A Novel Stroke Prediction Model Based on Clinical Natural Language Processing (NLP) and Data Mining Methods

by

Elham Sedghi
M.Sc., University of Victoria, 2012

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Computer Science

A Novel Stroke Prediction Model Based on Clinical Natural Language Processing (NLP) and Data Mining Methods

by

Elham Sedghi

M.Sc., University of Victoria, 2012

Supervisory Committee

Dr. Jens H Weber, Supervisor

(Department of Computer Science)

Dr. Alex Thomo, Supervisor

(Department of Computer Science)

Dr. Alex Kuo, Outside Member

(Department of Health Information Science)

Dr. Tony Sahama, External Examiner

(Queensland University of Technology)

Supervisory Committee

Dr. Jens H Weber, Supervisor

(Department of Computer Science)

Dr. Alex Thomo, Supervisor

(Department of Computer Science)

Dr. Alex Kuo, Outside Member

(Department of Health Information Science)

Dr. Tony Sahama, External Examiner

(Queensland University of Technology)

ABSTRACT

Early detection and treatment of stroke can save lives. Before any procedure is planned, the patient is traditionally subjected to a brain scan such as Magnetic Resonance Imaging (MRI) in order to make sure he/she receives a safe treatment. Before any imaging is performed, the patient is checked into Emergency Room (ER) and clinicians from the Stroke Rapid Assessment Unit (SRAU) perform an evaluation of the patient's signs and symptoms. The question we address in this thesis is: Can Data Mining (DM) algorithms be employed to reliably predict the

occurrence of stroke in a patient based on the signs and symptoms gathered by the clinicians and other staff in the ER or the SRAU? A reliable DM algorithm would be very useful in helping the clinicians make a better decision whether to escalate the case or classify it as a non-life threatening mimic and not put the patient through unnecessary imaging and tests. Such an algorithm would not only make the life of patients and clinicians easier but would also enable the hospitals to cut down on their costs.

Most of the signs and symptoms gathered by clinicians in the ER or the SRAU are stored in free-text format in hospital information systems. Using techniques from Natural Language Processing (NLP), the vocabularies of interest can be extracted and classified. A big challenge in this process is that medical narratives are full of misspelled words and clinical abbreviations. It is a well known fact that the quality of data mining results crucially depends on the quality of input data. In this thesis, as a first contribution, we describe a procedure to preprocess the raw data and transform it into clean, well-structured data that can be effectively used by DM learning algorithms. Another contribution of this thesis is producing a set of carefully crafted rules to perform detection of negated meaning in free-text sentences. Using these rules, we were able to get the correct semantics of sentences and provide much more useful datasets to DM learning algorithms.

This thesis consists of three main parts. In the first part, we focus on building classifiers to reliably distinguish stroke and Transient Ischemic Attack (TIA) from mimic cases. For this, we used text extracted from the "chief complaint" and "history of patient illness" fields available in the patients' files at the Victoria General Hospital (VGH). In collaboration with stroke specialists, we identified a well-defined set of stroke-related keywords. Next, we created practical tools to accurately assign keywords from this set to each patient. Then, we performed

extensive experiments for finding the right learning algorithm to build the best classifier that provides a good balance between sensitivity, specificity, and a host of other quality indicators.

In the second part, we focus on the most important mimic case, migraine, and how to effectively distinguish it from stroke or TIA. This is a challenging problem because migraine has many signs and symptoms that are similar to those of stroke or TIA. Another challenge we address is the imbalance that our datasets have with respect to migraine. Namely the migraine cases are a minority of the overall cases. In order to alleviate this rarity problem, we propose a randomization procedure which is able to drastically improve the classifier quality.

Finally, in the third part, we provide a detailed study on datamining algorithms for extracting the most important predictors that can help to detect and prevent Posterior circulation stroke. We compared our finding with the attributes reported by the Heart and Stroke Foundation of Canada, and the features found in our study performed better in accuracy, sensitivity and ROC.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

Great thanks to merciful GOD for all the countless gifts he has offered us, and thanks to my family for their endless love and support.

I would like to express my deepest thanks and sincere appreciation to my supervisors, Dr Weber and Dr Thomo whom without their guidance, support, and consistent help, this dissertation would not have been possible. Thank you so much for the knowledge you have passed on; it has been an honor to have been your PhD student.

SPECTRA project has contributed immensely to my studies; I wish to express my sincere thanks to Dr Votova, Dr Penn, and Dr Bibok for their valuable support and collaboration. I will always be grateful for having the opportunity to work with you.

"If we knew what we were doing, it wouldn't be called research, would it?"

Albert Einstein

DEDICATION

This dissertation dedicates to my parents who taught me to trust in GOD and supported me with endless love and caring.

Chapter 1

Main Points

1.1 Introduction

Medical data is often represented in semi-structured or unstructured form, including textual narrative. Natural Language Processing (NLP) methods help to locate and extract information within clinical narrative text and are useful to transform unstructured texts to data in a machine interpretable format.

After preprocessing with NLP, data mining techniques are helpful in analyzing and interpreting data and can be used to create appropriate models for predicting disease based on signs and symptoms. Several studies showed data mining as a successful approach for extracting information from electronic health records [59], [13], [26].

There were several challenges we needed to address in our work. Typically, clinical narratives contain misspelled terms and incorrect grammar; also, most of these

reports are full of abbreviations and clinical acronyms that are not found in dictionaries [6]. One of the challenges in this work was data pre-processing and implementing appropriate negation detection rules. More specifically, we identified medical problems in patient records by extracting pre-defined sign and symptom terms (or keywords) provided by stroke specialists. Afterwards, we reviewed different negation detection methods and adopted existing methods to fit our problem context. With negation detection rules, we determined whether each key sign or symptom is present, unmentioned, or declared absent (mentioned in a negated context) and generated structured (codified) data for the data mining process. After pre-processing, data mining algorithms were utilized to analyze the data and build models for predicting stroke or TIA in patients. We systematically evaluated different data mining algorithms and computed standard metrics, such as recall (sensitivity), specificity, precision, F-measure, and ROC for each algorithm.

A crucial product of the prediction models we learned from codified data was a list of keywords weighted by their importance in the prediction quality (as captured by sensitivity, specificity, etc). The top keywords (typically less than 30) of the list were usually responsible for more than 95% of the prediction quality.

In other words, considering only the top keywords gave us models that performed almost as well as their counterparts built on the full set of keywords. Having the top keywords allowed us to build a questionnaire-like, online, application for the triage staff to use. This was effective because the number of the top keywords was small. The backend part of the online application is a prediction model, which outputs the classification of mimic or stroke/TIA. Based on this output, the triage staff can better assess whether the patient needs to be hospitalized or can be discharged. The

workflow, from data preprocessing to stroke prediction, is shown in Figure 1.1.

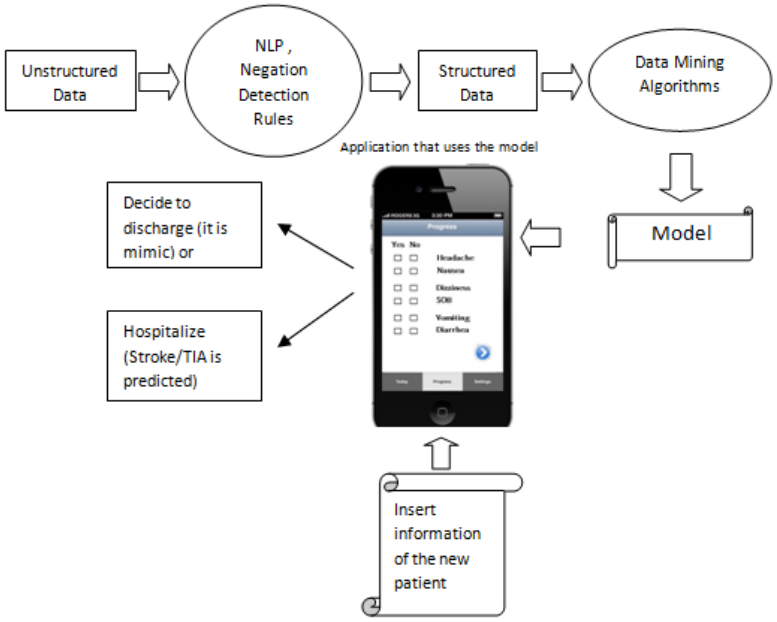


Figure 1.1: The workflow from data processing to stroke prediction.

As shown in Figure 1, the model (which represents the vector of features with specific weight for each feature) can be used in any type of application program (e.g. mobile, desktop, or online application) and clinicians can easily determine which feature is present, absent or unknown via a check-box driven form (e.g. Q1: Does patient have headache? yes , no , unknown) and once all the features are entered, the model can be called as a function and specifies whether the patient needs to be hospitalized (because of stroke/TIA) or can be discharged (mimic). [1]

In the second part of our study, we applied further analysis to establish a migraine detection model that distinguishes migraine patients from stroke cases. Distinguishing migraine from stroke is a challenge due to many common signs and

[1]Implementing the application is out of the scope of this thesis, but it is under progress.

symptoms. It is important to consider the cost of hospitalization and the time spent by neurologists and stroke nurses to visit, diagnose, and assign appropriate care to the patients; therefore, devising new ways to distinguish stroke, migraine and other types of mimic can help in improving decision making and saving time and expenses.

Our main challenge in this section was to tackle this data imbalance. Using the dataset in its original form to build classifiers led to a learning bias towards the majority class and against the minority (migraine) class. We used a sampling method to address the imbalance problem. First, different sources of data were preprocessed, and balanced datasets were generated; second, attribute selection algorithms were used to reduce the dimensionality of the data; third, a novel combination of data mining algorithms was employed in order to effectively distinguish migraine from other cases. We achieved a sensitivity and specificity of about 80% and 75 %, respectively, which is in contrast to a sensitivity and specificity of 15.7% and 97% when using the original imbalanced data for building classifiers.

To assess the correctness of our model, two sets of balanced training data and imbalanced test data were produced (The test set was imbalanced, reflecting the real distribution of instances). We compared the final evaluation result with the result obtained using alanced datasets and balanced test sets (as derived by cross-validation). Comparing the metrics shows that results on evaluation data were close to the results on the balanced data using cross-validation. A sensitivity above 72%, a specificity above 69%, and a ROC area above 80% were the acceptable levels for clinicians that both results met the criteria, and all the classifiers met these

three conditions. Further more, the cost-sensitive method was tested with different algorithms on the imbalanced dataset; the cost-sensitive method assigns misclassification costs to data from each class and forces the classifier to concentrate on the minority class. This method produced a good result, but the sensitivity did not meet the clinicians condition which was explain.

In the third part of our research, we extracted the most important predictors that can help to detect and prevent posterior circulation stroke. Posterior circulation transient ischemic attacks (POCS) are more difficult to diagnose than anterior circulation ischaemia. Delayed or incorrect diagnosis of POCS may have devastating consequences including death or severe disability, and early recognition of POCS may save lives.

The aim of this part of the research was to find the most important signs/symptoms related to POCS to create a model to distinguish POCS from other stroke-subtypes. In order to extract the most important signs/symptoms, we employed data mining methods. We also reviewed the signs/symptoms reported by the Heart and Stroke Foundation (HSF) of Canada and compared our finding with theirs. The HSF introduced 8 signs/symptoms, and we employed the BestFirst filtering method and extracted 22 attributes as the most important signs/symptoms from the VGH reports. We derived the top-8 attributes (based on their weights) and applied different classifiers over the HSF attributes and top-8 features.

Comparing the results gained from the both sets (HSF attributes and top8 features) showed that top8 features performed better in accuracy, sensitivity and ROC. For

evaluation, we set aside 2/3 of data for training (part of a balanced dataset) and 1/3 was used for testing (we used part of imbalanced data that was not used during training). We applied different classifiers on each set and again the top8 attributes provided better accuracy, sensitivity and ROC than HSF attributes for all classifiers.

1.2 Motivation

This thesis is focused on using NLP and data mining methods to analyze the data and build models for predicting stroke/TIA in patients. A product of the prediction models is a list of keywords weighted by their importance in the prediction quality. Having the top keywords allows building of a questionnaire-like, on-line application for triage staff to use and the triage staff can tell whether the patient needs to be hospitalized or can be discharged.

Besides, migraine can be distinguished from stroke and other non-stroke mimics via different types of tests such as CBC, Facial X-ray, CT scan, MRI, and EEG, which are all costly. Another motivation of this thesis is generating an effective model for migraine detection based on structured and unstructured data sources. This is a novel work that can be used in decision support systems and save time for GPs/family doctors and nurses.

The most important motivation of this thesis, is reducing costs for the healthcare system. The major costs to detect the stroke and migraine can be summarized as follows:

- MRI
- Specialist consultation

- Hospitalization

The cost for an MRI is approximately $500 in British Columbia, but the more comprehensive the brain scan, the more costly the MRI exam becomes. Besides, there are additional consultation fees for interpretation that are added on to the MRI scan. The neurological consultation fee can be found in [28]. The hospital fees in BC (for example at VIHA) can be found in [57].

In this research, we showed that text mining and data mining can offer a cheaper way to detect stroke, prevent costly tests, save time for GPs/specialists, and reduce cost for healthcare system.

1.3 My Contributions

My contributions are:

- I analyzed unstructured text to provide a model for stroke prediction, thus providing a first inexpensive screening for stroke/TIA detection prior to confirmation using MRI or CT scan. In addition, I introduced a new approach for data preparation and implemented a novel method for negation detection.
- I presented a detailed study using supervised machine learning to predict stroke/TIA vs. mimic based on visit descriptions and methodically compared several algorithms across a multitude of dimensions.
- I produced a set of carefully crafted rules to perform negation detection and detect negated meaning in free-text sentences.
- I introduced a approach to overcome rarity problem when dealing with highly imbalanced datasets with appropriate sensitivity and specificity result. In

addition, I introduced a novel and cheap method for detecting migraine.

- I extracted the most important signs/symptoms related to POCS to create a model to distinguish POCS from other stroke-subtypes and compared the result with the signs/symptoms reported by the Heart and Stroke Foundation of Canada (HSF). I showed that classifiers performed better with top-8 predictors (found in our research) than the HSF predictors to distinguish POCS from other stroke-subtypes.

1.4 Outline

The thesis is structured as follows:

Chapter 2 discusses challenging problems in clinical text analyzing and describes the approaches to process stroke reports.

Chapter 3 outlines the proposed approach and explains in detail all the steps of text processing and implementation of the negation detection rules. Then experiments and the methodology for them are fully described, and the evaluation of data is fully discussed.

Chapter 4 This chapter discusses the new method to analyze highly imbalanced data and outlines the new approach to detect migraine. A novel classification method was implemented by combining different data mining (classification) algorithms.

Chapter 5 Contains an overview about the POCS and extracts the most important signs/symptoms that can help to detect and prevent posterior circulation stroke. In this chapter, the extracted signs/symptoms are compared with those reported by the Heart and Stroke Foundation of Canada.

Chapter 6 Contains the conclusion of the dissertation. It also enumerates avenues of future work for further development of the concept and its applications.

Chapter 2

Challenging problems in clinical text analyzing

The clinical text analysis is completed in several steps. Each step has special challenge(s):

1. Data understanding and data preparation (overcoming the issues associated with data cleaning).
2. Identifying the important vocabularies and relationships (e.g. signs, symptoms, injured parts of body/brain, etc).
3. Identifying suitable rules for negation detection.
4. Considering limitation of data mining in healthcare (Challenges with imbalanced data and addressing the rarity problem).

2.1 Data understanding and data preparation

Clinical reports are full of noise; some particular features of noise are: misspelling (e.g. dizzyness for dizziness), deletion of character (e.g. lft for left), clinical

abbreviations (e.g. n/v for nausea/vomiting, afib for Arial Fibrillation), lack of standardization (e.g. ha or h/a for headache), and using foreign terms (e.g. Ptosis for eye droop).

Also, tokenization errors are generally related to incorrect punctuation. The vast majority of time is usually spent on data-cleaning and preparing appropriate data for analysis. According to a study, data cleaning should be described as an iterative exercise rather than a one-off procedure [33]; therefore, producing clean data is a challenge. As mentioned, one of the contribution of this thesis is to provide a model for stroke prediction and to provide inexpensive screening for stroke detection prior to confirmation using MRI or CT scan. Correct data cleaning is critically important to provide appropriate data for analysis. The way we prepared data for processing is described in detail in the next chapters.

2.2 Identifying the important vocabularies and relationships

Information extraction and identification of the relationship between features is another challenge that needs to be addressed. Sometimes, the list of important medical concepts is provided by GPs or specialist(s), sometimes we have to use computational methods to identify the important features/variables that provide specific information concerning an incident of a medical problem (e.g. stroke). In addition, providing the list of synonyms and similar phrases is important while the signs/symptoms are reported differently by different people.

Medical NLP systems are often limited to a certain medical domain and most of these systems rely on manually created lexicons for name entity recognition. A comprehensive data dictionary is required to map the words and synonyms to the medical concepts. Numerous Information Extraction (IE) systems have been implemented but in principle the used approaches can be categorized into three groups: expert designed rules, statistical methods, and manual interaction [1]. There are some cases where we do not have any specific relation types in mind but would like to discover relation types from a given corpus. A large amount of labeled training data is also required in order to learn a good named entity recognizer or relation extractor. However, both defining the structures for the information to be extracted and annotating documents require human expertise and are time consuming [4].

One of the contributions of this research was finding the most important predictors that can help to detect and prevent posterior circulation stroke. The aim of this part of the research was to find the most important signs/symptoms related to POCS to create a model to distinguish POCS from other stroke-subtypes. We analyzed the combination reports provided by stroke nurses and neurologists and we also reviewed the signs/symptoms reported by the Heart and Stroke Foundation (HSF) of Canada. We compared our finding with what HSF reported and the details of this study is discussed in Chapter 5.

2.3 Identifying appropriate rules for negation detection process

In most of the clinical texts, the signs and symptoms are negated. Identifying precise rules to determine whether a clinical concept is negated or not is challenging. Consider we define a rule as follows: if a medical concept preceded by a negation signal (e.g. no), then the concept is negated. In the following statement: he has no headache but nausea, the word no negates the concept headache, but it should not negate the word "nausea". Thus, defining appropriate rules is critical. Also, describing the domain of negation is important (should we use stop-words or punctuation to identify the domain? What if stop-words or some of the punctuation are dropped in the data-cleaning process?)

In this thesis, we implemented some rules, segmented the paragraphs to sentences, evaluated the status of each concept based on implemented rules, and finally accumulated the status of each concept and recorded the results. NegEx is one of the common methods to identify negation signals and negate all the medical terms within a window of five words of the negation signal [34]. In the context of our study, we found that clinicians often negate more than five words; therefore, several rules were implemented to determine whether a concept is positive or negated. Also, compound sentences are usually composed of two or more independent clauses that are joined by conjunctions (e.g, "but") which alter the context of the second clause [10].

One of the contribution of this thesis was implementing accurate negation detection rules and assigning appropriate value to each medical term to prepare suitable data

for applying data mining algorithms. This process is described in detail in chapter 3.

2.4 Considering limitation of data mining in healthcare (Challenges with imbalanced data and addressing the rarity problem)

The results of data mining (DM) are directly affected by the quantity and quality of the data [15]. There are about 56 classification methods in WEKA that can be used to classify medical data and create models. DM methods can be used to detect the occurrence of diseases and benefit healthcare providers to make effective decisions to enhance the patient health [54]. DM methods can be used in different healthcare domains, but no single DM method gives consistent results for all types of healthcare data [54]. Some of the limitations associated with data-mining in healthcare system are as follows:

- Data mining accuracy is not high enough because of low quality of patient data.
- Data is usually corrupted and sometimes missing [15] [14],
- The heterogeneity of data complicates the use of data mining techniques[15].
- Medical databases may contain data that is redundant, incomplete or imprecise which can affect the results of the data mining process [15].
- Different data-mining algorithms (or the mix of different methods) can be examined to create a useful model with appropriate sensitivity and specificity. Working with imbalanced data sets affect different measures such as accuracy.

One of the challenges in this study was to analyze the imbalanced data and overcome the rarity problem. This challenge is discussed and addressed in Chapter 4.

Chapter 3

Introducing a Novel Negation Detection Method and Using Clinical Reports to Detect Stroke.

One of the objectives of this thesis is to build an effective model for fast detection of stroke/TIA or mimic at the triage stage via the analysis of past clinical reports. The study we report in this chapter is part of a large-scale rapid stroke assessment project carried out at Victoria General Hospital (VGH) in British Columbia, Canada. The medical charts from 5,658 patients collected between 2008 and 2013 at the SRAU were analyzed and the most important signs and symptoms were extracted from these data.

