UNIVERSITY OF NAIROBI
SCHOOL OF COMPUTING AND INFORMATICS

A DIAGNOSTIC PATTERN DISCOVERY AND PREDICTION MODEL USING A HYBRID OF C4.5 AND ASSOCIATION RULE ALGORITHM IN A STANDARDIZED ELECTRONIC MEDICAL RECORDS IMPLEMENTATION

By
Kevin Okoth Ben

P58/75805/2012

Supervisor
Mr. Eric Ayienga

Submitted in partial fulfillment of the requirements for the degree of Masters of Science in computer science in the school of computing and informatics, University of Nairobi
Declaration
This project, as presented in this report, is my original work and has not been presented for a degree in any other university.

Kevin Okoth Ben
P58/75805/2012

Signed---------------------------------

Date-----------------------------------

This thesis has been submitted for examination with my approval as university supervisor.
Mr. Eric Ayienga
School of computing & Informatics

Signed -------------------------------

Date -------------------------------
Dedication
To all my best friends and co-workers Shadrack Onchaba, Daniel and Gladys who have kept me going throughout this project, their individual effort shall not go unnoticed for standing in for me when they were needed.

To my parents who have demanded more from me, supported and encouraged me, and to the almighty God, I will forever remain grateful.
ACKNOWLEDGEMENT

First and foremost I say thank you to school of Computing & Informatics for granting me the opportunity to pursue this course at this prestigious institution.

I would also like to thank my supervisor for the support in spite of his busy schedule, the panel here known as the judges, my gratitude for leading me all the way. Lastly I would like to thank all the medics who supported me whenever their expertise were required.
ABSTRACT

Data mining technologies have been used extensively in the commercial retail sectors to extract data from their “big data” warehouses. In healthcare, data mining has been used as well in various aspects which we explored. The voluminous amounts of data generated by medical systems form a good basis for discovery of interesting patterns that may aid decision making and saving of lives not to mention reduction of costs in research work and possibly reduced morbidity prevalence. It is from this that we set out to implement a concept using a hybrid of C4.5 and Apriori association rule mining technology to find out any possible diagnostic associations that may have arisen in patients’ medical records spanning across multiple contacts of care. The dataset was obtained from Practice Fusion’s open research data that contained over 98,000 patient clinic visits from all American states.

The research and prototype focuses majorly on development of an efficient and accurate hybrid algorithm out of the combination of C4.5 and the Apriori Algorithms. With the hybrid prototype, we were able to mine for patterns arising from medical diagnosis data. The diagnosis data was based on ICD-9 coding and this helped limit the set of possible diagnostic groups for the analysis. We then subjected the results to domain expert opinion. The panel of experts validated some of the most common associations with concurrence of 90% whereas others elicited debate amongst the medical practitioners. The results of our research showed that the hybrid of Apriori and C4.5 algorithms is more accurate, robust, efficient and effective and can be used to confirm what is already known from health data in form of comorbidity patterns while generating some very interesting disease diagnosis associations that can provide a good starting point and room for further exploration through studies by medical researchers to explain the patterns that are seemingly unknown to the concerned populations.
Table of Contents

DECLARATION .................................................................................................................................................. II

DEDICATION .................................................................................................................................................... III

ACKNOWLEDGEMENT ....................................................................................................................................... IV

ABSTRACT ............................................................................................................................................................ V

TABLE OF FIGURES ........................................................................................................................................... VIII

ACRONYMS ........................................................................................................................................................ IX

DEFINITION OF KEY TERMS ................................................................................................................................ X

CHAPTER ONE ..................................................................................................................................................... 1
  1.1 INTRODUCTION ............................................................................................................................................ 1
  1.2 BACKGROUND ............................................................................................................................................. 2
  1.3 PROBLEM STATEMENT ................................................................................................................................. 4
  1.4 SIGNIFICANCE OF THE STUDY .................................................................................................................. 4
  1.5 GOAL OF THE STUDY ................................................................................................................................... 5
  1.6 OBJECTIVES OF THE STUDY ....................................................................................................................... 5
    1.61 Project objective ........................................................................................................................................ 5
    1.62 System objectives ..................................................................................................................................... 5
  1.7 ASSUMPTION AND SCOPE OF THE RESEARCH ...................................................................................... 6
  1.8 CONCEPTUAL MODEL ............................................................................................................................... 7

