(i)}$.

2.7. Case Based Reasoning

The diagnosis and prognosis of cerebrovascular patients is a complex domain because of varying risk factors involved with each patient. Even more complex is the multitude of risk factors involved in the prediction of stroke. Although the data from electronic healthcare records (EHR) recorded with other diseases might help in recognizing stroke risk factors, it is complicated to keep track of all the patients during their treatment. A further complication is that stroke units and their databases are separated from EHR in some hospitals or centres. Consequently, it is usually a complex task for clinicians to predict a stroke occurrence using the available patient data. Case Based Reasoning (CBR), an artificial intelligent system, has been very useful in supporting healthcare systems for treatment and diagnosis for many years and still has significant role in healthcare applications (Fan et al. 2011). Their research uses a CBR application to develop a prediction model that they apply to predict the probability of stroke disease for new patients. It is based on the principle that a new case is solved by observing similar, previous problems and adapting their known solutions. It would be of help to medical staff in solving new problems based on previous experiences with cerebrovascular patients.

CBR can store explicit diagnosis, prognosis and subsequent rehabilitation information as a repository for cerebrovascular patients in EHR that contains demographic data and related risk factors. For a future cerebrovascular patient, whose diagnosis is assumed and has an indefinite prognosis, by applying CBR and prediction model, similar cases can be retrieved from the case base which may provide relevant information to the medical staff and hence assist them in reaching a diagnosis and care plan. Ideally it may help identify a new high-risk patient before a stroke occurs.

Another reason to use CBR system for cerebrovascular domain in particular is the assumption that patients who have high probability of having a stroke may have similar prognoses which may lead not only to prevention of the stroke but allows the clinician to setup a care plan for the patient as well.

2.7.1. What is CBR?

Problems are everyday life have various levels of challenge for humans to solve. These problems may have both complex and non-complex factors in case(s) such as a challenge to solve heavy traffic in rush hours; to solve complex air traffic control and also to correctly diagnosis complex diseases. In all scenarios the objective is to improve the performance and efficiency of finding a solution by utilizing previous experience. The CBR is an important concept of artificial intelligence for problem-solving. The basic idea behind CBR is to solve a new problem by remembering and reusing information from a previous similar experience or analogous experience. Based on the intended use of the reasoning, it can be applied in many ways such as to adapt and combine old solutions to solve a new problem, to critique new solutions based on old cases or to classify entities based on the criterion of similar features. Analogical reasoning also plays a significant role in human problem solving, decision making, perception, and communication which can also be performed in CBR. Two important contributors to the development of CBR are Gentner (1983) and Carbonell (1983). Gentner (1983) performed investigations that are attributed to analogical reasoning and developed a theoretical framework for solution by analogy. Carbonell (1983) explored the role of analogy in learning and plan generalization.

CBR has been proven to be a methodology suited to solve “weak theory” domains, which are the areas in which it is difficult or impossible to bring out or develop “first principle” rules to obtain solutions. Classification and regression applied within case-based reasoning is used in complex cases. For example, legal cases often have complex rulings, where the current case can use the solutions from previous cases to solve complex problems that require knowledge and expertise and involves multiple factors (Rissland 1983).

Yale University CYRUS system is a case-based reasoner, developed by Kolodner (2014), using Schank's dynamic memory model. It contains the previous case(s) that are represented in term of knowledge, and also contain a question-answering system stored with the previous information of travel and meetings of former US Secretary-of-State, Cyrus Vance. Since then, there have been an increasing number of research papers and diverse applications in CBR and it has become a field of widespread interest.

2.7.2. Architecture of CBR

The main concept of CBR cycle has four processes that are also referred to by the mnemonic, "the four REs" (Aamodt & Plaza 1994). It is started with *Retrieve* the similar case(s) ; *Reuse* the information from previous case(s) to solve the new problem, *Revise* the proposed solution to solve the new problem, and *Retain* that problem as a new case and store it in the case base (Richter & Weber 2013). This decomposition of the CBR cycle is derived from the contributions of Aamodt & Plaza (1994) and modified by Lu, Lu & Zhang (2009). (see in Figure 2-2).

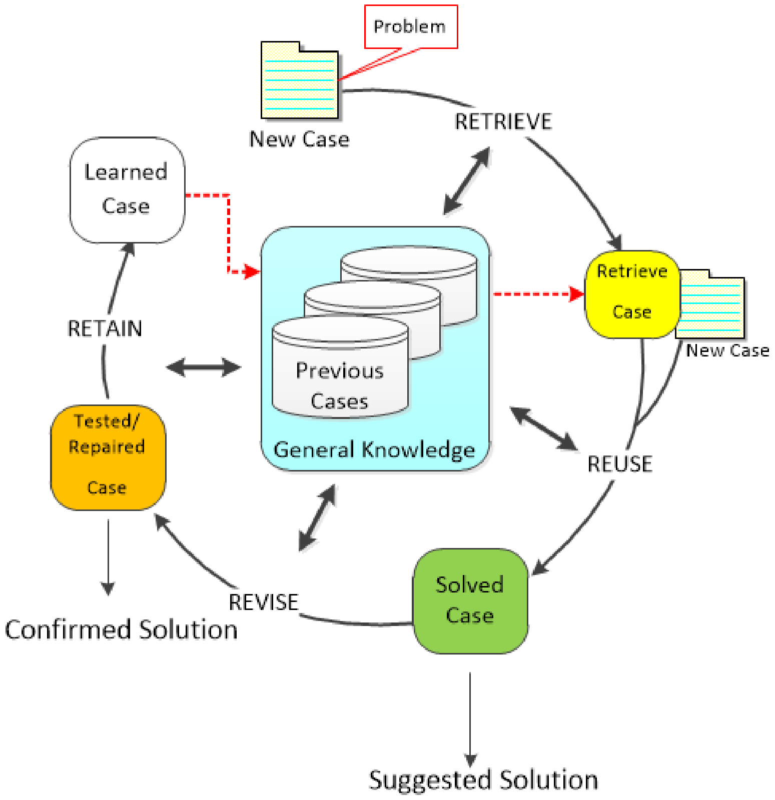


Figure 2-2 : The Case-based reasoning cycle introduced by Aamodt & Plaza (1994)

The new case starts at the top of stage, where an input is entered into the system. The previous case is compared to the new case and starts a *retrieve* step.

> "The retrieved case is combined with the new case – through *reuse* – into a solved case, i.e. a proposed solution to the initial problem. Through the *revise* process this solution is tested for success, e.g. by being applied to the real world environment or evaluated by a teacher, and repaired if failed. During *retain*, useful experience is retained for future reuse, and the case base is updated by a new learned case, or by modification of some existing cases."(Aamodt & Plaza 1994)

In practice the CBR system compares case base by comparing a similar case with a new case and all the cases in the system. A result will be a list in rank order based on the level of similarity of the cases.

2.7.3. Significance of CBR

Many problems are complex and difficult to correctly identify the best solution. Information Technology with the application of CBR can support humans in solving complex problems. Some processes are easier to perform by humans whereas others are more appropriate for computers using CBR. For instance, people can perform creative adaptation very well and expert knowledge can be created and adapted by humans but the complete range of applicable cases might be not remembered due to a bias in human memory. Novices in the field do not have adequate experience to solve a variety of problems. With the application of CBR, Humans and computers can interact in a productive manner in order to solve problems with some advantages, as follows:

- CBR can analyse the domain knowledge and compare (a) previous case(s) to new case(s) even when the reasoner does not have access to the entire domain knowledge (Gierl, Bull & Schmidt 1998).
- CBR can provide a solution when a previous case is similar to new case(s). It can be promptly proposed and reduce time in finding using reasoning (Ashley 2006; Choudhury & Begum 2016).
- A solution in CBR can help an improvement in capability in situations that most clearly involve specific case(s) and/or similar case(s). (Begum et al. 2006, 2009; Kolodner 2014) .
- CBR can avoid previous errors, showing solutions from previous cases and facilitating facilitating learning. The system can help by keeping records of each situation that occurred for future reference (D'Aquin, Lieber & Napoli 2006; Gierl, Bull & Schmidt 1998).

2.7.4. Case-based reasoning in healthcare

Case-based reasoning systems have many application areas in the healthcare sector by providing solutions for diagnosis and treatment of diseases based on past experiences. For example, the mixture of experts for case-based reasoning (MOE4CBR) (Arshadi & Jurisica 2005) is an application for high-dimensional biological domains for the prediction to disease. The data sets are used in ovarian mass spectrometry, leukemia and lung microarray data sets (Chuang 2011). The biomedical domains are complex, thus a CBR system is unsuitable for this research. Instead they used data-mining and a logistic regression method applied to the system and also improved the classification performance. A case is defined by a logistic regression approach that filters the important features in CBR. Similar cases are also grouped by the data-mining technique. The system also provides support for the "dimensionality" problem in this domain. For complex medical diagnoses, if patients have a complex disease, more medical domains have to be used for this. For example, the Premenstrual syndrome (PMS) relates both on gynaecology and psychiatry and thus needs a complex algorithm for diagnosis. The CBR-based expert system uses the k-Nearest Neighbours (k-NN) algorithm to search for k similar cases determined by focusing on the Euclidean distance measure (Chattopadhyay et al. 2013). CBR in treatment and management of diabetes is also represented in an application solving problems by using patient health records such as demographic data, laboratory test results, and physical examination. These are compared with previous cases by using the k-NN algorithm (Kiragu & Waiganjo 2016). For complex data, CBR based on gene expression profiles has been applied using machine-learning and data-mining techniques. This method used k-NN with a weighted feature-based technique to retrieve and compare between previous cases and new cases. Their proposed methodology used several data sets in this framework. The results indicate the percentage of gene expression profiles for new patients which have similarities with previous cases and thus help predict and identify those at risk of disease (Anaissi et al. 2015).

Sharaf-el-deen, Moawad & Khalifa (2014b) introduced an automated adaptation process, which applies adaptation rules for solving new cases. To evaluate the approach, the researchers develop a prototype for diagnosing breast cancer and thyroid diseases. They proposed a hybrid based medical diagnosis approach in order to enhance the performance of the CBR retrieval system. The main idea of the proposed approach is to combine both case-based and rule-based reasoning. In addition, Ahmed et al. (Ahmed, Banaee & Loutfi 2013) apply various data processing and feature extraction techniques by considering time and frequency domains for disease prediction. Given appropriate input data, the CBR system discovers the relevant cases and then creates an alarm based on the output. To evaluate the proposed system, the researchers compared their results with the classification results from experts in the field.

Furthermore, clinical decision support system for prediction and diagnosis of diseases are able to selection hidden patterns and relationships with the medical data providing ways for efficiently designing the decision support systems (Amin, Agarwal & Beg 2013a).

2.8. Data mining and predictive analysis in stroke

In this section, my review focuses on data mining techniques, predictive models and predictive analysis used in stroke diseases from the available literature. It runs in chronological order and tries to capture the evolution over time of data mining techniques and other technologies used in stroke disease diagnoses.

Lumley et al. (2002) reported the use of web-based application for forecasting a stroke rate in the aged. This research applied Cardiovascular Health Study (CHS) which provided data on 5,888 people who were at least 65 years old. Relevant information for stroke forecast consisted of common information (gender, nationality, age, and smoking record), coronary heart disease (CHD) record, antihypertensive medications usage, clinical record such as diabetes, peripheral vascular disease, congestive heart failure, atrial fibrillation, transient ischemic attack, and physical and biochemical measurements such as serum creatinine, systolic blood pressure, ratio of HDL to total cholesterol, etc. Thickness of carotid wall was measured by using an average value of internal and common wall. To measure the subject blood pressure, a doppler probe was applied in the subject's right arm. The American Diabetes Association classification described "diabetes" as fasting glucose 126 mg/dL (7.0 mmol/L) or treatment with insulin or oral hypoglycaemic agents, while "impaired fasting glucose" was described as fasting glucose between 110 and 126 mg/dL (6.1 and 7.0 mmol/L). The researchers applied a Java applet to generate predicted survival curves at a year interval.

Leira et al. (2004) presented a proper database of 1,266 stroke patients from the TOAST study. They also applied both univariate and stepwise regression approaches to evaluate patients who suffered either a transient ischemic attack (TIA), or a recurrent stroke within 3 months after the initial stroke. This work was based on probable relations which were created from 20 clinical variables.

Ohkubo et al. (2004) believed that stroke risk value was forecasted by considering home blood pressure which people took themselves at home. They collected data from 1491 people who lived in Japan, were more than 40 years old and had no stroke record. Each of them measured his/her blood pressures more than 14 times. To predict stroke risk, they applied the Cox proportional hazards regression model which is suited for working with confounding factors. The predictive value of home blood pressure measurements improved as the number of home measurements increased. The researchers mentioned that there was no recommended threshold of home blood pressure value. To precisely predict stroke risk, they recommended measuring blood pressure more than 14 times. The researchers concluded that home blood measurement is more effective in predicting stroke risk than the conventional (as measured by a clinician) blood pressure values.

König et al. (2008) attempted to forecast values of survival and functional recovery by using prognostic models and they focused on 5,419 patients from VISTA (the Virtual International Stroke Trials Archive). They adjusted interrupts and evaluated novel model parameters to enhance the correctness of the models. Regarding acute ischemic stroke patients, the researchers applied a simple age-based model. Based on NIHSS (National Institutes of Health Stroke Scale), the results presented show that they were able to precisely forecast survival and functional recovery after 3 months. This research showed that the predicted results were improved by using plain adjustment, especially when data set was huge.

Khosla et al. (2010) focused on approaches for predicting stroke based on the CHS dataset. The Cox proportional hazards model and a machine learning approach were analyzed in their work. Regular issues of feature selection, data imputation and stroke prediction by using medical datasets were examined. The researchers presented a new heuristic-based approach for automatically selecting features. The proposed approach also applied SVMs (Support Vector Machines). Comparing with the Cox proportional hazards model and L1 regularized Cox model, the proposed approach obtained the ROC curve with a bigger space. Moreover, the researchers also proposed a novel hybrid algorithm which merged a margin-based classification approach and a censored regression together. Comparing with the Cox model, concordance indexes of the proposed algorithm were preferable. Regarding AUC metrics and concordance indexes, the experimental results showed that the proposed approach performed better than the existing approaches. Furthermore, latent risk factors were also investigated. The proposed approach was able to be used for other disease predictions, although the data was incomplete and the factors were complex.

Mao et al. (2012) focused on warning the patients about the earlier deterioration based on monitored ICU (Intensive Care Unit) data. Regarding deterioration warning, data mining techniques were applied for analysing and extracting a huge feature set. The features consisted of first and second order time-series features, DFA (Detrended Fluctuation Analysis), spectral analysis, approximative entropy, and cross-signal features. Moreover, many data mining techniques were used to analyse and evaluate the features; namely forward feature selection, linear and nonlinear classification algorithms, and exploratory under sampling for class imbalance. A clinical warning system was proposed in this research and it was applied with the medical database at Barnes-Jewish Hospital.

Sudha, Gayathri & Jaisankar (2012 stated that many data mining techniques were applied for predicting several diseases. Regarding stroke disease, several classification approaches; such as Decision Tree, Artificial Neural Network, and Naïve Bayes, were applied with various relevant attributes. To reduce the data dimensions, they applied a component analysis algorithm which attempted to analyse relevant stroke attributes. Then the researchers predicted whether stroke disease hurt the patients.

Letham et al. (2015a) proposed a novel approach for predicting stroke. To obtain a prediction model, they applied the BRL (Bayesian Rule Lists) technique with a huge database. The database was the MarketScan Medicaid Multi-State Database (MDCD) which was collected from several states in the USA. Firstly, the researchers retrieved information of 12,586 patients; regarding records of atrial fibrillation diagnosis, one-year before diagnosis, and one-year after diagnosis, from the MDCD database. Fourteen percent of the patients suffered a stroke one year after they had the atrial fibrillation diagnosis. The researchers focused on aspect of drugs and conditions. They defined a binary variable to represent whether the patient received drug or condition before the atrial fibrillation diagnosis. In this research, there were 4,146 drugs and conditions in total. Age and sex features were also considered. The patients would be split into several groups by using typical split values of age such as 50, 60, 70 and 80 years old. The dataset they used in this research was much bigger than one which was applied for calculating CHADS2 score. To predict stroke, the cross-validation was processed five times. For each cross-validation, to gather feasible antecedents, the researchers applied frequent items-based mining technique, with a setting of a minimum threshold and a maximum cardinality as 10% and 2 respectively. To measure the efficacy of BRL point, ROC curve and AUC values were estimated.

A prominent point of their research was to use BRL method which meant that the proposed approach was precise, interpretable and flexible. Because of interpretability, domain experts were able to interpret the approach using credible rules, and this made it compact and reliable. The interpretable model assisted the domain experts to easier understand and solve problems in the sophisticated model. As a result, the model was more precise and useful. Although the interpretable model in this research focused on a medical domain, it is able to be applied in various domains such as science, engineering, and industry.

2.9. Deep Learning Method

2.9.1. Deep Learning

Presently, deep learning techniques have been applied for research in the area of prediction problems (Liang et al. 2014; Nie et al. 2015; Stier et al. 2015) The main idea of those techniques is to apply layer-based computational models to learn from data. A Back-propagation algorithm is employed in order to detect latent complicated structures which are necessary for prediction problems. Regarding the learning process, a machine can automatically adjust internal parameters in each layer to be proper for a previous layer representation. (LeCun, Bengio & Hinton 2015).

Typical deep learning architecture is a multi-layer architecture which mainly consists of an input layer, hidden layers and an output layer. Nodes in each layer connect to other nodes in the next layer with weighted links. The output value in each level will be conveyed to the next layer as an input value (see Fig 2-3).

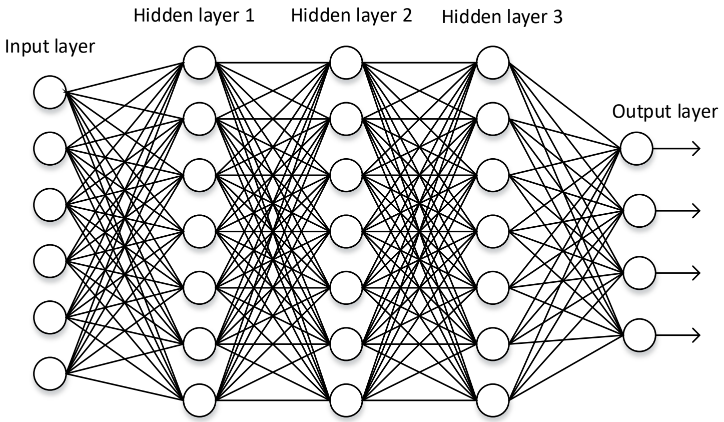


Figure 2-3 : A structure of deep neural network

2.9.2. Long Short -Term Memory - Recurrent Neural Networks (LSTM-RNN)

LSTM-RNN is one type of deep learning architecture. Like the typical architecture, LSTM-RNN consists of input, hidden and output layers. Input data is conveyed to the network through input *gates* which are multiplicative units. *Sigmoid* and *tanh* functions are applied to calculating and triggering network cells. Recurrent hidden layers of the network consist of memory blocks which comprise computation units. There are memory cells in each memory block. Those cells are able to collect instantaneous states of the network. Moreover, LSTM-RNN also includes a *forget gate* in the memory block in order to verify the input value before storing it in the memory cell. This is a distinctive point of this network because the serialized input streams sometimes are not divided with this arrangement. The *forget gate* assists the network to retune the memory cell value (Gers, Schmidhuber & Cummins 2000). Green lines in figure 2-4 represent links from the internal cells to all network gates for transmitting and learning information. A *Sigmoid* function is applied for the forget gate. Regarding an output layer, Sigmoid and tanh functions are computed the same as in the input layer (Gers, Schraudolph & Schmidhuber 2002) (see Fig 2-4).

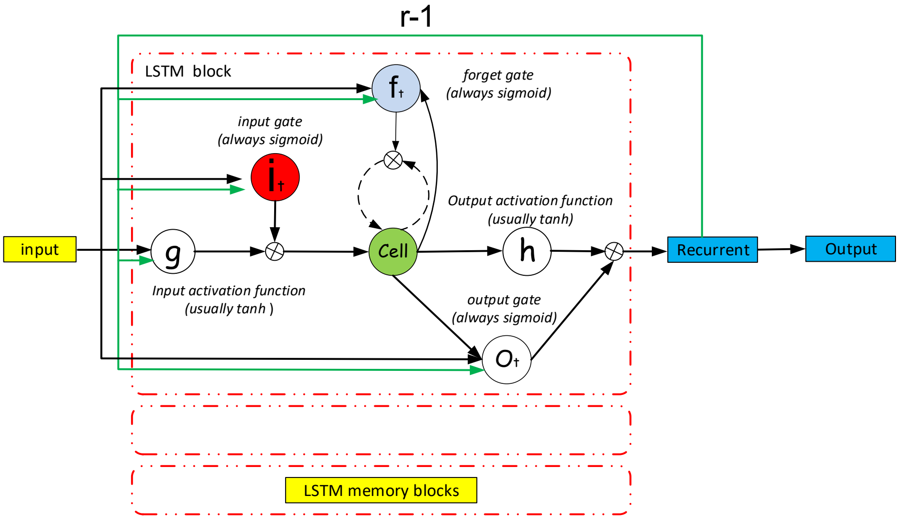


Figure 2-4 : LSTM-RNN memory architecture and a single memory block (Gers, Schraudolph & Schmidhuber 2002; Greff et al. 2017; Sak, Senior & Beaufays 2014).

2.9.3. Deep Learning in Healthcare sector

Different deep learning techniques are currently employed for predicting results in the area of healthcare. To learn and recognize patterns in a dataset without labelled data, either RBM (Restricted Boltzmann Machine) or an unsupervised learning approach using auto encoder techniques are recommended. In contrast, many researchers have focused on using labelling datasets and supervised learning approaches.

Supervised learning approaches also have been applied for solving problems in the area of text processing and image recognition (AU, AU & Hinton 2013; Kamijo & Tanigawa 1990). The RNTN (Recursive Neural Tensor Network), a recurrent network, is employed for text processing tasks such as phrase and sentence extraction, while the DBN (Deep Belief Network), a convolutional network, is employed for image pattern recognition (Brosch & Tam 2013; Gulshan et al. 2016; Li et al. 2014). Furthermore, those supervised learning approaches also have been applied in other problems, such as object recognition, speed recognition and time series analysis. Regarding classification problems, ReLUs (Rectified Linear Units) such as DBN and multi-layer perceptron, are recommended.

Medical datasets, for example EHRs data, has played an important role in disease analysis and treatment. Ordinarily, the dataset consists of observed clinical treatment records. Unfortunately, information about environments that affect diseases is not covered because it is complicated and sometimes ambiguous. This leads disease analysis to be difficult and challenging. Hammerla et al. (2015) developed a deep-learning based evaluation system which manipulated practical restrictions and also distinguished related diseases based on a given dataset with realistic settings. The dataset in this research was gathered from 34 Parkinson's disease patients.

Nie et al. (2015) applied a deep learning algorithm to devise a health seeker system. The system used health questions as its input and then discovered relevant diseases. It retrieved information which was necessary for disease analysis and prediction. Furthermore, evidence of symptoms were extracted and applied to the next process. There were two major part in this research. The first part was differentiation between medical signatures and raw features. The second part was to input data from the former part into the input and hidden layers. The interrelations among layers were adjusted based on the difference between pre-training output values and defined output values. To deeply learn dataset, the system was repeatedly computed for adjusting interrelation values.

Gulshan et al. (2016) developed a new approach which utilised retina images (RDR) for automatically detecting diabetes disease. Their proposed approach used a deep convolutional neural network to classify these retina images. The dataset which was used in their research contained 128,175 images. Their experimental results demonstrated that their approach can effectively recognise diabetes from the retina images.

The research about tissue survival detection is mainly applied for helping provide an immediate ischemic stroke remedy. When a clot is discovered, this research assists by analysing and estimating the hazards(Asadi et al. 2014). Stier et al. (2015) developed a deep learning-based model for detecting tissue survival. The dataset of this work was randomly collected from the hypo-perfusion (T-max) feature in MRI. The model was estimated based on the experience of the experts. The experimental results demonstrated that their proposed model outperformed a single-voxel-based regression model, showing that the use of deep learning methods was able to solve healthcare problems.