The data were generated in the SRAU and were unrelated to the Emergency Department (ED), even if the patient was referred to the unit from the ED. The fields were entered by keyboard and typed directly into the Stroke Guidance System (SGS) while consulting with a patient. Data elements included the patient medical

history, the time when the stroke signs and symptoms first appeared, changes in symptoms over time, recent injuries, and information from the patients or bystanders, such as time of appearance of each problem or the last time the patient was without symptoms. Also, history of stroke, Transient Ischemic Attack (TIA), Diabetes, hypertension and other clinical information were used in the analysis and the most important symptoms were extracted from these data.

3.1 Introduction

Statistics Canada reported Stroke as the third leading cause of death in Canada in 2012; six percent of all deaths were due to stroke, with women being the major victims [29]. Timely detection of stroke can contribute significantly in preventing long-term patient disability and can have a great impact in public health, reducing care costs and preventing expensive and potentially harmful neuro-imaging tests.

As mentioned above, one of the objectives of this thesis is building an effective model for fast detection of stroke/TIA or mimic at the triage stage via the analysis of past clinical reports. A TIA, or a mini-stroke, starts just like a stroke but then resolves leaving no noticeable symptoms or deficits [45]. For almost all TIAs, the symptoms go away within an hour and there is no way to tell whether it will be just a passing problem or persist and lead to death or disability [45]; therefore, all the signs and symptoms gathered by clinicians are valuable information for diagnosis.

One of the challenges in this work was data pre-processing and implementing appropriate negation detection rules. With negation detection rules, we determined whether each sign or symptom is present, unmentioned, or declared absent and

generated structured (codified) data for the data mining process.

We used DM algorithms to analyze the data and build models for predicting stroke/TIA in patients. A product of the prediction models was a list of keywords weighted by their importance in the prediction quality (as captured by sensitivity, specificity, etc). Having the top keywords allowed building of a questionnaire-like, online, application for triage staff to use and the triage staff can tell whether the patient needs to be hospitalized or can be discharged.

3.2 Related Work

As previously explained, one of the most important contributions of this thesis is providing a model to predict stroke and distinguishing stroke from mimics. There are several clinical scoring systems that are widely used in medical industry to assess the risk of medical outcomes. Some of the popular scoring systems and predictors that are used in hospitals are:

- "Acute Physiology and Chronic Health Evaluation II" or Apache II is a physiologically based classification system for measuring severity of illness in groups of critically ill patients [35]. This scoring system is reliable in classifying ICU admissions and used at ICUs. It scores from 0 to 71 and the higher scores correspond to more severe disease. The first APACHE model was introduced by Knaus et al. in 1981 [35].

- "Simplified Acute Physiology Score" or SAPS II is another ICU scoring system to measure the severity of disease. It scores between 0 and 163 to

predict mortality between 0% and 100%. This model was first presented by Le Gall et al. in 1993 [38].

- "Thrombolysis In Myocardial Infarction" or TIMI provides a basis for therapeutic decision making. Patients presenting with an acute coronary syndrome without ST-segment elevation are diagnosed as having unstable angina/nonST elevation myocardial infarction (MI) (UA/NSTEMI); TIMI identifies patients with different responses to treatments for UA/NSTEMI and categorizes a patient's risk of death and ischemic events. This score was provided by Antman et al. in 2000 [7].

- CHADS2 can quantify risk of stroke for patients with atrial fibrilation (AF) and may aid in selection of antithrombotic therapy [24]. CHADS2 scores between 0 and 6 and the high score corresponds to a greater risk of stroke. C stands for "Congestive heart failure", H for Hypertension, A for Age, D for "Diabetes Mellitus", and S for "prior Stroke or TIA/thromboembolism". CHADS2 was introduced by Gage et al. in 2001 [24].

- CURB criteria has been assessed to be used for predicting mortality in community-acquired pneumonia (CAP) [40]. A simple six point score, based on "Confusion of new onset", "Blood Urea nitrogen", "Respiratory rate", "Blood pressure" and "Age 65 or older" were used to stratify patients with CAP into different management groups [40].

- Ranson criteria is used for predicting the severity of acute pancreatitis and introduced in 1974 [47].

The above mentioned scoring/predicting systems are not optimized for accuracy; these scoring models involve very few calculations, allowing for quick prediction

during a doctor's visit. Our model is created based on text analysis and data-mining algorithms which learns from data. In this study, different metrics are calculated and described in detail.

A variety of medical language processing systems were developed to extract patient data from specific medical reports. Some prominent examples are SymText to identify pneumonia related concepts on chest x-ray [19], Regenstrief EXtraction Tool (REX) to extract pancreatic cyst patient data from medical text files [5], and MedLee [21] which was initially used to process radiological reports of the chest and then extended to other domains (e.g. mammography reports) [22].

There are also some studies that consider stroke prediction in particular by studying DNA and the number of single-gene disorders (cf. [48]). Another study by Amini et. al. focused on prediction and control of stroke, however, it did not involve text mining but rather used a predefined list of factors to predict stroke [6].

As it has been shown by Elkins et al. [17] and Hripcsak et al. [31], NLP is useful for rapidly performing complex data acquisition and can be as accurate as expert human coders. While [17] is a project in the stroke domain, it only deals with the neuroradiology reports of brain images, which are of a nature different from the medical description texts we consider in this study.

Another contribution in this chapter is implementing a rule-based negation detection method based on Chapman's negation detection algorithm (NegEx).

One of the ways to deal with negation is parsing the whole text. POS tagging is usually used for this purpose. Stanford NLP parser [3] is one of the accurate tools for POS tagging. In our study, we did not need to parse the whole text, because we had the list of pre-defined medical words that required to be assessed whether they were negated or not; therefore, POS tagging was not used in our study.

NegExpander is a regular expression-based approach that identifies conjunctive phrases [1] and defines negation boundaries [9]. NegExpander looks for negation phrases inside the conjuctive phrases, and if a negative phrase is found, all the medical phrases inside of the conjunctive phrases are negated. The weakness of this method is that some of the important negation indicators such as "deny" and "decline" are ignored because these verbs are not part of any conjunctive phrases.

Another way to detect the negation is using the online tools such as IBM alchemy [2]; this online program analyzes texts and evaluates the sentiment of the document. This online tool requires uploading the data which is against VIHA's confidential policy. Besides, this tool is good to evaluate document sentiment; for example, if you write: "patient is good.", it gives you the following result: Sentiment= positive, Score=0.246. Stroke nurses and neurologists usually mention different signs/symptoms in their reports. In one sentence, some signs/symptoms are presented; some are negated, and some are not mentioned at all; therefore, the IBM product was useless in our study.

NegEx is a simple negation detection algorithm implemented by Chapman et al.

[1] Conjunctive phrases are the noun phrases connected with conjunctions such as "and".

[63]. We implemented a pipline application consists of several components, and we implemented a rule-based negation detection method based on Negex which was not dependant on the domain.

In this study, we defined three sets of words: clinical words, negative words and stop words meaning “But”. In brief, the negative words included negative markers (e.g. not), negative quantifiers (e.g. no, nothing), negative adverbs (e.g. never), negative verbs (e.g. deny) and prepositions (e.g. without). The ”But” words were “but”, ”although” and “however”. We had list of 150 clinical words (defined by neurologists and stroke nurces) to be assessed for negation. We labeled all the clinical words with K (e.g. headache=K1), marked negative words with N (e.g. no=N1), and “But” words with B (e.g. But= B1). The labels were kept, and the rest of the words were dropped; then, the sentences were transformed to NBK format. Finally, we applied the rules that we defined in order to assess whether the medical words were negated or not. The details of these steps will be discussed in the following sections.

3.3 Data Preprocessing

A suite of regular expression scripts was developed to extract medical concepts and negative words. The extraction process is shown in Figure 3.1. As mentioned, the work is classified into several phases and each phase is described in detail in the following.

3.3.1 Data cleaning and splitting paragraphs into sentences

Patient medical history was stored in plain text in the system and records were distinguished from each other by the value of their primary key. In the first phase,

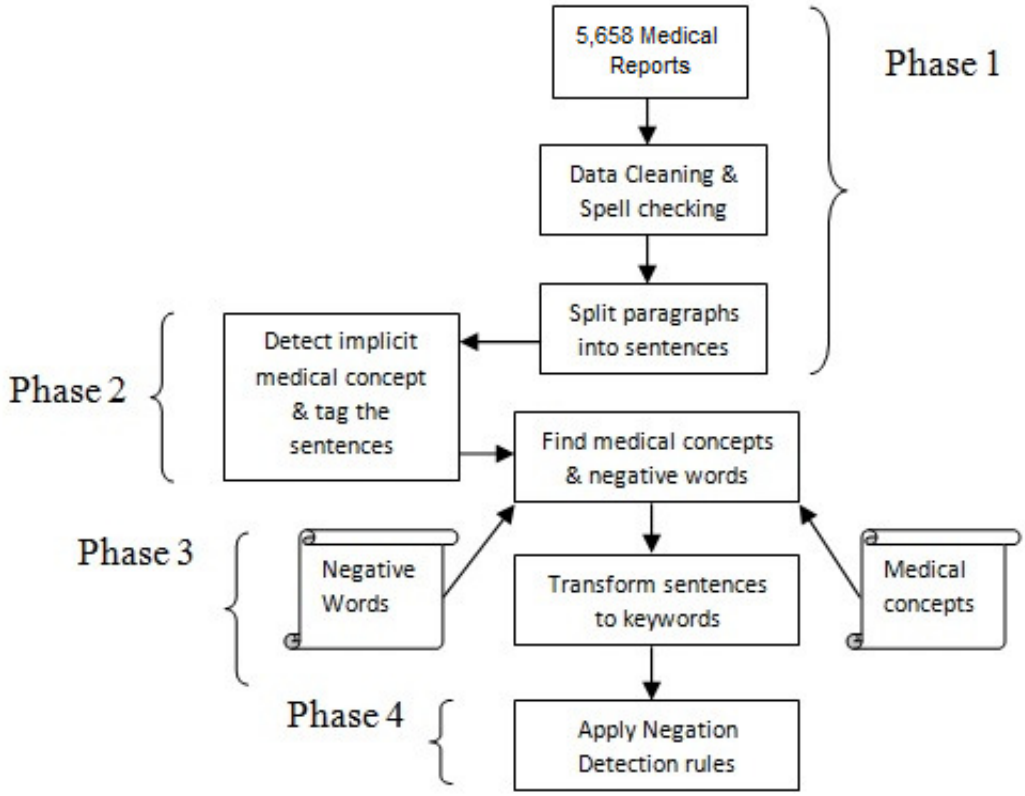


Figure 3.1: The process of medical concept extraction.

the plain texts (paragraphs) were spell-checked and a set of scripts were implemented to correct typographical errors and mistakes. The abbreviations were expanded to original form (e.g. N/V was changed to Nausea / Vomiting, SZ was changed to Seizure, and etc) and finally, a procedure was implemented to break paragraphs into sentences.

Tokenization is an important component of language processing and there is no widely accepted method to tokenize English texts, including biomedical texts [11]. If the text contains well formed sentences, then it may be possible to use existing software (e.g. Punkt) to segment text into sentences with few errors [11]. In this study, regular expression commands were utilized to evaluate each period to determine whether it is the sentence terminator or not. Periods followed by title prefix (e.g. Dr., Mr., Mrs., Miss.), periods used as a decimal point (e.g. 1.2 mg ASA), a row of three periods (...) that means "and so forth", and periods followed by date (e.g. September.21.2009) were all replaced by white space. Finally, a procedure was implemented to break the paragraphs into sentences using the remaining sentence separators (periods). We randomly selected 50 records consisting of 400 sentences and evaluated them manually to make sure the segmentation was done properly. The evaluation process is described in Phase 6.

3.3.2 Detection of implicit medical concepts and tag sentences

In this project, signs and symptoms were target of the feature extraction process. The stroke terms (variables) used in this study were selected based on an exploratory data analysis of historical stroke data by attending neurologists and stroke clinicians at VGH. Overall, 126 signs and symptoms were defined as stroke

signs and symptoms (terms) that were used as attributes to be extracted from the raw data set. Finding the top terms by data mining was the goal of our study to detect stroke/TIA or mimic and help triage decide for the next step of treatment or discharge the patient.

The negative words utilized in this project were partially borrowed from negation words/phrases used in NegEx (e.g. no, not, without, denies, etc) [63] and the negative words in the Wiktionary website [66].

Once the list of negative words and medical concepts was compiled, a unique id was assigned to each term. For example, letter "n" was assigned to negative words (e.g. n1="no", n2="not", etc.), letter "k" to medical symptom terms (e.g. k1="HTC", k2="numbness", etc), letter b to conjunctions meaning BUT (e.g. b1="but", b2= "although", b3="however") and letter "s" to the laterality (e.g. s1=left, s2=right, etc).

We implemented a program to find and tag phrases or sentences that explicitly referred to a predefined term. We defined a key for each term and the tagging application finds the appropriate terms or phrases for a given symptom and tags the sentence with the key assigned to that term. For example, diplopia means double vision and "k48" was assigned to this concept as a key or concept ID. Whenever the tagger application encounters the word diplopia or any phrase or keyword that means diplopia (e.g. "double vision", "everything went double"), it tags that sentence with k48.

3.3.3 Transforming sentences to sequences of negation signals and medical symptom terms

In this phase, the concepts were mapped to unique ids and each sentence was translated to a string containing n, k, b and s. For example, if s2= right, k15= hand, and k42=numb, the following sentence:

"Saturday afternoon she had just finished her lunch when her right hand went numb"

was translated to

"s2 k15 k42"

Another example is: "No leg weakness / numbness." It was translated to "n1 k26 k40 k42"

Figure 3.2 shows how sentences are transformed to KNB codes.

SENT_NUM	SENTENCE
1	saturday afternoon she had just finished her lunch when her right hand went numb.
2	she had right facial droop.
3	no leg weakness / numbness.
4	headache all night after this.

SENT_NUM	KNB
1	s2 k15 k42
2	s2 k73 k41
3	n1 k26 k40 k42
4	k11

Figure 3.2: Mapping sentences to KNB codes.

3.3.4 Applying negation detection rules

In this phase, different negation detection methods were reviewed and the list of the negative words were borrowed from NegEx. NegEx identifies negation signals and negates all the medical terms within a window of five words of the negation signal [34]. In the context of our study, we found that clinicians often negate more than five words; therefore, several rules were implemented to determine whether a concept is positive or negated. Also, compound sentences are usually composed of

two or more independent clauses that are joined by conjunctions (e.g. “but”) which alter the context of the second clause [10]. The rules that we defined, assigned the appropriate value to each medical term and suitable data was prepared for applying data mining algorithms.

The medical concepts were extracted and inserted into a table of concept vectors. The concepts were divided into two categories: *single concepts* that included the body parts (e.g. face, eye, arm, leg, etc.) and the *compound concepts* that described the symptoms, signs (e.g. loss of consciousness, shortness of breath, etc) and problems with specific parts of body (e.g. left eye droop, right arm numbness, etc). -1 was assigned when the concept was negated, +1 when the concept was present, and 0 if the concept was absent or not mentioned in the sentence. As mentioned, several rules were defined to assign the correct value to the medical concepts. Rules were defined based on the order of k, n, and b in each sentence. The rules are described in the following.

- Rule 1: The value of concept K is +1 if K is not preceded by a negation signal N or if there is no N in the sentence. In the following sentence: “he has headache”, the word “headache” is a medical concept which is translated to “k11” and its value is +1 because it is not negated.

- Rule 2: If a negative word, N, comes first, it negates the subsequent medical concept. In the following sentence: “he has no headache”, headache is negated by “no”, so the sentence is translated to “n1 k11” where “no” negates concept “headache”. Thus, for this sentence, the value of headache is -1.

- Rule 3: If there is a conjunction indicating an exception B, such as “but”, “although”, “however”, and the order of words in the sentence is NKBK, the value of the last concept is +1. For example: “The patient reports no diplopia but feels his left eye a bit droopy” is translated to “n1 k48 b1 s1 k31 k41” Here, “diplopia” or k48 is negated and its value is -1, but the value of the concept “left-eye-droop” is +1.

- Rule 4: If an exception B is indicated in the sentence and the order of words is KBNK, then the value of the last concept is -1. In the following sentence, “The patient has headache, but no dizziness”, the value of headache is +1, but dizziness is negated, so the value of concept dizziness is -1.

At the end of the process, the sentences were combined into paragraphs and the results were summed up for each record. A given concept might appear more than once in the same paragraph. Unlike Negex, if a concept was positive at least once, then all occurrences of it in that paragraph were considered to be positive.

3.4 Data Mining process

WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software and it contains a collection of algorithms and visualization tools for data analysis and predictive modeling [56]. WEKA was employed in this study to apply different data mining algorithms on the stroke dataset.

To enhance the data set for analysis, fourteen attributes were added from SGS which were provided by the unit staff after the patient was examined. These data carried other patient information, such as age, gender, blood pressure, ABCD score[2]

[2]The ABCD score alone did not give us acceptable levels of sensitivity and specificity.

[64], smoking status, diabetes status, and so on. By the end of this process, we had a data set with 140 attributes (variables) and 5,658 records. Recall that each sign and symptom term was represented by an attribute with values -1 (for negated), 0 (for absent or not mentioned), +1 (for present).

The class values we considered for classification were "stroke/TIA" and "mimic". In other words, we perform binary classification of "stroke/TIA" vs. "mimic". Differentiating between full stroke and TIA was not important for the clinicians in the project. This is because at triage stage, once someone is identified to have potentially suffered a stroke or TIA, the distinction is not crucial in the further process, i.e. in both the stroke and TIA cases, the patient will be immediately admitted to the hospital and proper tests will be done.

3.5 Evaluation

To evaluate the correctness of term mapping and the quality of negation detection rules, we randomly selected 50 records from the set of 5658 reports and a human expert manually segmented and translated the sentences into structured terms. The expert detected the negation manually and assigned an appropriate value to each medical term. The results gained from automated negation rules were compared against the results provided manually for the selected records. The expert used Kappa statistics and the level of agreement between the negation detection rules and human evaluation was 0.9[3].

To better understand the quality of classification methods, the data were divided

[3]If the estimate is 0.8 or above, there is excellent agreement between the algorithm and the human assessment, the score between 0.6 to 0.8 is considered good agreement [27]

into two sets: a training set (3520 records) and a test set (2138 records). The training set contained the data from four years (2008 to 2011) and the test set contained the data from two years (2012 and 2013). In total, the number of cases who experienced stroke/TIA were 3275 and the number of negative cases (mimic) was 2383. Ten fold cross validation was also run; the results obtained were similar to those on the aforementioned test set.

3.6 Experimental Result

The results on the full set of 5658 records with ten-fold cross validation showed that a parameter-tuned SVM with RBFKernel provides the highest accuracy (75.5 %) followed by logistic regression (73.67 %), NaiveBayes (72.48 %) and J48-Decision Tree (70.91 %). Normalizing the value of some of the continuous attributes, such as age and blood pressure (systolic and diastolic), did not provide significant change in the results.

Accuracy, however, is not considered most important in medical studies. Therefore, in the following we focus on recall (sensitivity), specificity, precision, F-measure and ROC area. The results shown here were obtained using the test set of 2138 records as described earlier.

To establish a baseline for the algorithms, we first considered raw text data. SVM, logistic regression, and neural network provided the highest recall (sensitivity), about 80 %, among the rest of the methods. The detailed results are presented in Table 3.1.

Next the analysis was performed on codified data obtained as described in detail in the previous section. Table 3.2 shows the results of different classification methods on codified data. Again, SVM and logistic regression provided the highest recall (sensitivity) of over 83 %. The results show that training classifiers on codified data significantly outperformed training them on raw data. The main added benefit of using codified data is of course the ranked list of terms we obtain as a side effect of the data mining algorithms. Having a ranked list of terms allows building a user-friendly software application for use in the triage phase.

Besides recall (sensitivity) in Figure 3.3, specificity, precision, F-measure and ROC area were also computed for both approaches and the results are shown in separate figures. In each figure, each classifier is represented by two bars, the first showing the performance achieved using raw data and the second showing the performance achieved using codified data.

Figure 3.4 presents the specificity gained from different methods using raw text and codified data. Specificity is the number of true negatives (TN) divided by the total number of negatives (N). A sensitivity level of 79 % or greater and a specificity level of 60 % or greater were deemed suitable by the clinicians in this project. SVM provided 84 % sensitivity and 63.3 % specificity, and logistic regression provided 83.2 % sensitivity and 63.7 % specificity using the codified data.

F-measure and precision are depicted in Figure 3.5 and 3.6 respectively. SVM and logistic regression provided the highest F-measure (78.1 and 77.8 % respectively) using codified data. The differences in precision between the two approaches did not turn out to be significant. The values provided by different algorithms varied

between 58.6 % and 79.2 % for raw text classification and 64.4 % to 76.4 % with codified approach.

Receiver Operating Characteristic (ROC) curves are an important outcome measure as they display the trade-off between sensitivity and specificity [20]. The ROC curve results are shown for both approaches in Figure 3.7.