CHAPTER TWO-LITERATURE REVIEW ........................................................................................................... 8
  2.1 EMR SYSTEMS ADOPTION IN THE KENYAN AND GLOBAL CONTEXT ......................................................... 8
    2.11 The Kenyan context .................................................................................................................................. 9
    2.12 The Global Context .................................................................................................................................. 9
  2.2 MAJOR STANDARDS ...................................................................................................................................... 10
    2.2.1 International Statistical Classification of Disease (ICD) ....................................................................... 10
    2.2.2 Health Level 7 standards (HL7) ............................................................................................................ 11
  2.3 ASSOCIATION RULE MINING AND THE APRIORI ALGORITHM .................................................................. 11
  2.4 HYBRID OF ASSOCIATION RULE ALGORITHM (APRIORI) AND THE C4.5 ........................................... 14
  2.5 ASSOCIATION RULE MINING AND C4.5 IN HEALTH CARE ..................................................................... 15
  2.6 RELATED WORK .......................................................................................................................................... 16
  2.7 LIMITATIONS AND RESEARCH GAPS ...................................................................................................... 17

CHAPTER THREE ............................................................................................................................................... 18
  3.0 METHODOLOGY ......................................................................................................................................... 18
  3.1 RESEARCH DESIGN ..................................................................................................................................... 19
  3.2 PROBLEM IDENTIFICATION AND SELECTION ....................................................................................... 19
  3.3 LITERATURE REVIEW AND CONCEPT DEVELOPMENT .......................................................... 19
  3.4 DATA COLLECTION, PREPARATION AND PROCESSING ........................................................................ 20
  3.5 PROTOTYPE DEVELOPMENT .................................................................................................................... 21
  3.6 PROTOTYPE TESTING AND IMPLEMENTATION ...................................................................................... 21
  3.7 ANALYSIS AND PRESENTATION OF RESULTS ....................................................................................... 22

CHAPTER IV .................................................................................................................................................... 23
  4.0 SYSTEM ANALYSIS AND DESIGN ........................................................................................................ 23
  4.1 KEY SYSTEM PROTOTYPE ELEMENTS .................................................................................................. 24
    4.1.1 Input .......................................................................................................................... 24
Table of Figures
Figure 1: Conceptual Model ........................................................................................................... 7
Figure 2: section of the raw dataset .............................................................................................. 20
Figure 3: illustration of research context ...................................................................................... 23
Figure 4: Database Schema .......................................................................................................... 25
Figure 5: Patient’s Demographics records table(Training syncDiagnosis). .................... 26
Figure 6: Clinic visits per patients table (Processed ICD9 Codes per Patient). .................. 26
Figure 7: Rules Table ................................................................................................................... 26
Figure 8: Decisiontree_c45transdata table, Decision tree transactional data ................. 27
Figure 9: RulesModified table ..................................................................................................... 27
Figure 10: Resultant association ................................................................................................. 28
Figure 11: Hybrid algorithm Model .............................................................................................. 29
Figure 12: Login screen .............................................................................................................. 30
Figure 13: Landing page .............................................................................................................. 30
Figure 14: Resultant page before user input .............................................................................. 31
Figure 15: Hybrid Screen, After user input .............................................................................. 31
Figure 16: Sample association rules ........................................................................................... 33
Figure 17: ICD9 Descriptions ....................................................................................................... 33
Figure 18: Likert scale scores ..................................................................................................... 34
Acronyms
CCR Continuity of Care Record
CDC Center for Disease Control
CDO Care Delivery Organization
EHR Electronic Health Records
EMR Electronic Medical Records
HIS Health Information System
HITECH Health Information Technology for Economic and Clinical Health Act (2009)
HL7 Health Level 7 International
ICD International Statistical Classification of Diseases
ISO International Organization for Standardization
WHO World Health Organization
Definition of Key Terms

**Multimorbidity** - The co-occurrence of multiple chronic or acute diseases and medical conditions within one person without any reference to an index condition. (Van den Akker et al., 1998). Also comorbidity.