Several researchers have studied and applied deep-learning based models for analysing and solving medical problems which are sophisticated and cannot be solved by using conventional models (Hung et al. 2017; Lyu et al. 2017). Various deep learning models for medical image analysis have been recently developed (Brosch & Tam 2013; Gulshan et al. 2016). Liang et al. (2014) presented an unsupervised-based deep network for retrieving medical features, and they then applied SVM, a supervised learning approach, for analysing data. For EMR (Electronic Medical Record), Electronic medical claims (EMCs) and HIS (Hospital Information System) datasets, the experimental results demonstrated that the proposed medical-based deep network model performed well (Hung et al. 2017), confirming that deep learning techniques are able to efficiently analyse and diagnose stroke disease.

2.10. Summary

My review of available literature focuses on predictive models and predictive analysis used in healthcare and stroke symptoms to accomplish capturing the machine learning techniques and other techniques used in healthcare (Table 2-1 and Table 2-2). The diagnosis and prognosis of stroke patients is a complex domain because of the varying factors involved with each patient. Even more complex is the multitude of risk factors involved in the prediction of stroke. Although the data from Electronic Healthcare Records (EHR) recorded with other diseases might help in recognizing stroke risk factors, it is complicated to keep track of all the patients during their treatment. Case Based Reasoning (CBR) concept is also reviewed with a focus on CBR in healthcare. It is based on the principle that a new case is solved by observing similar, previous problems and adapting their known solutions. It would be of help to medical staff in solving new problems based on previous experiences with stroke patients. While many researchers have applied machine learning and deep learning to medical data, the Deep learning technique employs learning from data with multiple level of abstraction by computational models that are associated with multiple processing layers. This method is intended to discover complex structures in a huge data set by using the backpropagation algorithm and other algorithms to predict the result. However, stroke symptoms have many risk factors and complexity. Therefore, datasets in EHRs are quite significant for decision-making in treatment. In general, a realistic dataset contains useful records for clinical practice and uncovers realistic environments for the analyses of diseases because it had included ambiguous and incomplete values that contribute to errors and are unsuitable for annalistic data sets and present a very challenging analysis.

Table 2-1: Disease prediction using Case-based reasoning system and Machine Learning.

Categories	Author	Diseases	Case-based reasoning	Machine Learning	Risk factors	Outcomes
Diagnosis and decision support systems	Kolodner & Kolodner (1987)	Psychiatry	✓			
	(Koton 1988)	Heart failure	✓		Symptoms, Test results (EKG), Medical history, and solution data	
	Turner (1988)_	Dyspnoea	✓			
	López & Plaza (1993)	Pneumonia	✓			BOLERO-RBS Reactive System
	Bichindaritz (1995)	Psychiatry	✓			
	Haddad, Adlassnig & Porenta (1997)	Detection of coronary Heart diseases	✓			SCINA based on Image processing
	Bichindaritz, Kansu & Sullivan (1998)	Stem cell transplantation	✓			Care-Partner based on Web technology
	Marling & Whitehouse (2001)	Alzheimer's disease	✓		Medical history, current physical, emotional, behaviour, and cognitive status	AUGUSTE application
	Chattopadhyay et al. (2013)	Premenstrual syndrome (PMS)	✓	K-nearest neighbour		
	Alexopoulos, Dounias & Vemmos (1999)	Post-stroke	×	Machine learning : learning from examples		
	Kononenko (1993)	Thyroid diseases Rheumatology Tumor Breast cancer	×	Machine learning : Bayesian Learning (classification)		

Categories	Author	Diseases	Case-based reasoning	Machine Learning	Risk factors	Outcomes
	Hsu & Ho (2004)	Multiple Diseases	✓	CBR, Neural networks, Fuzzy theory, Induction, Utility theory, and Knowledge-based planning technology		
	Kwiatkowska & Atkins (2004)	Obstructive sleep apnoea	✓	CBR, Fuzzy logic, and Semiotics		Somnus system
	Chang (2005)	Development delay in children	✓	CBR		
	Shi & Barnden (2005)	Multiple disorders	✓	CBR and Induction (adaptation with rules)		
Classification, Knowledge acquisition/ management	Macura & Macura (1995)	N/A	✓	CBR	Radiology image	MacRad system
	LeBozec et al. (1998)	N/A	✓	CBR	Radiology image	IDEM system
	Gierl, Bull & Schmidt (1998)	Epidemics	✓	CBR, Rule-based reasoning, and Model-based reasoning		TeCoMED system (forecasting)
	Perner (1999)	N/A	✓	CBR, Image processing, and Data mining	Medical image analysis	
	Schmidt, Pollwein & Gierl (1999)	Liver transplantation	✓	CBR		COSYL system
	Golobardes et al. (2002)	Breast cancer	✓	CBR		CaB-CS system
	Nilsson & Funk (2004)	Respiratory sinus arrhythmia	✓	CBR and Rule-based reasoning		
	Perner et al. (2004)	Recognition of Airborne Fungi Spores	✓	CBR and Image processing		

Categories	Author	Diseases	Case-based reasoning	Machine Learning	Risk factors	Outcomes
	Sharaf-el-deen, Moawad & Khalifa (2014b)	Thyroid disease Breast cancer	✓		Electronic Health Records	New hybrid Case-based reasoning (integrated CBR and rule based reasoning)
	van den Branden et al. (2011)	Lung cancer	✓		Electronic patient records	ExcelicareCBR Tools
	Koton (1989)	Coronary Disease	✓	CBR, Rule-based domain, and Model-based reasoning		CASEY System
	Bareiss, Porter & Wier (1987)	Hearing Disorders	✓	CBR		Protos system
	Gierl & Stengel-Rutkowski (1994)	Dysmorphic syndromes	✓	CBR		GS.52 system
	Reategui, Campbell & Leao (1997)	Congenital Heart Diseases	✓	CBR and Neural networks		
	Chien-Chang & Cheng-Seen (1998)	General	✓	CBR, Fuzzy logic, Neural networks, Induction, and Knowledge-based technology (adaptation with rule-based)		
	Goodridge, Peter & Abayomi (1999)	Hematological Diseases	✓	CBR and Neural networks		MED2000
	Phuong, Thang & Hirota (2000)	Lung Diseases	✓	CBR and Fuzzy logic		
	Montani et al. (2003)	Type 1 diabetes	✓	CBR, Rule-based reasoning, and Model-based reasoning		
	Vorobieva, Gierl & Schmidt (2003)	Endocrinology	✓	CBR		

Categories	Author	Diseases	Case-based reasoning	Machine Learning	Risk factors	Outcomes
	Brien, Glasgow & Munoz (2005)	Attention-deficit hyperactivity disorder	✓	CBR		
Treatment and management	Kiragu & Waiganjo (2016)	Diabetes	✓			
Healthcare Planning	Bradburn & Zeleznikow (1994)	N/A	✓			FLORENCE
	Marling & Whitehouse (2001)	Alzheimer’s Disease	✓	CBR and Rule-based reasoning		Auguste system
Prediction	Khosla et al. (2010)	Stroke	×	Machine learning: Compare the Cox proportional hazards regression model / approach: Heuristic (conservative mean) combined with Support Vector Machines (SVM)	Cardiovascular Health study and General factor	
	Palaniappan & Awang (2008)	Heart disease	×	Data mining: Namely, Decision trees, Naïve Bayes and Neural Network	Medical records	Intelligent Heart Disease Prediction Systems (web-based platform)
	Srinivas, Rani & Govrdhan (2010)	Heart attack	×	Data mining: Naïve Bayes	Medical records	Application
	Sudha, Gayathri & Jaisankar (2012)	Stroke	×	Data mining: Classification algorithm	Patient history	
	Alotaibi & Sasi (2016)	Transferring Stroke In-patients to ICU	×	Predictive model: Artificial Neural Network, Decision Tree, SVM, and Logistic regression		Comparing

Categories	Author	Diseases	Case-based reasoning	Machine Learning	Risk factors	Outcomes
	Letham et al. (2015a)	Stroke	×	Predictive model Using rules and Bayesian Analysis that compared with $CHAD_2$, $CHAD_2$-VAS_2	Atrial fibrillation	The medical scoring systems currently in use
	Mcheick et al. (2016)	Stroke	×	Stroke prediction in Emergency room	Transient ischemic attack (TIA)	Mobile Application
Warning/ Real-time monitoring	Mao et al. (2012)	N/A / use in ICU and RDS	×	Data mining (linear and nonlinear classification algorithm)	N/A	Performance, Time series

Table 2-2: Feature Selection technique in healthcare.

Author	Type of Healthcare datasets	Methods	Diseases	Outcomes
Chow et al. (2008)	Large Healthcare Datasets	SVM-GA feature selection model	Chromosome caching	Reduction of runtime and enhancement of classification performance.
Huang et al. (2007)	Diabetic patients' information	A feature selection technique and supervised model	Diabetes	Feature selection and classification model.
Khemphila & Boonjing (2011)	Heart disease dataset	Classification using Neural Network and Feature Selection	Heart disease	Reduction of the number of attributes in patients' information.
Anbarasi, Anupriya & Iyengar (2010)	Heart disease dataset	Feature selection using Genetic Algorithm	Heart disease	Reduction of the number of attributes in patients' information.
Balakrishnan et al. (2008)	Type II Diabetes database	SVM ranking with backward search for feature selection	Type II Diabetes	Improvement of feature selection and enhancement classification accuracy.
Bermingham et al. (2015)	HDL and BMI in UK dataset	High-Dimensional feature selection	High density lipoprotein cholesterol (HDL) and body mass index (BMI) in man	Improvement of the performance of a mixed model (G-BULP) and a Bayesian prediction method (Bayes C).
Hossain et al. (2013)	Disease Profile	Naïve Bayes, Multilayer perceptron (MLP) and Decision Tree J48	Disease profile	Using feature selection for accuracy benchmarking of clinical data.
Huda et al. (2016)	Brain Tumor dataset	Morphological feature and classification	Brain Tumor diagnosis	A hybrid feature selection with ensemble classification for imbalance healthcare dataset.
Doquire & Verleysen (2012)	Electronic Healthcare Records	Mutual information techniques	N/A	Feature selection with missing data by used mutual information estimators.
Khalilia, Chakraborty & Popescu (2011)	National Inpatient Sample data in US government agencies	Healthcare Cost and Utilization Project (HCUP) and Random forest classifiers	N/A	The prediction of disease risk from imbalanced dataset by used random forest and HCUP.
Kamkar et al. (2015)	Electronic Medical Records	t-test, Information Gain, ReliefF, Classification methods, and	N/A	Exploiting ICD Tree structure by using Tree-Lasso.

Author	Type of Healthcare datasets	Methods	Diseases	Outcomes
		Hierarchical features and Tree-Lasso		
Baxter, Williams & He (2001)	Elderly patients with diabetes	Three alternative feature vectors	Diabetes	Clustering patients and visualising the features devised to highlight interesting patterns of care.
Fialho et al. (2010)	Septic shock patient database	Wrapper methods	Septic Shock	Reduction of the number of features and accurately predicting the outcome for septic shock patients.
Selvakuberan et al. (2011)	PIMA Indian Diabetes Dataset (PIDD)	A combination of ranker search method in classification methods	Diabetes	Using feature selection method for classification for combination of ranker search method.
Tazin, Sabab & Chowdhury (2016)	Chronic Kidney Disease dataset collection from UCI repository	Effective Classification and Ranking algorithm	Kidney disease	Diagnosis of chronic kidney disease by using classification and feature selection.
Jafer, Matwin & Sokolova (2014)	Personal Health Information	Top_Diff algorithm	N/A	Improvement of the utility of differentially private data publishing.
Rajeswari, Vaithiyanathan & Pede (2013)	Medical datasets	Association and correlation mechanism	Heart, Breast cancer and Diabetes diseases	Selecting the correlated feature or attributes of medical dataset for clinical decision support system.
Hall (2000)	Large Database	Correlation-based filter algorithm	N/A	Improvement in identification of discrete problems.
Vieira et al. (2013)	MEDAN database	Binary particle swarm optimization (MBPSO) method	Septic shock	Modification of MBPSO for feature selection with the simultaneous optimization of SVM kernel
Xu et al. (2014)	Foetal heart rate dataset	Genetic algorithms	Heart	Reduction of the number of attributes and applied genetic algorithm for decision-making process.
Tsang-Hsiang, Chih-Ping & Tseng (2006)	Cardiovascular disease dataset	C4.5 and Correlation-based feature selection	Atherosclerosis disease	Enhancement in the predictive power of a classifier.

Chapter 3.

Electronic Healthcare Records

3.1. Introduction

This chapter presents an Electronic Healthcare Records in hospital database and the standard of International classification of diseases for medical data. Section 3.1 introduces the Electronic Healthcare Records of cerebrovascular disease patients and structure of patient record. Section 3.2 introduces the standard of International Classification of Disease that applied in Thai's public health.

3.2. Electronic Healthcare Records

The Electronic Healthcare Records (EHR) of cerebrovascular disease patients contain various information, including demographic data, potential risk factors, and non-potential risk factors that are recorded in the hospital database (See fig 3-1).

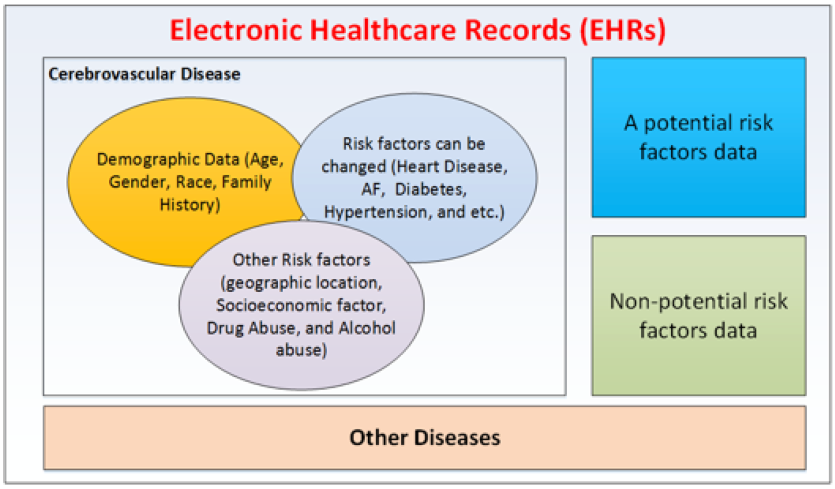

Electronic Healthcare Records (EHRs)
Cerebrovascular Disease
Demographic Data (Age, Gender, Race, Family History)
Risk factors can be changed (Heart Disease, AF, Diabetes, Hypertension, and etc.)
Other Risk factors (geographic location, Socioeconomic factor, Drug Abuse, and Alcohol abuse)
A potential risk factors data
Non-potential risk factors data
Other Diseases

Figure 3-1 :Electronic Healthcare Records of Stroke Patients (Chantamit-o-pas & Goyal 2018b).

The data dictionary of this EHR record consists of Hospital Code (HCODE); Hospital Number (HN); Gender; date of birth (DOB); clinic operation (CLINIC_OPD); date operation (DATEOPD); date diagnosis (DATEDX); clinic diagnosis (CLINIC_ODX); diagnosis code (DIAG); diagnosis type (DXTYPE); and Medical license (DRDX). (see Table 3-1). In term of the data dictionary of EHRs with stroke's risk factor, it consists of gender; date of birth (DOB); clinic operation (CLINIC_OPD); date operation (DATEOPD); date diagnosis (DATEDX); clinic diagnosis (CLINIC_ODX); diagnosis code (DIAG); and diagnosis type (DXTYPE). Gender was identified and represented by either the code 1 (Man) or the code 2 (Woman) (see Table 3-2). DOB field record contains patient's birthdate and some records have null value or error in term of date. We eliminated the null value or error and converted the value into age in the data preparation process. CLINIC_OPD field indicates the clinic number of hospital where patient has treatment. DATEOPD field demonstrates the data of service. DATEDX field indicates the date of diagnosis. CLINIC_ODX field shares similar code with CLINIC_OPD field. Code of diagnoses were obtained from doctors or medical experts who used ICD-10 codes and inputted in DIAG field. DXTYPE field specifies types of disease (see Table 3-1). The demographic data, disease type, and other information were recorded once each patient visited. For the prediction process, we integrated the multiple value dependencies into EHRs. Further details will be described in chapter 4.

Table 3-1 : The data structure of EHRs records

Attributes	TYPE	LENGTH	DECIMAL	Value
HCODE	C	5	0	Hospital Code
HN	C	9	0	Hospital Number
Gender	C	1	0	Male (1), Female (2)
DOB	D	8	0	Patient’s birthday
CLINIC_OPD	C	4	0	clinic number of hospitals
DATEOPD	D	8	0	data of service
DATEDX	D	8	0	date of diagnosis
CLINIC_ODX	C	4	0	clinic number of hospitals
DIAG	C	5	0	ICD-10th codes
DXTYPE	C	1	0	(1) primary disease (2) comorbidity disease (3) complication (4) other diseases
DRDX	C	6	0	Medical license

Table 3-2 : A partial sample of the data structure of EHRs records with stroke risk factors.

Attributes	Value
Gender	Male (1), Female (2)
DOB	Patient's birthday
CLINIC_OPD	clinic number of hospitals
DATEOPD	data of service
DATEDX	date of diagnosis
CLINIC_ODX	clinic number of hospitals
DIAG	ICD-10th codes
DXTYPE	(1) primary disease (2) comorbidity disease (3) complication (4) other diseases

3.3. The International Classification of Diseases, 10th Revision (ICD-10th)

There is significant growth in the amount of medical or patient data being generated in hospitals or clinics all over the world. In most cases, Electronic Health Records are used for storing most of this medical or patient data. In practice, the International Classification of Diseases (ICD) is an International Statistical Classification of diseases and signs, abnormal finding symptoms circumstances, complaints, and external causes of injury as classified by World Health Organization (WHO) whose codes become the standard of codification in the Electronic Medical Record system (EMR) (World Health Organization 2004).

In 1983, the Tenth revision of the International Statistical Classification of Diseases and Related Health problems was formalized in Geneva. The background of the International Classification of Diseases (ICD), 10th Revision i.e. ICD-10 and ICD-10-CM (Clinical

Modification) codes was reviewed and classified as given in volume 2. While the title has been amended to make clearer the content and purpose and to reflect the progressive extension of the scope of the classification beyond diseases and injuries, the “ICD” name has been retained. The program of work was guided by the expertise of Heads of WHO collaborating Centres for Classification of Disease. In addition, they contributed comments and suggestions. It was clear that many users wished the ICD would encompass types of data rather than the “diagnostic information” that it had always covered.

The Electronic Healthcare Records (EHRs) adopted ICD-10 and ICD-10-CM codes and these codes came into effect in the Electronic Medical Record system (EMR) from 1998. Currently, ICD-10th codes are used by hospitals and health professionals, which are retrieved from the Electronic Medical Records (EMR) system. The standard provides a very convenient platform for primary and secondary data analysis of these records for diagnosis and prediction of diseases, as well as for the improvement of medical and patient care. Normally, the structure of ICD-10 code presented in the classification of diseases (Diseases of the circulatory system, eye and adnexa, etc.) and represent the concept of a “family” of diseases in three- and four-character classification such as I65 (Occlusion and stenosis of vertebral artery) is the main category, and I65.1 (Occlusion and stenosis of basilar artery) is the sub-category. The ‘core’ three-character code of classification of ICD-10 is the mandatory level of coding for international reporting. It also has four character sub-categories which are not mandatory for international reporting (World Health Organization 2004).

A Thai modification to the WHO international Statistical Classification of Diseases and Related Health problems, tenth revision and the development of an accompanying Thai procedural coding system was funded by the WHO, South East Asia Region Office for introduction as the Thai standard for morbidity coding in health services and mortality statistics. ICD-10 TM is based on the WHO ICD-10 and has been modified in Thai by The Bureau of Policy and Strategy, Ministry of Public Health that has been given the responsibility for the development, introduction and maintenance of ICD-10 TM, which now has 39,220 codes (The Bureau of Policy and Strategy & Ministry of Public Health of Thailand 2016).

3.4. Summary

The historical data of patient had recorded in many sections that included the detail of diseases which are complex and hard to classify. Thus, the World Health Organization (WHO) used the International Classification of Diseases (ICD), 10th Revision for identifying the diseases and this has become the standard codification in the Electronic Medical Record system (EMR). ICD-10 has adapted the ICD codes to Thailand's health condition, based on the WHO ICD-10. Furthermore, the medical service department under the Ministry of Public Health of Thailand modified them and also applied them to patients' records. This dataset is a resource for this research and will be utilised in the following chapters.

Chapter 4.

Feature Selection from Electronic Healthcare Records

4.1. Introduction

In this chapter, we propose a feature selection method and to eliminate anomalous data for EHRs that focus on stroke risk factors. Section 4.1 presents the sample size and feature selection in machine learning to compare between machine learning and a statistical model. Section 4.2 introduce the FconvertEHRs algorithm to eliminate anomalous data and combine with ICD-10 codes for filtering stroke risk factors as pre-processing for the prediction process. This method concept is also described with feature selection for stroke symptoms.

4.2. Sample size and Feature Selection of risk factors

4.2.1. Sample size

With a traditional Statistical approach, we need to compute the sample size. Many researchers used inferential statistics from the problem domain applied to the sample population but a machine learning approach uses all data for finding a pattern and for modelling. For example, this research uses all the data, the "population". However, the minimum number of data records should be at least 1000 records for the medical sector to obtain a good accuracy and precision. If we used more than 5000 records, the result will have a higher precision. In Figure 4-1 we compare a traditional statistical approach with Machine Learning (ML) /Artificial Intelligent (AI) approach that is Uncertainty sampling with probabilistic classifier.

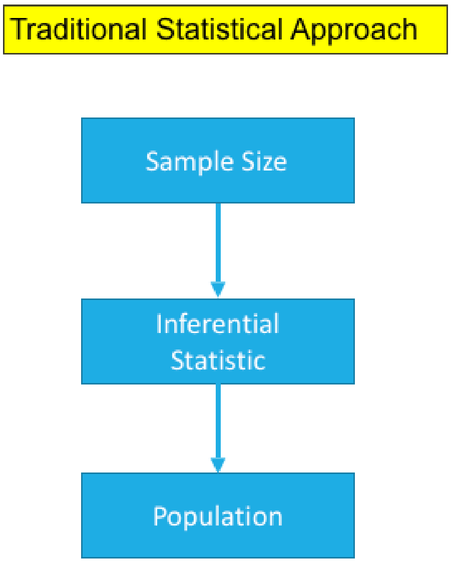


Figure 4-1 : Traditional Statistical Approach for calculate sample size.

$$P(C|w) = \frac{\exp(a + b\ \sum_{i=1}^{d} log\frac{P(W_i|C)}{P(W_i|C')})}{1 + \exp(a + b\ \sum_{i=1}^{d} log\frac{P(W_i|C)}{P(W_i|C')})}$$

C indicates class membership, and w_i is the i^{th} of d attribute values in the vector w for an instance. The instance is assigned to class C if P(c|w) exceed 0.5.

4.2.2. Feature Selection

With the Feature selection technique, the traditional statistical approach is to have inclusion and exclusion criteria for the case or problem domain depending on the observed data or intervention. In ML/AI approach, “Feature selection” will be selected by machine. The retrieval process and feature selection will select data into the algorithm for patients who have risk factors for stroke (see in Figure 4-2).

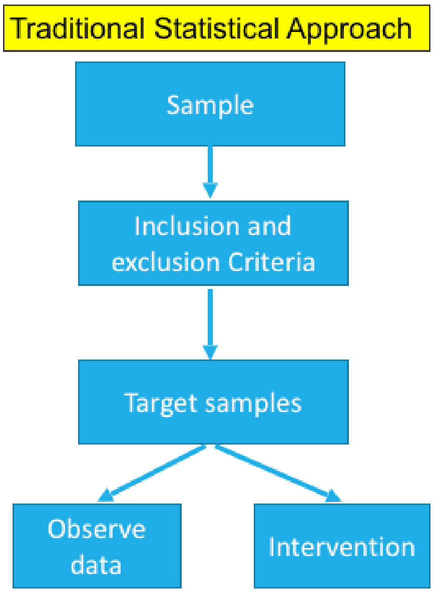

Traditional Statistical Approach
Sample
Inclusion and exclusion Criteria
Target samples
Observe data
Intervention

Figure 4-2 : The traditional Statistical approach for feature selection.