Remark. We would expect that working with codified data will be better than working with raw text. However, in reality the process of codifying data from raw text may also be a source of error. It is therefore to be expected that the observed improvements are not necessarily consistent across all measures. Nonetheless, we argue that still, working with codified data in this study, gives better performance overall even in those cases when some measures score better for raw text.

Specifically, let us focus, for example, on the Naive Bayes (NB) classifier, which shows the biggest positive difference in specificity for raw text compared to the codified approach. This is indeed true for the default classification threshold of 0.5. [Recall that Naive Bayes produces in fact a probability score, which is compared to a threshold in order to produce a binary classification.] However, NB with raw text scores quite poorly with respect to recall (sensitivity) compared to the codified approach.

For different classification thresholds, both the recall (sensitivity) and specificity will vary. We can better see the sensitivity/specificity behavior by building an ROC curve for each approach (raw text and codification). An ROC curve plots sensitivity vs (1-specificity) and each point in it corresponds to a different classification threshold. In order to not rely on visual inspection of different points in the ROC

curves, we use the well-known measure of the area under the curve (AUC). The bigger the AUC, the better in general the classifier. With respect to AUC (see Figure 3.7), NB performs better when using codification than raw text. In fact, we see that all the classifiers we consider perform better with respect to AUC when using codification.

A similar discussion can be made for the recall/precision combination. Naive Bayes scores better with respect to precision when using raw text. However, it is worst in terms of recall. Typically, we combine recall and precision into their harmonic mean, which is the F-measure. In terms of the latter, NB does worst using raw text than codification.

Finally, we report here few of the most important concepts discovered by our data mining process. The important warning signs and symptoms to detect stroke/TIA were face droop, visual loss, diplopia, language disturbance, speech disturbance, swallow difficulty, drunk, drag and etc. In addition, some of the most important concepts to detect mimic were headache, seizure, migraine, anxiety, fatigue, amnesia, photophobia, tremor, stress and etc.

3.7 Conclusion

Natural Language Processing (NLP) methods were used to identify and extract information from medical charts of patients collected between 2008 and 2013 at SRAU of Victoria General Hospital (VGH). The unstructured texts narratives were transformed to codified data in a computable format and data mining methods were utilized to build models for stroke/TIA prediction. Our clinical NLP-based system

consists of several components to extract and analyze data.

Various algorithms were utilized on codified data and compared against baselines on raw data. Evaluation metrics were computed for both approaches and showed that the codification approach outperformed the approach using raw data. The recall (sensitivity) provided by SVM and logistic regression showed that these two classifiers can provide reliable models to predict stroke/TIA based on patients' signs and symptoms instead of immediately using costly tests, such as MRI or CT scan.

The list of symptoms that play the most important role in stroke detection can be identified via the machine learning process. This list can be used in stroke assessment forms to help stroke nurses to decide on the next step of treatment in a more timely fashion. [4]

In this study, we used the history of patient illness and chief complaint information of the patients who experienced stroke/TIA or discharged with the conditions that mimic the symptoms of stroke to implement a model for stroke prediction in new patients with the same symptoms. In this analysis, we considered two class values "stroke/TIA" and "mimic", but future analysis will contain more possible values to determine different types of stroke (e.g. PACS, POCS, TACS, LACS, and etc) and different types of mimics such as migraine, TGA, BPV, and etc.

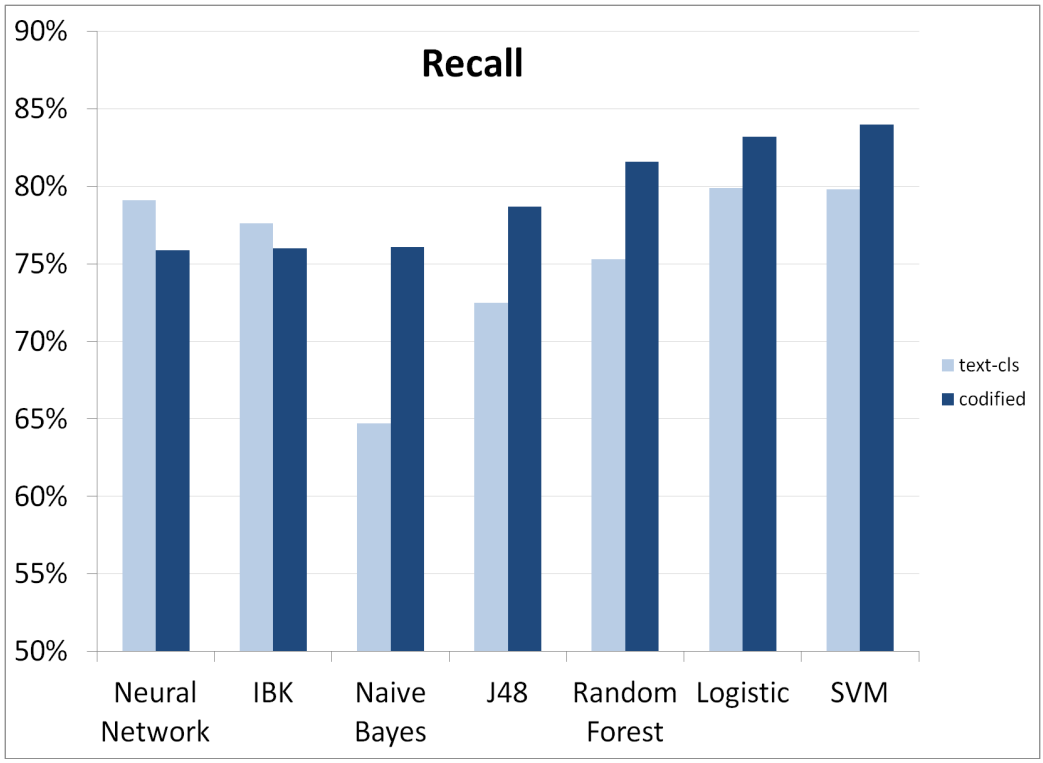
[4]Due to IP restrictions, we cannot provide here the list of important terms and the weights we derived for them. However, the interested readers can contact Dr Andrew Penn on how to obtain this information. Also, the front-end application is not part of this study. Again, details on the front-end application can be obtained from Dr Andrew Penn at andrew.penn@viha.ca.

Table 3.1: Results using raw text data

	Logistic	Naive Bayes	SVM	Neural Network	IBK	J48	Random Forest
Recall (Sensitivity)	**79.9%**	**64.7%**	**79.8%**	**79.1%**	**77.6%**	**72.5%**	**75.3%**
Specificity	**57.2%**	**80.0%**	**61.9%**	**64.8%**	**35.3%**	**52.5%**	**54.5%**
Precision	**68.7%**	**79.2%**	**71.2%**	**72.6%**	**58.6%**	**64.3%**	**66.1%**
F-Measure	**73.9%**	**71.2%**	**75.2%**	**75.7%**	**66.8%**	**68.1%**	**70.4%**
ROC	**76.3%**	**78.4%**	**70.8%**	**79.0%**	**58.0%**	**62.1%**	**70.7%**

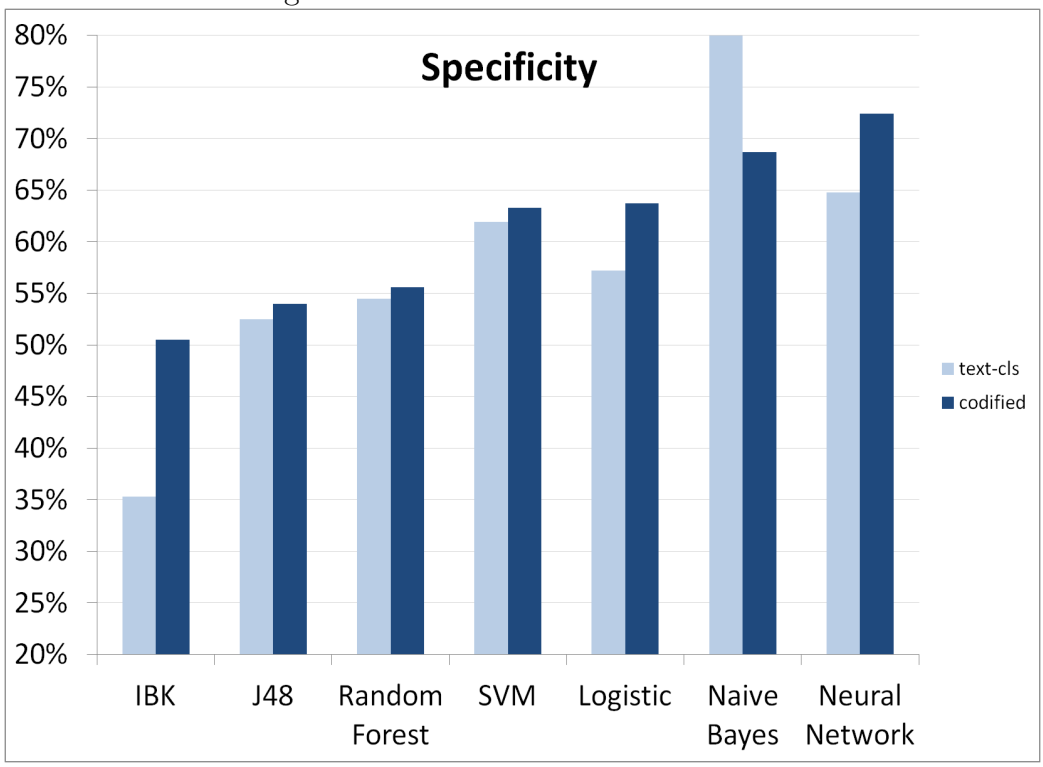
Table 3.2: Results using codified data

	Logistic	Naive Bayes	SVM	Neural Network	IBK	J48	Random Forest
Recall (Sensitivity)	**83.2%**	**76.1%**	**84.0%**	**75.9%**	**76.0%**	**78.7%**	**81.6%**
Specificity	**63.7%**	**68.7%**	**63.3%**	**72.4%**	**50.5%**	**54.0%**	**55.6%**
Precision	**73.0%**	**74.1%**	**73.0%**	**76.4%**	**64.4%**	**66.8%**	**68.4%**
F-Measure	**77.8%**	**75.1%**	**78.1%**	**76.1%**	**69.7%**	**72.3%**	**74.4%**
ROC	**82.1%**	**79.9%**	**73.7%**	**80.3%**	**63.2%**	**67.5%**	**76.1%**



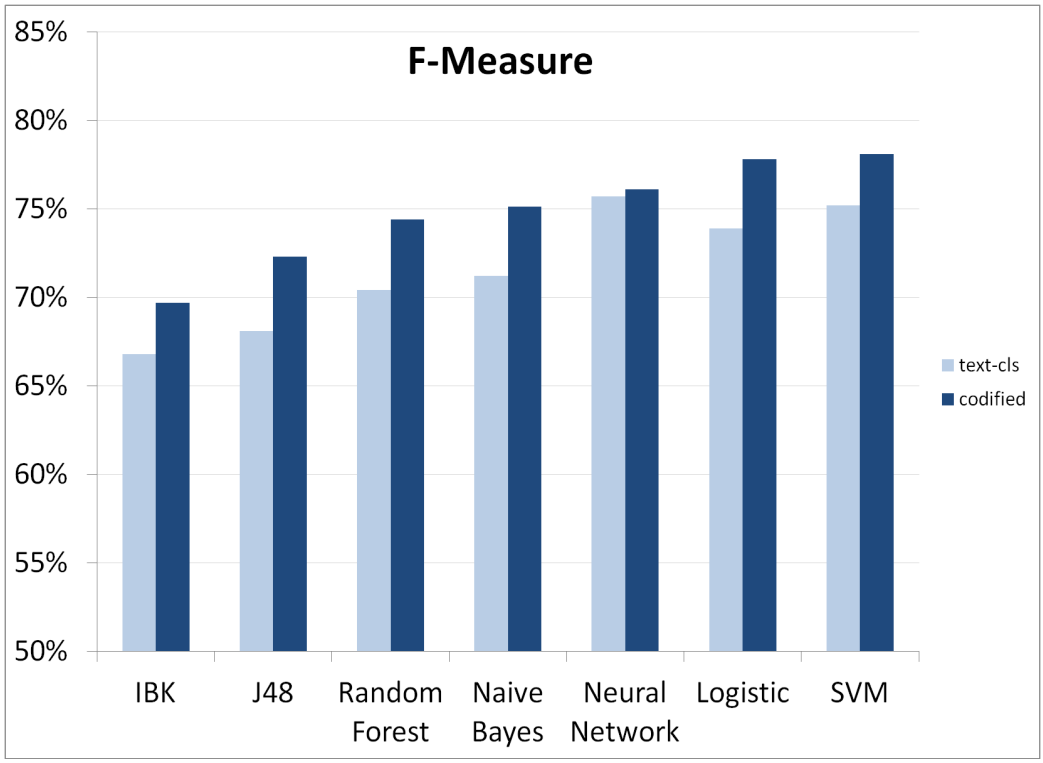
0.5

Figure 3.3: Recall for different methods.



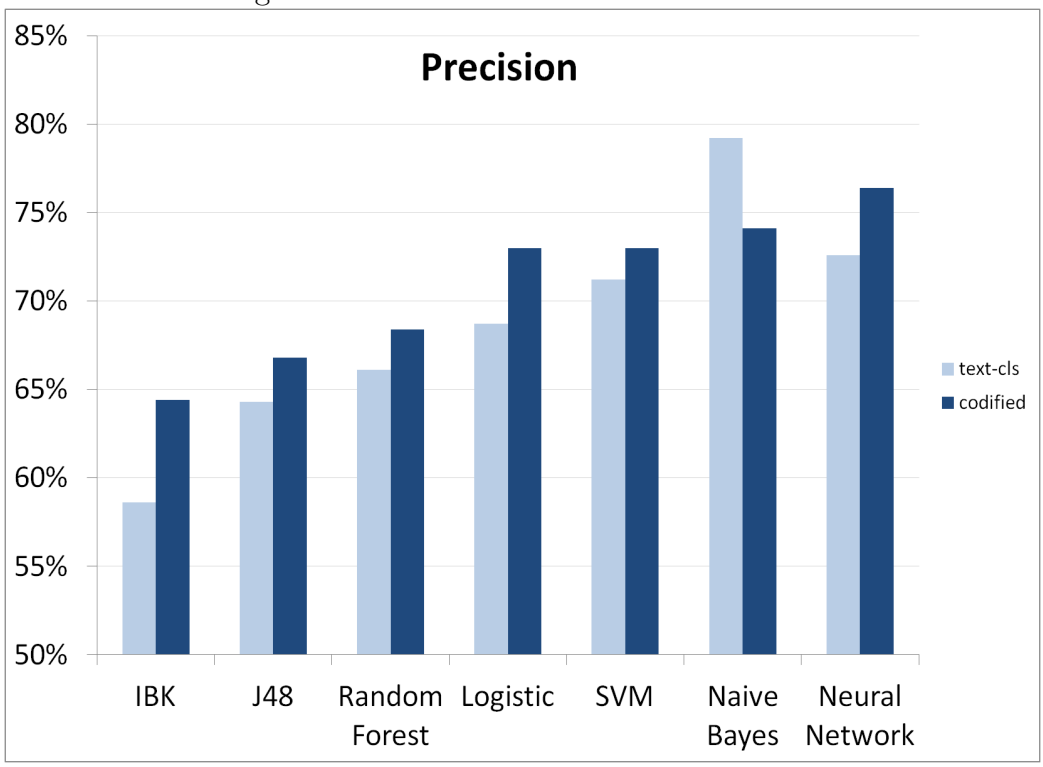
0.5

Figure 3.4: Specificity for different methods.



0.5

Figure 3.5: F-measure for different methods.



0.5

Figure 3.6: Precision for different methods.

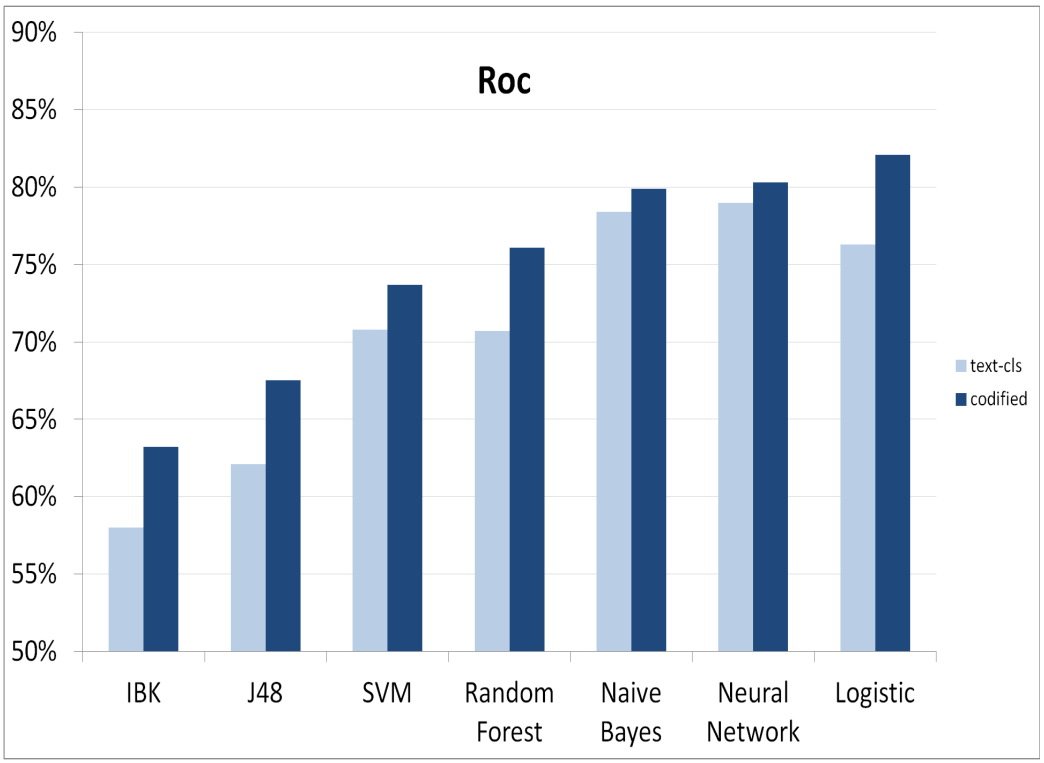


Figure 3.7: ROC area plotted for different methods.

Chapter 4

Introducing an Approach to Overcome the Rarity Problem

Distinguishing migraine from stroke is a challenge due to many common signs and symptoms. It is important to consider the cost of hospitalization and the time spent by neurologists and stroke nurses to visit, diagnose, and assign appropriate care to the patients; therefore, devising new ways to distinguish stroke, migraine and other types of mimic can help in saving time and expenses, and improve decision making.

Some of the contributions in this chapter are as follows:

- A novel classification method was implemented by merging different structured and unstructured data-sources. We extracted knowledge from the combination of high-level reports (provided by neurologists) and reports provided by stroke nurses.
- Introducing an approach to overcome the rarity problem in analysis of highly imbalanced data-sets.

- Introducing a low-cost solution to detect migration and distinguish it from stroke and other types of mimic

In previous chapter, we analyzed the stroke nurses' reports to build an effective model for fast detection of stroke. In this chapter, we analyzed the combination of clinical reports provided by neurologist and stroke nurses and a novel combination of data mining algorithms is employed in order to effectively distinguish migraine from other cases.
We utilized text and data mining methods to extract the most important predictors from clinical reports in order to establish a migraine detection model and distinguish migraine patients from stroke or other types of mimic (non-stroke) cases. The available data for this study was a heterogeneous mix of free-text fields, such as triage main-complaints and specialist final-impressions, as well as numeric data about patients, such as age, blood-pressure, and so on. After a careful combination of these sources, we obtained a highly imbalanced dataset where the migraine cases were only about 6% of the dataset.

Our main challenge in this section is copping with imbalance data; the migraine cases are a minority of the overall cases. In order to alleviate this rarity problem, we propose a randomization procedure which is able to improve the classifiers quality.

4.1 Introduction

About one in ten people suffers from migraine in North America [41]. Migraine and stroke share many common signs and symptoms; for example, headache is similar in both stroke and migraine. Also, migraine aura can mimic transient ischaemic attacks (TIAs) [51]. Migraine can be diagnosed via different types of tests. Some of

the common ones are blood test (e.g. Complete Blood Count (CBC)), X-ray (e.g. Facial X-ray), Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Electroencephalogram (EEG) [61]. Unfortunately, all these tests are costly. (For more information, please see Motivation section in Chapter 1.)

In addition, according to some studies, the risk of stroke is increased in people with migraine [18] [55]. Therefore, migraine detection can help clinicians offer appropriate treatment to migranous patients, prevent costly tests, and if required, monitor the patients' condition to prevent the problem from turning serious (e.g. become stroke).