**Standardized EMR** - In our context, this refers to an EMR that complies with the guidelines developed by the Kenya government contained in the document *(Standards and Guidelines for Electronic Medical Record Systems in Kenya, 2009)*, as well as other guidelines particularly those of the world health governing body WHO.

**Differential Diagnosis**- is a systematic diagnostic method used to identify the presence of an entity where multiple alternatives are possible (and the process may be termed differential diagnostic procedure), and may also refer to any of the included candidate alternatives (which may also be termed candidate condition) (“Differential diagnosis,” 2014).
CHAPTER ONE

1.1 Introduction

EHR systems are designed to store data accurately and to capture the state of a patient across time. It eliminates the need to track down a patient's previous paper medical records and assists in ensuring data is accurate and legible. It can reduce risk of data replication as there is only one modifiable file, which means the file is more likely up to date, and decreases risk of lost paperwork. Due to the digital information being searchable and in a single file, EMR's are more effective when extracting medical data for the examination of possible trends and long term changes in a patient.

In recent times however, world governing institutions like WHO and ISO have embraced the advent of Health Information Systems (HIS) and spearheaded the development of standards that were hitherto unavailable to implementers of health systems. These standards make it easy not only to capture and share data across multiple and seemingly disparate implementations, but to also query, analyze and extract useful statistics from data entered in the same systems.

Data mining technologies have been used extensively particularly in the commercial retail sectors to extract data from their “big data” warehouses. In healthcare, data mining has been used as well in various aspects which we will explore later. The voluminous amounts of data generated by these systems form a good basis for discovery of interesting patterns that may aid decision making and saving of lives not to mention reduction of costs in research work and possibly reduced morbidity prevalence. It is from this that we seek to implement a concept using a hybrid of association rule mining (ARM) and C4.5 technology to find out any possible diagnostic associations that may arise in patients’ medical records spanning across multiple contacts of care (visits).
1.2 BACKGROUND

The term EMR stands for Electronic Medical Records. In their work, “Electronic Medical Records vs. Electronic Health Records: Yes, There Is a Difference”, (Garets and Davis, 2006) define an EMR system as an application environment composed of the clinical data repository, clinical decision support, controlled medical vocabulary, order entry, computerized provider order entry, pharmacy, and clinical documentation applications. This environment supports the patient’s electronic medical record across inpatient and outpatient environments, and is used by healthcare practitioners to document, monitor, and manage health care delivery within a care delivery organization (CDO). The data in the EMR is the legal record of what happened to the patient during their encounter at the CDO and is owned by the CDO. This is to be differentiated with Electronic Health Record systems (EHR) which they again define as a subset of each care delivery organization’s EMR, presently assumed to be summaries like ASTM’s Continuity of Care Record (CCR) or HL7’s Continuity of Care Document (CCD), is owned by the patient and has patient input and access that spans episodes of care across multiple CDOs within a community, region, or state (or in some countries, the entire country). The terms are often used interchangeably though the difference, subtle as it may seem, may be of particular significance in this research.

EMRs have been in use in several countries by different health facilities over the years but standardization of the different electronic medical records implementation has been left as an individual task for different governments to pursue. The US, research has shown, had been lagging behind other (particularly Scandinavian countries) in the adoption of EHRs (Schoen et al., 2009). This is changing as they have aggressively embarked with the implementation of the Health Information Technology for Economic and Clinical Health Act (HITECH) of 2009, which provides $27 billion over 10 years for federal incentive payments to hospitals and clinicians for adopting EHRs (Gray et al., 2011). The use of certified electronic health records (EHR) and pertinent objectives to be achieved over several stages are known as “Meaningful Use”.