The ML/ AI approach requires finding the best estimate for β in the equation $h(t|x) = h0(t) \exp(\beta T\ x)$ (1) and $h(x) = (1 + \exp(-\beta T\ x))-1$ (2) which is typically computationally difficult, particularly given a large number of features. By introducing a complexity-based penalty term, we can identify irrelevant features and remove them from our model. The L1 regularized sparse learning problem has the following general form:

$$min_{\beta}\ g(\beta) + \ \lambda||\beta||1$$

Where $g(\cdot)$ is a convex function, β is a vector of length d, and $\lambda > 0$ is a regularization parameter. In this study, we evaluated both L1 regularized Cox model and L1 regularized logistic regression. We found that the L1 regularized feature selection gives better performance over the baselines (i.e., selecting features manually) by reducing the feature set to the most relevant ones.

For the retrieval process, the dataset of interest is compared with information stored in "*knowledge containers*" (Richter & Weber 2013). ML/AL approach for stroke patients includes a cased-based knowledge. Given output from the previous process, this process uses k-Nearest Neighbours (k-

NN) approach based on knowledge to classify each patient group into risk factors. To calculate the distance between p and q in k-NN algorithm, Equation is applied (Han, Pei & Kamber 2011).

$$dist = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

where p = (p_1, p_2, p_3,…, p_n) , q = (q_1, q_2, q_3,…, q_n) and n represents the number of dimensions. The stroke's patients are initial as p value, q as input data for comparison, and n is the size of the sample data.

4.3. EHRs management for Prediction of Stroke

In this section, a feature selection framework is applied to EHRs for the prediction of stroke. It is a process of the selection of Stroke patients that has two main steps i.e. selection of EHRs based on risk factors and the prediction process. In the first step, the null values and anomalous data were eliminated via JAVA programming. The ICD-10 codes were filtered by stroke risk factors. For EHRs, the ICD-10 codes that had been arranged by the previous process were processed and filtered again. This process combined all EHRs together and eliminated the irregular data such as negative values, null values etc. The EHRs files now consist of demographic data and group of symptom codes with risk factors. In the preparation phase, the EHRs were later converted to zero or one for defining diseases that the patients suffered (see in Table 4-1).

For the prediction algorithm, the dataset was trained by means of feature selection and retrieval process and then the LSTM-RNN prediction formula is applied. The input layer calculated the weight values based on ICD-10 codes and EHRs with risk factors of stroke. The ICD-10 codes that represent stroke risk factors are selected by using AHA guideline (Goldstein et al. 2006; Goldstein et al. 2001; The American Heart Association 2008).

This group was a knowledge-based reference that was used for computing the weight for embedding at hidden layer as input. The weight values and EHRs were integrated into LSTM-RNN layer. The output layer of prediction model represented the prediction value in a form of percentage risk (see in Figure 4-3).

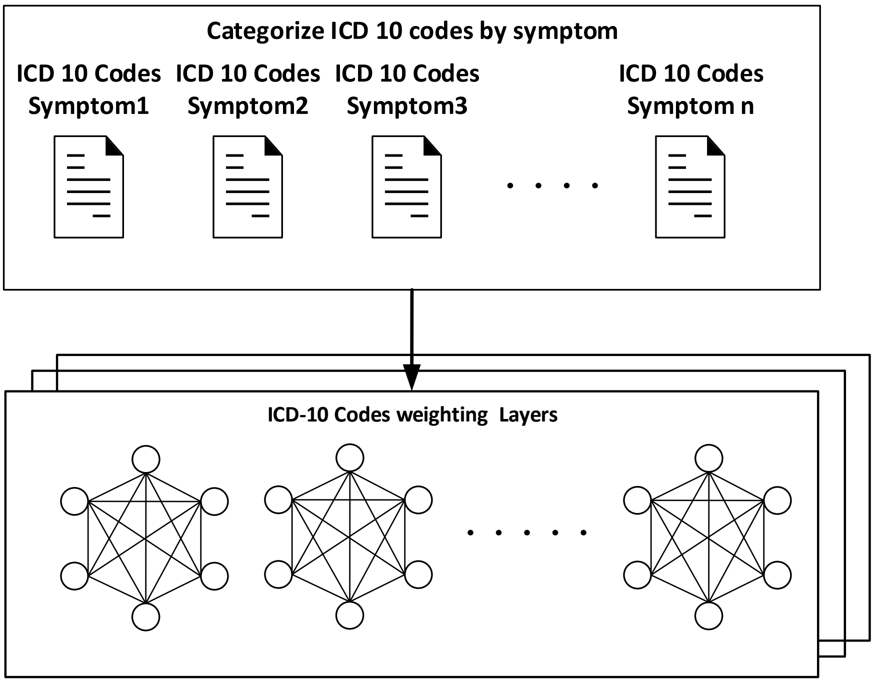


Figure 4-3 : The categorized ICD-10th codes by symptoms.

4.4. ICD-10th Complaint Electronic Healthcare Records

This framework has three steps for analytical prediction system in stroke disease as follows: manage incomplete data, classification, and prediction. First, the system will be active when it receives data from electronic healthcare records (EHRs). Incomplete data will be revised by statistical techniques to make sure that data is correct especially those data that are risk factors for stroke disease; this step is called “pre-process”. The techniques being used are as follows (Khosla et al. 2010):

- Column mean: enter the mean of the feature's observed values in each missing value.
- Column median: enter the median of the feature's observed values in each missing value.
- Imputation through linear regression
- Regularized Expectation Maximization (EM)

The ICD-10th code that had been applied in EHRs can be used for training probabilistic classifiers from the large data sets of EHRs. Specifically, we consider multilabel classification of stroke symptoms and risk factors for training and modelling by selection based on the AHA list of stroke factors (Goldstein et al. 2006; Goldstein et al. 2001; The American Heart Association 2008). Normally, ICD-10th code presented in main categories and sub-categories such as I65 (Occlusion and stenosis of vertebral artery) is the main category, and I65.1 (Occlusion and stenosis of basilar artery) is the sub-category. The code risk factor that has been chosen consists of 70 main-categories and about 200 sub-categories, all together there are 227 factors.

Table 4-1 :EHRs records with stroke's risk factors.

Age	Gender	B980	E108	E119	……..	I64	Z721	Z920
74	1	1	0	0	……..	0	0	0
65	2	1	0	0	……..	0	0	0
61	1	1	0	1	……..	0	0	0
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮
50	1	0	1	1	…….	0	0	0
62	2	0	1	0	…….	0	0	0
76	1	0	1	1	…….	1	0	0
83	2	0	0	1	…….	0	0	0

In the prediction models, ICD-10th codes in each symptom are converted to a group for the prediction method. The group of symptoms will be applied for training and testing in CBR and deep learning algorithm (see in Figure 6-1). Those symptoms are related to other diseases and the models can be used for calculating weight values in stroke.

This research starts with reorganizing and normalizing the EHRs dataset, which has anomalous values. We started with the FConvertEHRs algorithm. This algorithm can be combined with ICD-10th codes in order to eliminate anomalous values in patient records. In the elimination process, it converts and deletes anomalous data such as null values and multiple records. Next, each record is compared with ICD-10th codes. The outcome of this step is zero or one, representing whether or not the patient in the record has this symptom (see in Table 4-1). Then, the size of the output file becomes smaller than that of the original and is now suitable for calculation in the next step (see in Algorithm 1).

Algorithm 1: Filtering and Convert EHRs.

function FCovertEHRs **returns** a EHRs dataset

inputs: *examples*, a set of Electronic Healthcare Records, each with input patient records (**x1,x2,x3,….,xn)** and output patient record (**y**), a multi-patient records compared with ICD-10 code, activation function ConvertEHRs

```
repeat
  for each Patient records in EHRs dataset do $Patient\{age \in Patient|0 > age < 120\}$
    for each ICD-10 node $j$ in the input records do $DiagCode_j \leftarrow ICD10\ [j]$
      write Patient records to new patients dataset
until end of file
return Stroke patients' datasets
```

4.5. Summary

In this chapter, we combined the EHRs with ICD-10th code by using FConvertEHRs algorithm for feature selection. It prepared a dataset for the prediction process discussed in the next chapter. The diagnoses of diseases are represented by International Classification of Diseases, 10th Revision (ICD-10th) code in each patient record. The data contain records of all patients and their disease codes. However, this research is interested in only some particular disease codes and therefore the data must be filtered to remove irrelevant data. The filtering process takes a considerable amount of time due to the large size of the dataset as well as the large number of the disease codes.

Chapter 5.

Case Based Reasoning Framework for Stroke

5.1. Introduction

This chapter describes the case-based reasoning framework for stroke. Section 5.1 presents the main components of the framework and techniques applied in CBR for stroke domain. It gives a brief description of case base management issues, data collection and the prediction functional implementations. CBR can be utilized to create a repository of the information of the electronic healthcare records (EHRs) that includes stroke patients and non-stroke patients who have risk factors for stroke. For a new case where the diagnosis and prognosis are yet to be determined, a similar case can be retrieved from the previous case database, to provide useful information or suggestions for a care plan.

5.2. A CBR framework of prediction model in stroke patients

The research conducted a CBR framework of prediction model using Support Vector Machine (SVM), k-Nearest Neighbours (k-NN), and Supervised learning techniques applied by randomly selecting 80% of the data for training, leaving 10% for testing and used in prediction for all technique.

The prediction technique called "Case-based reasoning (CBR)" uses previous cases for decision-making for new cases. CBR is a methodology for solving problems that uses previous data or memorized problem situations called cases. The processes of CBR system follow four main steps as *retrieve, reuse, revise, and retain* (Aamodt & Plaza 1994) . The new case starts at the top of stage, where an input is entered into the system. The previous case is compared to the new case and starts *retrieve* step. In a practical CBR system a comparison is made between all the cases in the system and the new case and the output will list the ranking of similar cases.

The CBR systems have many application areas in the healthcare sector which have provided solutions for diagnosis and treatment of diseases based on past experiences (Ahmed, Banaee & Loutfi 2013; Anaissi et al. 2015; Arshadi & Jurisica 2005; Chattopadhyay et al. 2013; Kiragu & Waiganjo 2016; Sharaf-El-Deen, Moawad & Khalifa 2014a). The CBR-based expert system used the k-Nearest Neighbours (k-NN) algorithm to search k similar cases that focus on the Euclidean distance measure. This research has complex data and the CBR was applied by using machine-learning and data-mining techniques based on the Electronic Healthcare Records, solving problems by using patient health records including demographic data, stroke symptoms, and risk factors. These are compared with previous cases by using k-NN and SVM algorithm. The weight-feature technique calculates and retrieved and compared among previous cases and new cases. The results shown what is the probability of stroke symptoms for new patients using similar previous cases to help predict the risk of a stroke for this new patient.

The prediction outcomes from k-NN and SVM are compared with the observed outcome for validation. The outcomes using this approach not only assist in stroke disease decision-making, but will also be very useful for prevention and early treatment of patients.

With the Support Vector Machine algorithm, we used the backpropagation with weight or without weight backtracking for predictors that assumed independence of the predictor variables and repetition for the neural networks' training. The cross-validation operand consisted of two components, training and testing. The training component contained a Neural Network algorithm and k-NN algorithm, defined by arguments in R-programming.

Initially, the patient's historical data consists of a dataset of stroke patients in the core system. It contains a record of patients who have stroke disease. In addition, the system consists of 6 processes for prediction of a stroke patient which are clustering process, retrieval process, reusing process, prediction, review process and store process. The first process is where the new case or new problems are input for this stage, we start with clustering data from EHRs that are separately grouped by age, gender, race which are risk factors that cannot be changed. This process uses k-mean clustering techniques. Next, the retrieval process uses k-NN classification techniques. This process will retrieve previous cases that entered general information from EHR as personal information, diagnosis, disease, laboratory results, clinical notes, and medical knowledge respectively for comparing with the new case. The result may be a new case problem or similar case problem, from patients' healthcare records from the hospital's database system. The reuse process aims to match cases that are relevant to the given risk factors from the previous process. Finally, the prediction process uses machine learning technique for comparing and prediction for medical case-based reasoning. The result shows the likelihood of patient to suffer a stroke represented as a percentage risk from this result. It is then sent to the review processes for rechecking the result before it is stored in case-based reasoning. It means that any results are knowledge-based to compare with future cases (see in Figure 5-1 and 5-2).

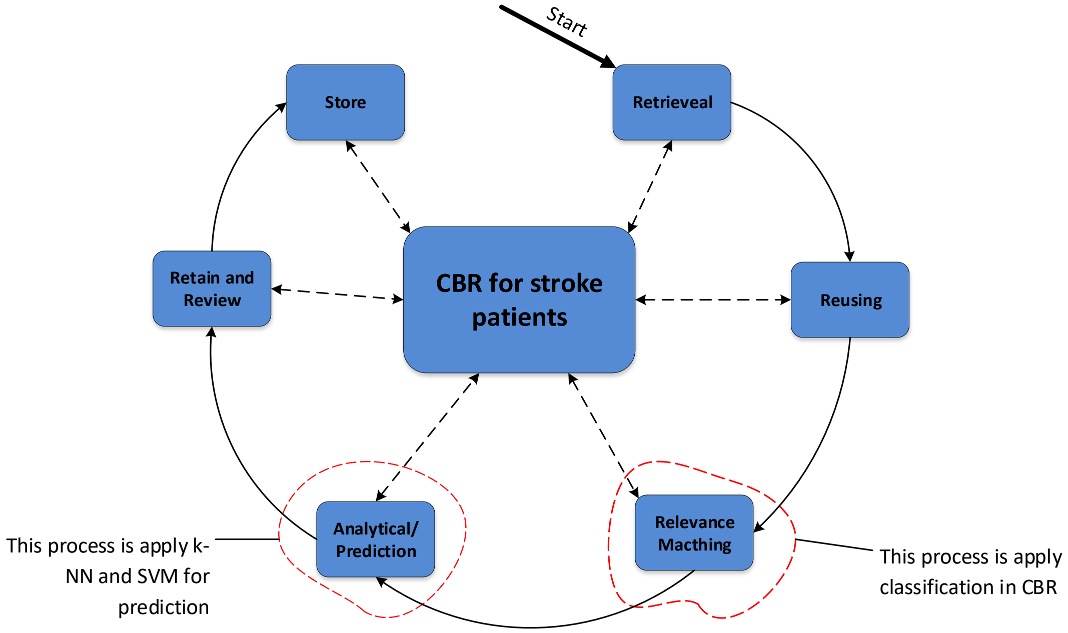


Figure 5-1: An overview of the case-based reasoning framework for the prediction analytical system in stroke patients (Chantamit-o-pas & Goyal 2018a).

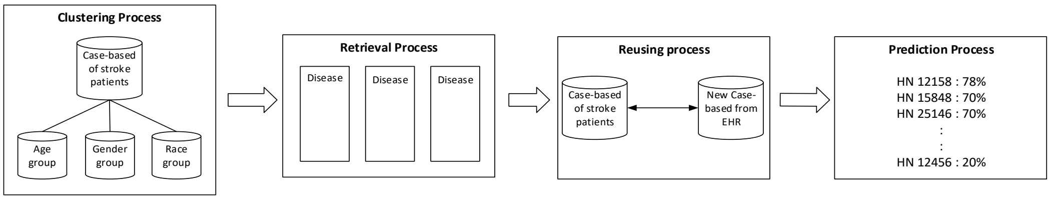


Figure 5-2 : Flowchart of the case-based reasoning for prediction of stroke patients (Chantamit-o-pas & Goyal 2018a).

A flowchart of the proposed framework is presented in figure 5-2. The detail of the processes described are as follows:

a) Clustering process –this process aims to cluster stroke patient records, based on age, gender, and race of patients. Those clusters are important factors to predict stroke disease. K-Mean clustering technique is applied for finding groups to partition n observations into k clusters. The basic algorithm is given by equation (1)

   (MacQueen 1967):

$$j = \sum_{i=1}^{k} \sum_{x \in S_i} \|x_i - c_i\|^2 \qquad (\mathbf{1})$$

   We assume that (x_1, x_2, x_3,…., x_n) is a collection of observations; where x_i is the i^{th} dimensional real vector. The observations are partitioned into k groups; s = {s_1, s_2, s_3,…, s_k}, and c_j is mean of s_j.

b) Retrieval process – this process is retrieved in which electronic healthcare records (EHRs) are compared with information stored in "*knowledge containers*" (Richter & Weber 2013). A CBR system for stroke patients includes a case-based knowledge system. Given the output from the previous process, this process uses FconvertEHRs function and ICD-10 code based on knowledge to classify each patient group into risk factors (see in Figure 5-3).

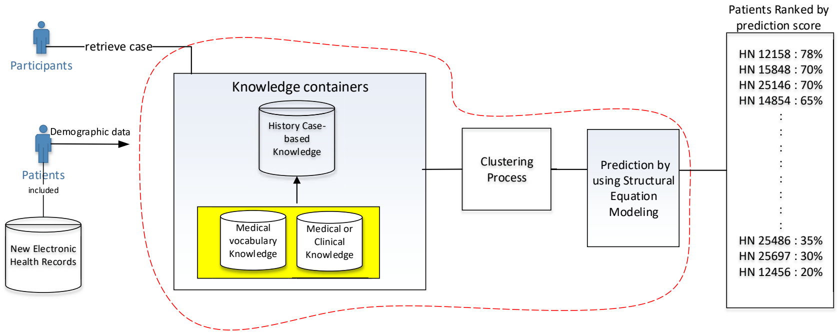

Participants
retrieve case
Demographic data
Patients
included
New Electronic Health Records
Knowledge containers
History Case-based Knowledge
Medical vocabulary Knowledge
Medical or Clinical Knowledge
Clustering Process
Prediction by using Structural Equation Modeling
Patients Ranked by prediction score
HN 12158 : 78%
HN 15848 : 70%
HN 25146 : 70%
HN 14854 : 65%
HN 25486 : 35%
HN 25697 : 30%
HN 12456 : 20%

Figure 5-3: The detail of the case-based reasoning for prediction of stroke patients (Chantamit-o-pas & Goyal 2018a).

c) Reusing process – this process aims to match cases that are relevant to the given risk factors from the previous process. As we mentioned above, cases are collected from the real cases in the hospital and stored in the knowledge container. In this research, we use those cases for stroke prediction in the next process.

d) Prediction process - data mining and statistical methods are well-known for dealing with medical data analysis and prediction. To properly select tools and develop prediction models, general and simple guidelines are necessary and required (Bellazzi & Zupan 2008). This process uses Support Vector Machine (SVM) and k-Nearest Neighbour (k-NN) because the data type has to be shown to multiple groups from risk factors such as diabetes data set, heart disease data set, behaviour data set, and so on, which are data sets that depend on other diseases and are related to stroke. In term of stroke disease, there exist various risk factors that are useful for effectively predicting disease. The analysis process identifies variables. The age values are independent variables (called “primary variables”) and risk factors are dependent variables. The algorithm then analyses and predicts based on previous cases and current patient records. It processes case-by-case with other disease groups that relate to risk factors. After that, the output presents in terms of stroke risk estimation. For data sets of stroke three main groups ae used: the risk factors that cannot be changed, the risk factors that can be changed, treated or controlled, and other risk factors that are less well-documented. The first group is demographic data such as age, race, gender, and prior stroke. The second group consists of behaviour and historical disease from EHR such as Hypertension, Heart Disease, Atrial Fibrillation, Peripheral Artery Disease, Carotid, Diabetes Mellitus, Obesity, High Blood Cholesterol, Sickle Cell Disease, First Stroke, Alcohol Abuse, Poor diet, Physical Inactivity, Drug Abuse, and Smoking. The last group includes the hometown of the patient, socioeconomic factor, alcohol abuse, and drug abuse. These are loaded from current patient records. For stroke patients, incidence of stroke is required and loaded from historical records.

The CBR-based expert system used the k-Nearest Neighbours (k-NN) algorithm to search k similar cases that focus on the Euclidean distance. The weight-feature technique calculates retrieves and compares between previous cases and new cases. In term of validation, a cross-validation operand is used for training and testing and 10-epochs parameter in k-NN algorithm. The training model does not include cerebrovascular code (I64) that is initial formula in factor variable for testing process (shown in Algorithm 2).

Algorithm 2: k-Nearest Neighbours (k-NN) algorithm for stroke's prediction process.

```
function k-NN returns k-NN_values
inputs: Stroke dataset : Age, Gender, and Stroke's risk factors , each with input (X) to k-NN training pattern, Test vector to test vector, output store a value, initial weight value to k and bias vector

    $sData \leftarrow$
$\begin{cases} f\ feature\ with\ \text{Stroke's patient and nonStroke's patient who have risk's factor} \\ \text{Selected From }\ FConvertEHRs\ \text{function} \end{cases}$

    $i \leftarrow number\ of\ epoch$
    $x \leftarrow percent\ of\ sample$

$Selected\ a\ sample\ size\ in\ \ x\%\ \ and\ test\ size\ in\ (total - x\%) from\ sData$
$\boldsymbol{Assign\ \ matrix\ \ to}\ x_{vector} from\ sample\ without\ Stroke's\ factor$
$\boldsymbol{Assign\ \ matrix\ \ to}\ test_{vector} from\ sample\ with\ risk's\ factor$
        $\boldsymbol{Prediction} \leftarrow Stroke\ pattern\ with\ test_{vector}$ a k to i $\boldsymbol{from}\ k - NN\ WITH\ CROSS\ VALIDATION()$
$\boldsymbol{Save}\ Prediction\ \boldsymbol{to}\ external\ file$
Return k-NN prediction value
```

In Support Vector Machine algorithm, we used backpropagation with weight or without weight backtracking for predictors that assumed independence of the predictor variables and repetition for the neural networks' training. This method used for sampling method in 10-fold cross validation for prediction, and the SVM-

Kernel was used for linear analysis. The cost of constraints or 'C'-constant of the regularization term is 0.1 (shown in Algorithm 3).

Algorithm 3: Support Vector Machine (SVM) algorithm for stroke's prediction process.

```
function SVM returns $SVM\_values$
inputs: Stroke dataset : Age, Gender, and Stroke's risk factors , each with input (X) to SVM training pattern, Test vector to test vector, output store a value
  $sData \leftarrow \begin{cases} f\ feature\ with \text{ Stroke's patient and nonStroke's patient who have risk's factor} \\ \text{Selected From } FConvertEHRs \text{ function} \end{cases}$
        $i \leftarrow number\ of\ epoch$
        $x \leftarrow percent\ of\ sample$
    $Selected\ a\ sample\ size\ in\ x\%\ and\ test\ size\ in\ (total - x\%) from\ sData$
    $\boldsymbol{Assign\ matrix\ to}\ x_{vector} from\ sample\ without\ Stroke's\ factor$
    $\boldsymbol{Assign\ matrix\ to}\ y_{vector} from\ sample\ with\ risk's\ factor$
    $Find\ the\ model\ prediction\ (Stroke\ pattern) =$
$\boldsymbol{Training}\ x_{vector}\ \boldsymbol{and}\ y_{vector}\ \boldsymbol{from}\ SVM()$
    $\boldsymbol{Prediction} \leftarrow Stroke\ pattern\ with\ a\ sample\ size\ in\ sData\ \boldsymbol{from}\ predict()$
    $\boldsymbol{Save}\ Prediction\ \boldsymbol{to}\ external\ file$
    Return SVM prediction value
```

e) Review process - After that, the output will be verified and sent to participants or nurses. The result shows percentage of stroke for individual patient.

f) Store process - The prediction results of patients who have risk factors for stroke disease will be stored in CBR system for reuse in the future. This information can help in decision-making for participants in order to make suggestions and warnings for patients as part of a care plan, life style, quality of life, and behaviour and so on. Finally, the outputs are updated in historical case-based knowledge.

5.3. Summary

This chapter proposed a CBR framework for stroke. This framework contains a record of patients who have stroke disease. In addition, the system consists of 6 processes for prediction in stroke patient as clustering process, retrieval process, reusing process, prediction, review process and store process. The first process where the new case or new problems are input for this stage, it started with clustering data from EHRs that are separately grouped by age, gender, race which are risk factors that cannot be changed. This process uses k-mean clustering techniques. Next, the retrieval process used Support Vector Machine (SVM), k-Nearest Neighbours (k-NN), and Supervised learning techniques. These techniques have been widely applied in clinical decision making such as predicting occurrence of a disease or diagnosis, evaluating prognosis or outcome of diseases and assisting clinicians to recommend further treatment of diseases. It will retrieve previous case that entered general information from EHR as personal information, diagnosis, disease, laboratory results, clinical notes, and medical knowledge respectively for comparing with the new case. The result may be a new case problem or similar case problem, from patients' healthcare records from hospital's database system. The reuse process aims to match cases that are relevant to the given risk factors from the previous process. Finally, the prediction process used machine learning technique for comparing and prediction for medical case-based reasoning. The result shows the likelihood of a patient suffering a stroke.