In the chapter 3, we analyzed the reports provided by stroke nurses and generated a model to distinguish between stroke and general mimic cases (e.g. migraine, seizure, benign positional vertigo, transient global amnesia, etc). In that section the dataset was balanced, i.e. the number of stroke cases was almost equal to the number of mimic cases.

In the current chapter, we focus on distinguishing migraine from stroke or other mimic (non-stroke) cases. However, when considering migraine as the "positive" class to predict, we are faced with a severe class imbalance problem. This is because the migraine cases make up only about 6% of the total cases. In this chapter, we use not only stroke nurses' reports (as in previous chapter), but also neurologists' reports.

More specifically, we used four heterogeneous data sources: (A) the history of the

patient illness, (B) the chief-complaint, (C) Stroke Guidance System (SGS)[1] data, and (D) the impression neurologists' reports. The data were collected between 2008 and 2013 at Victoria General Hospital (BC, Canada).

There were several challenges we addressed in this work. The first challenges came from the fact that the data were not clean or well structured. We used a variety of scripts and performed a careful data integration activity. The main challenge, however, was tackling the severe class imbalance with respect to the migraine cases. Using the imbalanced data to create machine learning classifiers gave us low sensitivity; the best value we recorded was only 15.7%.

In order to overcome the imbalance challenge, we employed careful sampling to create balanced datasets for training our models. For evaluation, we used a part of the original imbalanced data, so that the evaluation could be performed on realistic data. Naturally, this part of the data was not used during training.

After combining several classification processes and performing the aforementioned sampling, our results were significantly better than the baseline approach of using the original data. Namely, we were able to accurately predict migraine with a sensitivity of about 80% (which is in stark contrast with 15.7% we obtained previously) while maintaining a specificity of 75.5%.

This chapter is structured as follows. Section II discusses related work. Section III and IV outlines the data description and proposed approach respectively; it discusses the data mining process, how data was sampled, different classification methods,

[1]SGS was produced by Synapse Publishing in year 2001 and version 2.0.6 is currently used by VGH.

and feature selection for building the final model. Section V discusses the evaluation process. Section VI concludes this chapter and outlines future work directions.

4.2 Related Work

There are three main approaches to address the class imbalance problem: the sampling methods, data mining algorithms and feature selection methods [60].

Oversampling and undersampling the minority class are techniques that provides a dataset with balanced distribution and can improve the result over the original dataset, but each approach has also drawbacks [60] [62]. Undersampling can cause elimination of valuable samples and oversampling (e.g. duplicate existing sample) can cause a classifier to overfit data [60].

Cost-sensitive learning which assigns misclassification costs to data from each class, forces the classifier to concentrate on the minority classes [50]. AdaBoost changes the underlying data distribution and classifies in the re-weighted data space iteratively; incorporating cost-sensitivite learning to Adaboost focuses on the minority data set and a satisfactory result can be obtained [50].

Because the class imbalance problem is commonly accompanied by the issue of high dimensionality of the data set, applying feature selection techniques is a necessary course of action [50]. Filter, wrapper and embeded methods are common feature selection approaches [65]. Filter methods use the data to rank the variables and choose the top ranked attributes before the classification process [16, 8]. In wrapper approaches, the selection of variables is performed by the classifier [16, 8]. Embeded

methods are similar to wrappers and rely on a classifier to evaluate candidate subsets [16, 44]. Filter methods suppress the least interesting variables and it is robust to overfitting whereas wrapper methods evaluate subsets of variables and detect the possible interactions between them [65].The goal of these approaches is to find irrelevant and redundant features to remove them from the feature domain [65]. WEKA, the popular data mining tool [56] has implemented the mentioned methods and is used in this study.

Alternative choice for a tool that implements the described methods is Microsoft Azure Machine Learning Studio [43]. This tool requires data to be uploaded to Microsoft server which is against the hospitals policy. Another alternative would be to use special software environment such as R language [52] and use machine learning libraries that can extend it.

Our decision to adopt WEKA in our study was based on the fact that the tool is mature and contains highly tunable implementations of many popular algorithms. All algorithms in WEKA are open source and the tool provides a command line interface in addition to its GUI. If an organization would prefer to use the R environment, the WEKA algorithms can also be used from R by loading the RWeka library [30].

In this study, we introduce an approach that utilizes sampling and feature selection methods in a different way to overcome the rarity problem. We also utilized a cost-sensitive approach [39] and compared the result with the sampling method.

Detecting migraine via text mining and data mining is a novel work. In particular, electroencephalogram (EEG) is often prescribed as a first-line study in migraine patients [58]. One of the studies implemented a classification system to classify the patients into three different states of migraine (i.e. inter-ictal, pre-ictal, ictal)[36]. In this study, the EEG information was utilized and only migrainous patients in different migraine states participated.

In another study, the researchers developed a migraine-stage classification system based on migraineurs' resting-state EEG power[12]; they used migraineurs' O1 and O2 EEG activities during closing eyes from occipital lobe to identify pre-ictal and non-pre-ictal stages. Again, the EEG data was utilized to detect migraine.

Another study discussed about a tool generated based on likelihood ratios and headache specialists' experiences [25]; the study was conducted in headache clinic and the tool was utilized to detect migraine headache (one type of the migraine). In this study, only headache patients participated.

In our study, we generated a model that not only finds the most important attributes to detect migraine, but also identifies the important stroke predictors. In addition, different types of patients were involved in this study (stroke cases, migrainous cases and patients with other types of mimic such as bell's palsy, seizure, BPV, and etc). Detecting migraine via text mining and data mining is a novel work that can offer a cheaper way to detect migraine, save time for GPs/specialists, and reduce cost for healthcare system.

In summary, our main contribution in this chapter is to introduce an inexpensive

data mining method to distinguish migraine from stroke and stroke-mimics (other than migraine) and our second contribution is introducing an approach to overcome the rarity problem in highly imbalanced data.

4.3 Data Description

In our study, we used heterogeneous data reported by different groups of clinicians (nurses and specialists). Figure 4.1 shows the data sources that were used in our study. More specifically, we used four heterogeneous data sources: (A) the history of the patient illness, (B) the chief-complaint, (C) Stroke Guidance System (SGS) data, and (D) the impression neurologists' reports. The data were collected between 2008 and 2013.

The history of patient illness (A) and chief complaint (B) were provided by stroke nurses. The steps for data-cleaning and data-transformation of (A) and (B) are explained in detail in our previous study [49]. The terms (variables) in those datasets were selected based on an exploratory data analysis of historical stroke data by attending neurologists and stroke clinicians at Victoria General Hospital (VGH). Overall, 126 terms were defined as stroke/mimic signs and symptoms. These terms were used as attributes and extracted from the raw text data.

Source (C), SGS data, was provided by the unit staff after the patient was examined. This is a structured dataset, stored in a relational database (SQL-Server). It contains 14 attributes, such as age, gender, blood pressure (systolic, diastolic), smoking status, diabetes status, atrial fibrillation, hypertension, and so on. Some of the attributes including age, gender, and blood pressure were normalized. We found

the top terms in this study that these terms are easy to interpret and use by triage personnel. Hence, while our model is trained on expensive to obtain datasets (but also very accurate), its application at triage time is inexpensive.

Source (D) contains (text-based) impression reports from neurologists. Such reports include diagnosis information, prescription, and suggestions based on patients' MRI result. Class labels were extracted from source D by mining the reports. In source D, neurologists reported their diagnoses as follows: e.g. "Diagnosis: Stroke" or "Primary diagnosis: Stroke". To obtain this information, we implemented a script that mined the reports by searching for "diagnosis" and extracted the class label for each record.

As mentioned, a regular-expression script was developed to extract diagnosis information from this text field in order to label a patient's record with one of three different classes. These classes are: "non-migraine mimic" (class 0), "stroke" (class 1), and "migraine" (class 2). We would like to note that the class label is very reliable because it is extracted from source (D), (created after thorough examination of the patient by a specialist neurologist).

The data obtained from sources (A)-(D) were integrated using the episode-id (an id for the patient). By the end of the process, a structured dataset with 6,912 records and 142 attributes (features) was obtained. The dataset was very imbalanced; it contained 4373 stroke cases, 392 migraine cases, and 2147 non-migraine mimic cases.

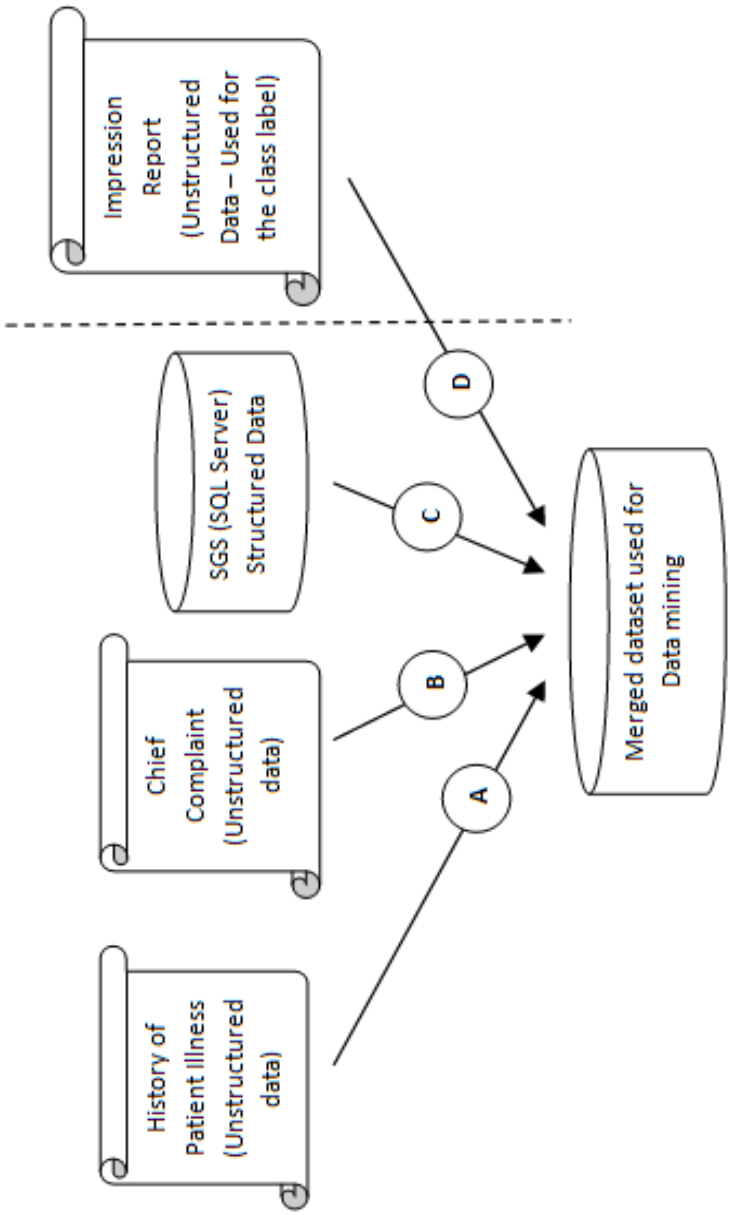


Figure 4.1: Heterogeneous datasets.

4.4 Proposed Approach

A novel classification method was implemented by combining different data mining (classification) algorithms. More specifically, the algorithms were organized in a directed acyclic graph (see Figure 4.2).

In our previous work [49], the stroke cases were distinguished from (general) mimic cases. In the current study, we label the data using three different classes (migraine, stroke, other-mimic).

To explain our methodology, please refer to Figure 4.2. Here, we implemented two classification nodes. Node 1 takes migraine and stroke cases as input and distinguishes migraine from stroke. Node 2 takes migraine and no-migraine-no-stroke cases as input and distinguishes migraine from these other types of stroke-mimic.

In each node, the most important features (predictors) are extracted and the final model is generated based on the combination of predictors found in nodes 1 and 2. The final model distinguishes migraine from all the cases (see Figure 4.3).

One of the main challenges in this work was tackling the imbalance of the data and overcoming the rarity problem of the migraine cases. To rectify this problem, we used a sampling method and generated multiple data sets with balanced class distributions. Each dataset included all minority class samples and a subset of the majority class samples.

For example, the number of stroke cases were 12 times more than the number of

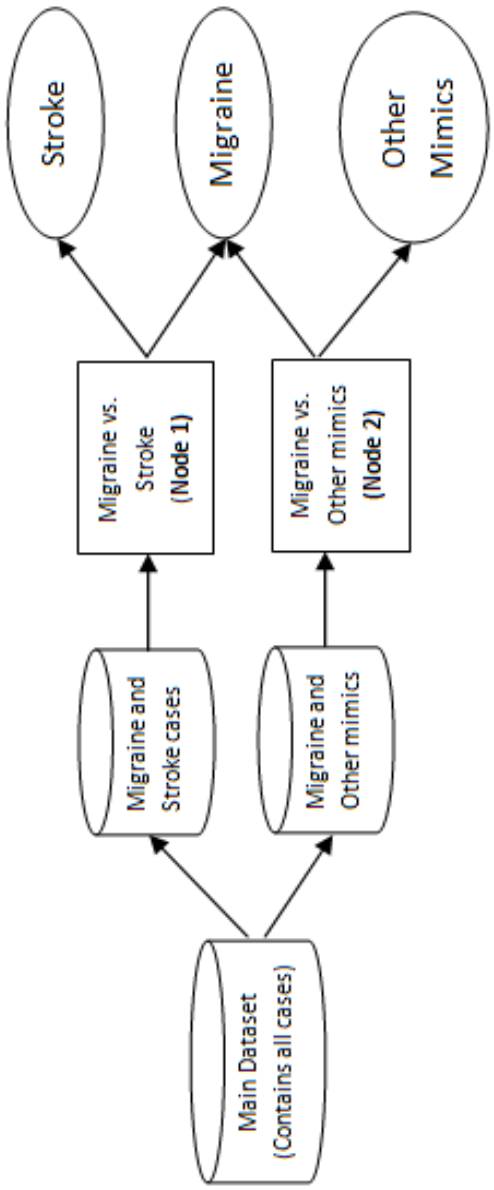


Figure 4.2: Node 1 is to distinguish migraine from stroke cases, and Node 2 is to distinguish migraine from other mimic sub-types

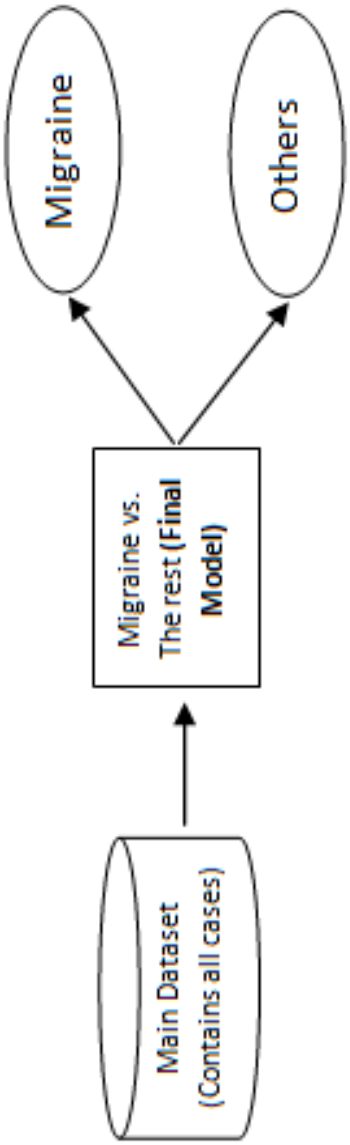


Figure 4.3: Final model distinguishes Migraine from other cases

migraine cases. So, 12 sets of data were created, such that each contained all the migraine cases and 1/12 of stroke records (split randomly); therefore, each majority class sample occurs in at least one set and no data is wasted. These 12 sets contained only stroke and migraine classes. They were used to generate the classification model for Node 1.

For each of the created datasets, we found the best predictor attributes (using a popular data mining tool - WEKA). WEKA provides an attribute selection tool which has two parts: "Attribute Evaluator" and "Search Method" [32]. These two sections are responsible for searching and assessing the attributes [32]. The "CfsSubsetEval" method, values subsets that correlate highly with the class value and have low correlation with each other. The "BestFirst" search method, uses a best-first search strategy to navigate attribute subsets [32]. CfsSubsetEval and BestFirst were used in this work to reduce the dimensionality of the data.

Afterwards, all the standard metrics such as sensitivity, specificity, accuracy and ROC area were calculated for each set. We only retain in the end the subset of attributes that are selected as best attributes for at least 50% of the datasets (i.e. 6 datasets).

A matrix was generated for this purpose. The columns of the matrix represent the datasets and rows represent the best selected attributes for the dataset (see Figure 4.4). In the figure, we show the best selected attributes chosen in each dataset. We selected those attributes chosen as best predictors for 50% or more datasets.

The selected attributes were used as features for building the final classification

model. We applied the same process for distinguishing migraine and other-mimic case in Node 2. This time, six sets of balanced datasets were created such that each one contained all the migraine cases and 1/6 of the other-mimic cases.

Figure 4.5 shows the chosen attributes for each dataset along with the final selected attributes. Figure 4.6 shows the list of features found in each node.

We would like to point out that the purpose of the chunks and classification nodes 1 and 2 was to find the best predictors (selected attributes) to use for the final model. We extracted the most important attributes from Node1 and Node2, merged them and generated the final model to distinguish migraine from the other two cases.

We also would like to indicate that the use of imbalanced data leads to unsatisfactory results and mostly ignores the minor cases; this was why we generated sets of balanced data (from imbalanced data), extracted the list of best predictors, trained the classifiers on balanced data and used imbalanced data for test.

4.5 Experimental Results

WEKA was employed in this study to apply different data mining algorithms on our datasets.

To establish a realistic baseline for the algorithms, we considered the (original) sets of imbalanced data with all the migraine and all the stroke cases (392 migraine, 4373 stroke) as well as all the migraine and all the other-mimic cases (392 migraine, 2147 other-mimic). Results gained from different classification methods are shown

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
ACCURACY	80.73	79.33	79.84	80.48	77.42	77.16	80.35	79.97	81.37	78.44	80.73	89.18
HEADACHE	x	x	x	x	x	x	x	x	x	x	x	x
DRAG	x					x	x	x				x
PHOTOPHOBIA	x		x			x	x		x	x		
POSITIVE VISUAL	x	x	x	x	x	x	x	x	x	x	x	x
VISUAL DISTURBANCE	x	x	x			x	x	x	x	x	x	x
DROOL	x		x		x		x		x	x		
TINNIT	x											
HISTORY OF MIGRAINE	x	x	x	x	x	x	x	x	x	x	x	x
GENDER	x	x	x	x	x	x	x	x	x	x	x	x
DIABETES	x			x	x		x		x	x	x	
HYPERTENTION		x										
FALL		x	x		x	x		x	x		x	
NARROWING		x				x						
BLUR		x							x			
SUDDEN ONSET		x				x						
PALLOR		x	x	x		x	x	x	x		x	
WEAK FACE LEFT		x						x			x	
WEAK ARM RIGHT		x					x					
CHEST TIGHTNESS			x									
CONCENTRATION			x	x	x	x			x	x		
MARCH			x	x		x	x		x	x		
ANGST(ANEXIETY, STRESS)			x	x	x	x	x	x		x	x	
AGE	x	x	x	x	x	x	x	x	x	x	x	x
AMNESIA											x	
FACE DROOP				x		x						
UNSTEADINESS				x			x					
HEARING LOSS								x	x			
VALASALVA								x				
FREEZE								x				
AMNESIA											x	
CHEST PAIN												x
NECK PAIN												x
SONOPHOBIA		x									x	
WEAK ARM LEFT					x							

Figure 4.4: Best selected attributes in migraine vs stroke (balanced datasets)

	D1	D2	D3	D4	D5	D6
ACCURACY	**74.74**	**73.59**	**75**	**71.55**	**71.17**	**81**
HEADACHE	x	x	x	x	x	x
AMNESIA	x	x	x	x	x	
POSITIVE_VISUAL	x	x	x	x	x	x
VISUAL DISTURBANCE	x	x	x	x	x	x
DROOL	x	x	x	x	x	x
SUDDEN_ONSET	x	x	x	x	x	
MIGRAINE	x	x	x	x	x	x
HEARING LOSS	x	x		x	x	x
NUMB_FACE_RIGHT	x					
DIABETES	x		x		x	
PHOTOPHOBIA		x				
FLUSH		x				
COGNITIVE		x				
HYPERTEN		x				
CHEST TIGHTNESS			x	x		
FALL		x	x		x	x
BLPRS_NORM			x			
VF				x		
DRAG				x		
NYSTAGMUS				x		
PALLOR				x		x
TINNIT					x	
FREEZE					x	
AGE				x	x	x
OTHCARD					x	
WEAK FACE RIGHT						x
VISION LOSS					x	

Figure 4.5: Best selected attributes in migraine vs other mimic (balanced datasets)

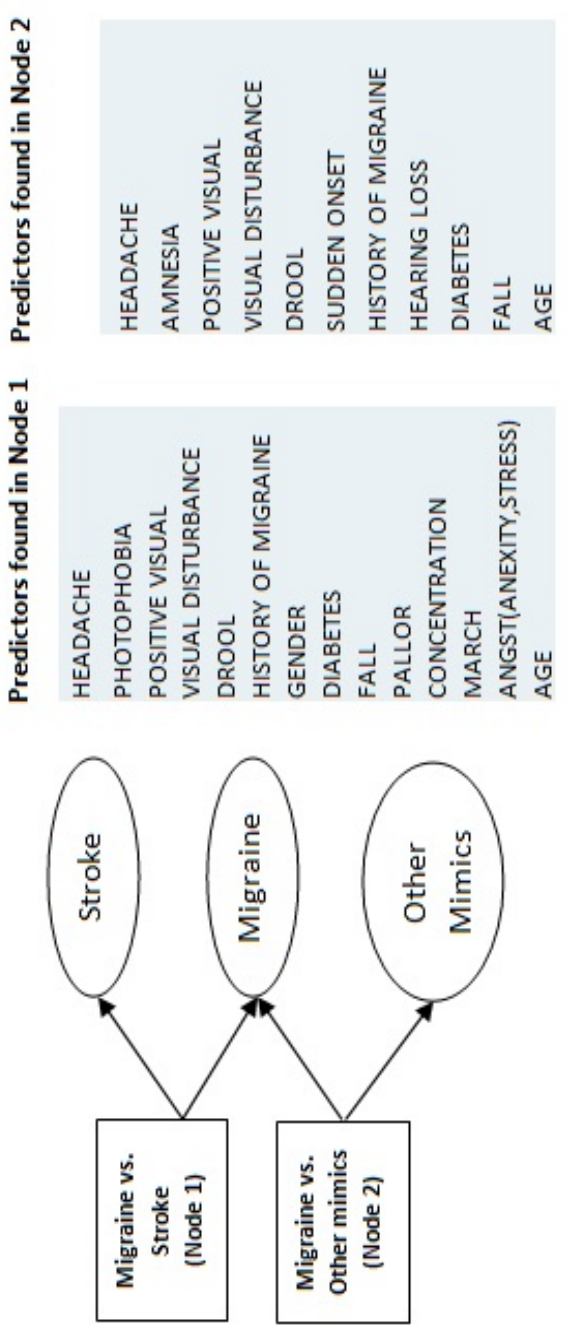


Figure 4.6: List of predictors found in each node

in Table 4.1.