Closer home, the government of Kenya openly admits to challenges of obtaining health data, due to the weak health information infrastructure, a poor information culture that does not spur demand for information, multiple and parallel information systems, a thin and stretched human
resource to support data collection, transformation, presentation and archiving among others. This is in its health information policy and strategic plan (Standards and Guidelines for Electronic Medical Record Systems in Kenya, 2009). The Division of Health Information Systems (HIS), in this policy document continues to say: “It is with this background that the ministries through the Division of Health Information system (HIS) undertook to develop a health information policy and strategic plan (2009- 2014) to guide the health information strengthening agenda in the country. In its Strategic Plan, the HIS has planned to improve data management and strengthen the use and application of information technology in data management. To effectively do this, there is need to develop standards that will ensure quality of software, compatibility of data sharing, ease of maintenance and common understanding among the workforce.

Data complexity, volumes of patients served and the desire to have efficient health information systems have defined the need for Electronic Medical Record (EMR) Systems in Kenya. EMR systems, when well developed and implemented, can improve the process of data collection resulting in better quality and more reliable health information. These systems can also greatly improve aggregation and reporting of data from facilities. EMR systems support provision of health care through the integrated clinical decision support functions and by ensuring that patient information is available across facilities for continuity of care.”

The policy document goes ahead to lay a regulatory framework that is based on international standards from institutions such as WHO, ISO and CDC. The standards that will be of particular interest in our research are the International Classification of Diseases (ICD) standards, and HL7 health information interchange standards.

This only demonstrates the significance of standardized EMR systems and the evolutionary role they are likely to play not only in the world in general but in Kenya in particular.

It is with this standardized data capture and storage, that there emanates useful data which can not only be analyzed but that can also have useful patterns discovered that could aid
governments and medical practitioners alike in improving healthcare services to the public and their patients respectively.

1.3 Problem statement

Medical practitioners generate data with a wealth of hidden and potentially useful information present and it is not properly being used effectively for predictions or for any medical gain, therefore there is a need to employ adequate methods to enable the realization of full benefits of such data.

Using a combination of Association rule mining and C4.5 algorithms, we unraveled the hidden diagnosis patterns that could be present within such data and use the information gained to help in a disease prediction. The algorithms were combined on a strength basis with focus on efficiency, accuracy and on overcoming their individual weaknesses.

1.4 Significance of the study

The research is based on a model that can be replicated across multiple EMR implementations as long as they adhere to the stipulated standards and have the same output format. This means that we have the ability to apply this technique to data that could span to say the entire continent in a bid to discover hidden diagnostic patterns. The newly discovered strong associations shall be utilized in clinical diagnosis to predict multiple related targets attributes for a specified disease.

Based on the newly discovered strong associations, we are able to know which diseases tend to appear together from amongst the patients, this can help the government place intervention measures in advance. This can include putting in place health measures requiring pathological tests for a certain disease given that another closely associated one has been diagnosed.

Based on the findings, policy makers can also focus on conducting health campaigns for certain diseases with the knowledge that the success of such will essentially have a certain related effect on the other associated diseases.
It also helps the health industry to finally take advantage of years’ worth of input, in the sense that it is possible to utilize multiple sources of medical data to aid in decision making given certain morbidity patterns.

1.5 **Goal of the Study**

To build and evaluate an efficient, fast and easy diagnostic pattern discovery and a prediction model using a combination of C4.5 and association rule algorithms, that when passed to patients’ past medical records then the model predicts the patient’s vulnerability factor, with these knowledge individuals can easily take it as an early warning and therefore do a health work around to prevent the disease from occurring in real life. The model can also be used as an early detection tool.

1.6 **Objectives of the Study**

1.6.1 Project objective

i. Identify the most efficient prediction techniques that can be derived from a combination of C4.5 and Association rule algorithms to help in discovery of trends in patients’ data.

ii. Determine a prediction methodology that can design a model for inference of the characteristics of a predicted class from a combination of other data

iii. Identify data mining classification Techniques to generate and discover strong rules (relationship) and patterns of the risk factors.

iv. Match these newly discovered associations to patients’ demographics and extract new knowledge from them.

1.6.2 System objectives

i. To adopt and design a combination of C4.5 and the association rule algorithm to patient diagnostic data.
Research Questions

At the end of the research we shall seek to answer the following questions:

(a) What are the prediction techniques in data mining that can be used in discovery of trends and patterns in patients’ data?
(b) What methodologies can be used to design a model to infer characteristics of a predicted class from a combination of other data?
(c) What are the classification techniques that can be used to generate and discover strong rules and patterns of the risk factors?
(d) How to design a hybrid combination of C4.5 and association rule algorithm.