Chapter 6.

Deep Learning Framework for Stroke

6.1. Introduction

This chapter describes the Deep Learning framework for stroke (DL). This technique employs learning from data with multiple levels of abstraction by computational models that are associated with multiple processing layers. This method is intended to discover complex structure in big data sets by using the backpropagation algorithm, Recurrent Neural Network (RNN) and a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to predict the result. It contains the main components of the framework and techniques applied in DL for the stroke domain. It gives a brief description of framework management issues and the prediction functional implementations. In addition, the stroke algorithm can be utilized to predict using the information from the electronic healthcare records (EHRs) that includes stroke patients and non-stroke patients who have risk factors for stroke. For a new case where the diagnosis and prognosis are yet to be determined, a similar case can be retrieved from the previous case database, to provide useful information or suggestions for a care plan.

6.2. Deep Learning Algorithm

The Deep Learning method is intended to discover complex structure in huge data sets by using advanced mathematical algorithms to predict the result. The machine can learn from source and change its internal parameters by computing the representation in each layer to form the representation in the previous layer.

Algorithm 4: The backpropagation algorithm for learning in multi-layer networks.

```
function Deep Learning for prediction returns a neural network
inputs: examples, a set of data, each with input vector x and output vector y network, a multi-layer network with L layers, weights $W_{j,i}$, activation function g
repeat
  for each e in examples do
    for each node j in the input layer do $a_j \leftarrow x_j[e]$
    for $\ell$ = 2 to M do
        $in_i \leftarrow \sum_i W_{j,i} a_j$
        $a_i \leftarrow g(in_i)$
    for each node i in the output layer do
        $\Delta_j \leftarrow g'(in_j) \times (y_i[e] - a_i)$
    for $\ell$ = M − 1 to 1 do
        for each node j in layer $\ell$ do
        $\Delta_j \leftarrow g'(in_j) \sum_i W_{j,i} \Delta_i$
          for each node i in layer $\ell$ + 1 do
            $W_{j,i} \leftarrow W_{j,i} + \alpha \times a_j \times \Delta_i$
until some stopping criterion is satisfied
return NEURAL NETWORK
```

This model is applied in this study to the main deep learning model for learning in multi-layer networks. This model is supervised concept extractor for the original dataset samples. A backpropagation is treated as a multi-layer feedforward neural network with hidden layers or multiple-layer neural network (see in Figure 6-1a). Each hidden unit can be considered as multiple outputs perceptron network and can be considered a soft-threshold linear combination of the hidden units which are equivalent to the output unit perceptron.

We need to consider a multiple output unit for multi-layer networks. Let (x, y) be a single sample with its desired output labels $y = \{y_1, \dots \dots .., y_i\}$. The error at the output unit is just $y - h_W(x)$, which can be used this to adjust the weights between the hidden layers and the output layers. It is this process that produces the error at the hidden layers in terms of equivalence to the error at the hidden layers. This is subsequently used to update the weights between the input units and the hidden layers as in algorithm (see in Figure 6-1a).

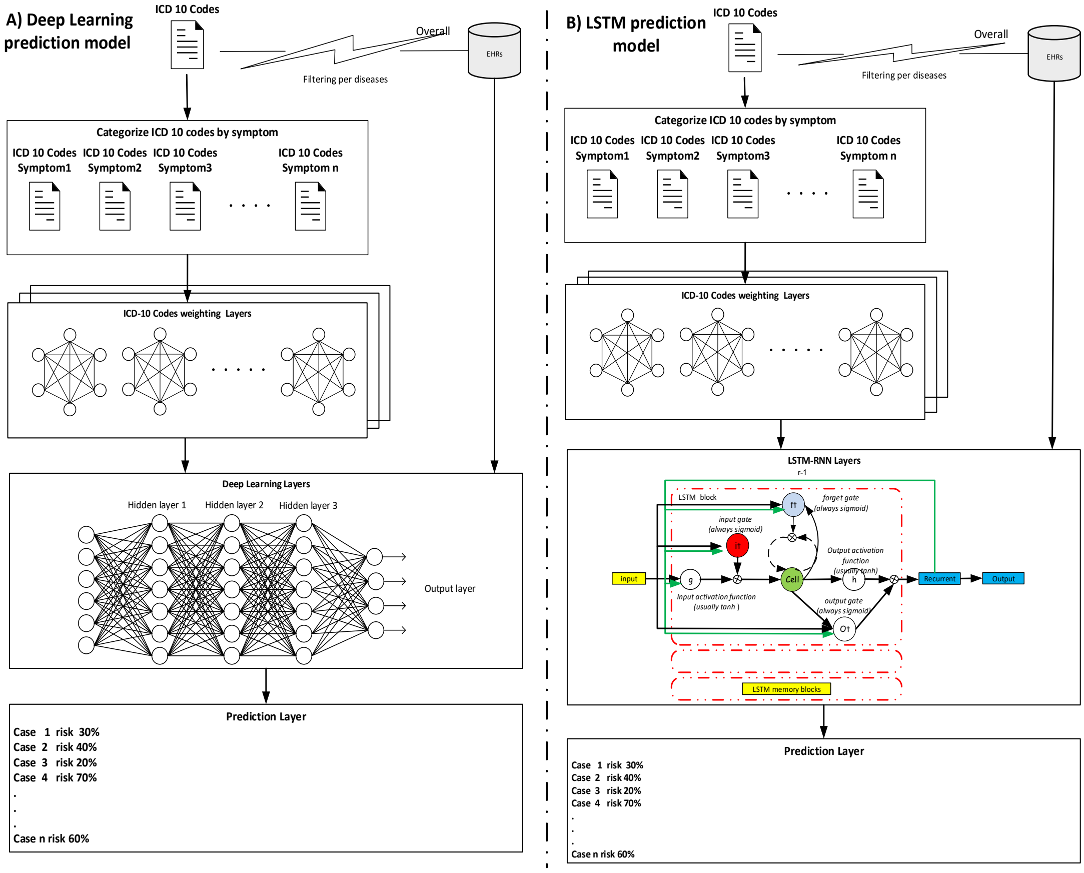


Figure 6-1: Prediction model; A) Deep Learning prediction model and B) LSTM prediction model.

In the prediction models, we applied two models; deep learning algorithm and LSTM algorithm (see in Algorithm 4 and 5). The ICD-10th codes structure have relationships with other diseases that are represented in a star topology. This layer shares similarity between deep learning and LSTM methods.

Next, this algorithm relies on a Recurrent Neural Network (RNN) and on a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) that are used in prediction (see in Figure 6-1b). When applied to a large-scale aggregated file from the Electronic Healthcare Records, the prediction model for training to recognize stroke symptoms and risk factors is based on ICD-10th standard. The equations of the model appear as follows:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}{}^{x_t}) \tag{1}$$

$$a_i \ = \ \sigma(W_a x_i + W_a h_{i-1} + b_a) \tag{2}$$

$$g_i \ = \ \sigma\big(W_g x_i + W_g h_{i-1} + b_g\big) \tag{3}$$

$$\tilde{h}_i \ = \ tanh(W_h x_i + g_i \circ W_h h_{i-1} + b_h) \tag{4}$$

$$h_i \ = \ a_i \circ \ h_{i-1} + (\,1 - \ g_i) \ \circ \ \tilde{h}_i \tag{5}$$

Where the W terms are weight matrices value, W_h, W_b, and W_a, are diagonal weight values for next layer to layer connections. The b terms are bias vectors. The logistic sigmoid function is represented by σ. The input gate, forget gate, and output gate are represented by a, g, and n respectively. All of them are the same size as the cell output activation vectors h_i, ∘ is the element product of the vector, $\tilde{h}_i$ is the cell input and cell output activation function, generally and in this research network is *tanh*.

$$h_t = \tanh\begin{pmatrix} W(I64{\sim}Age) + W(I64{\sim}Gender) + \\ W(I64{\sim}Stroke's\;risk\;factors) \end{pmatrix} \tag{6}$$

Overall, the machine learned from the model and pattern. The group of codes was computed for finding the weight value in each node. The learning rate term is 0.1 and epoch is 10 and the network types are Deep learning, RNN and LSTM.

Algorithm 5: Long Short-Terms Memory – Recurrent Neural Network (LSTM-RNN) algorithm for stroke's prediction process.

```
function LSTM-RNN returns LSTM_values
inputs: examples (x), a set of patient records, each with input ($\boldsymbol{a_i}$ ) to LSTM unit, each forget gate's ($g_i$) activation to vector, output gate's ($\tilde{h}_i$) activation vector, output vector $\tilde{h}_t$ to LSTM unit, a Memory Cell ($C_t$ ) store a value, weights $W_{a,i}$ and bias vector

repeat
  for each x in patient records do
      Step 1: input gate for the states of the memory cells at time t:

          $C_t =$
      tanh ( $W(I64{\sim}Age) + W(I64{\sim}Gender) + W(I64{\sim}Stroke's\ risk\ factors)$)
          $a_i = \sigma(W_a patients_i + W_a h_{i-1} + b_a)$

      Step 2: forget gate for the activation of the memory cells at time t:

          $g_i = \sigma\left(W_g patients_i + W_g h_{i-1} + b_g\right)$

    Step 3: output gate for the new state of the memory cells with their output:

          $\tilde{h}_i = tanh(W_h patinets_i + g_i \circ W_h h_{i-1} + b_h)$

          $h_i = a_i \circ h_{i-1} + (1 - g_i) \circ \tilde{h}_i$

until some stopping criterion is satisfied
return LSTM_values
```

6.3. Stroke Algorithm

In term of a stroke prediction process, we started with the FConvertEHRs algorithm that produced the new dataset for initiating the process. Next, we used two algorithms – Deep Learning with backpropagation algorithm and LSTM algorithm - in the training and prediction process (see in Algorithm 4 and 5). Both algorithms are learning with any risk factors (see in Algorithm 6). All techniques demonstrated the results of training using a randomized 80% of the dataset. For testing, 10% of the dataset is used. The learning rate is 0.1 and a number of iterations (10-epochs) were used for prediction. The prediction process is computed in both techniques. After that, the results are stored in external files and returned as output variables.

Algorithm 6: Stroke prediction (DL or LSTM technique)

inputs: Stroke dataset : Age, Gender, and Stroke's risk factors , each with input $\boldsymbol{i}$ to LSTM unit, each forget gate's (f_t) activation to vector, output gate's (o_t) activation vector, output vector h_t to LSTM unit, a Memory Cell (C_t) store a value, weights $W_{j,i}$ and bias vector

output: **Prediction:** a percentage of chance of stroke symptoms

```
$sData \leftarrow \begin{cases} f\ feature\ with \text{ Stroke's patient and nonStroke's patient who have risk's factor} \\ \text{Selected From } FConvertEHRs \text{ function} \end{cases}$
$i \leftarrow number\ of\ epoch$
$n \leftarrow number\ of\ iterations$
$x \leftarrow percent\ of\ sample$

$Selected\ a\ sample\ size\ in\ \ x\%\ from\ sData$
$\boldsymbol{Assign\ \ matrix\ \ to}\ x_{vector} from\ sample\ without\ Stroke's\ factor$
$\boldsymbol{Assign\ \ matrix\ \ to}\ y_{vector} from\ sample\ with\ risk's\ factor$
$\boldsymbol{for}\ i\ \leftarrow 1\ \ \boldsymbol{to}\ n\ \boldsymbol{and}\ batch_{size}\ \ \boldsymbol{do}\ //\ \boldsymbol{training\ process}$
    $\begin{cases} \boldsymbol{Training}\ x_{vector} \boldsymbol{and}\ \ y_{vector} with\ learningate\ =\ 0.1 \\ \boldsymbol{from}\ LSTM - RNN()\ \boldsymbol{or}\ \ DL() \\ Find\ the\ model\ prediction\ (Stroke\ pattern) \end{cases}$
$\boldsymbol{end\ for}$
$\boldsymbol{Save}\ Stroke\ pattern\ \ \leftarrow Find\ the\ best\ model\ from\ training\ loop$
$\boldsymbol{Assign\ matrix\ to}\ test_{vector}\ from\ sample$
$\boldsymbol{Prediction}\ \leftarrow\ \ Stroke\ pattern\ with\ test_{vector}\ \boldsymbol{from}\ LSTM - RNN()\ \boldsymbol{or}\ DL()$
$\boldsymbol{Save}\ Prediction\ \boldsymbol{to}\ external\ fil$
$\boldsymbol{return}\ \ Prediction$
```

6.4. Summary

We have produced a deep learning framework for stroke. This method is intended to discover complex structure in a huge data set by using an advanced mathematical algorithm to predict the result. The machine can learn from source and change its internal parameters by computing the representation in each layer to form the representation in the previous layer. In prediction models, we applied two models; deep learning algorithm and LSTM algorithm. The ICD-10 codes structure have relationships with other diseases that are represented in star topology. This layer shares a similarity between deep learning and LSTM methods. Next, the Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) architecture contains computation units in each memory block in the recurrent hidden layer. It employs the use of data obtained from patient health records and a comparison between previous cases, observation, or inspection. Stroke has complex risk factors. In term of stroke prediction algorithm, we started with FConvertEHRs algorithm that produced the new dataset for preparing process. Next, we used two algorithms – Deep Learning with backpropagation algorithm and LSTM algorithm - in training and prediction process.

Chapter 7.

Experiments and results

7.1. Introduction

This chapter used CBR framework and Deep Learning framework with two datasets: heart dataset; and Electronic Healthcare Records (EHRs) for analysis and prognosis. Section 7.1 describe the data source that is used in the two frameworks. Section 7.2 presents the experiment with the heart dataset. Section 7.3 presents an analysis and prognosis with EHRs. Section 7.4 compares the result between CBR and deep learning techniques in EHRs.

7.2. Data Source

7.2.1. Heart datasets

A deep learning model was applied using the heart disease dataset (available at UCI Machine learning website) for testing the algorithm. It has 899 records and 76 attributes per record. It contains Patient Number, Social Security Number, Age, Gender, Blood pressure, type of chest pain, Cigarettes, Family history, Hypertension, Cholesterols, Years, EKG (day/month/year), Heart rate, Nitrates, calcium channel blocker, and so on. It covers four hospitals at medical centres in Hungary, Switzerland, Cleveland and Long Beach, Virginia.

7.2.2. Electronic Healthcare Records dataset

This research used aggregated files of Electronic Healthcare Records (EHRs) from the Department of Medical Services, Ministry of Public Health of Thailand collected between 2015 and 2016 (326,134 records). This research was granted the ethics approval by

University of Technology Sydney, Australia (The ethics approval number UTS HREC ETH17-1406). EHRs consisted of demographic data, diseases codes (ICD-10 codes), Dates of diagnosis, clinic types, and types of diagnosis (see in Table 4-1). According to the source, EHRs data had multiple value dependencies that were cleaned of anomalous data and filtered by ICD-10 codes for risk factors of stroke. Incomplete (missing) data were removed from the dataset so as not to affect the performance of the prediction process. Consequently, our new EHRs dataset (see in Table 4-1) actually had 96,190 records of the stroke patients and non-stroke patients with potential risk factors (see in Figure 7-1).

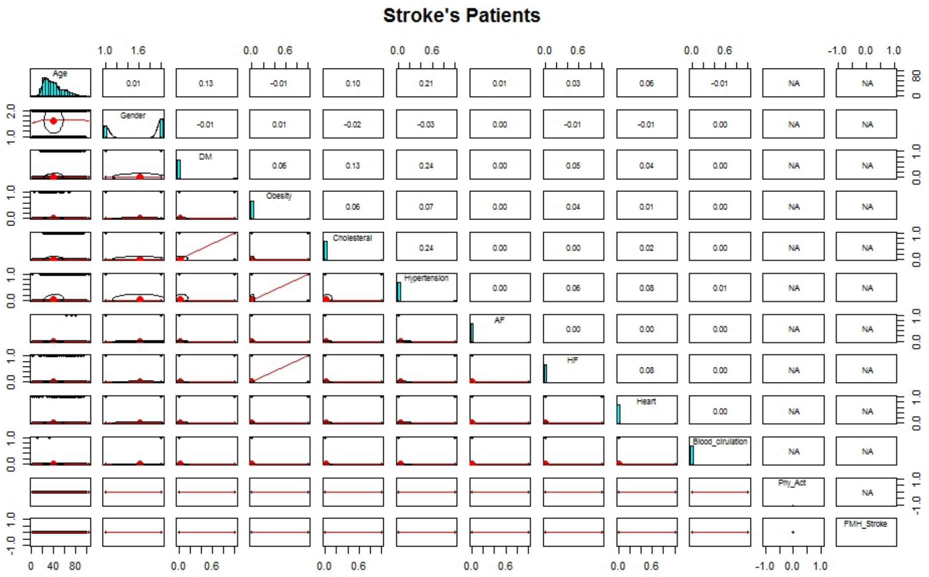


Figure 7-1: The main risk factors of stroke patients in electronic healthcare record.

7.3. Experiments in Heart Datasets

Heart dataset have been used in this research for validation, error estimation and experiments. The algorithms Naïve Bayes and SVM are widely used in prediction. The Heart dataset was used for comparison of the three models: Naïve Bayes; Support Vector Machine (SVM); and Deep Learning. All algorithms used a learning rate of 10-epochs. All techniques demonstrated the results of training 80%. For testing, 10% of the dataset is used. For the Deep Learning method was applied with a learning rate of 0.1 and the number of iterations (10-epochs) were used for prediction.

7.3.1. Choosing number of attributes in Heart datasets

In the Heart dataset, we selected eleven attributes from the original dataset such as Age, Gender, Blood Pressure (Low and High), Chest Pain, Cigarettes, Family History, Hypertension, Cholesterols, Heart Rate, and blood vessels. These attributes are related to risk factors for stroke used for prediction, as described in AHA guideline (Goldstein et al. 2006; Goldstein et al. 2001). We initialize an attribute for classification of stroke patients by using the stroke attributes. Thus 12 attributes were used for prediction and the dataset now has 899 patients that included both stroke patients and non-stroke patients (see in Figure 7-2).

All algorithms used cross validation in training in order to avoid over fitting and to achieve better generalized results. The prediction process also used training and test sample and used a 10-fold cross validation.

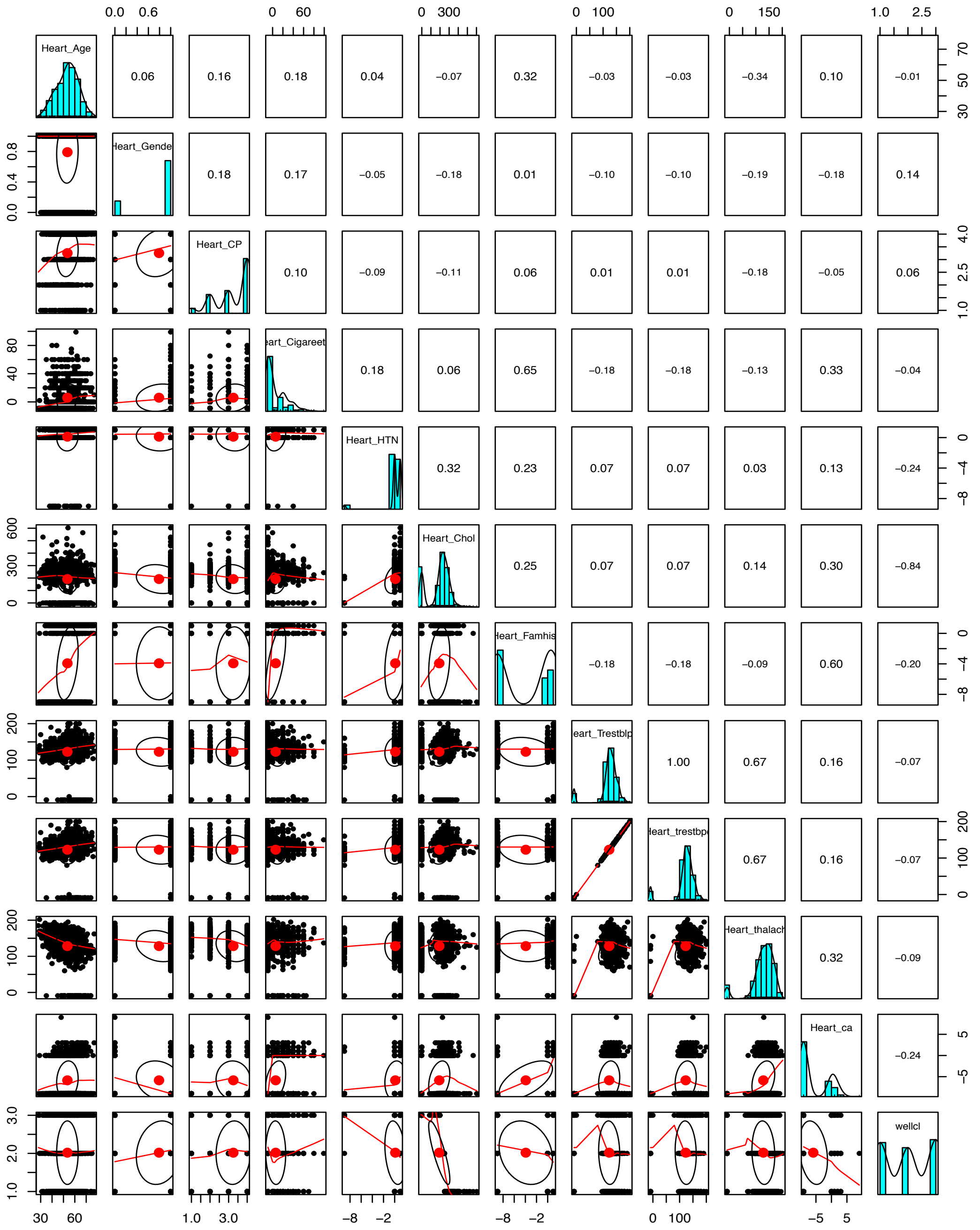


Figure 7-2: The main risk factors for stroke patients in the Heart dataset

7.3.2. Validation of prediction in Naïve Bayes Algorithm

In order to demonstrate the result of a prediction algorithm, experiments are preformed using the Naïve Bayes algorithm with the Heart dataset. We used the classifier for discrete predictors that assumed independence of the predictor variables, and Gaussian distribution for metric prediction. The attributes with missing values and the corresponding table entries were omitted for the prediction. In order to test, we calculated the values of the *a priori* probabilities of 0.4784 (Non-Stroke) and 0.5215 (Stroke). This result show that patients who have risk factors for stroke, have a 50% chance of suffering stroke symptoms. In term of medical prognosis, this does not mean that you will necessarily have a stroke in the future.

7.3.3. Validation of prediction in Support Vector Machine Algorithm

The SVM method was used for sampling in 10-fold cross validation for prediction, and the SVM-Kernel was used for linear analysis. The cost of constraints or 'C'-constant of the regularization term is 0.1, and also their best performance is 0.465. The prediction values of Naïve Bayes and SVM of 0 (Non-stroke) or 1 (Stroke) indicates whether the patients are suffering from a stroke or not.

7.3.4. Validation of prediction in Deep Learning

In terms of Deep Learning, we used a Deep Neural Network model by using a feedforward multi-layer artificial neural network. The computation shows that Mean Square Error (MSE) is 0.2596. This value indicates the confidence has best performance for prediction of stroke. The Mean Value and Standard Deviation are different for each technique, so the percentages of predicted stroke with the three models are shown in table 1indicating that during training procedure, the mean values for prediction in Deep Learning is higher than for the other two techniques (see in Table 7-1). The Deep Learning technique prediction shown as percentage probability of having a stroke is different from Naïve Bayes and SVM. The number of edges with Deep Learning are plotted (see in Figure 7-3). An edge of weight value in each layer is the predictor with risk factors. A prediction of stroke implied some concepts of a medical domain that preferred for good performance and explanation.

Table 7-1 : Comparison of three techniques used for prediction of Stroke.

