

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.

Production Note:

Signature removed prior to publication.

Signature:

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

ACKNOWLEDGEMENT

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

CERTIFICATE OF ORIGINAL AUTHORSHIP.....	I
ACKNOWLEDGEMENT	II
LIST OF PUBLICATIONS.....	III
TABLE OF CONTENTS	IV
LIST OF FIGURES.....	VIII
LIST OF TABLES	IX
ABSTRACT.....	X
CHAPTER 1. INTRODUCTION.....	1
1.1. INTRODUCTION	1
1.2. MOTIVATION	6
1.3. RESEARCH QUESTION	8
1.4. OBJECTIVE AND AIMS	9
1.5. PLAN OF THE THESIS	10
CHAPTER 2. LITERATURE REVIEW.....	13
2.1. INTRODUCTION	13
2.2. WHAT IS STROKE?.....	13
2.3. RISK FACTORS EFFECTING STROKE PATIENTS.....	15
2.3.1. <i>The unchangeable risk factors of stroke</i>	15
2.3.2. <i>The changeable risk factors of stroke</i>	17
2.3.3. <i>Other risk factors that are less well-documented</i>	20
2.4. WHY IS PREDICTION REQUIRED?	22
2.5. PREDICTIVE MODELS IN HEALTHCARE	22
2.5.1. <i>Predictive Data mining</i>	22
2.5.2. <i>Decision Support systems</i>	24

2.5.3.	<i>Standard medical prediction score</i>	28
2.6.	FEATURE SELECTION IN HEALTHCARE.....	29
2.6.1.	<i>Variable ranking in feature selection</i>	36
2.6.2.	<i>Principle of the Method and Notations</i>	37
2.6.3.	<i>Correlation Criteria</i>	37
2.7.	CASE BASED REASONING.....	37
2.7.1.	<i>What is CBR?</i>	39
2.7.2.	<i>Architecture of CBR</i>	40
2.7.3.	<i>Significance of CBR</i>	42
2.7.4.	<i>Case-based reasoning in healthcare</i>	43
2.8.	DATA MINING AND PREDICTIVE ANALYSIS IN STROKE.....	44
2.9.	DEEP LEARNING METHOD.....	49
2.9.1.	<i>Deep Learning</i>	49
2.9.2.	<i>Long Short -Term Memory - Recurrent Neural Networks (LSTM-RNN)</i>	50
2.9.3.	<i>Deep Learning in Healthcare sector</i>	51
2.10.	SUMMARY.....	53
CHAPTER 3. ELECTRONIC HEALTHCARE RECORDS.....		62
3.1.	INTRODUCTION.....	62
3.2.	ELECTRONIC HEALTHCARE RECORDS.....	62
3.3.	THE INTERNATIONAL CLASSIFICATION OF DISEASES, 10 TH REVISION (ICD-10 TH).....	65
3.4.	SUMMARY.....	67
CHAPTER 4. FEATURE SELECTION FROM ELECTRONIC HEALTHCARE RECORDS.....		68
4.1.	INTRODUCTION.....	68
4.2.	SAMPLE SIZE AND FEATURE SELECTION OF RISK FACTORS.....	68
4.2.1.	<i>Sample size</i>	68
4.2.2.	<i>Feature Selection</i>	69
4.3.	EHRs MANAGEMENT FOR PREDICTION OF STROKE.....	71
4.4.	ICD-10 TH COMPLAINT ELECTRONIC HEALTHCARE RECORDS.....	72
4.5.	SUMMARY.....	75

CHAPTER 5. CASE BASED REASONING FRAMEWORK FOR STROKE	76
5.1. INTRODUCTION	76
5.2. A CBR FRAMEWORK OF PREDICTION MODEL IN STROKE PATIENTS	76
5.3. SUMMARY	85
CHAPTER 6. DEEP LEARNING FRAMEWORK FOR STROKE.....	86
6.1. INTRODUCTION	86
6.2. DEEP LEARNING ALGORITHM.....	86
6.3. STROKE ALGORITHM	92
6.4. SUMMARY	94
CHAPTER 7. EXPERIMENTS AND RESULTS.....	95
7.1. INTRODUCTION	95
7.2. DATA SOURCE.....	95
7.2.1. <i>Heart datasets</i>	95
7.2.2. <i>Electronic Healthcare Records dataset</i>	95
7.3. EXPERIMENTS IN HEART DATASETS.....	96
7.3.1. <i>Choosing number of attributes in Heart datasets</i>	97
7.3.2. <i>Validation of prediction in Naïve Bayes Algorithm</i>	98
7.3.3. <i>Validation of prediction in Support Vector Machine Algorithm</i>	99
7.3.4. <i>Validation of prediction in Deep Learning</i>	99
7.4. EXPERIMENTS IN EHRs DATASET	100
7.4.1. <i>Choosing number of attributes in EHRs dataset</i>	101
7.4.2. <i>Validation of Case-Based Reasoning</i>	101
7.4.3. <i>Validation of Deep Learning approach</i>	102
7.5. COMPARISON OF CASE-BASED REASONING AND DEEP LEARNING TECHNIQUES	106
7.6. SUMMARY	109
CHAPTER 8. CONCLUSION	110
8.1. INTRODUCTION	110
8.2. CONTRIBUTIONS OF THIS THESIS TO THE EXISTING LITERATURE	110