1.7 Assumption and Scope of the Research

The research exists within a well governed health domain and as such makes several basic assumptions. Key to this is that the diagnosis codes used are from the ICD coding standards. This is to limit the number of diagnosis groups and to ensure that we have a consistent pool of data to draw our comparisons from. We also make an assumption that the practitioners observed the guidelines in (ICD-9-CM Official Guidelines for Coding and Reporting, 2011). This, amongst others, provides for recording of the most accurate diagnosis describing the patient’s condition and avoiding “provisional” or “working” diagnosis and where diagnosis is “probable” or “questionable”, it is coded as if it existed (ICD-9-CM Official Guidelines for Coding and Reporting, 2011, p. 92). The conceptual model is as shown on Figure 1 below.
1.8 Conceptual Model

![Conceptual Model Diagram]

Figure 1: Conceptual Model
CHAPTER TWO-LITERATURE REVIEW

1.0 Introduction

In their work, *Fast algorithms for mining association rules in large databases*, (Agrawal and Srikant, 1994), the authors presented an algorithm, known as Apriori, for discovering association rules within large, primarily transactional, sales databases. This algorithm was a development of previously known algorithms for itemset mining and association rules discovery. We have a brief look at how this algorithm works and its known uses in the commercial, particularly retail sales databases, for which the authors admit the algorithm was originally conceived. We will also explore the benefits accrued by using this algorithm over other known algorithms for association rules mining. We shall also look at the Decision tree algorithm particularly the C4.5; its strengths and weaknesses and how we can combine it with the apriori algorithm to increase efficiency and speed in our research.

Over the years, numerous works have been done related to health systems with different data mining algorithms by different authors. They tried to achieve efficient methods and accuracy in finding out and predicting diseases by their work including datasets and different algorithms along with the experimental results and future work that can be done on the system to achieve more efficient results. We shall analyze different data mining techniques that have been introduced in recent years and applied in EMR system by different authors.

We will also highlight the efforts that have been put in the realization of electronic health records not just by the global community but also by Kenya as well. We also look at several standardization efforts like ICD and HL7 that are making it easier by the day to implement these systems and to share data across multiple system implementations. We highlight the importance that this plays in the realization and enabling role it plays in our research.

We will also go further to look at how data mining has been used by other researchers in the health industry and in particular the use of the C4.5 and the association Rule mining algorithms in the same. We will look at how and why the algorithms best suits our research in comparison with other data mining techniques. We delve deeper into our specific realm, the electronic medical records use of the same and highlight a few works on what others have done in this field.
2.1 EMR Systems Adoption in the Kenyan and Global Context

2.11 The Kenyan context
Earlier in this report, we made a distinction between Electronic Medical Records (EMR) and Electronic Health Records (EHR) as defined by (Garets and Davis, 2006, p. 2). EHR systems play a much wider role as they span multiple EMR’s whether integrated or not, several Care Delivery Organizations (CDOs) which could include hospitals, emergency or ambulatory care services, psychological or mental institutions and other health organizations. The government of Kenya in this context has been keen to strategically spearhead the adoption of EMRs in the country. This is with the establishment of guidelines through the Division of Health Information System (HIS). Though the government’s main aim, as it indicates in the policy, is to improve health data management, use of ICT in health and ease sharing of health data, there will be a lot of other benefits in the long run that will be as a result of this, part of which we wish to exploit in this research.

The government has chosen to adopt standards adopted by other “partner” countries the world over, mainly guided by WHO and ISO standards. Of particular interest to our case, the guideline requires any system being implemented to have the ability to maintain a coded list of problems/diagnoses (Standards and Guidelines for Electronic Medical Record Systems in Kenya, 2009, p. 23). It also goes ahead to indicate that the EMR ought to maintain a “Problem List” associated with the patient, its status, and the coded list of problems/diagnoses. These are some of the developments that have developed our interest in this research.