Techniques	Mean Value	Standard Deviation
Naïve Bayes	49%	0.038
Support Vector Machine	47.78%	0.1106
Deep Learning	36.73%	0.084

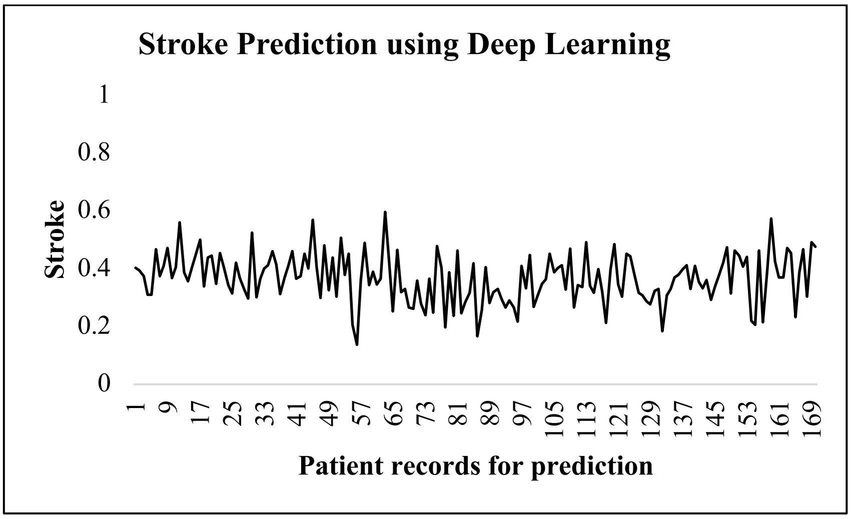


Figure 7-3: Stroke prediction using deep learning with heart dataset (Chantamit-o-pas & Goyal 2017).

7.4. Experiments in EHRs dataset

We conducted a comparison between five techniques: Support Vector Machine (SVM); k-nearest neighbour (k-NN); Backpropagation; RNN; and LSTM- RNN. All techniques demonstrated the results of training 80%. For testing, 10% of the dataset is used. In terms of the Deep Learning method, backpropagation, RNN, and LSTM-RNN algorithm were applied using a learning rate of 0.1 and the number of iterations (10-epochs) for prediction. For the CBR method, we used cross-validation operand for training and testing and 10-epochs parameter in k-NN algorithm. The parameters of SVM were used SVM-type (C-classification), SVM-Kernel (radial), and gamma (0.1) was applied for modelling and prediction.

7.4.1. Choosing number of attributes in EHRs dataset

We used a dataset from FConvertEHRs function that consists of stroke patients and non-stroke patients represented in binary format. Each record contains the risk factors that cannot be changed, the risk factors that can be changed (treat or control), and other risk factors that are less well-documented. These resulted in 227 risk factors for stroke prediction process. The variable used all stroke's risk factors without a stroke code (I64) for learning process or modelling. For the I64 code issue, patients in the dataset can have ischemic code, haemorrhagic code, brain stem stroke syndrome, cerebellar stroke syndrome and other familiar stroke syndromes. A patient may have more than one of the syndromes mentioned above. Some may have none. I64 covers all types of stroke syndromes.

7.4.2. Validation of Case-Based Reasoning

The result shows that during the training procedure, the prediction value, the accuracy value, precision value, recall value, and F1 scores for prediction in k-NN shows that accuracy is at 0.9766, F1 score as 0.9869 (see in Table 7-3). This algorithm show that accuracy value is higher than other algorithms in CBR. Cross validation was applied during the training and testing process that managed negative values and positive values at the prediction process. Additionally, Figure 7-4 and Figure 7-5 show the number of edges by using k-NN and SVM respectively. An edge of weight value in each technique is the predictor with stroke's risk factors. A prediction of stroke implied some concepts of a medical domain.

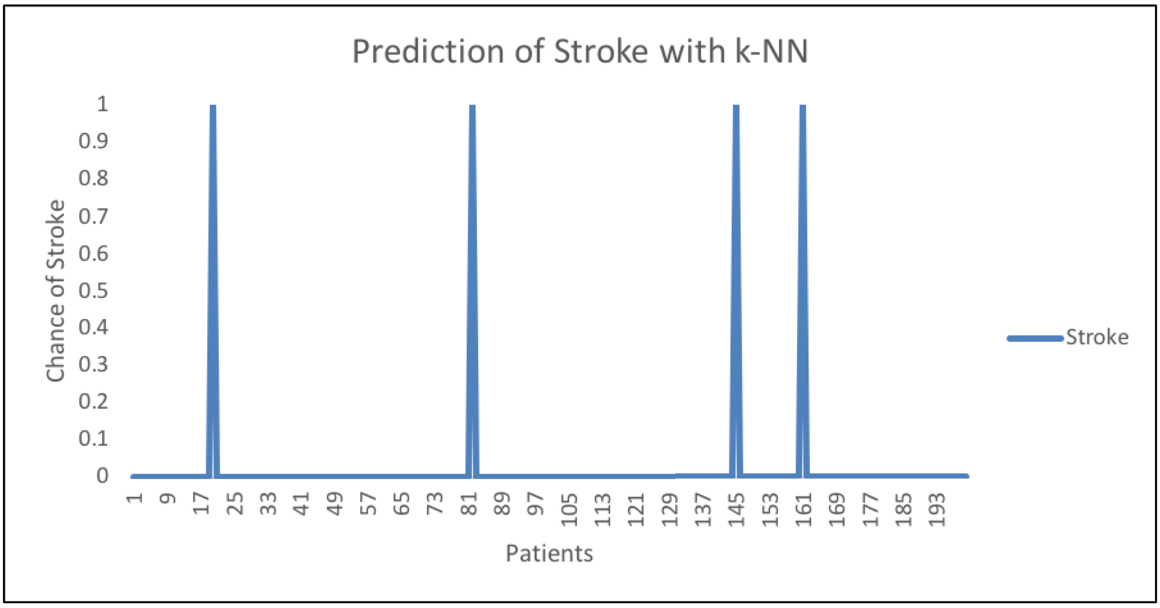


Figure 7-4: The prediction result using k-NN.

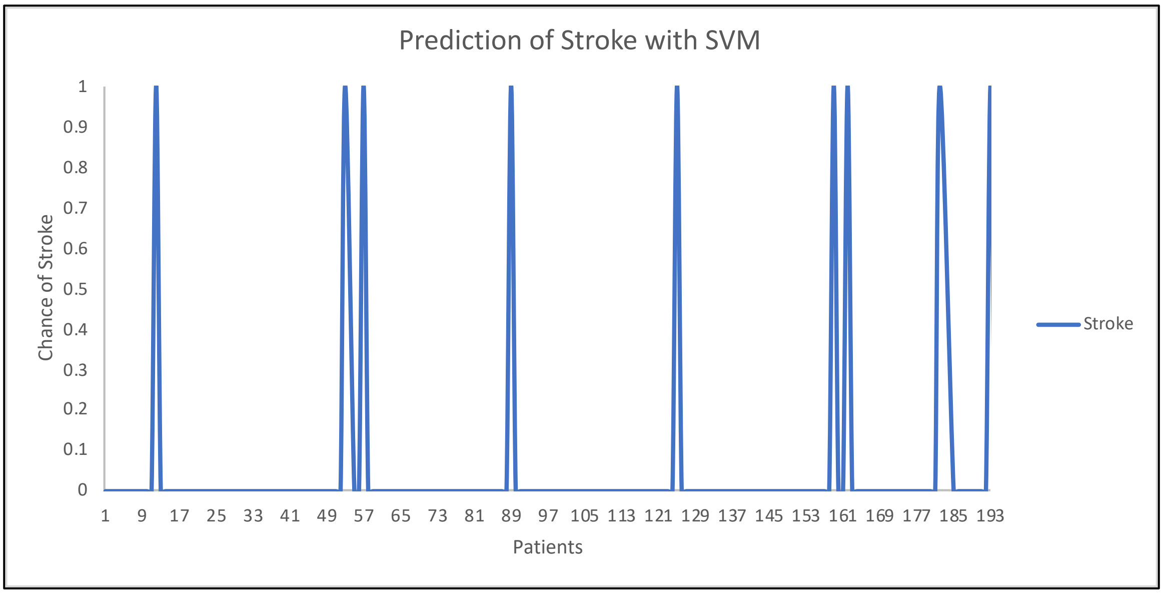


Figure 7-5: The prediction result using SVM.

7.4.3. Validation of Deep Learning approach

We conducted a comparison between three models: Backpropagation; RNN; and LSTM- RNN. All techniques demonstrated the results of training 30%, 50%, and 80%

respectively. For testing, 10% of the dataset is used. A learning rate is 0.1 and the number of iteration (10-epochs) were used for prediction. The variable for calculation used 227 risk factors for stroke.

The result shows that during the training procedure, the accuracy value, precision value, recall value, and F1 scores for prediction in LSTM-RNN are higher than those obtained from the other two techniques. The results for RNN shown for all value is lowest at 50% of the sample size (0.3570; 0.3612; 0.6476; 0.5456). Additionally, the number of edges with backpropagation, RNN, and LSTM are plotted as shown in Figure 7-6, Figure 7-7, and Figure 7-8 respectively. An edge of weight value in each technique is the predictor with 227 risk factors. A prediction of stroke implied some concepts of a medical domain that identifies patients who have chance of stroke in future.

In backpropagation, we used a feedforward multilayer artificial neural network. The computation shows that accuracy is at 0.8912, 0.8917, and 0.8914, F1 score as 0.3857, 0.3857, and 0.3860 respectively (shown in Table 7-2). This method shows that all values are the same in the three sample sizes used for prediction. Only, the accuracy shown here is marginally changed when the sample data are slightly increased.

By using the same parameters as in previous techniques, the LSTM-RNN show the best performance for prediction of stroke. The accuracy is 0.9279, 0.9493, and 0.9998 and F1 score are 0.9626, 0.9738, and 0.9999 respectively. The prediction for stroke shown as a percentage indicates the probability of having a stroke. In the medical domain, the good performance of LSTM-RNN is the preferred algorithm for use with a large dataset.

We used cross validation in training in order to avoid over fitting and to achieve better generalized results. The prediction process also used training and test samples and used a 10-fold cross validation. In term of positive prediction value and negative prediction value, these are calculated using the positive argument. The prevalence of the ICD-10 code is computed for the dataset, the detection rate and the detection prevalence.

Table 7-2 : Metrics of Stroke Prediction

Train 30% and Test 10%

Techniques	Precision	Recall	F1	Accuracy
Backpropagation	1	1	0.3857	0.8912
RNN	0.9371	0.9375	0.9371	0.8825
LSTM-RNN	**0.0219**	**0.0216**	**0.9626**	**0.9279**

Train 50% and Test 10%

Techniques	Precision	Recall	F1	Accuracy
Backpropagation	1	1	0.3857	0.8917
RNN	0.3570	0.3612	0.6476	0.5456
LSTM-RNN	**0.9741**	**0.9738**	**0.9738**	**0.9493**

Train 80% and Test 10%

Techniques	Precision	Recall	F1	Accuracy
Backpropagation	1	1	0.3860	0.8914
RNN	0.9996	0.9996	0.9996	0.9992
LSTM-RNN	**0.9999**	**0.9999**	**0.9999**	**0.9998**

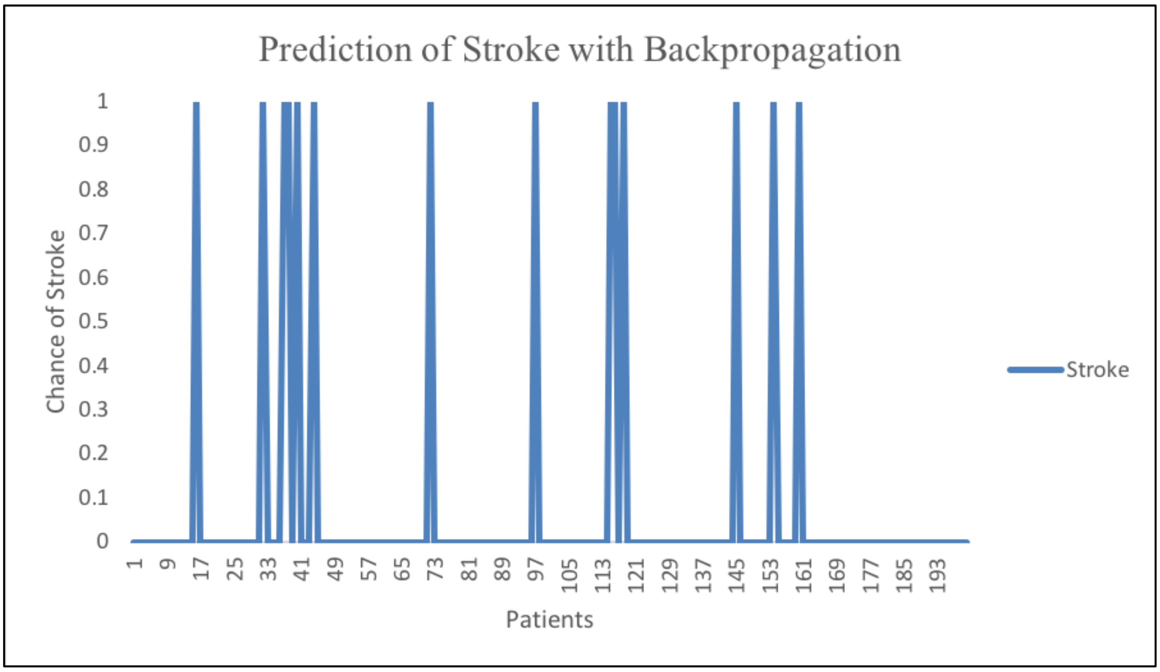


Figure 7-6: The prediction result using backpropagation.

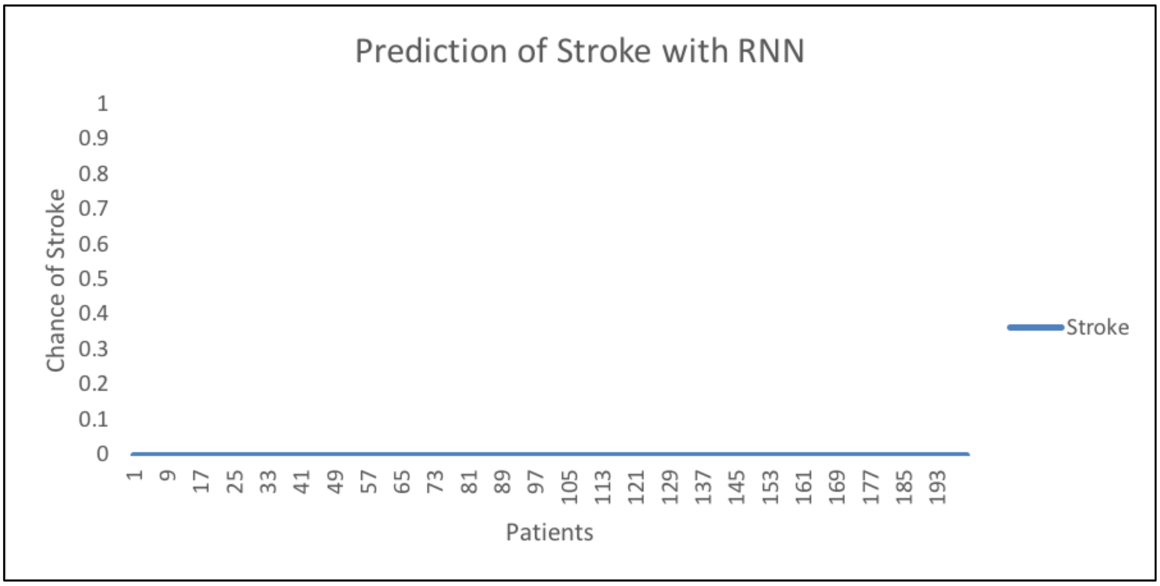


Figure 7-7: The prediction result using RNN.

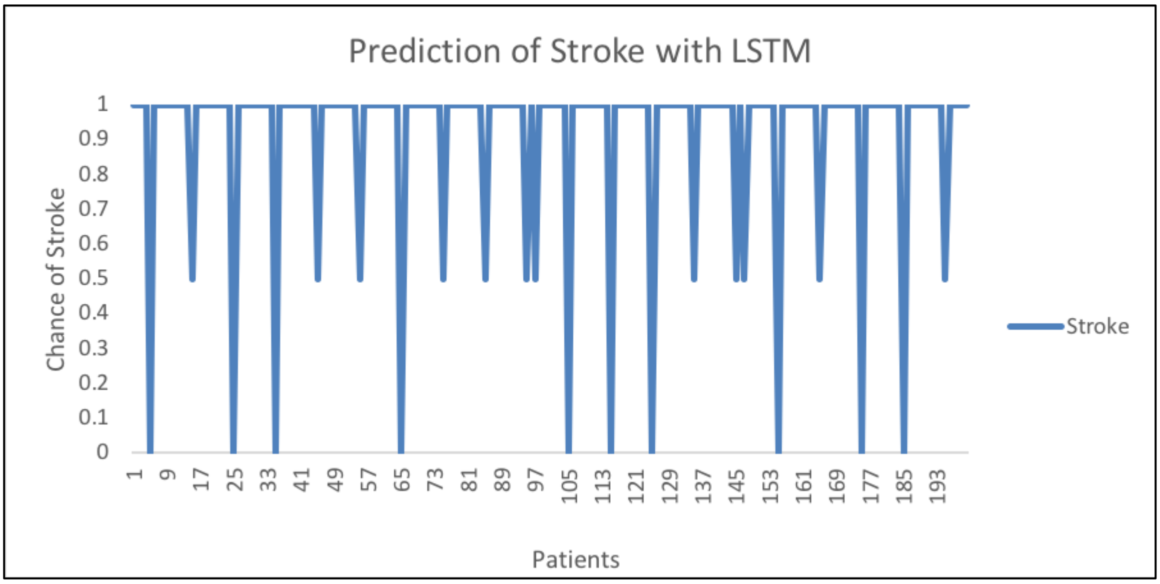


Figure 7-8: The prediction result using LSTM.

7.5. Comparison of case-based reasoning and deep learning techniques

The result shows that during training procedure, the prediction value, the accuracy value, precision value, recall value, and F1 scores for prediction in LSTM-RNN are higher than those obtained from the other four techniques (see in Table 7-3).

Table 7-3 : Comparison of metrics for Stroke Prediction

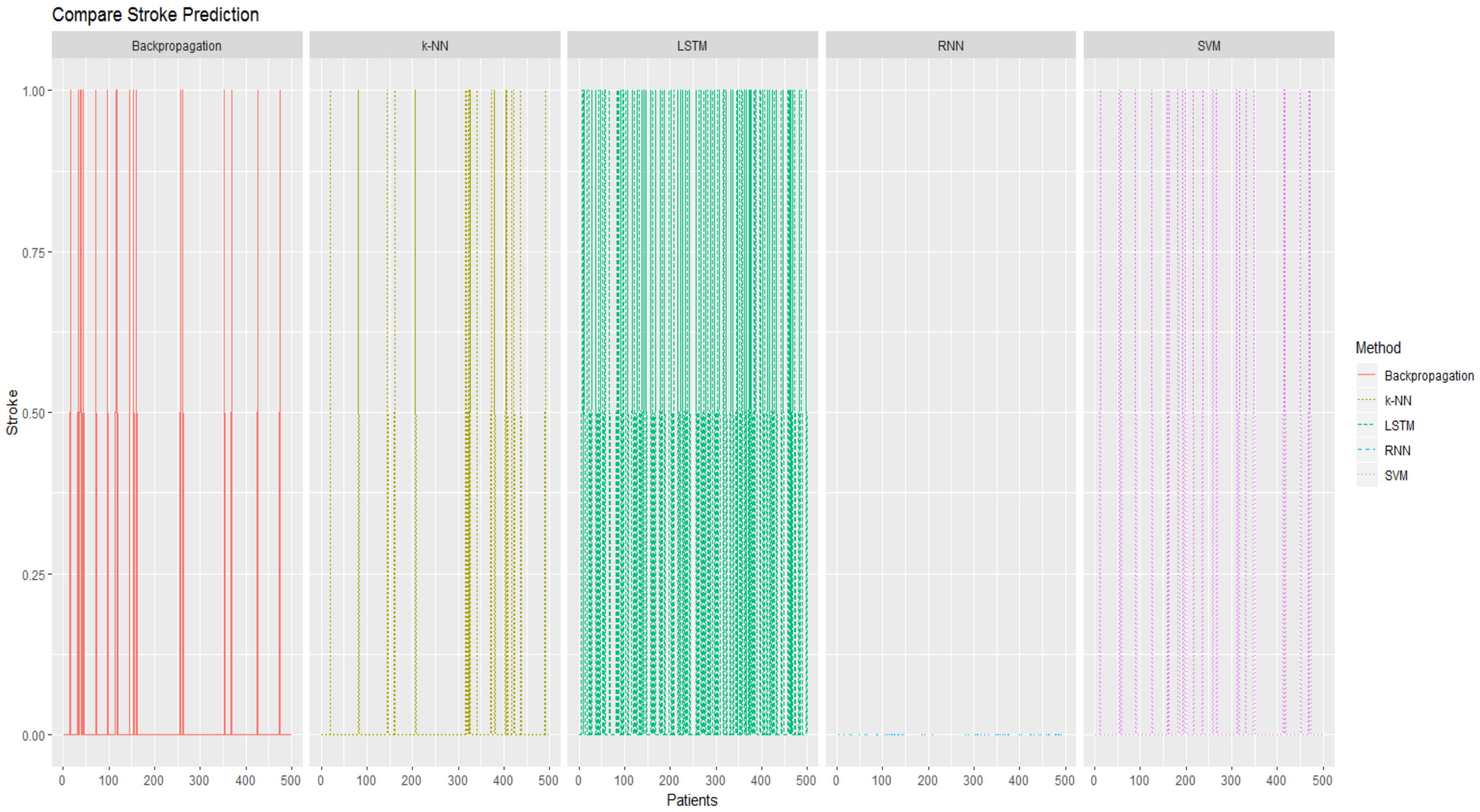
Methods	Techniques	Precision	Recall	F1	Accuracy
Case-Based Reasoning	k-NN	1	0.9744	0.9869	0.9766
	SVM	0.9923	0.9327	0.9615	0.9295
Deep Learning	Backpropagation	1	1	0.3860	0.8914
	RNN	0.9996	0.9996	0.9996	0.9992
	LSTM-RNN	**0.9999**	**0.9999**	**0.9999**	**0.9998**

Table 7-4 : Chance of stroke in EHRs.

Methods	Techniques	Stroke	Non-Stroke
Case-Based Reasoning	k-NN	843	8,776
	SVM	552	9,097
Deep Learning	Backpropagation	1,737	7,933
	RNN	0	9,619
	LSTM-RNN	**1,797**	**7,700**

In term of comparing the CBR and Deep Learning techniques, we selected five techniques: k-NN, SVM, Backpropagation, RNN, and LSTM-RNN. The accuracy values differ for each technique (see in Table 7-3). The LSTM-RNN show the best performance for accuracy.

For the prediction result k-NN and Backpropagation consists of the stroke patients and non-stroke patients. Backpropagation has 1,737 patients, Non-Stroke has 7,933 patients. Patients likely to suffer stroke in k-NN is 843 patients, Non-Stroke is 8,776 patients. Likely Stroke numbers are 552 for SVM, but Non-Stroke are 9,097 patients. The LSTM method has the highest prediction for stroke symptoms with 1,797 patients, with Non-Stroke at 7,700 patients (see in Table 7-4). The five techniques show that all values for prediction differ (see in Figure 7-9). The results for RNN and SVM shown for non-stroke patient are the highest when compared with the other techniques.



Source: Department of Medical Service in Thailand

Figure 7-9: The prediction result using five different methods.

7.6. Summary

This chapter aims at the investigation of five techniques: Support Vector Machine (SVM); k-Nearest Neighbours (k-NN); Backpropagation; Recurrent Neural Network (RNN); and Long Short-Term Memory - Recurrent Neural Network (LSTM-RNN). Those are powerful and widely used techniques in machine learning and bioinformatics. The empirical research is intended to evaluate the ability of machine learning and deep learning to recognize patterns in multi-label classification of stroke. Heart dataset and EHRs dataset have been used in this research for validation, error estimation and experiments. The predictive models or techniques applied in clinical decision making such as predicting occurrence of a disease or diagnosis, evaluating prognosis or outcome of diseases and assisting clinicians to recommend treatment of diseases. For EHRs dataset, we used aggregated files of EHRs from Department of Medical Services, The Ministry of Public Health of Thailand between 2015 and 2016. We decoded ICD-10 codes into the health records, as well as other potential risk factors within EHRs into the pattern and model for prediction. The results show several strong baselines that include accuracy, recall, and F1 measure score.

Chapter 8.