Rare cases (migraine) cover only a small region of the instance space, therefore migraine is more difficult to be detected. We can see that the best sensitivity value can be achieved when trying to distinguish migraine from stroke is only 57.7% by Naive Bayes. In terms of accuracy, the best value we get is 91.8% for support vector machine. This is good, but it ignored the minority of cases (migraine).

Regarding the other imbalanced dataset (migraine vs. other-mimic), the classifiers fare again quite poorly. For example, the best sensitivity we can get is only 49.2%.

In other words, using imbalanced datasets is ineffective in distinguishing migraine (the minority of classes) from the other classes. The accuracy of classifiers, albeit good, is a measure that is always biased towards the majority classes, in our case, stroke and other-mimic.

Now we show the results using our extracted balanced datasets (see Table 4.2). As explained earlier, for Node 1, twelve balanced datasets were extracted that contained an equal number of migraine and stroke cases (784 records).

The highest sensitivity we achieved in distinguishing migraine from stroke has much improved, 80.1% (using PART algorithm). The specificity is good as well, 75.5% (PART). The other classifiers gave similar results. The best specificity we achieved was 81.1% (Naive Bayes). Also, the average of accuracy with ten-fold cross-validation shows that Naive Bayes provides the highest accuracy of 81.8 %

Table 4.1: Different measures calculated for different methods using the imbalanced migraine-stroke dataset, migraine other-mimic dataset, and migraine-The Rest dataset

		NB	Logistic	SVM	PART	J48
Migraine - Stroke	Sensitivity	57.7	37.2	29.3	32.7	24.2
	Specificity	91	96.5	97.5	95	91.5
	ROC	88	78.7	63.4	69.9	65.6
	Accuracy	88.2	91.6	91.8	89.9	91.5
		NB	Logistic	SVM	PART	J48
Migraine - Other Mimic	Sensitivity	49.2	27.6	20.4	33.4	23.2
	Specificity	85.3	92.7	95.2	88.9	93.9
	ROC	80.6	73.2	57.8	64.2	59.8
	Accuracy	79.7	82.7	83.65	80.3	82.9
		NB	Logistic	SVM	PART	J48
Migraine - The Rest	Sensitivity	46.4	15.6	0	14.5	0
	Specificity	90.8	97.9	100	96.2	100
	ROC	84.2	76.5	50	68.4	49.9
	Accuracy	88.29	93.22	94.3	91.57	94.3

followed by Support Vector Machines, 80.64 %, and Logistic Regression, 80.4 %.

Next the analysis was performed on Node 2, migraine vs. other-mimic, on six balanced datasets. Table 4.1 shows the result for the baseline (all the imbalanced migraine-other mimic cases). Table 4.3 shows how the results outperformed using the extracted balanced datasets for six chunks.

Reviewing Table 4.3 shows that the highest sensitivity and specificity were 76% and 75.3%, respectively. The best accuracy we can get is 74.7% and 74.1% for Naive Bayes and Logistic Regression respectively. Again Naive Bayes and Logistic Regression provided the highest accuracy among other methods. The summary of accuracy for each node is depicted in Figure 4.7 and 4.8.

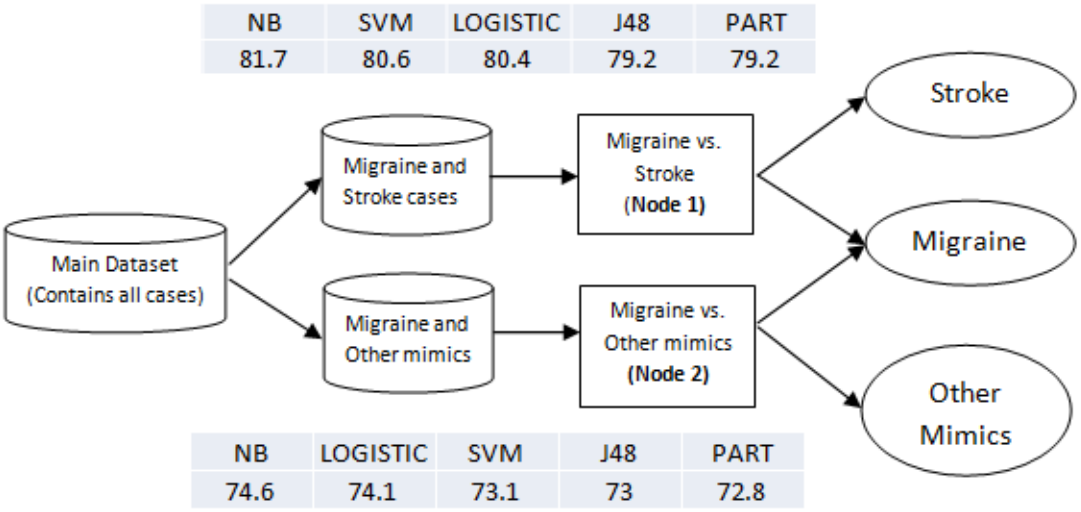


Figure 4.7: Accuracy gained in each node by different classifiers

The list of predictors (best attributes) found for Node 1 and Node 2 were combined and a balanced dataset was created that contained all the migraine cases along with

Table 4.2: All measures calculated for different methods using 12 chunks of the Migraine-Stroke datasets

		D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	AVG
Naive Bayes	Sensitivity	77	79.1	76.5	78.8	77.8	77.3	79.8	78.3	79.8	76.5	80.1	95.7	79.7
	Specificity	87.2	84.2	84.9	82.4	80.6	82.4	83.7	84.2	86	82.7	85.5	49.2	81.1
	ROC	89.3	89.1	88.5	89.2	85.2	87.9	89.4	89	90.2	86.9	88.2	89.8	88.6
	Accuracy	82.1	81.3	80.4	80.6	79.3	80.1	81.8	81.3	82.9	79.6	82.8	89.2	81.8
		D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**	**D11**	**D12**	**AVG**
SVM	Sensitivity	79.1	78.6	75.3	80.4	80.1	75.8	80.1	80.1	75.3	76	79.3	95.7	79.7
	Specificity	84.7	82.9	84.9	80.1	74.5	81.1	80.9	80.9	86.5	82.7	84.4	31.1	77.9
	ROC	81.9	80.7	80.1	80.2	77.3	78.4	80.5	80.5	80.9	79.3	81.9	63.4	78.8
	Accuracy	81.9	79.8	80.0	80.4	77.3	78.4	80.5	80.5	80.9	79.3	81.9	87.0	80.6
		D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**	**D11**	**D12**	**AVG**
J48	Sensitivity	80.6	78.8	77.8	78.6	78.3	76	77.8	79.1	74.2	78.8	80.6	93.9	79.5
	Specificity	80.9	79.6	78.3	80.4	73.7	77	77.3	77.6	84.7	79.3	77.3	50.8	76.4
	ROC	83.9	82.7	80.3	82.8	78.4	78	80	81.4	82.6	82.7	82.1	80.6	81.3
	Accuracy	80.7	78.8	78.1	79.5	76.0	76.5	77.6	78.3	79.5	79.1	79.0	88.1	79.3
		D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**	**D11**	**D12**	**AVG**
PART	Sensitivity	83.4	77	80.4	79.3	76.5	77.6	76.5	82.7	78.3	78.3	75	96.7	80.1
	Specificity	75.5	79.1	74.7	78.3	74	76.5	77.8	75.8	83.2	78.1	82.4	50.8	75.5
	ROC	84.8	82.8	82.1	83.1	78.6	81.7	84.1	85.2	86.5	82.7	82.6	82.7	83.1
	Accuracy	79.5	76.7	77.7	78.8	75.3	77.9	77.2	79.2	80.7	78.2	78.7	90.5	79.2
		D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**	**D11**	**D12**	**AVG**
LOGISTIC	Sensitivity	76.5	77.3	75.8	77.8	76.3	74.2	78.6	78.6	76.5	74.2	79.3	96.2	78.4
	Specificity	84.9	81.6	83.9	83.9	78.6	80.1	82.1	81.4	86.2	82.7	82.1	44.3	79.3
	ROC	88.5	87.2	87.2	88.1	83.9	85.6	88	87.9	88.3	86.5	87.3	87.5	87.2
	Accuracy	80.7	79.3	79.8	80.5	77.4	77.2	80.4	80.0	81.4	78.4	80.7	89.2	80.4

Table 4.3: All measures calculated for different methods using 6 chunks of the Migraine- other mimic datasets.

		D1	D2	D3	D4	D5	D6	AVG
Naive Bayes	Sensitivity	72.2	71.2	70.7	70.9	69.4	85.5	73.3
	Specificity	77.8	75.3	78.1	76.5	74.7	69.5	75.3
	ROC	81.6	79.5	79.9	80.2	79.8	83.6	80.8
	Accuracy	75	72.45	74.36	73.72	72.06	80.31	74.7
		D1	D2	D3	D4	D5	D6	AVG
SVM	Sensitivity	74.7	75	72.2	67.6	70.9	85.2	74.3
	Specificity	73.7	68.6	73.5	77.6	72.7	61	71.2
	ROC	74.2	71.8	72.8	72.6	71.8	73.1	72.7
	Accuracy	74.2	70.3	72.6	72.6	71.8	77.4	73.1
		D1	D2	D3	D4	D5	D6	AVG
J48	Sensitivity	74.5	69.9	80.4	67.3	76.3	87.8	76.0
	Specificity	73.5	70.7	66.8	73.2	65.3	61	68.4
	ROC	77.7	71.9	75.7	73.7	73.5	74.6	74.5
	Accuracy	74.0	70.2	73.6	70.3	70.8	79.1	73.0
		D1	D2	D3	D4	D5	D6	AVG
PART	Sensitivity	77.8	66.1	73	65.6	75.3	86.5	74.1
	Specificity	71.4	72.4	68.4	71.4	68.6	63.1	69.2
	ROC	79.3	74.3	74.3	74.2	76.9	76.4	75.9
	Accuracy	74.6	72.5	70.7	68.5	71.9	78.9	72.8
		D1	D2	D3	D4	D5	D6	AVG
LOGISTIC	Sensitivity	72.4	73.7	77.3	71.7	70.7	88.8	75.8
	Specificity	77	73.5	72.7	71.4	71.7	64.7	71.8
	ROC	81	78.8	79	80.1	79.1	82.4	80.1
	Accuracy	74.7	72.1	74.4	71.6	71.2	81.0	74.1

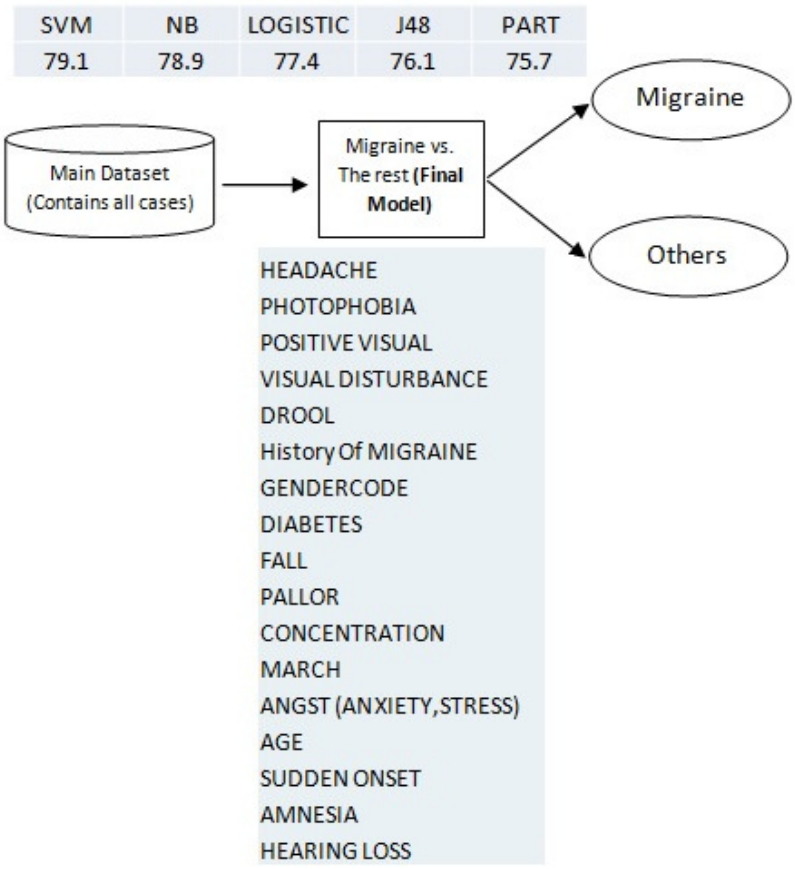


Figure 4.8: List of predictors to distinguish migraine from other cases

Table 4.4: Different measures for the final model using Migraine-Other dataset

	Naive bayes	Logistic	SVM	PART	J48
Sensitivity	81.1	79.3	79.3	76	75.8
Specificity	77.6	75.5	77.6	78.1	75.5
ROC	86.4	85.6	78.4	79.6	79.4
Accuracy	79.3	77.4	78.4	77	75.6

an equal number of other cases. Figure 4.8 shows the list of predictors to distinguish migraine from other cases. Based on the neurologist, the following factors are well recognized: Headache, photophobia, positive visual, visual disturbance, history of migraine, gender, pallor, concentration, march, anxiety, stress, age, and sudden onset. The most unexpected factors for distinguishing migraine from other mimics/stroke cases were (in order): Diabetes, fall, drool, amnesia, and hearing loss.

Note: the purpose of the chunks was to find the best selected attributes to use them for the final model and distinguish migraine from all the other cases.

Results gained from different classification methods on migraine-other cases dataset are shown in Table 4.4. The highest sensitivity we get to distinguish migraine from other cases is 81.1% for Naive Bayes and 79.3% for Logistic and SVM. The best specificity we get is 78.1% and 77.6% for Part and Naive Bayes respectively. Also, the average of accuracy with ten-fold cross-validation shows that Naive Bayes provides the highest accuracy of 79.3% followed by Support Vector Machines, 78.4%, and Logistic Regression, 77.4%.

4.6 Evaluation of the final model on the imbalanced dataset

As mentioned, the aim of this study is to provide a model to distinguish migraine from other cases, and because migraine is a minor class, one of the challenges is overcoming the rarity problem. Besides, differentiating between migraine and stroke is important for the clinicians to provide immediate care for stroke patients (Node 1).

To evaluate the correctness of Node 1, Node 2 and the final model, two sets of balanced train and imbalanced test were generated for each node; therefore, 2/3 of records were put aside for training and 1/3 of records were used for testing (for each node). The result is depicted in Table 4.5.

In order to build the final classifier for the evaluation, we only used the best attributes derived in the previous section when using balanced datasets (chunks).

As stated above, the test set is imbalanced reflecting the real distribution of instances; comparing Table 4.5 with the baseline (Table 4.1) which also contains imbalanced data shows that sensitivity is improved significantly. In Table 4.1 that distinguishes **migraine from stroke**, the highest sensitivity is 57.7% gained by Naive Bayes, but in Table 4.5 (the final evaluation result), the sensitivity reached 73.1% (Naive Bayes). Also, reviewing the specificity achieved by Naive Bayes in both tables shows that specificity did not suffer too much (it was 91% in the baseline and it became 86.1% in evaluation).

Next we would like to compare the final evaluation result with the result obtained using balanced datasets and balanced test sets (as derived by cross-validation, Table 4.2). In Table 4.5 (final evaluation table), the model was tested with imbalanced dataset and the sensitivity and specificity were between 68.5% and 73.1% and 80.6% and 86.1% respectively; comparing these numbers with average column in Table 4.2, indicates that results on imbalanced data (evaluation data) are very close to the results on the balanced data using cross-validation (Table 4.2). The accepted level for clinicians was a sensitivity above 72% and a specificity above 69%; both results in Table 4.2 and Table 4.5 met the criteria.

Comparing Table 4.5 with Table 4.1 (the baseline) shows that sensitivity is also improved significantly for migraine and other mimics. In Table 4.1 (baseline) that distinguishes **migraine from other mimics**, the highest sensitivity is 49.2% gained by Naive Bayes, but in Table 4.5 (the final evaluation), the sensitivity reached to 76.9% for Naive Bayes. Support vector machine, J48 and PART provided the highest sensitivity (79.2%). Also, the specificity of equal or above 69% was ideal that logistic and Naive Bayes provided specificity of 72.8% and 71.4% respectively.

As mentioned, the ROC above 80% was ideal for clinicians; Naive Bayes and logistic regression provided the highest ROC of 79.4% and 79.5% respectively; reviewing the ROC achieved by Naive Bayes in both tables shows that the ROC didn't change too much (it was 81.6% in the baseline and it became 79.4% in the evaluation result and it improved extremely for logistic (it was 73.2% in baseline and became 79.5%).)

All the metrics performed great for the final model that distinguished **migraine from the other cases**. Comparing Table 4.5 with Table 4.1 shows that sensitivity

Table 4.5: Evaluation result with balanced and imbalanced data (balanced train and imbalanced test dataset)

		NB	Logistic	SVM	J48	PART
Migraine - Stroke	Sensitivity	73.1	70.8	68.5	68.5	70.8
	Specificity	86.1	83.8	80.6	85.1	82.1
	ROC	87.5	86.8	74.6	78	81.9
	Accuracy	85	82.73	79.64	83.74	81.16
		NB	Logistic	SVM	J48	PART
Migraine - Other Mimics	Sensitivity	76.9	74.6	79.2	79.2	79.2
	Specificity	71.4	72.8	66.3	64.9	68.9
	ROC	79.4	79.5	72.8	75.4	77.9
	Accuracy	72.2	73	68.3	67.1	70.4
		NB	Logistic	SVM	J48	PART
Migraine - The rest	Sensitivity	76.2	73.8	72.3	78.5	76.2
	Specificity	81.8	79.9	79	76.2	77.8
	ROC	85.2	84.9	75.6	81.1	81.7
	Accuracy	81.5	79.5	78.59	76.3	77.7

is improved for all the classifieres specially for logistic regression (it improved significantly from 15.6% to 73.8%). As stated before, a sensitivity above 72%, a specificity above 69%, and a ROC area above 80% were the acceptable levels for clinicians, and all the classifiers met these three conditions. The highest sensitivity is provided by J48 (78.5%) followed by Naive Bayes (76.2%); the highest specificity is provided by Naive Bayes and Logistic Regression for 81.8% and 79.9% respectively and again Naive Bayes and Logistic Regression provided the highest ROC area (85.2% and 84.9% respectively).