2.12 The Global Context
The federal government of United States of America has been on the headlines in recent times in the adoption of EMRs in the country. A survey of eleven countries carried out by (Schoen et al., 2009) found USA to be lagging behind other mainly Scandinavian countries. As (Gray et al., 2011) argue, the availing of $27 billion over 10 years to health providers by the US government for not just adopting EHRs but attaining “Meaningful Use” in improving patient care is set to change the outlook. This has an effect of ensuring that EHR implementers trample on each other to deliver systems to their clients seeking to attain Meaningful Use objectives in order to benefit from the financial incentives. According to the US government’s official health IT website, the
Meaningful Use objectives are divided into three stages, each with a target date as follows ("Meaningful Use Definition & Objectives," n.d.):

i. Stage 1: Data capture and sharing (2011-2012).


Meaningful use is defined as using certified electronic health record (EHR) technology to:

- Improve quality, safety, efficiency, and reduce health disparities.
- Engage patients and family.
- Improve care coordination, and population and public health.
- Maintain privacy and security of patient health information.

They go on to outline that the meaningful use compliance will result in:

- Better clinical outcomes
- Improved population health outcome
- Increased transparency and efficiency
- Empowered individuals
- More robust research data on health systems

2.2 Major Standards

There are several EMR standards in use today but our study shall only focus on those that relates to medical diagnosis.

2.2.1 International Statistical Classification of Disease (ICD)

ICD is the international "standard diagnostic tool for epidemiology, health management and clinical purposes". The ICD is maintained by the World Health Organizations (WHO), the directing and coordinating authority for health within the United Nations System.

The ICD is designed as a health care classification system, providing a system of diagnostic codes for classifying diseases, including nuanced classifications of a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease. This system is designed to map health conditions to corresponding generic categories together with specific variations, assigning for these a designated code, up to six characters long.
Thus, major categories are designed to include a set of similar diseases.

The ICD is published by the WHO and used worldwide for morbidity and mortality statistics, reimbursement systems, and automated decision support in health care. This system is designed to promote international comparability in the collection, processing, classification, and presentation of these statistics. As in the case of the analogous (but limited to mental and behavioral disorders) Diagnostic and Statistical Manual of Mental Disorders, currently in version 5), the ICD is a major project to statistically classify health disorders, and provide diagnostic assistance. The ICD is a core statistically based classificatory diagnostic system for health care related issues of the WHO Family of International Classifications (WHO-FIC)

2.2.2 Health Level 7 standards (HL7)
HL7 and its members provide a framework (and related standards) for the exchange, integration, sharing, and retrieval of electronic health information. These standards define how information is packaged and communicated from one party to another, setting the language, structure and data types required for seamless integration between systems. HL7 standards support clinical practice and the management, delivery, and evaluation of health services, and are recognized as the most commonly used in the world.

2.3 Association Rule Mining and the Apriori algorithm
Association rule mining has been used extensively in the commercial industry particularly in the retail sector. It has mainly been used to do market basket analysis where the focus is on analyzing the contents of the customer’s “basket”. As (Berry and Linoff, 2004, p. 287) explain, Market basket analysis provides insight into the merchandise by telling us which products tend to be purchased together and which are most amenable to promotion. Association rules identify strong relations that exist in databases using several measures of interestingness (usually based on minimum support and minimum confidence) (Matheus et al., 1993).
The patterns discovered may have different uses in nature and they may be categorized as actionable rules (contain high-quality, actionable information), trivial rules (already known by anyone at all familiar with the business) or inexplicable rules (these seem to have no explanation and do not suggest a course of action), (Berry and Linoff, 2004, pp. 296–298). When large databases are involved, an efficient algorithm to find frequently items that exist together (frequent itemsets) and find any patterns amongst these is needed. (Agrawal and Srikant, 1994) present an algorithm (Apriori) that aims at discovering association rules between items in a large database of sales transactions. The algorithm is simple in concept and is split into two main sub problems:

a. Find all sets of items (item_sets) that have transaction support above minimum support.
   The support for an item set is the number of transactions that contain the item_set.
   Item_sets with minimum support are called large item_sets, and all others small item sets.

b. Use the large item_sets to generate the desired rules.