Conclusion

8.1. Introduction

The contribution of this thesis is discussed in this final chapter. The primary contribution of the thesis is to propose a Case Based Reasoning System and Deep Learning for stroke prediction. Based on existing literature, the research identified the problem of stroke symptom-based method for use in a prediction model which meets medical standards. This research proposed two frameworks and five techniques and also developed and applied them for the prediction process. To manage the dataset, the filtering algorithm was developed by using feature selection. To validate the proposed two frameworks, all prediction algorithms were developed and tested. Finally, the result tested the performance of all frameworks by using four performance measures as follows: precision; recall; f-measure; and accuracy rate.

This chapter is organized as follows: Section 8.2 summarises the contribution of the thesis. The discussion section is presented in Section 8.3. A concluding discussion and suggestions for future research are in Section 8.4 and 8.5 respectively.

8.2. Contributions of this thesis to the existing literature

The main contribution of this thesis is the development of a prediction analysis for stroke. The research demonstrates a good potential for providing effective warnings and predictions for all new patients who have risk factors related to the complex medical condition of stroke.

Stroke is complex disease which is hard to predict and also has no warning sign for new patients. Case-based Reasoning system (CBR), Naïve Bayes, Support Vector Machine (SVM), and k-Nearest Neighbours (k-NN) and Statistical Models have been widely applied in clinical decision making such as predicting occurrence of a disease or diagnosis, evaluating prognosis or outcome of diseases and assisting clinicians to recommend treatment of diseases. However, the conventional predictive models or techniques are still not effective enough in capturing the underlying knowledge because they are incapable of simulating the complexity of feature representation of problems in the medical domains.

This research proposed a stroke prediction approach using CBR, machine learning, and deep learning model based on common EHRs dataset. The dataset includes the major risk factor of stroke and share common variables to predict stroke. The outcomes of this research are more accurate than medical scoring systems currently in use for warning non-stroke patients if they are likely to develop stroke. Comprehensive CBR, Machine Learning, and Statistical models for predicting the likely occurrence of stroke were developed as predictive tools. The aim is to react to the prediction, warn the patient, and suggest treatment for the new patient. However, the high volume of medical data, which is both heterogeneous, and complex, has become the biggest challenges in diseases prediction. Algorithms with very high level of accuracy are therefore vital for reliable medical diagnosis. The development of appropriate algorithms, nevertheless, still remains obscure despite its importance and necessity for healthcare. Good performance comes along with specific favourable circumstances, for instance, when well designed and formulated inputs are guaranteed. Consequently, the deep learning allows the disclosure of some unknown or unexpressed knowledge during prediction procedure, which is beneficial for decision-making in medical practice and provides useful suggestions and warnings to patients about unpredictable stroke. Finally, we used 227 risk factors for prediction using a deep learning algorithm and compared its results to other techniques. This research has made several effective contributions, as follows:

8.2.1. Contribution 1: Identify and select stroke risk factors

Stroke is a complex disease, which is hard to predict and also has no warning sign for new patients. The EHRs, as the source of patients' data, contain major risk factors for stroke and share common variables that could be used to predict stroke.

8.2.2. Contribution 2: Application of Case-Based Reasoning Framework for Stroke

The Case-Based Reasoning framework (CBR) is a methodology for solving problems that uses a previous case(s) or memorized problem situations called cases. It uses previous case-based knowledge to predict stroke symptoms. For a future stroke patient, whose diagnosis is assumed and has an indefinite prognosis, by applying CBR and prediction model, similar cases can be retrieved from the case base which may provide information to the medical staff and hence assist them in reaching a potential diagnosis for stroke. This procedure would overcome the drawback of the existing method and show the chance of stroke symptoms in percentage likelihood to support their decision-making and suggestions for a new patient.

8.2.3. Contribution 3: Apply different machine learning and Deep Learning techniques

The high volume of heterogenous and complex medical data has become the biggest challenge in diseases prediction. Algorithms with a very high level of accuracy are therefore vital for medical diagnosis. The development of algorithms, nevertheless, still remains obscure despite their importance and necessity for healthcare. Good performance comes along with specific favourable circumstances, for instance, when well designed and formulated inputs are guaranteed. Consequently, the back propagation and deep learning techniques allow the disclosure of some unknown or unexpressed knowledge during prediction procedure, which is beneficial for decision-making in medical practice and can provide useful suggestions and warnings to patient about unpredictable stroke. All methods can be applied to learning previous case(s) in EHRs and testing a new incoming case(s). It would define the weight value in data groups. These values depend on high risk factors. Both methods have been applied for a prediction model and utilize data from Electronic Healthcare Records (EHRs). The result of the study are able to show percentages of the new patients that would be likely to suffer a stroke in the future and a new modelling pattern of risk factors is proposed based on backpropagation deep learning techniques. This technique imitates the experts' judgment that normally works by analogy from previous cases to analyse for clinical reasoning and clinical decision making to predict, diagnose and make a prognosis in terms of providing a satisfactory solution for resolving the patient's problem.

8.2.4. Contribution 4: Comparison of Case-Base Reasoning and Deep Learning

Many predictive techniques (Case-based Reasoning, Naïve Bayes, Support Vector Machine, etc.) have been widely applied in clinical decision making such as predicting occurrence of a disease or diagnosis, evaluating prognosis or outcome of diseases and assisting clinicians to recommend treatment of diseases. However, the conventional predictive models or techniques are still not effective enough in capturing the underlying knowledge because it is incapable of simulating the complexity of features representative of the medical domain. This research reports on predictive analytical techniques for stroke using a CBR and Machine Learning model that compares a EHRs dataset and a predictive algorithm relying on complex computational models. This algorithm is able to handle a reduction of complex data for prediction and to compare the existing groups with efficiency and accuracy. A comprehensive CBR and Machine Learning for stroke symptoms as a tool to predict stroke is created and used with risk factors to predict using a deep learning algorithm and compared with other techniques. The result aims to predict, warn and recommend treatment for the new patient.

8.3. Discussion

This research reports on predictive analytical techniques for stroke using a deep learning model applied on a heart disease dataset for the pilot steps of experiment. The atrial fibrillation symptoms in heart patients are considered also a major risk factor for stroke and share common variables in the prediction of stroke. The first outcomes of this pilot research are more accurate than the standard medical prediction scoring systems currently in use for warning heart patients if they are likely to develop stroke. Comprehensive CBR, Machine Learning, and Statistical models for stroke disease are used for this research as a tool for predicting stroke.

Using the heart dataset, we compared three techniques: the Deep Learning technique; Naïve Bayes; and Support Vector Machine for stroke prediction. All techniques used data gathered from heart patients for identification of risk factors and prediction of the chance of stroke. The results are significantly favourable and valid for decision-making by a medical practitioner and for providing warnings to patients. We compared the use of Deep Learning with Deep Neural Network, Naïve Bayes for discrete predictor, and linear performance in SVM. The results from of Naïve Bayes and SVM show whether patients are likely to suffer from a stroke or not, and the Deep Learning technique shows in percentage terms the probability of having a stroke. This confirmed that the Deep

Learning technique is most suitable for generating predictive analysis of a stroke based on the heart dataset. This result allows a medical practitioner to react to the prediction, provide a warning and recommend treatment for the newly identified patient. Algorithms with very high level of accuracy are vital for such medical diagnosis. However, the extremely high volume of medical data with its heterogeneity and complexity, have become the biggest challenges for prediction of disease.

Information Technology can support humans to solve problems. With CBR, some processes are easier to perform by humans whereas others are more appropriate for computers. For instance, people can demonstrate creative adaptation very well and expert knowledge can be created and adapted by humans but the complete range of applicable cases may not be remembered due to bias in human memories or for novices who do not yet have enough adequate experience to solve the complexity of some problems. Humans and computers can interact in a productive manner in order to solve problems using CBR, though some disadvantages may still occur. For example, the data in old cases may be poor, may be biased, and perhaps not all appropriate cases were retrieved. Retrieval or adaptation of the knowledge will still need to be applied when used in the future to aid diagnosis.

Case-based Reasoning systems (CBR), Support Vector Machine (SVM), k-Nearest Neighbours (k-NN) and Statistical models have been widely applied in clinical decision making such as predicting occurrence of a disease or diagnosis, evaluating prognosis or outcome of diseases and assisting clinicians to recommend further treatment of diseases. However, the conventional predictive models or techniques are still not effective enough in capturing the underlying knowledge because they are incapable of simulating the complexity of feature representation in the medical problem domains. This research thus used k-NN and SVM in the prediction process as other techniques for comparing algorithms for large datasets in EHRs.

EHRs using ICD-10th code have some issues and challenges in the data analyses of various diseases and health problems using Deep Learning. The excellent analysis using different predicting techniques requires the use of data obtained from patient health records and a comparison between previous cases, observation, or inspection. Stroke has complex risk factors, so algorithms with very high level of accuracy are therefore vital for medical diagnosis. The development of algorithms, nevertheless, still remains obscure despite its importance and necessity for healthcare. Good performance comes along with specific favourable circumstances, for instance, when well designed and formulated inputs are guaranteed. However, Deep Learning can disclose some unknown or unstated knowledge during the prediction procedure, which is beneficial for decision-making in medical practice and can provide useful suggestions and warnings to patients about unpredictable stroke.

A notable issue is the limitation of EHRs, where data is recorded per patient visit so multiple records of the same patient exist likely with the same risk factors as in earlier records. In 2015 and 2016, 10,518 of 96,190 patient records have stroke risk factors from the original EHRs. In 2015 alone, the figure is 7,276 of 76,311 while in 2016 it is 1,239 of 8,987. The total number of patients having a stroke risk factor in 2015 and 2016 combined is not equal to the overall figure due to the repeat of demographic data and health records of patient shown in 2015 and 2016. The EHRs dataset had over two million records and contained huge number of ICD-10 codes per record. It contained null values and anomalous values. We removed the anomalous records in the data preparation cleaning process. This research used R programming and Rstudio for predicting stroke symptoms. These software tools have issues in memory management when used with a very large (huge) dataset. R programming cannot allocate memory for all the records. Consequently, we needed to re-format the dataset and reduce its size before beginning the prediction process. The final EHRs dataset used to compare the five techniques was done using the FConvertEHRs algorithm. The data were then divided for training and testing purposes. In the predictive process, the results contain both negative and positive values. We managed them by using a confusion matrix algorithm for filtering the predicted values. The results then show a greater accuracy in the prediction process. However, there is still a problem when 50% or less of the data are randomized for training, The RNN result is lower than those in other cases.

The results of using LSTM-RNN show that accuracy rate, recall and F1 measure score are different from those of Support Vector Machine (SVM); k-NN, Backpropagation and RNN algorithms. The accuracy rate depends on the algorithm used. Unlike other techniques, the result is more reliable once there are large datasets for the prediction of stroke. This led us to the conclusion that the LSTM-RNN algorithm is most suitable for predictive analysis of any cerebrovascular disease or stroke. Furthermore, the LSTM-RNN algorithm predicts a higher probability of patients with stroke symptoms compared with the SVM, k-NN, and Backpropagation algorithms. The outcomes of all approaches show not only shown the estimated number of patients with future stroke symptoms, but also will be very useful in prevention and early treatment of patients. Specially, it can give suggestions and warnings to patients in spite of the fact that strokes do not have warning signs. Machine Learning and Deep Learning allow the disclosure of some unknown or unexpressed knowledge during prediction procedure, which is highly beneficial for decision-making in medical practice and provides useful suggestions and warnings to patient about unpredictable stroke.

The development of algorithms as used in this research convincingly applied Deep Learning for stroke prediction. However, stroke has complex risk factors and remains challenging even when an algorithm with high accuracy for prediction based on known risk factors is used. Focusing on unexpressed knowledge during the prediction procedure is the hope of using Big Data such as in EHRs. One warning is noted, EHRs using ICD-10 codes still have issues and challenges in data analysis.

8.4. Conclusion

Although, the original prediction score used various prediction models, the CHADS2 and CHA2DS2-VASc are standards that are now most widely used in practice for the prediction of stroke disease, atrial fibrillation, and heart disease. A score is input by physicians or medical staff. If patients get a score higher than a certain threshold, then there is a high risk for the disease. The main purpose of this method is often used for predicting which patients with atrial fibrillation will have other thromboembolic events, such as stroke disease. The major risk factors include stroke, TIA and age. This research differs from other computational predictions. The stroke dataset in this research used more risk factors for predicting the potential outcome for patients who have different risk factors. Some factors are common to both approaches. While CHADS2 and CHA2DS2-VASc score used just 9 attributes for their risk factors, stroke has many more attributes, around 140-150 and as high as 227 in this study. Some risk factors such as carotid or other artery diseases, ethnicity, poor diet or physical inactivity and obesity were not used. The Stroke dataset in this study has relatively more complexity and this results in better prediction and accuracy. We applied machine learning and deep learning techniques to achieve this goal.

Deep learning is widely used in prediction of diseases, especially in the prognosis and data analyses in the healthcare sector. The data obtained from patient health records are used compared with previous cases, observations, or inspection. Stroke has complex risk factors. Therefore, the concepts or decision-making techniques cannot be directly extracted from a source that requires the involvement of a number of human experts.

This research aims at the comparison of Case-based Reasoning and Deep Learning technique with EHRs based on risk factors for prediction of cerebrovascular disease. The Electronic Healthcare Records (EHRs) provide descriptive details about a patient's physical and mental health, diagnoses, lab results, treatments care plans and so forth. The data are difficult to mine effectively due to the sheer size of the data and further complicated by missing data. The diagnoses of disease are represented by International Classification of Diseases, 10th Revision (ICD-10th) code in each patient record. The data contain records of all patients and their disease codes. However, we are interested in only some particular disease codes and therefore the data must be filtered to remove irrelevant data. The filtering process takes a good amount of time due to the large size of the data as well as the large number of the disease codes.

This study is a step in a new direction using Machine Learning and Case-Based Reasoning (CBR) system for prevention and prediction of stroke for all patient's based on their healthcare records and ICD-10 codes. In this research Machine learning and CBR have been demonstrated as being very useful in improving public health.

This research can be applied for prediction with many other related diseases. As we have seen, prediction process is applied on EHRs dataset with ICD-10 codes. It is important to take advantage of the wealth of knowledge hidden in these datasets and create a domain knowledge which will lead to more informed treatment decision resulting in a new patient who has the potential of suffering stroke symptoms.

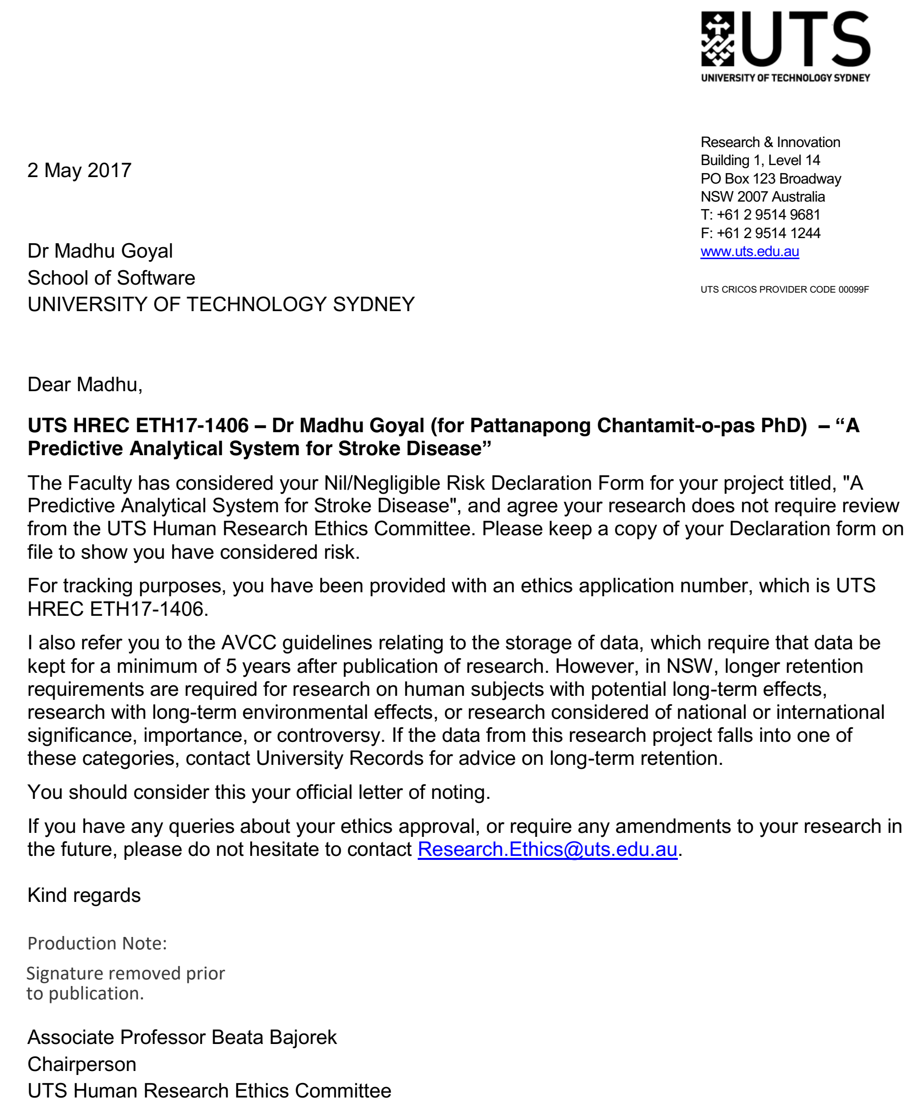
8.5. Future Work

The ultimate goal of this research is to be able to use Case-Based Reasoning and Machine Learning for prevention of stroke symptoms which can assist medical staff in their decision-making and setting care plans for future patients who have stroke risk factors. If future patients received a high-risk diagnosis, given a particular treatment plan and survived, then it would make sense for clinicians to prescribe this treatment for high risk patients and to warn new patients who have a possibility of stroke symptoms in the future.

CBR and Machine Learning retrieval were able to successfully identify the similarities between the previous cases and identify a new patient with high risk factors. In addition to achieving better results, our method can automatically identify potential risk factors without carrying out extensive medical studies to understand each one in detail. This would allow for a quick method of characterizing a new disease and identifying its predictors before other studies confirm them. Furthermore, this procedure could also be used to suggest risk factors that might have been previously unexplored.

In the future, several key milestones are ahead. We will use more related diseases, lab results, and MRI/CT images in order to predict future stroke patients and also create a performance test of the stroke prediction algorithm. This should enable us to potentially devise and implement an e-stroke application. One key milestone is to combine CBR and the Deep Learning algorithm to construct expectations of how attributes are related within the known cases. These expectations will be used to form the basis for retrieval and adaptation in CBR cycle. We also hope to overcome the limitations of existing software tools. A pilot e-stroke application could be developed without waiting for full scale implementation in the interim.

Appendix A UTS Human Research Ethics Committee



UTS
UNIVERSITY OF TECHNOLOGY SYDNEY

Research & Innovation
Building 1, Level 14
PO Box 123 Broadway
NSW 2007 Australia
T: +61 2 9514 9681
F: +61 2 9514 1244
www.uts.edu.au

UTS CRICOS PROVIDER CODE 00099F

2 May 2017

Dr Madhu Goyal
School of Software
UNIVERSITY OF TECHNOLOGY SYDNEY

Dear Madhu,

UTS HREC ETH17-1406 – Dr Madhu Goyal (for Pattanapong Chantamit-o-pas PhD) – "A Predictive Analytical System for Stroke Disease"

The Faculty has considered your Nil/Negligible Risk Declaration Form for your project titled, "A Predictive Analytical System for Stroke Disease", and agree your research does not require review from the UTS Human Research Ethics Committee. Please keep a copy of your Declaration form on file to show you have considered risk.

For tracking purposes, you have been provided with an ethics application number, which is UTS HREC ETH17-1406.

I also refer you to the AVCC guidelines relating to the storage of data, which require that data be kept for a minimum of 5 years after publication of research. However, in NSW, longer retention requirements are required for research on human subjects with potential long-term effects, research with long-term environmental effects, or research considered of national or international significance, importance, or controversy. If the data from this research project falls into one of these categories, contact University Records for advice on long-term retention.

You should consider this your official letter of noting.

If you have any queries about your ethics approval, or require any amendments to your research in the future, please do not hesitate to contact Research.Ethics@uts.edu.au.

Kind regards

Production Note:

Signature removed prior
to publication.

Associate Professor Beata Bajorek
Chairperson
UTS Human Research Ethics Committee

BIBLIOGRAPHY

Aamodt, A. & Plaza, E. 1994, 'Case-based reasoning: Foundational issues, methodological variations, and system approaches', *AI communications*, vol. 7, no. 1, pp. 39-59.

Addo, J., Ayerbe, L., Mohan, K.M., Crichton, S., Sheldenkar, A., Chen, R., Wolfe, C.D.A. & McKevitt, C. 2012, 'Socioeconomic Status and Stroke', *Stroke*, vol. 43, no. 4, pp. 1186-91.

Ahmed, M.U., Banaee, H. & Loutfi, A. 2013, 'Health monitoring for elderly: An application using case-based reasoning and cluster analysis', *ISRN Artificial Intelligence*, vol. 2013.

Alankus, G. 2011, 'Motion-based video games for stroke rehabilitation with reduced compensatory motions', Washington University in St. Louis.

Alexopoulos, E., Dounias, G. & Vemmos, K. 1999, 'Medical diagnosis of stroke using inductive machine learning', *Machine Learning and Applications: Machine Learning in Medical Applications*, pp. 20-3.

Alotaibi, N.N. & Sasi, S. 2016, 'Comparison of Predictive Models for Transferring Stroke In-Patients to Intensive Care Unit', *Transactions on Machine Learning and Artificial Intelligence*, vol. 4, no. 3, pp. 1-22.

Amin, S.U., Agarwal, K. & Beg, R. 2013a, 'Genetic neural network based data mining in prediction of heart disease using risk factors', *Information & Communication Technologies (ICT), 2013 IEEE Conference on*, pp. 1227-31.

Amin, S.U., Agarwal, K. & Beg, R. 2013b, 'Genetic neural network based data mining in prediction of heart disease using risk factors', *2013 IEEE Conference on Information & Communication Technologies (ICT)*, pp. 1227-31.

Anaissi, A., Goyal, M., Catchpoole, D.R., Braytee, A. & Kennedy, P.J. 2015, 'Case-Based Retrieval Framework for Gene Expression Data', *Cancer Informatics*, vol. 14, pp. 21-31.

Anbarasi, M., Anupriya, E. & Iyengar, N. 2010, 'Enhanced prediction of heart disease with feature subset selection using genetic algorithm', *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5370-6.

Arshadi, N. & Jurisica, I. 2005, 'Data mining for case-based reasoning in high-dimensional biological domains', *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 8, pp. 1127-37.

Asadi, H., Dowling, R., Yan, B. & Mitchell, P. 2014, 'Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy', *PLoS ONE*, vol. 9, no. 2, p. e88225.

Ashley, K. 2006, 'CASE-BASED REASONING', in A.R. Lodder & A. Oskamp (eds), *Information Technology and Lawyers: Advanced Technology in the Legal Domain, from Challenges to Daily Routine*, Springer Netherlands, Dordrecht, pp. 23-60.

AU, A.G., AU, A.r.M. & Hinton, G. 2013, 'Speech recognition with deep recurrent neural networks', paper presented to the *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 26-31 May 2013, <http://ieeexplore.ieee.org.ezproxy.lib.uts.edu.au/stamp/stamp.jsp?tp=&arnumber=6638947&isnumber=6637585>.

Baird, A.E., Dambrosia, J., Janket, S.-J., Eichbaum, Q., Chaves, C., Silver, B., Barber, P.A., Parsons, M., Darby, D., Davis, S., Caplan, L.R., Edelman, R.E. & Warach, S. 2001, 'A three-item scale for the early prediction of stroke recovery', *The Lancet*, vol. 357, no. 9274, pp. 2095-9.

Balakrishnan, S., Narayanaswamy, R., Savarimuthu, N. & Samikannu, R. 2008, 'SVM ranking with backward search for feature selection in type II diabetes databases', *2008 IEEE International Conference on Systems, Man and Cybernetics*, pp. 2628-33.

Banerjee, A., Fowkes, F.G. & Rothwell, P.M. 2010, 'Associations between peripheral artery disease and ischemic stroke: implications for primary and secondary prevention', *Stroke*, vol. 41, no. 9, pp. 2102-7.

Bareiss, E.R., Porter, B.W. & Wier, C.C. 1987, 'Protos: An exemplar-based learning apprentice', *Proceedings of the fourth international workshop on machine learning*, pp. 12-23.