In addition, the cost-sensitive method [39] was examined with different algorithms on the imbalanced dataset (containing migraine and the rest of the cases) and provided a good result, but the sensitivity did not meet the clinician's condition.

As stated before, the cost-sensitive method assigns misclassification costs to data from each class and forces the classifier to concentrate on the minority class [50]. When we increase the weight of errors on class migraine (false negatives, FN), for instance in a 10:1 relation, the classifier will then try to avoid false negatives, because each one is equivalent to 10 false positives (FP). The result gained from the cost-sensitive method is depicted in table 4.6. Cost-bf and cost-af show the amounts before and after increasing the weight of mistakes.

Comparing table 4.4 with table 4.6 shows that the sampling method provides better sensitivity than the cost-sensitive method for all the classifiers (almost 20% higher). The cost-sensitive method provides a slightly better specificity than the sampling method (about 10% higher) and the ROC is almost the same. As mentioned, the sensitivity above 72% was acceptable by clinicians, but the highest sensitivity

Table 4.6: Evaluation result with imbalanced data and using the cost-sensitive method

		Part	J48	SVM	Logistic	NB
cost-bf	**sensitivity**	4	0	0	6	31.6
cost-af		**57**	**59.2**	**55.4**	**60.2**	**64**
cost-bf	**specificity**	99	100	100	99.4	95.6
cost-af		**85**	**85.3**	**89**	**88**	**86.7**
cost-bf	**ROC**	76.7	49.8	50	85.5	85.8
cost-af		**73.4**	**70.8**	**72.2**	**85.6**	**85.9**
cost-bf	**Accuracy**	93.7	94.3	94.3	94	91.94
cost-af		**83.62**	**83.7**	**87**	**86.4**	**85.38**

provided by the cost-sensitive method was 64% which did not meet the requirement. Therefore, the model provided by the sampling method that meets all the requirements is reliable and can be used by clinicians.

4.7 Conclusions

Detecting migraine from stroke is a challenge due to many common signs and symptoms. In this part of the study, we propose an inexpensive method that can help to detect the migraine. Natural language processing, text-mining and data mining methods were utilized to analyze the data.

A novel classification method was implemented by merging different structured and unstructured data-sources. After a careful combination of data provided by neurologists and stroke nurses, we obtained a highly imbalanced dataset where the migraine cases were only about 6% of the dataset.

We implemented two classification nodes: node 1 that took migraine and stroke cases as input and distinguished migraine from stroke, and node 2 that took migraine and no-migraine-no-stroke cases as input and distinguished migraine from other types of stroke-mimic. In each node, the most important features (predictors) were extracted and the final model was generated based on the combination of predictors found in nodes 1 and 2.

One of the main challenges in this study was to overcome the rarity problem because of the highly imbalanced dataset. To rectify this issue, a sampling method was utilized and balanced chunks of data were created from the original imbalanced

dataset. We created 12 balanced chunks for extracting best attributes in distinguishing migraine from stroke, and six balanced chunks for extracting best attributes in distinguishing migraine from other (non-stroke) mimics. The cross-validation results on these balanced chunks were much improved compared to the baseline of using imbalanced data. This was encouraging because it gave us assurance that the best attributes selected from models trained on the balanced chunks would be of good quality. We achieved a sensitivity and specificity of about 80% and 75% respectively, which was in contrast to a sensitivity and specificity of 15.7% and 97% when using the original imbalanced data for building classifiers.

To evaluate the correctness of our model, two sets of balanced train and imbalanced test were generated. We compared the final evaluation result with the result obtained using balanced datasets and balanced test sets (as derived by cross-validation). Comparing the metrics indicates that results on evaluation data are very close to the results on the balanced data using cross-validation. A sensitivity above 72%, a specificity above 69%, and a ROC area above 80% were the acceptable levels for clinicians that both results met the criteria and all the classifiers met these three conditions.

In addition, the cost-sensitive method was examined with different algorithms on the imbalanced dataset (the cost-sensitive method assigns misclassification costs to data from each class and forces the classifier to concentrate on the minority class). This method provided a good result, but the sensitivity did not meet the clinicians condition.

Indeed, the final model based on the selected attributes was quite robust performing

well not only on balanced data, but more importantly performing almost equally well on imbalanced test data, which accurately reflect reality.

In conclusion, we would like to mention that typically migraine is distinguished from stroke and other non-stroke mimics via different types of tests such as CBC, Facial X-ray, CT scan, MRI, and EEG, which are all costly. In our study, we generated an effective model for migraine detection based on structured and unstructured data sources. This is a novel work that can be used in decision support systems and save time for GPs/family doctors and nurses plus reduce costs for the healthcare system.

Chapter 5

Discovering Signs and Symptoms Relating to Posterior Circulation Ischemic Stroke (POCS): Report analysis

5.1 Introduction

Posterior circulation transient ischemic attacks may include brief or minor brainstem symptoms and are more difficult to diagnose than anterior circulation ischaemia [42]. The risk of recurrent stroke after posterior circulation stroke is at least as high as anterior circulation stroke. Delayed or incorrect diagnosis may have devastating consequences including potentially preventable death or severe disability [37]. Early recognition of posterior circulation stroke or transient ischaemic attack (TIA) may save lives, but it remains more difficult to recognise and treat effectively than other stroke types [42]. New acute treatment options and stroke prevention strategies

specific to the posterior circulation are important areas of active research [42].

In this chapter, we tried to extract the most important signs/symptoms (that might be helpful to detect POCS) from the reports provided by stroke nurses and the neurologists. We also reviewed the signs/symptoms reported by the Heart and Stroke Foundation (HSF) of Canada and compared our predictors with theirs. The results gained from both sets of predictors are compared and evaluated in detail.

5.2 Method

Two types of reports were utilized in this study: reports provided by stroke nurses (before MRI) and reports provided by neurologists (after MRI).

Both reports were analyzed and transformed to structured format[1]. The dataset had two classes of POCS and non-POCS. When considering POCS as the positive class, we were faced with a class imbalance problem. This was because the POCS cases made up only about 3% of the total cases.

For the experiment, we employed a sampling method and generated a balanced dataset with 8537 records and 168 columns. Undersampling can cause elimination of valuable samples [60], so we used the over-sampling method to make the data balanced. For evaluation, we set aside 2/3 of data for training (part of a balanced dataset) and 1/3 was used for testing (we used a part of imbalanced data that was not used during training). The work is described in detail in each section.

[1]Transforming unstructured data (text) to structured data (table) is explained in detail in Chapter 3.

5.3 Experiment and Result

The Heart and Stroke Foundation of Canada (HSF) has provided a "pocket guide" that introduces common signs and symptoms that related to different types of stroke (Figure 5.1). Based on the HSF, the signs and symptoms reported for POCS are as follows: weakness, vertigo, ataxia, dysarthria, nausea, vomiting, hemianopsia, and sensory.

As mentioned, we used data reported by stroke nurses and neurologists and generated a structured dataset. The dataset was highly imbalanced (163 cases of POCS and 4299 cases of other stroke sub-types). Rare cases (POCS) covered only a small region of the instance space; therefore, POCS was more difficult detect. The use of imbalanced data ignored the minor cases; that was why we needed to generate a set of balanced data from imbalanced data. We used the over-sampling method and provided balanced data set that included all majority class samples (so that no case was missing) and the number of records became 8537.

We cloned this dataset, excluded 160 columns and kept only the 8 columns (signs and symptoms of POCS) reported by the HSF. This set was provided for our first experiment. We applied several classification algorithms and the result gained from this experiment is depicted in Table 5.1.

We provided another copy of the original dataset and applied the BestFirst filtering method (using WEKA) to extract the most important signs and symptoms. 22 attributes were selected, so the second dataset used in our experiment, contained 8537 records and 22 columns. We applied different data mining algorithms on this set and gained the results shown in Table 5.2.

Figure 5.1: The common signs and symptoms relate to different types of stroke provided by Heart and Stroke Foundation of Canada (HSF)

Table 5.1: Results gained from HSF set (8537 records and 8 HSF columns).

	PART	J48	NB	Logistic	SVM
Sensitivity	42.3	42.3	41.5	42	40.3
Specificity	86.3	86.2	84.8	85.2	86
ROC	67.5	67.3	66	66.3	63.1
Accuracy	64.49	64.43	63.32	63.74	63.28

Table 5.2: Results gained from the BestFirst algorithm (8537 records and 22 columns selected by BestFirst filtering).

	PART	J48	NB	Logistic	SVM
Sensitivity	89.7	89.7	73.6	76	75.9
Specificity	63.2	62.5	69.6	68.4	68
ROC	82.9	81.6	78.6	78.8	71.9
Accuracy	76.36	76.01	71.61	72.15	71.88

Figure 5.2 shows the list of 22 attributes selected by the BestFirst method. Only 3 out of 22 attributes (Nausea, Ataxia, Vertigo) were common in both sets. We utilized the logistic regression algorithm to check the weight of each attribute in this list to explore the subset of the best 8 attributes out of the 22 that the filtering process produces. We applied the same data-mining algorithms and gained the results shown in Table 5.3.

The list of attributes with the highest to the lowest weight is as follows: Unstead (Unsteadiness), Vertigo, Ataxia, Diplopia, Visual Disturbance, Dizz (Dizziness),

Table 5.3: Results based on 8537 records and 8 columns (Selected by BestFirst algorithm).

	PART	J48	NB	Logistic	SVM
Sensitivity	89.6	89.7	56.8	62.9	69.8
Specificity	60.8	60.5	80.5	74.5	68.9
ROC	81.5	81	77.4	77.3	69.3
Accuracy	75.09	74.97	68.72	68.74	69.33

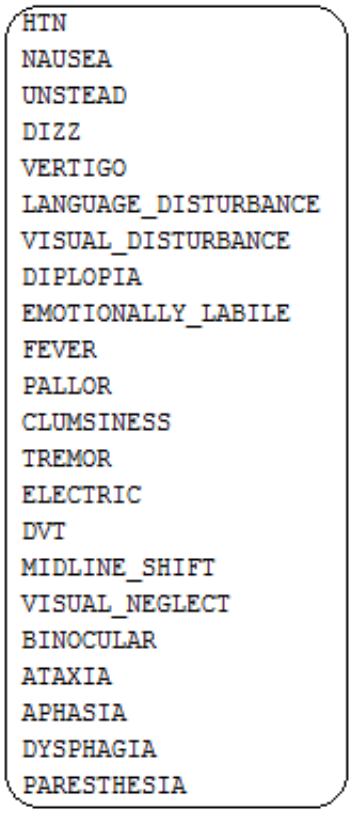


Figure 5.2: List of 22 attributes, selected by BestFrst filtering method.

Nausea, Language Disturbance, Aphasia, Midline shift, DVT (Deep vein thrombosis), Vusual Neglect, Emotionaly Labile, Pallor, Paresthesia, Binocular, Fever, Electric, Dysphagia, Clumsiness, HTN, Tremor.

5.4 Discussion

As mentioned, we employed the BestFirst search method and found the best attributes (22 predictors), then used the logistic regression algorithm to explore the top 8 features based on their weights. In this section, we would like to point out that attributes selected by BestFirst algorithm provided better results than those mentioned by the HSF.

Assume that set (A) contains only the attributes mentioned by the HSF pocket-guide, set (B) contains the 22 best-first attributes selected by the BestFirst algorithm and set (C) includes the top 8 features (which were derived from the 22 best-first attributes) and chosen based on their weights. We calculated the standard metrics such as sensitivity, specificity, accuracy and ROC for each set and compared the results in four graphs.

The accuracy calculated for each set is shown in figure 5.3. This graph shows that attributes selected by the Bestfirst method provided the highest accuracy among all. Also, the top 8 attributes provided higher accuracy than the HSF pocket-guide predictors. The sensitivity gained from each set is demonstrated in Figure 5.4. Again, the best first attributes provided the highest sensitiviy, and the HSF predictors generated the lowest one.

The specificity obtained from each dataset is shown in Figure 5.5. In this graph, HSF provided the highest specificity, but comparing the sensitivity and specificity gained from each set indicates that these two measures are totally imbalanced for HSF attributes (high specificity and low sensitivity). However these two measures gained from dataset B and C (that contain best-first attributes and top 8 features) are balanced.

Figure 5.6 compares the ROC obtained from all three sets (pocket guide, best selected attributes (22 attributes) and top 8 attributes). Obviously, the best first predictors provided the highest ROC and then come the top 8 and HSF respectively.

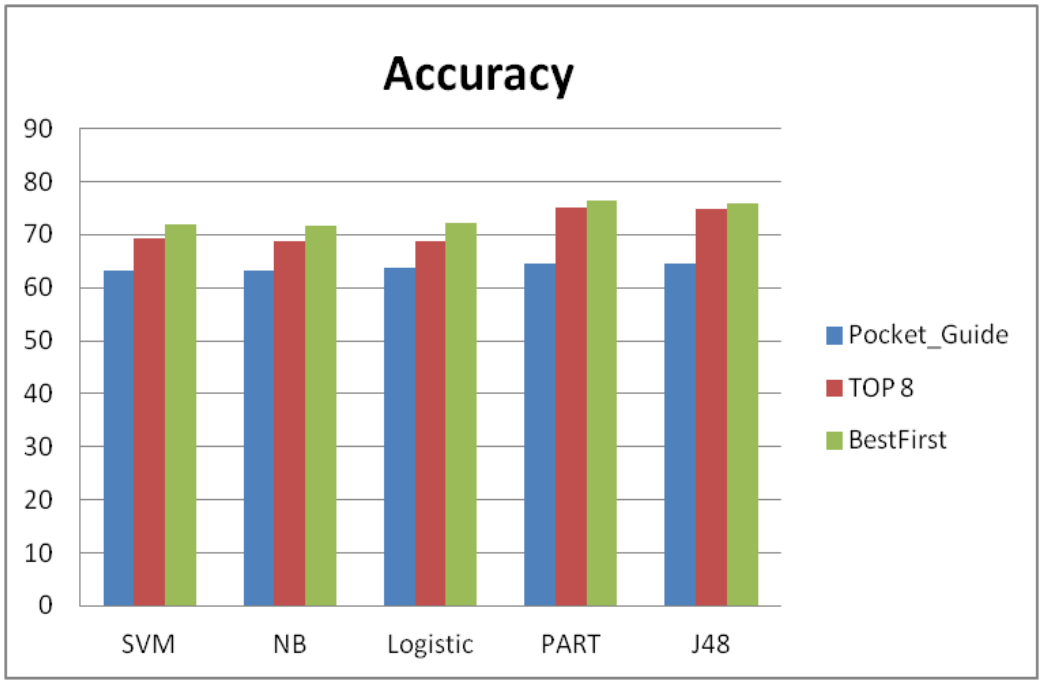


Figure 5.3: Accuracy for different methods.

Overall, comparing the results gained from these three sets showed that the best first attributes (22 attributes) provided the highest accuracy, sensitivity, and ROC and the features introduced in HSF pocket guide provided the lowest amount for the

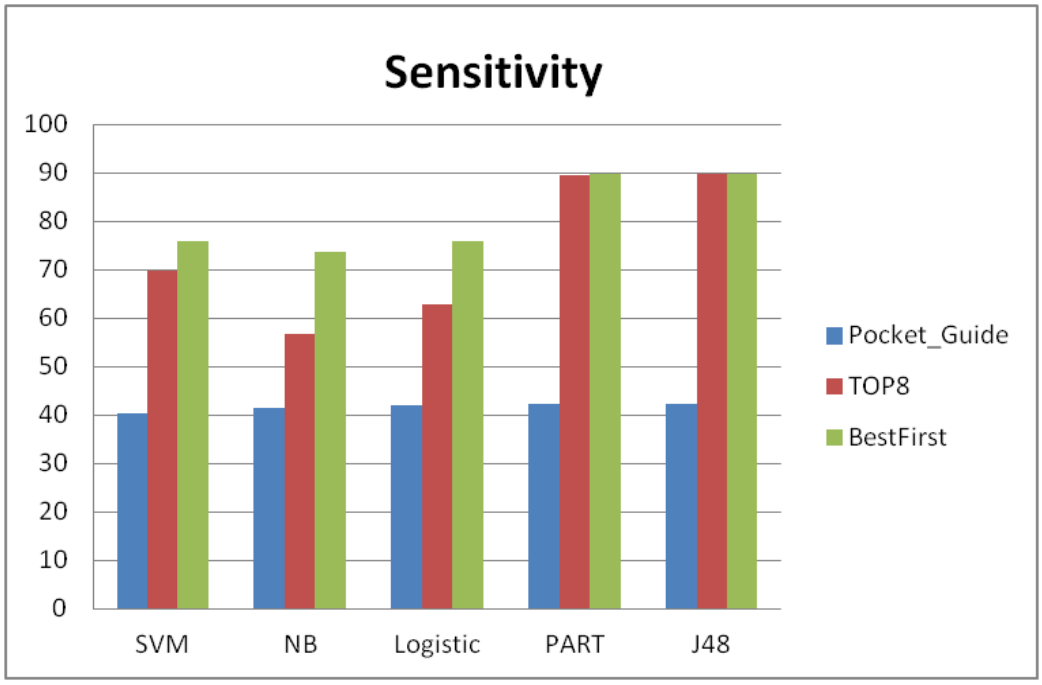


Figure 5.4: Sensitivity for different methods.

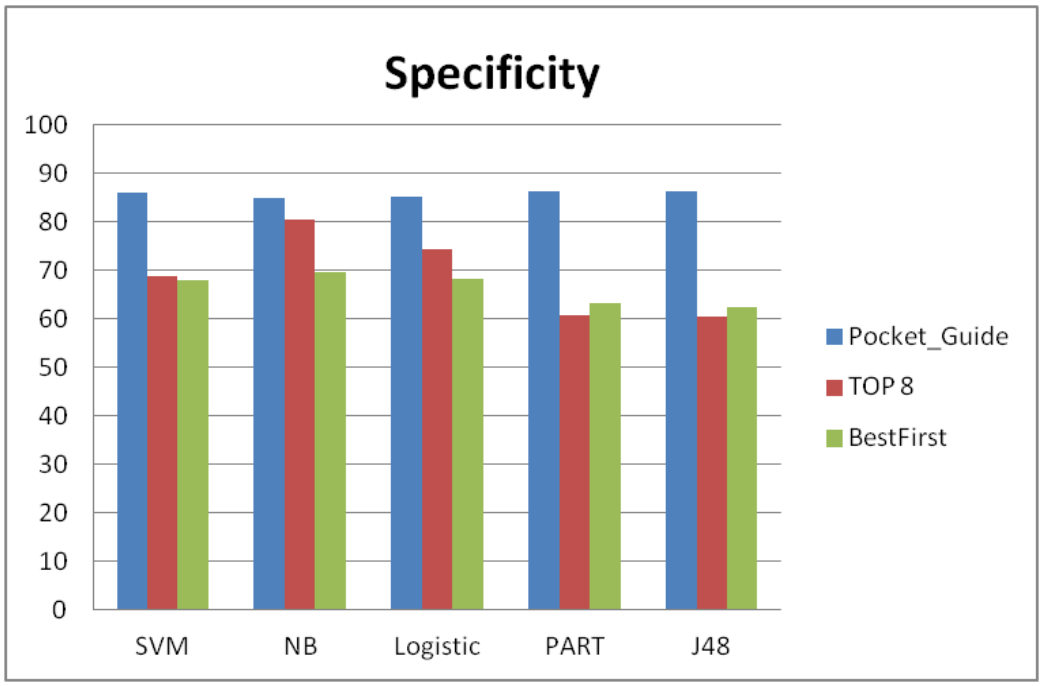


Figure 5.5: Specificity for different methods.

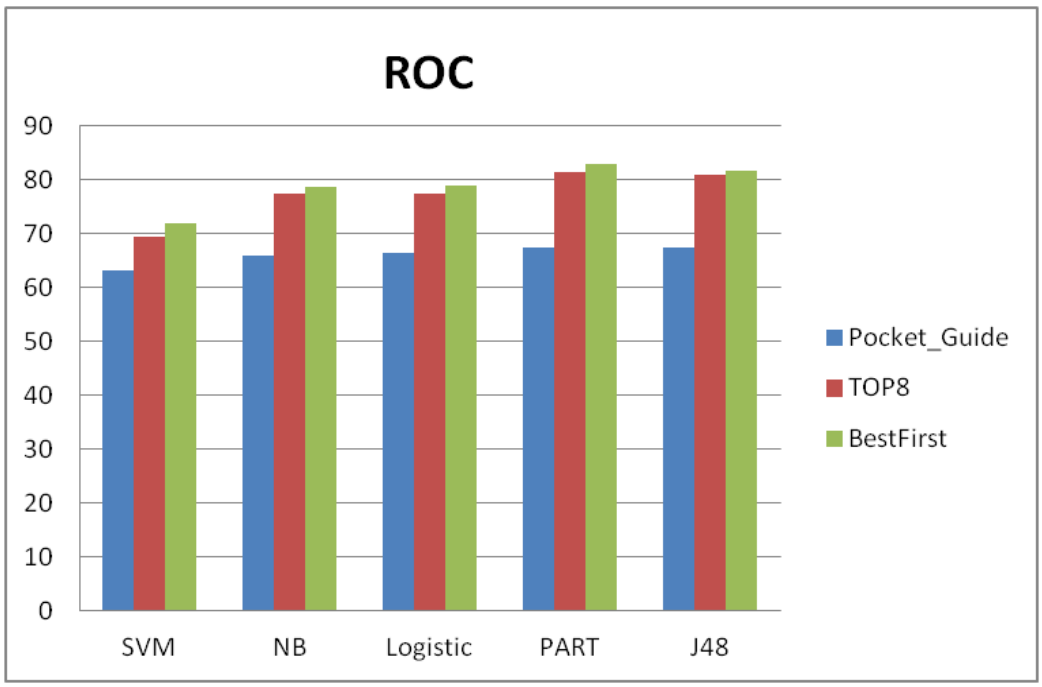


Figure 5.6: ROC for different methods.

mentioned measures.