The minimum support and confidence are given as follows (Bhargavi et al., 2013):

\[
\text{Support:} \quad \text{Supp}(x) = \frac{\text{No. Of transactions which contains the item_set } X}{\text{Total No. of transactions}}
\]

\[
\text{Confidence:} \quad \text{Conf}(X \rightarrow Y) = \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)}
\]

**Decision Tree and the C4.5 Algorithm**
Inductive inference is the process of moving from concrete examples to general models, where the goal is to learn how to classify objects by analyzing a set of instances (already solved cases) whose classes are known. Instances are typically represented as attribute-value vectors. Learning input consists of a set of such vectors, each belonging to a known class, and the output consists of a mapping from attribute values to classes. This mapping should accurately classify both the given instances and other unseen instances.

A decision tree [Quinlan, 1993] is a formalism for expressing such mappings and consists of tests or attribute nodes linked to two or more sub-trees and leafs or decision nodes labeled with a class which means the decision. A test node computes some outcome based on the attribute values of an instance, where each possible outcome is associated with one of the sub trees. An instance is classified by starting at the root node of the tree. If this node is a test, the outcome for the instance is determined and the process continues using the appropriate sub tree. When a leaf is eventually encountered, its label gives the predicted class of the instance.

This algorithm is an extension to ID3 developed by Quinlan Ross. It is also based on Hunt’s algorithm. C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute.

C4.5 has been used in real-world problems especially medical decision making because the methods can simultaneously provide high classification accuracy, simple representation of gathered knowledge. It is an improved algorithm of ID3 algorithm, presented by Quinlan [10]. C4.5 is a top-down algorithm and builds a decision tree model using a recursive process (also known as divides and conquer strategy). It uses information gain as splitting criteria to build a decision tree, an attribute with the most information which is computed on training set is first selected, the next one is selected the most informative from the remaining attributes, and so on. C4.5 algorithm can handle numeric attributes and missing values.
2.4 Hybrid of Association Rule Algorithm (Apriori) and the C4.5

One of the most obvious drawbacks of classical decision tree induction algorithms is poor processing of incomplete, noisy data. If some attribute value is missing, classical algorithms do not perform well on processing of such object. For example, in Quinlan’s algorithms before C4.5 such data objects with missing data have been left out of the training set – this fact of course resulted in decreased quality of obtained solutions (in this way the training set size and successively the information about the problem’s domain have been reduced). Algorithm C4.5 introduced the technique to overcome this problem, but is still not very effective. In real world problems, especially in medicine, missing data is very usual – therefore, effective processing of such data is of vital importance.

The next important drawback of classical induction methods is the fact that they can produce only one decision tree for a problem (when the same training set is used). In many real world situations it would be of a great benefit if more decision trees would be available and a user could choose the most appropriate one for a single case. As it is possible for training objects to miss some attribute value, the same goes also for a testing object – there can be a new case where some data is missing and is not possible to obtain it (unavailability of some medical equipment, for example, or invasiveness of a specific test for a patient). In this way another decision tree could be chosen that does not include a specific attribute test to make a decision.

Therefore, a goal is to build a decision tree is such a way that the accuracy of classification for those most important decisions is maximized. Once again, this problem is not solved adequately in classical decision tree induction methods.

When decision trees and ARM are compared, one can see that their advantages and drawbacks are almost complementary. For instance knowledge representation of decision trees is easily understood by humans, which is not the case for ARM; decision trees have trouble dealing with noise in training data, which is again not the case for ARM; decision trees learn very fast and ARM learn relatively slow, etc. Therefore, the idea is to combine decision trees and ARM in order to combine their advantages.
We also propose, the Apriori to be employed for important feature selection, and C4.5 method to be used as a fitness function of the Apriori in order to test for the efficiency of the set of selected features. First we propose to calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

2.5 Association Rule Mining and C4.5 in Health Care

In health informatics, a lot of work has gone in the use of data mining to previously commercial only applications. Key amongst the uses has been in matching patient diagnosis with symptoms which intertwines a lot with the use of knowledge based systems. It is difficult to induce reliable diagnostic rules from amongst a set of possibly infinite permutations of symptoms since the resulting hypotheses may have unsatisfactory prediction accuracy (Rajak and Gupta, 2008).