Baxter, R.A., Williams, G.J. & He, H. 2001, 'Feature Selection for Temporal Health Records', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 198-209.

Begum, S., Ahmed, M.U., Funk, P., Xiong, N. & Von Schéele, B. 2006, 'Using calibration and fuzzification of cases for improved diagnosis and treatment of stress', *8th European Workshop on Case-based Reasoning in the Health Sciences, Turkey 2006*, pp. 113-22.

Begum, S., Ahmed, M.U., Funk, P., Xiong, N. & Von Schéele, B. 2009, 'A CASE-BASED DECISION SUPPORT SYSTEM FOR INDIVIDUAL STRESS DIAGNOSIS USING FUZZY SIMILARITY MATCHING', *Computational Intelligence*, vol. 25, no. 3, pp. 180-95.

Bellazzi, R. & Zupan, B. 2008, 'Predictive data mining in clinical medicine: Current issues and guidelines', *International Journal of Medical Informatics*, vol. 77, no. 2, pp. 81-97.

Benamer, H.T.S. & Grosset, D. 2009, 'Stroke in Arab countries: A systematic literature review', *Journal of the Neurological Sciences*, vol. 284, no. 1–2, pp. 18-23.

Bermingham, M.L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A.F., Wilson, J.F., Agakov, F., Navarro, P. & Haley, C.S. 2015, 'Application of high-dimensional feature selection: evaluation for genomic prediction in man', *Scientific Reports*, vol. 5, p. 10312.

Bichindaritz, I., Kansu, E. & Sullivan, K.M. 1998, 'Case-based reasoning in CARE-PARTNER: Gathering evidence for evidence-based medical practice', in B. Smyth & P. Cunningham (eds), *Advances in Case-Based Reasoning: 4th European Workshop, EWCBR-98 Dublin, Ireland, September 23–25, 1998 Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 334-45.

Bradburn, C. & Zeleznikow, J. 1994, 'The application of case-based reasoning to the tasks of health care planning', in S. Wess, K.-D. Althoff & M.M. Richter (eds), *Topics in Case-Based Reasoning: First European Workshop, EWCBR-93 Kaiserslautern, Germany, November 1–5, 1993 Selected Papers*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 365-78.

Brien, D., Glasgow, J. & Munoz, D. 2005, 'The Application of a Case-Based Reasoning System to Attention-Deficit Hyperactivity Disorder', in H. Muñoz-Ávila & F. Ricci (eds), *Case-Based Reasoning Research and Development: 6th International Conference on Case-Based Reasoning, ICCBR 2005, Chicago, IL, USA, August 23-26, 2005. Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 122-36.

Brosch, T. & Tam, R. 2013, 'Manifold Learning of Brain MRIs by Deep Learning', in K. Mori, I. Sakuma, Y. Sato, C. Barillot & N. Navab (eds), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 633-40.

Carbonell, J.G. 1983, 'Learning by Analogy: Formulating and Generalizing Plans from Past Experience', in R.S. Michalski, J.G. Carbonell & T.M. Mitchell (eds), *Machine Learning: An Artificial Intelligence Approach*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 137-61.

Chait, A. & Bornfeldt, K.E. 2009, 'Diabetes and atherosclerosis: is there a role for hyperglycemia?', *Journal of Lipid Research*, vol. 50, no. Supplement, pp. S335-S9.

Chambless, L.E., Heiss, G., Shahar, E., Earp, M.J. & Toole, J. 2004, 'Prediction of Ischemic Stroke Risk in the Atherosclerosis Risk in Communities Study', *American Journal of Epidemiology*, vol. 160, no. 3, pp. 259-69.

Chang, C.-L. 2005, 'Using case-based reasoning to diagnostic screening of children with developmental delay', *Expert Systems with Applications*, vol. 28, no. 2, pp. 237-47.

Chantamit-o-pas, P. & Goyal, M. 2017, 'Prediction of Stroke Using Deep Learning Model', *International Conference on Neural Information Processing*, Springer, Guangzhou, China, pp. 774-81.

Chantamit-o-pas, P. & Goyal, M. 2018a, 'A Case-Based Reasoning Framework for Prediction of Stroke', in D.K. Mishra, A.T. Azar & A. Joshi (eds), *Information and Communication Technology : Proceedings of ICICT 2016*, Springer Singapore, Singapore, pp. 219-27.

Chantamit-o-pas, P. & Goyal, M. 2018b, 'Long Short-Term Memory Recurrent Neural Network for Stroke Prediction', *International Conference on Machine Learning and Data Mining in Pattern Recognition*, Springer International Publishing, Cham, pp. 312-23.

Chattopadhyay, S., Banerjee, S., Rabhi, F.A. & Acharya, U.R. 2013, 'A Case-Based Reasoning system for complex medical diagnosis', *Expert Systems*, vol. 30, no. 1, pp. 12-20.

Chien-Chang, H. & Cheng-Seen, H. 1998, 'A hybrid case-based medical diagnosis system', *Proceedings Tenth IEEE International Conference on Tools with Artificial Intelligence (Cat. No.98CH36294)*, pp. 359-66.

Choudhury, N. & Begum, S.A. 2016, 'A Survey on Case-based Reasoning in Medicine', *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, vol. 7, no. 8, pp. 136-44.

Chow, R., Zhong, W., Blackmon, M., Stolz, R. & Dowell, M. 2008, 'An efficient SVM-GA feature selection model for large healthcare databases', paper presented to the *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, Atlanta, GA, USA.

Chuang, C.-L. 2011, 'Case-based reasoning support for liver disease diagnosis', *Artificial Intelligence in Medicine*, vol. 53, no. 1, pp. 15-23.

Clifford, N. 2010, *Stroke Understanding the Disease*, S. Garner., Victoria Australia. Distributed by Video Education Australasia.

Crockett, D. 2013, '4 Essential Lessons for Adopting Predictive Analytics in Healthcare'.

D'Aquin, M., Lieber, J. & Napoli, A. 2006, 'ADAPTATION KNOWLEDGE ACQUISITION: A CASE STUDY FOR CASE-BASED DECISION SUPPORT IN ONCOLOGY', *Computational Intelligence*, vol. 22, no. 3-4, pp. 161-76.

Deo, R.C. 2015, 'Machine Learning in Medicine', *Circulation*, vol. 132, no. 20, p. 1920.

Doquire, G. & Verleysen, M. 2012, 'Feature selection with missing data using mutual information estimators', *Neurocomputing*, vol. 90, pp. 3-11.

Fan, C.-Y., Chang, P.-C., Lin, J.-J. & Hsieh, J.C. 2011, 'A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification', *Applied Soft Computing*, vol. 11, no. 1, pp. 632-44.

Fialho, A.S., Cismondi, F., Vieira, S.M., Sousa, J.M.C., Reti, S.R., Howell, M.D. & Finkelstein, S.N. 2010, 'Predicting Outcomes of Septic Shock Patients Using Feature Selection Based on Soft Computing Techniques', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 65-74.

Flaherty, M.L., Kissela, B., Khoury, J.C., Alwell, K., Moomaw, C.J., Woo, D., Khatri, P., Ferioli, S., Adeoye, O., Broderick, J.P. & Kleindorfer, D. 2012, 'Carotid Artery Stenosis as a Cause of Stroke', *Neuroepidemiology*, vol. 40, no. 1, pp. 36-41.

Fonseca, A.C. & Ferro, J.M. 2013, 'Drug Abuse and Stroke', *Current Neurology and Neuroscience Reports*, vol. 13, no. 2, pp. 1-325.

Francis, J., Raghunathan, S. & Khanna, P. 2007, 'The role of genetics in stroke', *Postgraduate medical journal*, vol. 83, no. 983, pp. 590-5.

Gage, B.F., Waterman, A.D., Shannon, W., Boechler, M., Rich, M.W. & Radford, M.J. 2001, 'Validation of clinical classification schemes for predicting stroke: Results from the national registry of atrial fibrillation', *JAMA*, vol. 285, no. 22, pp. 2864-70.

Genetic Alliance 2009, *Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals.*, Genetic Alliance, Washington (DC), viewed 15 Jan 2019, <https://www.ncbi.nlm.nih.gov/books/NBK115560/>.

Gentner, D. 1983, 'Structure-Mapping: A Theoretical Framework for Analogy*', *Cognitive Science*, vol. 7, no. 2, pp. 155-70.

Gers, F.A., Schmidhuber, J. & Cummins, F. 2000, 'Learning to forget: Continual prediction with LSTM', *Neural Computation*, vol. 12, no. 10, pp. 2451-71.

Gers, F.A., Schraudolph, N.N. & Schmidhuber, J. 2002, 'Learning precise timing with LSTM recurrent networks', *Journal of machine learning research*, vol. 3, no. Aug, pp. 115-43.

Gierl, L., Bull, M. & Schmidt, R. 1998, 'CBR in Medicine', in M. Lenz, H.-D. Burkhard, B. Bartsch-Spörl & S. Wess (eds), *Case-Based Reasoning Technology: From Foundations to Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 273-97.

Gierl, L. & Stengel-Rutkowski, S. 1994, 'Integrating consultation and semi-automatic knowledge acquisition in a prototype-based architecture: Experiences with dysmorphic syndromes', *Artificial Intelligence in Medicine*, vol. 6, no. 1, pp. 29-49.

Go, A.S., Mozaffarian, D., Roger, V.L., Benjamin, E.J., Berry, J.D., Borden, W.B., Bravata, D.M., Dai, S., Ford, E.S., Fox, C.S., Franco, S., Fullerton, H.J., Gillespie, C., Hailpern, S.M., Heit, J.A., Howard, V.J., Huffman, M.D., Kissela, B.M., Kittner, S.J., Lackland, D.T., Lichtman, J.H., Lisabeth, L.D., Magid, D., Marcus, G.M., Marelli, A., Matchar, D.B., McGuire, D.K., Mohler, E.R., Moy, C.S., Mussolino, M.E., Nichol, G., Paynter, N.P., Schreiner, P.J., Sorlie, P.D., Stein, J., Turan, T.N., Virani, S.S., Wong, N.D., Woo, D. & Turner, M.B. 2013, 'Executive Summary: Heart Disease and Stroke Statistics—2013 Update', *Circulation*, vol. 127, no. 1, p. 143.

Goldstein, L.B., Adams, R., Alberts, M.J., Appel, L.J., Brass, L.M., Bushnell, C.D., Culebras, A., DeGraba, T.J., Gorelick, P.B., Guyton, J.R., Hart, R.G., Howard, G., Kelly-Hayes, M., Nixon, J.V. & Sacco, R.L. 2006, 'Primary Prevention of Ischemic Stroke: A Guideline From the American Heart Association/American Stroke Association Stroke Council: Cosponsored by the Atherosclerotic Peripheral Vascular Disease Interdisciplinary Working Group; Cardiovascular Nursing Council; Clinical Cardiology Council; Nutrition, Physical Activity, and Metabolism Council; and the Quality of Care and Outcomes Research Interdisciplinary Working Group: The American Academy of Neurology affirms the value of this guideline', *Stroke*, vol. 37, no. 6, pp. 1583-633.

Goldstein, L.B., Adams, R., Becker, K., Furberg, C.D., Gorelick, P.B., Hademenos, G., Hill, M., Howard, G., Howard, V.J., Jacobs, B., Levine, S.R., Mosca, L., Sacco, R.L., Sherman, D.G., Wolf, P.A., del Zoppo, G.J. & Members 2001, 'Primary Prevention of Ischemic Stroke', *Circulation*, vol. 103, no. 1, pp. 163-82.

Golobardes, E., Llorà, X., Salamó, M. & Martí, J. 2002, 'Computer aided diagnosis with case-based reasoning and genetic algorithms', *Knowledge-Based Systems*, vol. 15, no. 1–2, pp. 45-52.

Goodridge, W., Peter, H. & Abayomi, A. 1999, 'The Case-Based Neural Network Model and Its Use in Medical Expert Systems', in W. Horn, Y. Shahar, G. Lindberg, S. Andreassen & J. Wyatt (eds), *Artificial Intelligence in Medicine: Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99 Aalborg, Denmark, June 20–24, 1999 Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 232-6.

Gorelick, P.B., Sacco, R.L., Smith, D.B., Alberts, M., Mustone-Alexander, L., Rader, D., Ross, J.L., Raps, E., Ozer, M.N. & Brass, L.M. 1999, 'Prevention of a first stroke: a review of guidelines and a multidisciplinary consensus statement from the National Stroke Association', *Jama*, vol. 281, no. 12, pp. 1112-20.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R. & Schmidhuber, J. 2017, 'LSTM: A search space odyssey', *IEEE transactions on neural networks and learning systems*.

Guerra, F., Scappini, L., Maolo, A., Campo, G., Pavasini, R., Shkoza, M. & Capucci, A. 2016, 'CHA2DS2-VASc risk factors as predictors of stroke after acute coronary syndrome: A systematic review and meta-analysis', *European Heart Journal: Acute Cardiovascular Care*.

Gulshan, V., Peng, L., Coram, M. & et al. 2016, 'Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs', *JAMA*, vol. 316, no. 22, pp. 2402-10.

Guyon, I. & Elisseeff, A. 2003, 'An introduction to variable and feature selection', *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157-82.

Haddad, M., Adlassnig, K.-P. & Porenta, G. 1997, 'Feasibility analysis of a case-based reasoning system for automated detection of coronary heart disease from myocardial scintigrams', *Artificial Intelligence in Medicine*, vol. 9, no. 1, pp. 61-78.

Hall, M.A. 2000, *Correlation-based feature selection of discrete and numeric class machine learning*, Working Paper, University of Waikato, Department of Computer Science, 00/08.

Hammerla, N.Y., Fisher, J., Andras, P., Rochester, L., Walker, R. & Plötz, T. 2015, 'PD disease state assessment in naturalistic environments using deep learning', *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1742-8.

Han, J., Pei, J. & Kamber, M. 2011, *Data mining: concepts and techniques*, Elsevier.

Hanchaiphiboolkul, S., Puthkhao, P., Towanabut, S., Tantirittisak, T., Wangphonphatthanasiri, K., Termglinchan, T., Nidhinandana, S., Suwanwela, N.C. & Poungvarin, N. 2014, 'Factors Predicting High Estimated 10-Year Stroke Risk: Thai Epidemiologic Stroke Study', *Journal of Stroke and Cerebrovascular Diseases*, vol. 23, no. 7, pp. 1969-74.

Hitman, G.A., Colhoun, H., Newman, C., Szarek, M., Betteridge, D.J., Durrington, P.N., Fuller, J., Livingstone, S., Neil, H.A.W. & on behalf of the, C.I. 2007, 'Stroke prediction and stroke prevention with atorvastatin in the Collaborative Atorvastatin Diabetes Study (CARDS)', *Diabetic Medicine*, vol. 24, no. 12, pp. 1313-21.

Hossain, J., FazlidaMohdSani, N., Mustapha, A. & SurianiAffendey, L. 2013, 'Using feature selection as accuracy benchmarking in clinical data mining', *Journal of Computer Science*, vol. 9, no. 7, p. 883.

Howard, V.J. & McDonnell, M.N. 2015, 'Physical Activity in Primary Stroke Prevention', *Stroke*, vol. 46, no. 6, pp. 1735-9.

Hsu, C.-C. & Ho, C.-S. 2004, 'A new hybrid case-based architecture for medical diagnosis', *Information Sciences*, vol. 166, no. 1–4, pp. 231-47.

Hu, D., Huang, J., Wang, Y., Zhang, D. & Qu, Y. 2014, 'Fruits and vegetables consumption and risk of stroke: a meta-analysis of prospective cohort studies', *Stroke*, vol. 45, no. 6, pp. 1613-9.

Huang, Y., McCullagh, P., Black, N. & Harper, R. 2007, 'Feature selection and classification model construction on type 2 diabetic patients' data', *Artificial Intelligence in Medicine*, vol. 41, no. 3, pp. 251-62.

Huda, S., Yearwood, J., Jelinek, H.F., Hassan, M.M., Fortino, G. & Buckland, M. 2016, 'A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis', *IEEE access*, vol. 4, pp. 9145-54.

Hung, C.Y., Chen, W.C., Lai, P.T., Lin, C.H. & Lee, C.C. 2017, 'Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database', *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3110-3.

Ibrahim-Verbaas, C.A., Fornage, M., Bis, J.C., Choi, S.H., Psaty, B.M., Meigs, J.B., Rao, M., Nalls, M., Fontes, J.D., O'Donnell, C.J., Kathiresan, S., Ehret, G.B., Fox, C.S., Malik, R., Dichgans, M., Schmidt, H., Lahti, J., Heckbert, S.R., Lumley, T., Rice, K., Rotter, J.I., Taylor, K.D., Folsom, A.R., Boerwinkle, E., Rosamond, W.D., Shahar, E., Gottesman, R.F., Koudstaal, P.J., Amin, N., Wieberdink, R.G., Dehghan, A., Hofman, A., Uitterlinden, A.G., DeStefano, A.L., Debette, S., Xue, L., Beiser, A., Wolf, P.A., DeCarli, C., Ikram, M.A., Seshadri, S., Mosley, T.H., Longstreth, W.T., van Duijn, C.M. & Launer, L.J. 2014, 'Predicting stroke through genetic risk functions: The CHARGE risk score project', *Stroke; a journal of cerebral circulation*, vol. 45, no. 2, pp. 403-12.

Jafer, Y., Matwin, S. & Sokolova, M. 2014, 'Using Feature Selection to Improve the Utility of Differentially Private Data Publishing', *Procedia Computer Science*, vol. 37, pp. 511-6.

Jauch, E.C., Saver, J.L., Adams, H.P., Bruno, A., Connors, J.J., Demaerschalk, B.M., Khatri, P., McMullan, P.W., Qureshi, A.I., Rosenfield, K., Scott, P.A., Summers, D.R., Wang, D.Z., Wintermark, M. & Yonas, H. 2013, 'Guidelines for the Early Management of Patients With Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association', *Stroke*, vol. 44, no. 3, pp. 870-947.

Kamel, H., Okin, P.M., Elkind, M.S.V. & Iadecola, C. 2016, 'Atrial Fibrillation and Mechanisms of Stroke: Time for a New Model', *Stroke*, vol. 47, no. 3, pp. 895-900.

Kamijo, K. & Tanigawa, T. 1990, 'Stock price pattern recognition-a recurrent neural network approach', paper presented to the *1990 IJCNN International Joint Conference on Neural Networks*, San Diego, CA, USA, <http://ieeexplore.ieee.org.ezproxy.lib.uts.edu.au/stamp/stamp.jsp?tp=&arnumber=5726532&isnumber=3745>.

Kamkar, I., Gupta, S.K., Phung, D. & Venkatesh, S. 2015, 'Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso', *Journal of Biomedical Informatics*, vol. 53, pp. 277-90.

Kassim, A.A., Galadanci, N.A., Pruthi, S. & DeBaun, M.R. 2015, 'How I treat and manage strokes in sickle cell disease', *Blood*, vol. 125, no. 22, pp. 3401-10.

Khalilia, M., Chakraborty, S. & Popescu, M. 2011, 'Predicting disease risks from highly imbalanced data using random forest', *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, p. 51.

Khemphila, A. & Boonjing, V. 2011, 'Heart Disease Classification Using Neural Network and Feature Selection', *2011 21st International Conference on Systems Engineering*, pp. 406-9.

Khosla, A., Cao, Y., Lin, C.C.-Y., Chiu, H.-K., Hu, J. & Lee, H. 2010, 'An integrated machine learning approach to stroke prediction', paper presented to the *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington, DC, USA.

Kiragu, M.K. & Waiganjo, P.W. 2016, 'Case based Reasoning for Treatment and Management of Diabetes', *Diabetes*, vol. 145, no. 4.

Kolodner, J. 2014, *Case-based reasoning*, Morgan Kaufmann.

Kolodner, J.L. & Kolodner, R.M. 1987, 'Using Experience in Clinical Problem Solving: Introduction and Framework', *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 17, no. 3, pp. 420-31.

König, I.R., Ziegler, A., Bluhmki, E., Hacke, W., Bath, P.M.W., Sacco, R.L., Diener, H.C., Weimar, C. & Investigators, o.b.o.t.V.I.S.T.A. 2008, 'Predicting Long-Term Outcome After Acute Ischemic Stroke: A Simple Index Works in Patients From Controlled Clinical Trials', *Stroke*, vol. 39, no. 6, pp. 1821-6.

Kononenko, I. 1993, 'Inductive and Bayesian learning in medical diagnosis', *Applied Artificial Intelligence an International Journal*, vol. 7, no. 4, pp. 317-37.

Koton, P. 1988, 'Reasoning about evidence in causal explanations', *Proceedings of the Seventh AAAI National Conference on Artificial Intelligence*, AAAI Press, pp. 256-61.

Koton, P. 1989, 'A medical reasoning program that improves with experience', *Computer Methods and Programs in Biomedicine*, vol. 30, no. 2, pp. 177-84.

Kwiatkowska, M. & Atkins, M. 2004, 'Case representation and retrieval in the diagnosis and treatment of obstructive sleep apnea: a semio-fuzzy approach', *Proceedings of the 7th European Conference on Case-Based Reasoning*, pp. 25-35.

Langhorne, P., Bernhardt, J. & Kwakkel, G. 2011, 'Stroke rehabilitation', *The Lancet*, vol. 377, no. 9778, pp. 1693-702.