Also, the top 8 features performed better than the HSF features and provided better accuracy, sensitivity and ROC.

The sensitivity and specificity gained from the pocket guide predictors were not balanced (high specificity and low sensitivity), but these two measures were balanced for the best-first attributes and the top 8 features.

5.5 Evaluation

As mentioned, the aim of this study was to find the most important signs/symptoms related to POCS to provide a model to distinguish POCS from other stroke-subtypes. We have reviewed and assessed the POCS signs/symptoms identified by HSF pocket-guide in previous section. We showed that the attributes selected in our experiment (using the BestFirst algorithm) provided better results than attributes reported by HSF. In this section, we evalute our work.

As mentioned, we had three sets of data: set (A) contained the predictors introduced by the HSF pocket-guide, set (B) contained the BestFrst attributes and set (C) included the top 8 features. In order to evaluate the predictors, we used a balanced training set and an imbalanced test set for set A and C that had equal number of attributes.

We set aside 2/3 of data for training (part of a balanced dataset) and 1/3 was used for testing; we used part of the imbalanced data that was not used during training. We applied different classifiers on each set and compared the results gained from set A (containing HSF predictors) and set C (containing the top 8 features). The results

gained from evaluation datasets are demonstrated in Table 5.4 and 5.5 respectively.

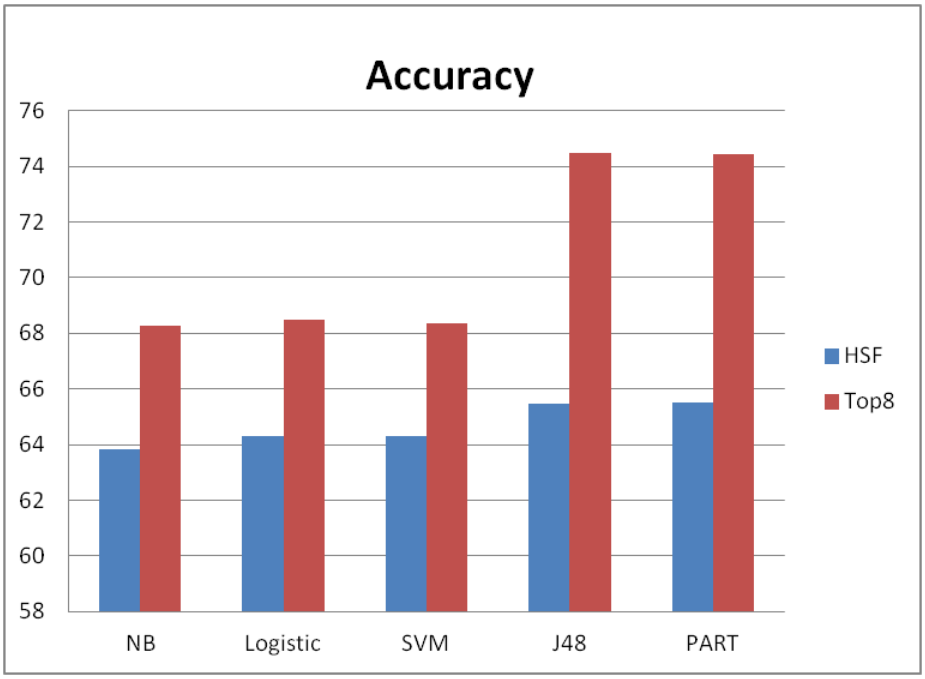


Figure 5.7: Accuracy for different methods.

We also demonstrated the metrics in separate graphs to show the differences between each set. Figure 5.7 and Figure 5.9 show that accuracy and sensitivity are improved significantly with Top8 attributes.

Based on Table 5.4 and 5.5, PART and J48 provided the highest accuracy in both sets. The top8 attributes provided about 74.5% accuracy whereas the highest accuracy provided by HSF predictors was 65.5% with same classifiers.

Comparing the sensitivity in both Tables (5.4, 5.5) and Figure 5.9 shows that all the classifiers provided almost the same sensitivity (44%) with HSF dataset. The sensitivity provided by HSF attributes was less than the lowest sensitivity provided

Table 5.4: Results gained from evaluation datasets (using HSF attributes).

Pocket Guide (HSF)	PART	J48	Naive Bayes	Logistic Regresion	SVM
Accuracy	65.50%	65.47%	63.82%	64.31%	64.31%
Sensitivity	44%	44%	44%	44%	44%
Specificity	86.70%	86.70%	83.40%	84.40%	84.40%
ROC	70.40%	70.40%	67.80%	68.20%	68.20%

Table 5.5: Results gained from evaluation datasets (using top8 attributes).

Top 8 Attributes	PART	J48	Naive Bayes	Logistic Regresion	SVM
Accuracy	74.43%	74.49%	68.26%	68.49%	68.35%
Sensitivity	89%	85.30%	56%	65.10%	64.20%
Specificity	60%	63.80%	80.40%	71.80%	72.40%
ROC	81.50%	80.10%	75.70%	76.40%	68.30%

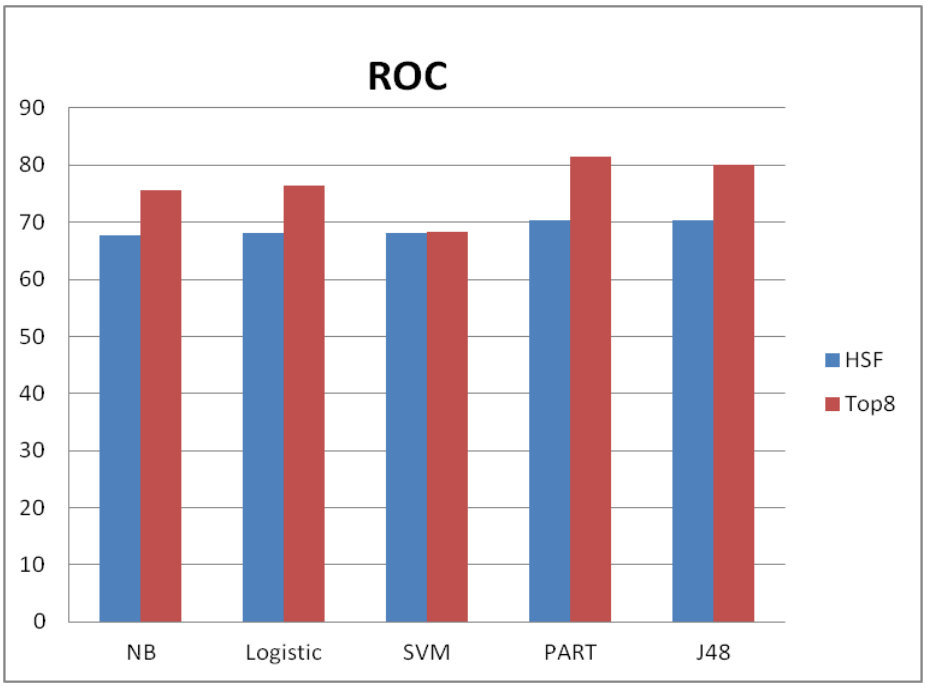


Figure 5.8: ROC for different methods.

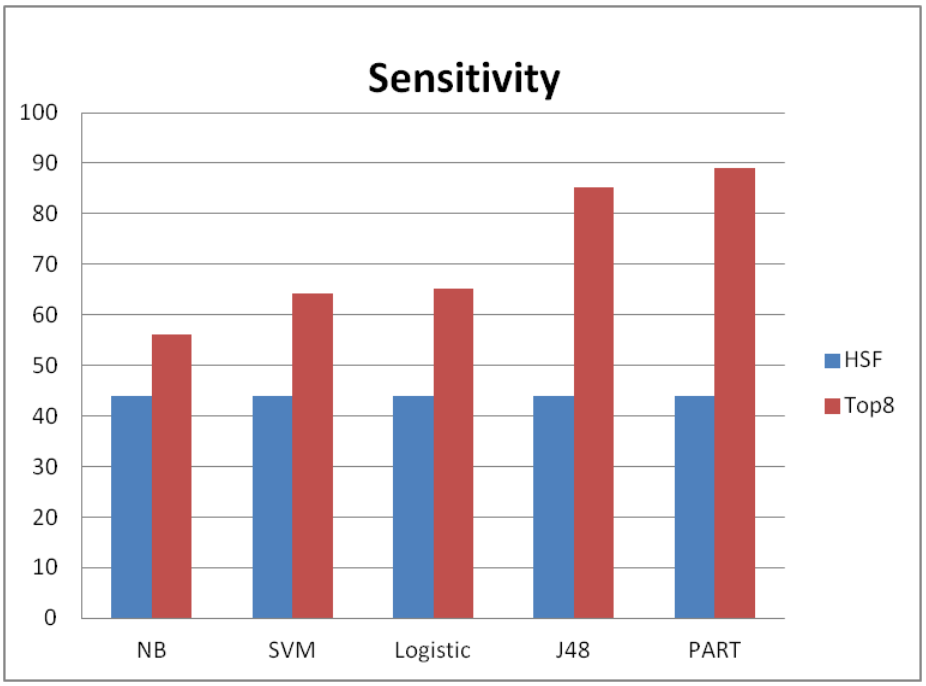


Figure 5.9: Sensitivity for different methods.

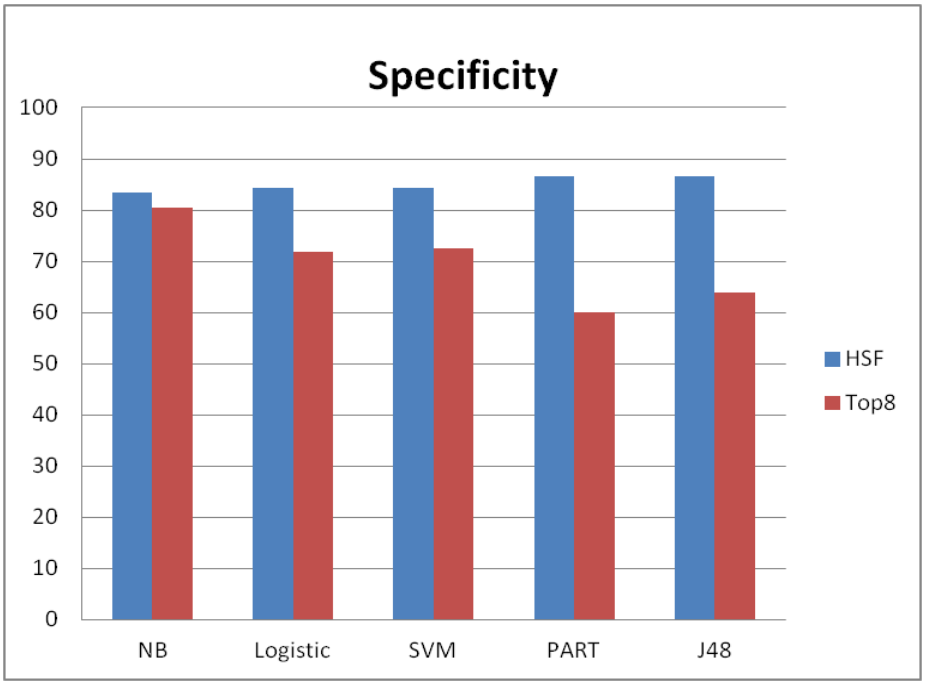


Figure 5.10: Specificity for different methods.

by Top8 attributes (56%). The highest sensitivity obtained from top8 attributes was 89% which was provided by PART.

Comparing HSF and top8 results in Table 5.4 and 5.5 shows that all classifiers (except SVM) provided a better ROC with the top8 dataset. The ROC provided by SVM classifier is almost the same in both sets. The ROC gained from each dataset is demonsterated in Figure 5.8.

Comparing the results in Table 5.4 and 5.5 and reviewing Figure 5.10 shows HSF attributes provided better specificity. Also this result shows that sensitivity and specificity gained from the HSF predictors were not balanced (high specificity and low sensitivity), but these two measures were balanced for the top8 features for all classifiers.

Overall, the top8 attributes provided better accuracy, sensitivity and ROC than HSF attributes for all classifiers. The classifiers performed better with top8 predictors than the HSF predictors to distinguish POCS from other stroke-subtypes.

How predictive is the model we learned? We know that both training and test data are representative samples of the underlying problem, but the test set contains independent instances that have played no part in formation of classifier [53]; therefore, we used test data for error estimation. We calculated 90% and 95% confidence interval for different classifiers and the results are demonstrated in table 5.6 and 5.7.

Table 5.6: With 90% confidence

	Part	J48	Naive	Logistic	SMO
N=1487	1487	1487	1487	1487	1487
S=920	920	980	1183	1083	1088
Mean p=S/N	0.62	0.66	0.8	0.73	0.73
q=1-p	0.38	0.34	0.2	0.27	0.27
variance(pq/n)	0.049	0.047	0.04	0.044	0.044
p+	0.70085	0.73755	0.866	0.8026	0.8026
p-	0.53915	0.58245	0.734	0.6574	0.6574

Table 5.7: With 95% confidence

	Part	J48	Naive	Logistic	SMO
N=1487	1487	1487	1487	1487	1487
S=920	920	980	1183	1083	1088
Mean p=S/N	0.62	0.66	0.8	0.73	0.73
q=1-p	0.38	0.34	0.2	0.27	0.27
variance(pq/n)	0.049	0.047	0.04	0.044	0.044
p+	0.718	0.754	0.88	0.818	0.818
p-	0.522	0.566	0.72	0.642	0.642

Let S be the success rate (the class of instance that is predicted correctly) and N the number of records in the test set (here N=1488). Let S/N be the random variable for the success rate. Let the true probability of success be p and the true probability of error be q=1-p [53].

When N is big, the probability distribution of the random variable f=S/N is approximated by a normal distribution with mean p and variance pq/N [53]. The probability that a random variable X, with 0 mean, lies within a certain confidence range of width 2z is Pr[- z$\leq X \leq z$] = c
(Called c% confidence interval)
To calculate 90% confidence interval, we used Pr[- 1.65$\leq X \leq$ 1.65] = 90%
To calculate 95% confidence interval, we used Pr[- 2$\leq X \leq$ 2] = 95%
Based on Table 5.6, with a 90% confidence we have that the true success rate p of the classifier (Naive Bayes) will be $0.734 \leq p \leq 0.866$
To calculate the 95% confidental interval, the following formulas are used:

$$P = \frac{S}{N} + 2.\sqrt{\frac{1}{N}.\frac{S}{N}.(1 - \frac{S}{N})}$$

And

$$P = \frac{S}{N} - 2.\sqrt{\frac{1}{N}.\frac{S}{N}.(1 - \frac{S}{N})}$$

Based on Table 5.7, with a 95% confidence we have that the true success rate p of the classifier (Naive Bayes) will be $0.72 \leq p \leq 0.88$

Statistical significance is attained when a p-value is less than the significance level

Table 5.8: p-value for Top8

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.247524303				
R Square	0.06126828				
Adjusted R Square	0.056187189				
Standard Error	0.181801734				
Observations	1487				
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	8	3.188340333	0.398542542	12.05809346	8.48037E-17
Residual	1478	48.85066437	0.03305187		
Total	1486	52.03900471			

(denoted $\alpha, alpha$)[23].

Fisher suggested a probability of one in twenty (0.05) as a convenient cutoff level to reject the null hypothesis; therefore the significance level (e.g. 0.05), which is also called alpha, be set ahead of time, prior to any data collection [46].

In this study, the p-value calculated for top8 and HSF attributes are both less than alpha. The p-value calculated for top8 is equal to 8.48 x 10 raised to the -17 power

Table 5.9: p-value for HSF

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.169337945				
R Square	0.02867534				
Adjusted R Square	0.023401998				
Standard Error	0.18486837				
Observations	1487				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	1.492236141	0.186529518	6.237553614	5.88294E-08
Residual	1479	50.54676857	0.034176314		
Total	1487	52.03900471			

(Table 5.8). The p-value calculated for HSF is equal to 5.8 x 10 raised to the -8 power (Table 5.9). The differences between these two amounts shows that the differences between the top8 and the pocket guide (HSF) are statistically significant.

5.6 Conclusion

Posterior circulation transient ischemic attacks (POCS) are more difficult to diagnose than anterior circulation ischaemia. Delayed or incorrect diagnosis of POCS may have devastating consequences including death or severe disability. Early recognition of posterior circulation stroke may save lives; therefore, stroke prevention strategies specific to the posterior circulation are important areas of active research.

The aim of this study was to find the most important signs/symptoms related to POCS to create a model to distinguish POCS from other stroke-subtypes. We analyzed the reports provided by stroke nurses and neurologists at Victoria General Hospital (VGH) and extracted 168 sings/symptoms. In order to extract the most important signs and symptoms (that might be helpful to detect POCS), we employed data mining methods. We also reviewed the signs/symptoms reported by the Heart and Stroke Foundation (HSF) of Canada and compared our finding with theirs.

The signs and symptoms reported by HSF for POCS were as follows: weakness, vertigo, ataxia, dysarthria, nausea, vomiting, hemianopsia, and sensory. We employed the BestFirst filtering method and extracted the most important signs/symptoms (22 attributes) from the VGH reports. We derived the top-8

attributes (based on their weights) and applied different classifiers over the HSF attributes and top-8 features.

Comparing the results gained from the both sets (HSF attributes and top8 features) showed that top8 features performed better in accuracy, sensitivity and ROC. For evaluation, we set aside 2/3 of data for train (part of a balanced dataset) and 1/3 was used for test (we used part of imbalanced data that was not used during training). We applied different classifiers on each set and again, the top8 attributes provided better accuracy, sensitivity and ROC than HSF attributes for all classifiers.

We showed that classifiers performed better with top-8 predictors than the HSF predictors to distinguish POCS from other stroke-subtypes. The list of top-8 features found in this study were Unsteadiness, Vertigo, Ataxia, Diplopia, Visual Disturbance, Dizziness, Nausea and Language Disturbance. Only 3 of them (Nausea, Ataxia, Vertigo) were common with the HSF predictors. Finally we showed that the differences between the top8 and the pocket guide (HSF) are statistically significant.

Chapter 6

Conclusions

Medical data is often represented in semi-structured or unstructured format. NLP is helpful to extract information within clinical narrative text and transform unstructured texts to data in a machine interpretable format. After preprocessing with NLP, data mining techniques are helpful in analyzing and interpreting data and can be used to create appropriate models for predicting disease based on signs and symptoms.

There were several challenges that we have addressed in our research. One of the main challenges in this work was data pre-processing and implementing appropriate negation detection rules. We have reviewed different negation detection methods, adopted existing methods to fit our problem context and implemented several rules that could determine whether each sign/symptom is present, absent or unmentioned. The quality of negation detection rules were evaluated and the results gained from automated negation rules were compared against the results provided manually. The expert used Kappa statistics and the level of agreement between the negation detection rules and human evaluation was 0.92.

After pre-processing, data mining algorithms were utilized to analyze the data and build models for predicting stroke or TIA in patients. The unstructured texts narratives were transformed to codified data in a computable format and data mining methods were utilized to build models for stroke/TIA prediction. Various algorithms were utilized on codified data and compared against baselines on raw data.

We systematically evaluated different data mining algorithms and computed standard metrics. A crucial product of the prediction models we learned from codified data was a list of keywords weighted by their importance in the prediction quality (as captured by sensitivity, specificity, etc). The top keywords (typically less than 30) of the list were usually responsible for more than 95% of the prediction quality. Having the top keywords allowed building of a questionnaire-like, online application for the triage staff to use. This was effective because the number of the top keywords was small. The backend part of the online application is a prediction model, which outputs the classification of mimic or stroke/TIA. Based on this output, the triage staff can better assess whether the patient needs to be hospitalized or can be discharged.