8.2.1.	<i>Contribution 1: Identify and select stroke risk factors</i>	112
8.2.2.	<i>Contribution 2: Application of Case-Based Reasoning Framework for Stroke</i>	112
8.2.3.	<i>Contribution 3: Apply different machine learning and Deep Learning techniques</i>	112
8.2.4.	<i>Contribution 4: Comparison of Case-Base Reasoning and Deep Learning</i>	113
8.3.	DISCUSSION	114
8.4.	CONCLUSION	118
8.5.	FUTURE WORK	120
APPENDIX A UTS HUMAN RESEARCH ETHICS COMMITTEE		122
BIBLIOGRAPHY		123

LIST OF FIGURES

FIGURE 2-1 : RISK FACTORS EFFECTIVE STROKE SYMPTOMS	21
FIGURE 2-2 : THE CASE-BASED REASONING CYCLE INTRODUCED BY AAMODT & PLAZA (1994)	41
FIGURE 2-3 : A STRUCTURE OF DEEP NEURAL NETWORK	49
FIGURE 2-4 : LSTM-RNN MEMORY ARCHITECTURE AND A SINGLE MEMORY BLOCK (GERS, SCHRAUDOLPH & SCHMIDHUBER 2002; GREFF ET AL. 2017; SAK, SENIOR & BEAUFAYS 2014).	50
FIGURE 3-1 :ELECTRONIC HEALTHCARE RECORDS OF STROKE PATIENTS (CHANTAMIT-O-PAS & GOYAL 2018B).	62
FIGURE 4-1 : TRADITIONAL STATISTICAL APPROACH FOR CALCULATE SAMPLE SIZE.	69
FIGURE 4-2 : THE TRADITIONAL STATISTICAL APPROACH FOR FEATURE SELECTION.	70
FIGURE 4-3 : THE CATEGORIZED ICD-10 TH CODES BY SYMPTOMS.	72
FIGURE 5-1: AN OVERVIEW OF THE CASE-BASED REASONING FRAMEWORK FOR THE PREDICTION ANALYTICAL SYSTEM IN STROKE PATIENTS (CHANTAMIT-O-PAS & GOYAL 2018A).	79
FIGURE 5-2 : FLOWCHART OF THE CASE-BASED REASONING FOR PREDICTION OF STROKE PATIENTS (CHANTAMIT-O-PAS & GOYAL 2018A).	79
FIGURE 5-3: THE DETAIL OF THE CASE-BASED REASONING FOR PREDICTION OF STROKE PATIENTS (CHANTAMIT-O-PAS & GOYAL 2018A).	81
FIGURE 6-1: PREDICTION MODEL; A) DEEP LEARNING PREDICTION MODEL AND B) LSTM PREDICTION MODEL.	88
FIGURE 7-1: THE MAIN RISK FACTORS OF STROKE PATIENTS IN ELECTRONIC HEALTHCARE RECORD.	96
FIGURE 7-2: THE MAIN RISK FACTORS FOR STROKE PATIENTS IN THE HEART DATASET	98
FIGURE 7-3: STROKE PREDICTION USING DEEP LEARNING WITH HEART DATASET (CHANTAMIT-O-PAS & GOYAL 2017).	100
FIGURE 7-4: THE PREDICTION RESULT USING K-NN.	102
FIGURE 7-5: THE PREDICTION RESULT USING SVM.	102
FIGURE 7-6: THE PREDICTION RESULT USING BACKPROPAGATION.	105
FIGURE 7-7: THE PREDICTION RESULT USING RNN.	105
FIGURE 7-8: THE PREDICTION RESULT USING LSTM.	106
FIGURE 7-9: THE PREDICTION RESULT USING FIVE DIFFERENT METHODS.	108

LIST OF TABLES

TABLE 2-1: DISEASE PREDICTION USING CASE-BASED REASONING SYSTEM AND MACHINE LEARNING.....	55
TABLE 2-2: FEATURE SELECTION TECHNIQUE IN HEALTHCARE.....	60
TABLE 3-1 : THE DATA STRUCTURE OF EHRs RECORDS	64
TABLE 3-2 : A PARTIAL SAMPLE OF THE DATA STRUCTURE OF EHRs RECORDS WITH STROKE RISK FACTORS.	65
TABLE 4-1 :EHRs RECORDS WITH STROKE’S RISK FACTORS.	73
TABLE 7-1 : COMPARISON OF THREE TECHNIQUES USED FOR PREDICTION OF STROKE.....	100
TABLE 7-2 : METRICS OF STROKE PREDICTION	104
TABLE 7-3 : COMPARISON OF METRICS FOR STROKE PREDICTION.....	106
TABLE 7-4 : CHANCE OF STROKE IN EHRs.....	107

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.