Langhorne, P., O'Donnell, M.J., Chin, S.L., Zhang, H., Xavier, D., Avezum, A., Mathur, N., Turner, M., MacLeod, M.J., Lopez-Jaramillo, P., Damasceno, A., Hankey, G.J., Dans, A.L., Elsayed, A., Mondo, C., Wasay, M., Czlonkowska, A., Weimar, C., Yusufali, A.H., Hussain, F.A., Lisheng, L., Diener, H.-C., Ryglewicz, D., Pogosova, N., Iqbal, R., Diaz, R., Yusoff, K., Oguz, A., Wang, X., Penaherrera, E., Lanas, F., Ogah, O.S., Ogunniyi, A., Iversen, H.K., Malaga, G., Rumboldt, Z., Magazi, D., Nilanont, Y., Rosengren, A., Oveisgharan, S., Yusuf, S., O'Donnell, M., Yusuf, S., Rangarajan, S., Rao-Melacini, P., Zhang, X.M., Islam, S., Kabali, C., Casanova, A., Chin, S.L., DeJesus, J., Dehghan, M., Agapay, S., McQueen, M., Hall, K., Keys, J., Wang, X., Devanath, A., Gupta, R., Prabhakaran, D., Diaz, R., Schygiel, P., Garrote, M., Rodriguez, M.A., Caccavo, A., Duran, R.G., Sposato, L., Molinos, J., Valdez, P., Cedrolla, C.M., Nofal, P.G., Huerta, M.F., Desmery, P.M., Zurru, M.C., Della Vedova, B., Varigos, J., Hankey, G., Kraemer, T., Gates, P., Bladin, C., Herkes, G., Avezum, A., Pereira, M.P., Minuzzo, L., Oliveira, L., Teixeira, M., Reis, H., Carvalho, A., Ouriques Martins, S., Carvalho, J.J., Gebara, O., Minelli, C., Oliveira, D.C., Sobral Sousa, A.C., Ferraz de Almeida, A.C., Hernandez, M.E., Friedrich, M., Mota, D.M., Ritt, L.E., Correa Vila Nova, D., Teal, P., Gladstone, D., Shuaib, A., Silver, F., Dowlatshahi, D., Lanas, F., Carcamo, D., Santibañez, C., Garces, E., Liu, L.S., Zhang, H.Y., Fang, H.P., Lian, M.F., Shen, F., Luo, F.X., Wen, X.X., Xu, Z.Q., Liu, Z.Z., Yan, W., Yu, J.F., Wang, W.K., Liu, L.H., Sun, Y.H., Zhou, L.C., Zhang, Z.F., Lv, J., Zhang, C.S., Chen, G., Wang, H.L., Chen, Y., Zheng, H., Huang, J.J., Li, W.Z., Wang, L.J., Shi, J.X., Hu, C.Y., Song, H.F., Ji, R.Y., Wang, D.L., Meng, L.H., Meng, Q.W., Duan, L.J., Liu, H.F., Luo, Y.C., Zhang, Q.Y., Wu, Y.B., Wang, C.R., Zhao, J.G., Liu, S.G., Shi, C.L., Wang, X.Y., Lopez-Jaramillo, P., Martinez, A., Sanchez-Vallejo, G., Molina, D.I., Espinosa, T., Garcia Lozada, H., Gomez-Arbelaez, D., Camacho, P.A., Rumboldt, Z., Lusic, I., Iversen, H.K., Truelsen, T., Back, C., Pedersen, M.M., Peñaherrera, E., Duarte, Y.C., Cevallos, S., Tettamanti, D., Caceres, S., Diener, H.C., Weimar, C., Grau, A., Rother, J., Ritter, M., Back, T., Winter, Y., Pais, P., Xavier, D., Sigamani, A., Mathur, N., Rahul, P., Murali, A., Roy, A.K., Sarma, G.R.K., Matthew, T., Kusumkar, G., Salam, K.A., Karadan, U., Achambat, L., Singh, Y., Pandian, J.D., Verma, R., Atam, V., Agarwal, A., Chidambaram, N., Umarani, R., Ghanta, S., Babu, G.K., Sathyanarayana, G., Sarada, G., Navya Vani, S., Sundararajan, R., Sivakumar, S.S., Wadia, R.S., Bandishti, S., Gupta, R., Agarwal, R.R., Mohan, I., Joshi, S., Kulkarni, S., Partha Saradhi, S., Joshi, P., Pandharipande, M., Badnerkar, N., Joshi, R., Kalantri, S.P., Somkumar, S., Chauhan, S., Singh, H., Varma, S., Singh, H., Sidhu, G.K., Singh, R., Bansal, K.L., Bharani, A., Pagare, S., Chouhan, A., Mahanta, B.N., Mahanta, T.G., Rajkonwar, G., Diwan, S.K., Mahajan, S.N., Shaikh, P., Devendrappa, H.R., Agrawal, B.K., Agrawal, A., Khurana, D., Thakur, S., Jain, V., Oveisgharan, S., Bahonar, A., Kelishadi, R., Hossienzadeh, A., Raeisidehkordi, M., Akhavan, H., Walsh, T., Albaker, O., Yusoff, K., Chandramouli, A., Shahadan, S., Ibrahim, Z., Husin, A., Damasceno, A., Lobo, V., Loureiro, S., Govo, V.A., Ogah, O.S., Ogunniyi, A., Akinyemi, R.O., Owolabi, M.O., Sani, M.U., Owolabi, L.F., Iqbal, R., Wasay, M., Raza, A., Malaga, G.G., Lazo-Porras, M., Loza-Herrera, J.D., Acuña-Villaorduña, A., Cardenas-Montero, D., Dans, A., Collantes, E., Morales, D., Roxas, A., Villarruz-Sulit, M.V.C., Czlonkowska, A., Ryglewicz, D., Skowronska, M., Restel, M., Bochynska, A., Chwojnicki, K., Kubach, M., Stowik, A., Wnuk, M., Pogosova, N., Ausheva, A., Karpova, A., Pshenichnikova, V., Vertkin, A., Kursakov, A., Boytsov, S., Al-Hussain, F., DeVilliers, L., Magazi, D., Mayosi, B., Elsayed, A.S.A., Bikhari, A., Sawaraldahab, Z., Hamad, H., ElTaher, M., Abdelhameed, A., Alawad, M., Alkabashi, D., Alsir, H., Rosengren, A., Andreasson, M., Kembro Johansson, J., Cederin, B., Schander, C., Elgasen, A.C., Bertholds, E., Boström Bengtsson, K., Nilanont, Y., Nidhinandana, S., Tatsanavivat, P., Paryoonwiwat, N., Poungvarin, N., Suwanwela, N.C., Tiamkao, S., Tulyapornchote, R., Boonyakarnkul, S., Hanchaiphiboolkul, S., Muengtaweepongsa, S., Watcharasaksilp, K., Sathirapanya, P., Pleumpanupat, P., Oguz, A., Akalin, A.A., Caklili, O.T., Isik, N., Caliskan, B., Sanlisoy, B., Balkuv, E., Tireli, H., Yayla, V., Cabalar, M., Culha, A., Senadim, S., Arpaci, B., Dayan, C., Argun, T., Yilmaz, S., Celiker, S., Kocer, A., Asil, T., Eryigit, G., Mondo, C., Kayima, J., Nakisige, M., Kitoleeko, S., Yusufali, A.M., Zuberi, B.J., Mirza, H.Z., Saleh, A.A., BinAdi, J.M., Hussain, F., Langhorne, P., Muir, K., Walters, M., McAlpine, C., Ghosh, S., Doney, A., Johnston, S., Mudd, P., Black, T., Murphy, P., Jenkinson, D., Kelly, D., Whiting, R., Dutta, D., Shaw, L., McFarlane, C., Ronald, E. & McBurnie, K. 2018, 'Practice patterns and outcomes after stroke across countries at different economic levels (INTERSTROKE): an international observational study', *The Lancet*, vol. 391, no. 10134, pp. 2019-27.

LeBozec, C., Jaulent, M.C., Zapletal, E. & Degoulet, P. 1998, 'Unified modeling language and design of a case-based retrieval system in medical imaging', *Proceedings of the AMIA Symposium*, pp. 887-91.

LeCun, Y., Bengio, Y. & Hinton, G. 2015, 'Deep learning', *Nature*, vol. 521, no. 7553, pp. 436-44.

Leira, E.C., Ku-Chou, C., Davis, P.H., Clarke, W.R., Woolson, R.F., Hansen, M.D. & Adams, H.P., Jr. 2004, 'Can We Predict Early Recurrence in Acute Stroke?', *Cerebrovascular Diseases*, vol. 18, no. 2, pp. 139-44.

Letham, B., Rudin, C., McCormick, T.H. & Madigan, D. 2015a, 'Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model', pp. 1350-71.

Letham, B., Rudin, C., McCormick, T.H. & Madigan, D. 2015b, 'Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model', *The Annals of Applied Statistics*, vol. 9, pp. 1350-71.

Li, R., Zhang, W., Suk, H.-I., Wang, L., Li, J., Shen, D. & Ji, S. 2014, 'Deep Learning Based Imaging Data Completion for Improved Brain Disease Diagnosis', in P. Golland, N. Hata, C. Barillot, J. Hornegger & R. Howe (eds), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part III*, Springer International Publishing, Cham, pp. 305-12.

Liang, Z., Zhang, G., Huang, J.X. & Hu, Q.V. 2014, 'Deep learning for healthcare decision making with EMRs', *IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2014*, pp. 556-9.

Lip, G.Y.H., Frison, L., Halperin, J.L. & Lane, D.A. 2010, 'Identifying Patients at High Risk for Stroke Despite Anticoagulation', *Stroke*, vol. 41, no. 12, p. 2731.

Lip, G.Y.H., Nieuwlaat, R., Pisters, R., Lane, D.A. & Crijns, H.J.G.M. 2010, 'Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: The euro heart survey on atrial fibrillation', *Chest*, vol. 137, no. 2, pp. 263-72.

Lisabeth, L. & Bushnell, C. 2012, 'Stroke risk in women: the role of menopause and hormone therapy', *The Lancet Neurology*, vol. 11, no. 1, pp. 82-91.

López, B. & Plaza, E. 1993, 'Case-based planning for medical diagnosis', in J. Komorowski & Z.W. Raś (eds), *Methodologies for Intelligent Systems: 7th International Symposium, ISMIS'93 Trondheim, Norway, June 15–18, 1993 Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 96-105.

Lu, N., Lu, J. & Zhang, G. 2009, 'An Integrated Knowledge Adaption Framework for Case-Based Reasoning Systems', in J.D. Velásquez, S.A. Ríos, R.J. Howlett &

L.C. Jain (eds), *Knowledge-Based and Intelligent Information and Engineering Systems: 13th International Conference, KES 2009, Santiago, Chile, September 28-30, 2009, Proceedings, Part II*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 372-9.

Lumley, T., Kronmal, R.A., Cushman, M., Manolio, T.A. & Goldstein, S. 2002, 'A stroke prediction score in the elderly: validation and Web-based application', *Journal of Clinical Epidemiology*, vol. 55, no. 2, pp. 129-36.

Lyu, C., Chen, B., Ren, Y. & Ji, D. 2017, 'Long short-term memory RNN for biomedical named entity recognition', *BMC Bioinformatics*, vol. 18, no. 1, p. 462.

MacQueen, J. 1967, 'Some methods for classification and analysis of multivariate observations', *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, University of California Press, Berkeley, Calif., pp. 281-97.

Macura, R.T. & Macura, K.J. 1995, 'MacRad: Case-based retrieval system for radiology image resource', *Proceedings of ICCBR*, vol. 95.

Manuel, D.G., Tuna, M., Perez, R., Tanuseputro, P., Hennessy, D., Bennett, C., Rosella, L., Sanmartin, C., van Walraven, C. & Tu, J.V. 2015, 'Predicting Stroke Risk Based on Health Behaviours: Development of the Stroke Population Risk Tool (SPoRT)', *PLoS ONE*, vol. 10, no. 12, p. e0143342.

Mao, Y., Chen, W., Chen, Y., Lu, C., Kollef, M. & Bailey, T. 2012, 'An integrated data mining approach to real-time clinical monitoring and deterioration warning', paper presented to the *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Beijing, China.

Marling, C. & Whitehouse, P. 2001, 'Case-Based Reasoning in the Care of Alzheimer's Disease Patients', in D.W. Aha & I. Watson (eds), *Case-Based Reasoning Research and Development: 4th International Conference on Case-Based Reasoning, ICCBR 2001 Vancouver, BC, Canada, July 30 – August 2, 2001 Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 702-15.

Mayo Clinic Staff Oct 2018, *Stroke*, Mayo Clinic, viewed 15 Jan 2019, <https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113?p=1>.

Mcheick, H., Nasser, H., Dbouk, M. & Nasser, A. 2016, 'Stroke Prediction Context-Aware Health Care System', *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 30-5.

Md Mahfuj Ul, A., Shah Md Sarwer, J., Afrin, S. & Md Zakir, H. 2017, 'Diabetic and Non-diabetic Subjects with Ischemic Stroke: Risk Factors, Stroke Topography and Hospital Outcome', *Journal of Medicine*, vol. 18, no. 2, p. 75.

Medeiros, F., Casanova, M.d.A., Fraulob, J.C. & Trindade, M. 2012, 'How can diet influence the risk of stroke?', *International journal of hypertension*, vol. 2012, pp. 763507-.

Montani, S., Magni, P., Bellazzi, R., Larizza, C., Roudsari, A.V. & Carson, E.R. 2003, 'Integrating model-based decision support in a multi-modal reasoning system for managing type 1 diabetic patients', *Artificial Intelligence in Medicine*, vol. 29, no. 1–2, pp. 131-51.

Morales-Vidal, S. & Biller, J. Dec 16, 2003, *Hormonal contraception and stroke*, MedLink Corporation, San Diego, viewed 10 Oct 2018, <http://www.medlink.com/article/hormonal_contraception_and_stroke>.

Morillas, P., Pallarés, V., Fácila, L., Llisterri, J.L., Sebastián, M.E., Gómez, M., Castilla, E., Camarasa, R., Sandin, M. & García-Honrubia, A. 2015, 'The

CHADS<sub>2</sub> Score to Predict Stroke Risk in the Absence of Atrial Fibrillation in Hypertensive Patients Aged 65 Years or Older', *Revista Española de Cardiología (English Edition)*, vol. 68, no. 06, pp. 485-91.

Mullen, M.T. 1996, *Stroke associated with drug abuse*, MedLink Corporation, San Diego, viewed 10 Sep 2018, <http://www.medlink.com/article/stroke_associated_with_drug_abuse>.

Nichols, F.T. 2018, *Stroke associated with sickle cell disease*, MedLink Corporation, San Diego, viewed 1 Oct 2018, <http://www.medlink.com/article/stroke_associated_with_sickle_cell_disease>.

Nie, L., Wang, M., Zhang, L., Yan, S., Zhang, B. & Chua, T.S. 2015, 'Disease Inference from Health-Related Questions via Sparse Deep Learning', *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 8, pp. 2107-19.

Nilsson, M. & Funk, P. 2004, 'A Case-Based Classification of Respiratory Sinus Arrhythmia', in P. Funk & P.A. González Calero (eds), *Advances in Case-Based Reasoning: 7th European Conference, ECCBR 2004, Madrid, Spain, August 30 - September 2, 2004. Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 673-85.

O'Donnell, M.J., Chin, S.L., Rangarajan, S., Xavier, D., Liu, L., Zhang, H., Rao-Melacini, P., Zhang, X., Pais, P., Agapay, S., Lopez-Jaramillo, P., Damasceno, A., Langhorne, P., McQueen, M.J., Rosengren, A., Dehghan, M., Hankey, G.J., Dans, A.L., Elsayed, A., Avezum, A., Mondo, C., Diener, H.-C., Ryglewicz, D., Czlonkowska, A., Pogosova, N., Weimar, C., Iqbal, R., Diaz, R., Yusoff, K., Yusufali, A., Oguz, A., Wang, X., Penaherrera, E., Lanas, F., Ogah, O.S., Ogunniyi, A., Iversen, H.K., Malaga, G., Rumboldt, Z., Oveisgharan, S., Al Hussain, F., Magazi, D., Nilanont, Y., Ferguson, J., Pare, G. & Yusuf, S. 2016, 'Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study', *The Lancet*, vol. 388, no. 10046, pp. 761-75.

O'Donnell, M.J., Xavier, D., Liu, L., Zhang, H., Chin, S.L., Rao-Melacini, P., Rangarajan, S., Islam, S., Pais, P., McQueen, M.J., Mondo, C., Damasceno, A., Lopez-Jaramillo, P., Hankey, G.J., Dans, A.L., Yusoff, K., Truelsen, T., Diener, H.-C., Sacco, R.L., Ryglewicz, D., Czlonkowska, A., Weimar, C., Wang, X. & Yusuf, S. 2010, 'Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study', *The Lancet*, vol. 376, no. 9735, pp. 112-23.

Ohkubo, T., Asayama, K., Kikuya, M., Metoki, H., Hoshi, H., Hashimoto, J., Totsune, K., Satoh, H. & Imai, Y. 2004, 'How many times should blood pressure be measured at home for better prediction of stroke risk? Ten-year follow-up results from the Ohasama study', *Journal of Hypertension*, vol. 22, no. 6, pp. 1099-104.

Olesen, J.B., Lip, G.Y.H., Hansen, M.L., Hansen, P.R., Tolstrup, J.S., Lindhardsen, J., Selmer, C., Ahlehoff, O., Olsen, A.-M.S., Gislason, G.H. & Torp-Pedersen, C. 2011, 'Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study', *BMJ*, vol. 342.

Owolabi, M.O., Sarfo, F., Akinyemi, R., Gebregziabher, M., Akpa, O., Akpalu, A., Wahab, K., Obiako, R., Owolabi, L., Ovbiagele, B., Owolabi, M.O., Sarfo, F.S., Akinyemi, R., Gebregziabher, M., Akpa, O., Akpalu, A., Wahab, K., Obiako, R., Owolabi, L., Ovbiagele, B., Tiwari, H.K., Arnett, D., Lackland, D., Adeoye, A.M., Akin, O., Ogbole, G., Jenkins, C., Arulogun, O., Ryan, I.M., Armstrong, K., Olowoyo, P., Komolafe, M., Osaigbovo, G., Obiabo, O., Chukwuonye, I., Adebayo, P., Adebayo, O., Omololu, A., Otubogun, F., Olaleye, A., Durodola, A., Olunuga, T., Akinwande, K., Aridegbe, M., Fawale, B., Adeleye, O., Kolo, P., Appiah, L., Singh, A., Adamu, S., Awuah, D., Saulson, R., Agyekum, F., Shidali, V., Ogah, O., Oguntade, A., Umanruochi, K., Iheonye, H., Imoh, L., Afolaranmi, T., Calys-Tagoe, B., Okeke, O., Fakunle, A., Akinyemi, J., Akpalu, J., Ibinaiye, P., Agunloye, A., Sanni, T., Bisi, A., Efidi, C., Bock-Oruma, A., Melikam, S., Olaniyan, L., Yaria, J., Odo, C.J., Lakoh, S., Ogunjimi, L., Salaam, A., Oyinloye, L., Asaleye, C., Sanya, E., Olowookere, S., Makanjuola, A., Oguntoye, A., Uvere, E., Faniyan, M., Akintunde, A., Kehinde, I., Diala, S., Adeleye, O., Ajose, O.A., Onyeonoro, U., Amusa, A.G., Owusu, D. & Mensah, Y. 2018, 'Dominant modifiable risk factors for stroke in Ghana and Nigeria (SIREN): a case-control study', *The Lancet Global Health*, vol. 6, no. 4, pp. e436-e46.

Palaniappan, S. & Awang, R. 2008, 'Intelligent heart disease prediction system using data mining techniques', *2008 IEEE/ACS International Conference on Computer Systems and Applications*, pp. 108-15.

Palem, G. April 2013, *The Practice of Predictive Analytics in Healthcare*, viewed 5 Auguest 2016, <https://www.researchgate.net/publication/236336250_The_Practice_of_Predictive_Analytics_in_Healthcare>.

Perner, P. 1999, 'An architecture for a CBR image segmentation system', *Engineering Applications of Artificial Intelligence*, vol. 12, no. 6, pp. 749-59.

Perner, P., Perner, H., Janichen, S. & Buhring, A. 2004, 'Recognition of airborne fungi spores in digital microscopic images', *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3, pp. 566-9 Vol.3.

Phuong, N.H., Thang, V.V. & Hirota, K. 2000, 'Case Based Reasoning for Medical Diagnosis using Fuzzy Set Theory', *Biomedical fuzzy and human sciences : the official journal of the Biomedical Fuzzy Systems Association*, vol. 5, no. 2, pp. 1-7.

Poçi, D., Hartford, M., Karlsson, T., Herlitz, J., Edvardsson, N. & Caidahl, K. 2012, 'Role of the chads2 score in acute coronary syndromes: Risk of subsequent death or stroke in patients with and without atrial fibrillation', *Chest*, vol. 141, no. 6, pp. 1431-40.

Rahman, A.S., Akhtar, S.W., Jamal, Q., Sultana, N., Siddiqui, M.A. & Hassan, Z. 2017, 'Ischaemic stroke and peripheral artery disease', *Journal of the Pakistan Medical Association*, vol. 67, no. 8, pp. 1138-43.

Rajeswari, K., Vaithiyanathan, V. & Pede, S.V. 2013, 'Feature selection for classification in medical data mining', *International Journal of Emerging Trends and Technology in Computer Science (IJETTCS)*, vol. 2, no. 2, pp. 492-7.

Reategui, E.B., Campbell, J.A. & Leao, B.F. 1997, 'Combining a neural network with case-based reasoning in a diagnostic system', *Artificial Intelligence in Medicine*, vol. 9, no. 1, pp. 5-27.

Richter, M.M. & Weber, R. 2013, *Case-Based Reasoning: A Textbook*, Springer Science & Business Media.

Rissland, E.L. 1983, 'Examples in Legal Reasoning: Legal Hypotheticals', *Proceedings of the Eighth International Joint Conference on Artificial Intelligence, IJCAI*, pp. 90-3.

Sak, H., Senior, A. & Beaufays, F. 2014, 'Long short-term memory recurrent neural network architectures for large scale acoustic modeling', *Fifteenth Annual Conference of the International Speech Communication Association*, Singapore, pp. 338-42.

Schmidt, R., Pollwein, B. & Gierl, L. 1999, 'Medical multiparametric time course prognoses applied to kidney function assessments', *International Journal of Medical Informatics*, vol. 53, no. 2–3, pp. 253-63.

Selvakuberan, K., Kayathiri, D., Harini, B. & Devi, M.I. 2011, 'An efficient feature selection method for classification in health care systems using machine learning techniques', *2011 3rd International Conference on Electronics Computer Technology*, vol. 4, pp. 223-6.

Sharaf-El-Deen, D.A., Moawad, I.F. & Khalifa, M.E. 2014a, 'A New Hybrid Case-Based Reasoning Approach for Medical Diagnosis Systems', *Journal of Medical Systems*, vol. 38, no. 2, p. 9.

Sharaf-el-deen, D.A., Moawad, I.F. & Khalifa, M.E. 2014b, 'A New Hybrid Case-Based Reasoning Approach for Medical Diagnosis Systems', *Journal of Medical Systems*, vol. 38, no. 2, pp. 1-9.

Shi, W. & Barnden, J.A. 2005, 'Using Inductive Rules in Medical Case-Based Reasoning System', in A. Gelbukh, Á. de Albornoz & H. Terashima-Marín (eds), *MICAI 2005: Advances in Artificial Intelligence: 4th Mexican International Conference on Artificial Intelligence, Monterrey, Mexico, November 14-18, 2005. Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 900-9.

Soni, J., Ansari, U., Sharma, D. & Soni, S. 2011, 'Predictive data mining for medical diagnosis: An overview of heart disease prediction', *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43-8.

Srinivas, K., Rani, B.K. & Govrdhan, A. 2010, 'Applications of data mining techniques in healthcare and prediction of heart attacks', *International Journal on Computer Science and Engineering (IJCSE)*, vol. 2, no. 02, pp. 250-5.

Stier, N., Vincent, N., Liebeskind, D. & Scalzo, F. 2015, 'Deep learning of tissue fate features in acute ischemic stroke', *IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2015*, pp. 1316-21.

Stroke Risk in Atrial Fibrillation Working Group 2008, 'Comparison of 12 Risk Stratification Schemes to Predict Stroke in Patients With Nonvalvular Atrial Fibrillation', *Stroke*, vol. 39, no. 6, pp. 1901-10.

Sudha, A., Gayathri, P. & Jaisankar, N. 2012, 'Effective analysis and predictive model of stroke disease using classification methods', *International Journal of Computer Applications*, vol. 43, no. 14, pp. 26-31.

Tazin, N., Sabab, S.A. & Chowdhury, M.T. 2016, 'Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique', *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)*, pp. 1-6.

The American Heart Association 2008, 'Comparison of 12 Risk Stratification Schemes to Predict Stroke in Patients With Nonvalvular Atrial Fibrillation', *Stroke*, vol. 39, no. 6, pp. 1901-10.

The American Heart Association 2016, *Understanding Stroke Risk*, The American Heart Association, viewed 12 October 2016, <http://www.strokeassociation.org/STROKEORG/AboutStroke/UnderstandingRisk/Understanding-Stroke-Risk_UCM_308539_SubHomePage.jsp>.

The Bureau of Policy and Strategy & Ministry of Public Health of Thailand 2016, *ICD-10-TM 2016 Tabular List of Diseases*, vol. 1, Thailand.

Tsang-Hsiang, C., Chih-Ping, W. & Tseng, V.S. 2006, 'Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches', *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pp. 165-70.

Tun, N.N., Arunagirinathan, G., Munshi, S.K. & Pappachan, J.M. 2017, 'Diabetes mellitus and stroke: A clinical update', *World journal of diabetes*, vol. 8, no. 6, pp. 235-48.

van den Branden, M., Wiratunga, N., Burton, D. & Craw, S. 2011, 'Integrating case-based reasoning with an electronic patient record system', *Artificial Intelligence in Medicine*, vol. 51, no. 2, pp. 117-23.

Vieira, S.M., Mendonça, L.F., Farinha, G.J. & Sousa, J.M.C. 2013, 'Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients', *Applied Soft Computing*, vol. 13, no. 8, pp. 3494-504.

Vorobieva, O., Gierl, L. & Schmidt, R. 2003, 'Adaptation methods in an endocrine therapy support system', *Workshop Proceedings of the Fifth International Conference on Case-Based Reasoning. NTNU, Trondheim, Norway*, pp. 80-8.

Weimar, C., Ziegler, A., König, R.I. & Diener, H.-C. 2002, 'Predicting functional outcome and survival after acute ischemic stroke', *Journal of Neurology*, vol. 249, no. 7, pp. 888-95.

Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. & Qureshi, N. 2017, 'Can machine-learning improve cardiovascular risk prediction using routine clinical data?', *PLoS ONE*, vol. 12, no. 4, pp. 1-14.

Westover, A.N., McBride, S. & Haley, R.W. 2007, 'Stroke in Young Adults Who Abuse Amphetamines or Cocaine: A Population-Based Study of Hospitalized Patients', *Archives of General Psychiatry*, vol. 64, no. 4, pp. 495-502.

World Health Organization 2004, *ICD10: International Statistical Classification of Disease and Related Health Tenth Revision*, World Health Organization, Geneva.

Xu, L., Redman, C.W.G., Payne, S.J. & Georgieva, A. 2014, 'Feature selection using genetic algorithms for fetal heart rate analysis', *Physiological Measurement*, vol. 35, no. 7, pp. 1357-71.