We showed that classifiers can provide reliable models to predict stroke/TIA based on patients signs and symptoms instead of immediately using costly tests, such as MRI or CT scan. We also showed that the list of signs/symptoms that play the most important role in stroke detection can be identified via the machine learning process and this list can be used in stroke assessment forms to help stroke nurses to decide on the next step of treatment in a more timely fashion. The experiment,

evaluation and results are all available in detail in Chapter 3.

Detecting migraine from stroke was another challenge due to many common signs and symptoms. It is important to consider the cost of hospitalization and the time spent by neurologists and stroke nurses to visit, diagnose, and assign appropriate care to the patients; therefore, devising new ways to distinguish stroke, migraine and other types of mimics can help in saving time and cost, and improve decision-making.

In this part of the research, first, different sources of data were preprocessed and balanced datasets were generated; second, attribute selection algorithms were used to reduce the dimensionality of the data; and third, a novel combination of data mining algorithms was employed in order to effectively distinguish migraine from other cases.

A novel classification method was implemented by merging different structured and unstructured data-sources. We extracted knowledge from the combination of high-level reports (provided by neurologists) and reports provided by stroke nurses. After a careful combination of these sources, we obtained a highly imbalanced dataset where the migraine cases were only about 6% of the dataset.

We implemented two classification nodes: node 1 that took migraine and stroke cases as input and distinguished migraine from stroke, and node 2 that took migraine and no-migraine-no-stroke cases as input and distinguished migraine from other types of stroke-mimic. In each node, the most important features (predictors) were extracted and the final model was generated based on the combination of

predictors found in nodes 1 and 2.

Using the dataset in its original form to build classifiers led to a learning bias towards the majority class and against the minority (migraine) class. One of the main challenges in this work was to overcome the rarity problem because of the highly imbalanced dataset. To rectify this issue, a sampling method was utilized.

We generated 12 balanced chunks (via sampling) for extracting best attributes in distinguishing migraine from stroke, and six balanced chunks for extracting best attributes in distinguishing migraine from other (non-stroke) mimics. The cross-validation results on these balanced chunks were much improved compared to the baseline of using imbalanced data. We achieved a sensitivity and specificity of about 80% and 75% respectively, which was in contrast to a sensitivity and specificity of 15.7% and 97% when using the original imbalanced data for building classifiers.

To evaluate the correctness of our model, two sets of balanced train and imbalanced test were generated (The test set was imbalanced reflecting the real distribution of instances). We compared the final evaluation result with the result obtained using balanced datasets and balanced test sets (as derived by cross-validation). Comparing the numbers indicates that results on evaluation data are close to the results on the balanced data using cross-validation. A sensitivity above 72%, a specificity above 69%, and a ROC area above 80% were the acceptable levels for clinicians that both results met the criteria and all the classifiers met these three conditions.

In addition, the cost-sensitive method was examined with different algorithms on the imbalanced dataset (the cost-sensitive method assigns misclassification costs to data from each class and forces the classifier to concentrate on the minority class). This method provided a good result, but the sensitivity did not meet the clinicians condition.

We showed that the model provided by the sampling method met the clinicians condition and the final model based on the selected attributes was quite robust performing well not only on balanced data, but more importantly performing almost equally well on imbalanced test data, which accurately reflect reality.

We should mention that typically migraine is distinguished from stroke and other non-stroke mimics via different types of tests such as CBC, Facial X-ray, CT scan, MRI, and EEG, which are all costly. In our research, we generated an effective model for migraine detection based on structured and unstructured data. This work is novel and can be used in decision support systems and save time for GPs/family doctors and nurses plus reduce costs for the health-care system. Details related to this part of the research are available in Chapter 4.

Another contribution of this research was finding the most important predictors that can help to detect and prevent Posterior circulation stroke. Posterior circulation transient ischemic attacks (POCS) are more difficult to diagnose than anterior circulation ischaemia. Delayed or incorrect diagnosis of POCS may have devastating consequences including death or severe disability. Early recognition of posterior circulation stroke may save lives; therefore, stroke prevention strategies specific to the posterior circulation are important areas of active research.

The aim of this part of the research was to find the most important signs/symptoms related to POCS to create a model to distinguish POCS from other stroke-subtypes. We analyzed the combination reports provided by stroke nurses and neurologists at Victoria General Hospital (VGH) and extracted 168 sings/symptoms. In order to extract the most important signs/symptoms, we employed data mining methods.

We also reviewed the signs/symptoms reported by the Heart and Stroke Foundation (HSF) of Canada and compared our finding with theirs. The signs and symptoms reported by HSF for POCS were: weakness,vertigo, ataxia, dysarthria, nausea, vomiting, hemianopsia, and sensory. We employed the BestFirst filtering method and extracted 22 attributes as the most important signs/symptoms from the VGH reports. We derived the top-8 attributes (based on their weights) and applied different classifiers over the HSF attributes and top-8 features.

Comparing the results gained from the both sets (HSF attributes and top8 features) showed that top8 features performed better in accuracy, sensitivity and ROC. The list of top-8 features were Unsteadiness, Vertigo, Ataxia, Diplopia, Visual Disturbance, Dizziness, Nausea and Language Disturbance. Only 3 of them (Nausea, Ataxia, Vertigo) were common with the HSF predictors.

For evaluation, we set aside 2/3 of data for train (part of a balanced dataset) and 1/3 was used for test (we used part of imbalanced data that was not used during training). We applied different classifiers on each set and again, the top8 attributes provided better accuracy, sensitivity and ROC than HSF attributes for all classifiers.

We showed that classifiers performed better with top-8 predictors than the HSF predictors to distinguish POCS from other stroke-subtypes. Finally we showed that the differences between the top8 and the pocket guide (HSF) are statistically significant.

In summary, we addressed the following challenges:

- We analyzed unstructured text to provide a model for stroke prediction, thus providing a first inexpensive screening for stroke/TIA detection prior to confirmation using MRI or CT scan. In addition, we introduced a new approach for data preparation and implemented a novel method for negation detection.

- We presented a detailed study using supervised machine learning to predict stroke/TIA vs. mimic based on visit descriptions and methodically compare several algorithms across a multitude of dimensions.

- A novel classification method was implemented by combining different data mining(classification) algorithms. More specifically, the algorithms were organized in a directed acyclic graph. We introduced a new approach to overcome the rarity problem when having highly imbalanced datasets with appropriate sensitivity and specificity result. In addition, we introduced a novel and cheap method to detect migraine.

- We found the most important attributes that can help to detect and prevent Posterior circulation stroke. We compared our fining with features reported by the Heart and Stroke Foundation (HSF) of Canada. The results gained from the both sets, HSF attributes and our finding indicated that top8 features performed better in accuracy, sensitivity and ROC.

6.1 Future Research

Clinical reports are full of hidden information, and extracting useful knowledge from this unstructured data is hard and time consuming. Medical concept co-occurrences can be used by clinical decision support (CDS) systems, and knowledge of such relatedness can be useful for outcome prediction. Discovering information pertaining stroke sub-types and related injured parts of brain and body can be a potential future research that may help in preventing and predicting different types of stroke such as Partial Anterior Circulation Stroke (PACS), Lacunar Stroke Syndrome (LACS) and Total Anterior Circulation Syndrome (TACS). Research of this nature requires suitable amount of data from different resources.

Appendix A

List of Abbreviations

ABCD score Age, Blood Pressure, Clinical Features, Duration of TIA

ACVS ..... Acute Cerebrovascular Syndrome

AF ..... Atrial Fibrilation

ANGST .... Anexiety, Stress

Apache ..... Acute Physiology and Chronic Health Evaluation

ASA ..... Acetylsalicylic Acid

AUC ..... Area Under the Curve

BC ..... British Columbia

BPV ..... Benign Positional Vertigo

CAP ..... Community Acquired Pneumonia

CBC ..... Complete Blood Count

CDS ..... Clinical Decision Support

CHADS2 ... C (Congestive heart failure), H (Hypertension), A (Age), D (Diabetes mellitus), S2 (Prior Stroke or TIA)

CT scan .... Computed Tomography Scan

DM ..... Data Mining

DNA ..... Deoxyribonucleic Acid

EEG ..... Electroencephalogram

ER/ED ..... Emergency Room/Emergency Department

etc ..... et cetera

FN ..... False Negative

FP ..... False Positive

GP ..... General Practitioner

GUI ..... Graphical User Interface

HSF ..... Heart and Stroke Fundation

IBM ..... International Business Machines

ICU ..... Intensive Care Unit

ID ..... Identification

IE ..... Information Extraction

LACS ..... Lacunar Stroke Syndrome

MRI ..... Magnetic Resonance Imaging

NB ..... Naive Bayes

NLP ..... Natural Language Processing

PACS ..... Partial Anterior Circulation Stroke

POCS ..... Posterior Circulation Ischemic Stroke

POS ..... Part-Of-Speech

RBF Kernel Radial basis function kernel

REX ..... Regenstrief EXtraction Tool

ROC ..... Receiver operating characteristic

SAPS ..... Simplified Acute Physiology Score

SGS ..... Stroke Guidance System

SRAU ..... Stroke Rapid Assessment Unit

SVM ..... Support Vector Machine

TACS ..... Total Anterior Circulation Syndrome

TGA ..... Transient Global Amnesia

TIA ..... Transient Ischemic Attack

TIMI ..... Thrombolysis In Myocardial Infarction

TN ..... True Negative

TP ..... True Positive

VGH ..... Victoria General Hospital

VIHA ..... Vancouver Island Health Authority

WEKA ..... Waikato Environment for Knowledge Analysis

Bibliography

[1] Das-discussion: Information extraction. http://www.iapr-tc11.org/mediawiki/index.php/DASDiscussion:Information_Extraction.

[2] Alchemy language. https://alchemy -language -demo.mybluemix.net, 2016.

[3] The stanford parser: A statistical parser. http://nlp.stanford.edu/software/lexparser.shtml, 2016.

[4] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[5] Mohammad A Al-Haddad, Jeff Friedlin, Joe Kesterson, Joshua A Waters, Juan R Aguilar-Saavedra, and C Max Schmidt. Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. *HPB*, 12(10):688–695, 2010.

[6] Leila Amini, Reza Azarpazhouh, Mohammad Taghi Farzadfar, Sayed Ali Mousavi, Farahnaz Jazaieri, Fariborz Khorvash, Rasul Norouzi, and Nafiseh Toghianfar. Prediction and control of stroke by data mining. *International journal of preventive medicine*, 4(Suppl 2):245, 2013.

[7] Elliott M Antman, Marc Cohen, Peter JLM Bernink, Carolyn H McCabe, Thomas Horacek, Gary Papuchis, Branco Mautner, Ramon Corbalan, David

Radley, and Eugene Braunwald. The timi risk score for unstable angina/non–st elevation mi: a method for prognostication and therapeutic decision making. *Jama*, 284(7):835–842, 2000.

[8] Antonio Arauzo-Azofra, Jose Manuel Benitez, and Juan Luis Castro. Consistency measures for feature selection. *Journal of Intelligent Information Systems*, 30(3):273–292, 2008.

[9] David B Aronow, Feng Fangfang, and W Bruce Croft. Ad hoc classification of radiology reports. *Journal of the American Medical Informatics Association*, 6(5):393–411, 1999.

[10] Mordechai Averbuch, Tom Karson, Benjamin Ben-Ami, Oded Maimon, and Lior Rokach. Context-sensitive medical information retrieval. In *Proc. of the 11th World Congress on Medical Informatics (MEDINFO-2004)*, pages 1–8. Citeseer, 2004.

[11] Neil Barrett and Jens Weber-Jahnke. Building a biomedical tokenizer using the token lattice design pattern and the adapted viterbi algorithm. *BMC bioinformatics*, 12(Suppl 3):1, 2011.

[12] Ze-Hong Cao, Li-Wei Ko, Kuan-Lin Lai, Song-Bo Huang, Shuu-Jiun Wang, and Chin-Teng Lin. Classification of migraine stages based on resting-state eeg power. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–5. IEEE, 2015.

[13] Patricia Cerrito. Application of data mining for examining polypharmacy and adverse effects in cardiology patients. *Cardiovascular toxicology*, 1(3):177–179, 2001.

[14] PB Cerrito and JC Cerrito. Clinical data mining for physician decision making and investigating health outcomes: Methods for prediction and analysis. hershey, 2010.

[15] Dursun Delen, Glenn Walker, and Amit Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–127, 2005.

[16] Béatrice Duval, Jin-Kao Hao, and Jose Crispin Hernandez Hernandez. A memetic algorithm for gene selection and molecular classification of cancer. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 201–208. ACM, 2009.

[17] Jacob S Elkins, Carol Friedman, Bernadette Boden-Albala, Ralph L Sacco, and George Hripcsak. Coding neuroradiology reports for the northern manhattan stroke study: a comparison of natural language processing and manual review. *Computers and biomedical research*, 33(1):1–10, 2000.

[18] Mahyar Etminan, Bahi Takkouche, Francisco Caamaño Isorna, Ali Samii, et al. Risk of ischaemic stroke in people with migraine: systematic review and meta-analysis of observational studies. *Bmj*, 330(7482):63, 2005.

[19] M Fiszman, Wendy W Chapman, Scott R Evans, and Peter J Haug. Automatic identification of pneumonia related concepts on chest x-ray reports. In *Proceedings of the AMIA Symposium*, page 67. American Medical Informatics Association, 1999.

[20] Christopher M Florkowski. Sensitivity, specificity, receiver-operating characteristic (roc) curves and likelihood ratios: communicating the

performance of diagnostic tests. *The Clinical Biochemist Reviews*, 29(Suppl 1):S83, 2008.

[21] Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

[22] Carol Friedman, Lyudmila Shagina, Socrates A Socratous, and Xiao Zeng. A web-based version of medlee: A medical language extraction and encoding system. In *Proceedings of the AMIA Annual Fall Symposium*, page 938. American Medical Informatics Association, 1996.

[23] Lawrence M Friedman. Clinical significance versus statistical significance. *Encyclopedia of Biostatistics*, 2005.

[24] Brian F Gage, Amy D Waterman, William Shannon, Michael Boechler, Michael W Rich, and Martha J Radford. Validation of clinical classification schemes for predicting stroke: results from the national registry of atrial fibrillation. *Jama*, 285(22):2864–2870, 2001.

[25] Kavian Ghandehari, Farah Ashrafzadeh, Zahra Izadi Mood, Saeed Ebrahimzadeh, and Khatereh Arabikhan. Development and validation of the asian migraine criteria (amc). *Journal of Clinical Neuroscience*, 19(2):224–228, 2012.

[26] Justin M Glasgow and Peter J Kaboli. Detecting adverse drug events through data mining. *American journal of health-system pharmacy*, 67(4):317–320, 2010.

[27] Sergey Goryachev, Margarita Sordo, Qing T Zeng, and Long Ngo. Implementation and evaluation of four different methods of negation detection. *Boston, MA: DSG*, 2006.

[28] Government of British Columbia. Msc payment schedule index, neurology. http://www2.gov.bc.ca/assets/gov/health/practitioner-pro/medical-services-plan/msc-payment-schedule-june-2016.pdf, 2016.

[29] Heart and Stroke foundation. Statistics. http://www.heartandstroke.com/site/c.ikIQLcMWJtE/b.3483991/k.34A8/Statistics.htm, 2015.

[30] Kurt Hornik, Christian Buchta, Torsten Hothorn, Alexandros Karatzoglou, David Meyer, and Achim Zeileis. Rweka: R/weka interface. https://cran.r-project.org/web/packages/RWeka, 2016.

[31] George Hripcsak, John HM Austin, Philip O Alderson, and Carol Friedman. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports 1. *Radiology*, 224(1):157–163, 2002.

[32] Jason Brownlee. Feature selection to improve accuracy and decrease training time. http://machinelearningmastery.com/feature-selection-to-improve-accuracy-and-decrease-training-time/, 2016.

[33] Paul Jermyn, Maurice Dixon, and Brian J Read. Preparing clean views of data for data mining. *ERCIM Work. on Database Res*, pages 1–15, 1999.

[34] John D Kelleher and Brian Mac Namee. A review of negation in clinical texts: Dit technical report: Soc-aig-001-08. 2001.

[35] William A Knaus, Jack E Zimmerman, Douglas P Wagner, Elizabeth A Draper, and Diane E Lawrence. Apache-acute physiology and chronic health evaluation: a physiologically based classification system. *Critical care medicine*, 9(8):591–597, 1981.

[36] Li-Wei Ko, Kuan-Lin Lai, Pei-Hua Huang, Chin-Teng Lin, and Shuu-Jiun Wang. Steady-state visual evoked potential based classification system for detecting migraine seizures. In *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*, pages 1299–1302. IEEE, 2013.

[37] Abraham Kuruvilla, Pratik Bhattacharya, Kumar Rajamani, and Seemant Chaturvedi. Factors associated with misdiagnosis of acute stroke in young adults. *Journal of Stroke and Cerebrovascular Diseases*, 20(6):523–527, 2011.

[38] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

[39] Len Trigg. Class costsensitiveclassifier. http://weka.sourceforge.net/doc.dev, 2016.

[40] WS Lim, MM Van der Eerden, R Laing, WG Boersma, N Karalus, GI Town, SA Lewis, and JT Macfarlane. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax*, 58(5):377–382, 2003.

[41] MediResource. C.health, migraine (migraine headache). http://chealth.canoe.com/channel_condition_info_details.asp?disease_id=88, 2015.

[42] Áine Merwick and David Werring. Posterior circulation ischaemic stroke. *BMJ*, 348:g3175, 2014.

[43] Microsoft. Microsoft azure machine learning studio. https://azure.microsoft.com, 2016.

[44] Amir Navot. *On the role of feature selection in machine learning*. PhD thesis, Hebrew University, 2006.

[45] NIH. Stroke, hope through research. http://www.ninds.nih.gov/disorders/stroke/ detail_stroke.htm, 2015.

[46] Gerry P Quinn and Michael J Keough. *Experimental design and data analysis for biologists*. Cambridge University Press, 2002.

[47] JHC Ranson. Prognostic signs and the role of operative management in acute pancreatitis. *Surg Gynecol Obstet*, 139:69–81, 1974.

[48] Michael Regnier. Focus on stroke: Predicting and preventing stroke. http://blog.wellcome.ac.uk/2012/05/07/focus-on-stroke-predicting-and-preventing-stroke/.

[49] Elham Sedghi, Jens H Weber, Alex Thomo, Maximilian Bibok, and Andrew Penn. Mining clinical text for stroke prediction. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 4(1):1–9, 2015.

[50] Yanmin Sun, Mohamed S Kamel, and Yang Wang. Boosting for learning multiple classes with imbalanced class distribution. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 592–602. IEEE, 2006.

[51] TheMigraineTrust. Stroke and migraine. http://www.migrainetrust.org/factsheet-stroke-and-migraine-10891, 2015.

[52] The_R_Foundation. What is r? https://www.r-project.org/about.html.

[53] Alex Thomo. Data mining course - evaluation. www.engr.uvic.ca/ seng474/eval.ppt, 2009.

[54] Divya Tomar and Sonali Agarwal. A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5):241–266, 2013.

[55] Christophe Tzourio, Alain Tehindrazanarivelo, Serge Iglesias, Annick Alperovitch, Francois Chedru, Jacques d'Anglejan Chatillon, and Marie-Germaine Bousser. Case-control study of migraine and risk of ischaemic stroke in young women. *Bmj*, 310(6983):830–833, 1995.

[56] University of Waikato, New Zealand. Weka (machine learning). http://en.wikipedia.org/wiki/Weka(machine learning), 2014.

[57] VIHA. cost of hospital. http://www.viha.ca/visit/fees/hospital_fees.htm, 2016.

[58] G Viticchi, L Falsetti, M Silvestrini, S Luzzi, L Provinciali, and M Bartolini. The real usefulness and indication for migraine diagnosis of neurophysiologic evaluation. *Neurological Sciences*, 33(1):161–163, 2012.

[59] Pernille Warrer, Ebba Holme Hansen, Lars Juhl-Jensen, and Lise Aagaard. Using text-mining techniques in electronic patient records to identify adrs from medicine use. *British journal of clinical pharmacology*, 73(5):674–684, 2012.

[60] Mike Wasikowski and Xue-wen Chen. Combating the small sample class imbalance problem using feature selection. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1388–1400, 2010.

[61] WebMD. Tests for diagnosing migraines. http://www.webmd.com/migraines-headaches/migraine-diagnosing-tests, 2015.

[62] Gary M Weiss. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19, 2004.

[63] W Wendy. Chapman, will bridewell, paul hanbury, gregory f. cooper, and bruce g. buchanan. 2001. a simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[64] Wikipedia. Abcd score. http://en.wikipedia.org/wiki/ABCD_score, 2014.

[65] Wikipedia. Feature selection. https://en.wikipedia.org/wiki/Feature_selection, 2016.

[66] Wiktionary. Category:english words suffixed with -n't. http://en.wiktionary.org/, 2013.