However, other researchers have come up with further refinements by using association rules to improve the prediction level claimed at 90% by (Serban et al., 2006) by combining it with supervised learning methods. The researchers applied their work to cancer but they claim that this can be extended to other disease diagnosis.

Association analysis as it is also called has been used to give probabilistic statements such as “If patients undergo treatment A, there is a 0.35 probability that they will exhibit symptom Z” (Koh and Tan, 2011). These can be useful when establishing relationships that affect effectiveness of particular patient treatment plans.
2.6 RELATED WORK

In this research’s specific field, some work has been done to take advantage of association rule mining in general and the Apriori algorithm in particular. Most of it centers on mining patterns in relation to a specific disease or diagnostic factor.

More work was also done (Kim et al., 2012) to analyze comorbidity in patients with type 2 diabetes mellitus (T2DM). The data was obtained from a medical center in Korea with an EMR that uses ICD-10 coding for the clinical diagnosis. The researchers developed a tool that uses Apriori algorithm to generate the strongest rules (diagnosis) that are associated with the T2DM. They then published the results of their findings with the resultant support and confidence levels.

The database used ICD-9 diagnosis coding and drew a sample of about 18,000 patients aged 18 and below with a diagnosis of ADHD. The researchers then made comparisons using Apriori algorithm to check the strength of associations amongst comorbidity rates and relative risk (RR) ratios of both groups of each diagnosis which were compared to one another. The results were published along with the resultant levels of interestingness.

(Tai and Chiu, 2009) used association rule mining to discover associations from data obtained from the National Health Insurance Database of Taiwan. Their work was intended not only to discover the comorbidity patterns of Attention Deficiency Hyperactive Disorder (ADHD) but to also examine the application of association rule mining in clinical databases.

Another prototype namely Clinical State Correlation Prediction (CSCP) was developed in order to predict the correlation(s) amongst the primary disease (the disease for which the patient visits the doctor) and secondary disease/s (which is/are other associated disease/s carried by the same patient having the primary disease (Rashid et al., 2014). The system developed uses the Apriori algorithm as well and checks the correlation between the primary disease and other secondary diseases. The CSCP is built on top of the transaction based health system which they base on and refer to as the OLTP. The diagnoses are not based on any diagnosis group like ICD. They also use data from this health OLTP, and pass the algorithm over data selected for different age
groups and sex. The results of the top two-item itemsets are then analyzed for any meaningful information.

2.7 Limitations and Research Gaps

These research papers are some that show the extent to which data mining in general has been used to discover interesting patterns in health systems.

Most advances in data mining and health systems research is based on diagnosis prediction systems that try to map symptoms with as small a differential diagnosis list as possible.

There are several other research works that follow similar approaches where either the same Apriori or another algorithm is used to mine associations between already known diseases with the intent of knowing which other conditions are most associated with those in question.

In the case of discovering previously unknown or thought disease patterns as in the work of (Rashid et al., 2014), they do not use a standardized EMR and the results are captured in the health system by free-text entered diagnosis which do not seem to adhere to any standard coding practice as presented in both their methodology and their results. This makes it difficult to statistically analyze or potentially group broader categories of diagnosis in order to get a wide variety of actionable rules/diagnosis.

This research aims at filling in these gaps by discovering associations that are accurate and efficient from EMR systems that are built on this standard model and exploit the presence of data that is generated by multiple health providers that use EMRs governed by the same standards.
CHAPTER THREE

3.0 METHODOLOGY

Computer science applies several research methods over and above the traditional quantitative and qualitative methods used in other disciplines like social sciences. (Glass et al., 2004), define over nineteen methods applied in the related fields of Computer Science, information Systems and Software Engineering. They conduct for each discipline, an analysis of the most dominant research topics, research approaches, research methods, reference disciplines, and levels of analysis. In previous work by (Ramesh et al., 2004), formulative research was found to be the most widely used research approach in computer science disciplines (79.15%) as compared to descriptive and evaluative research approaches. Among the top three research methods were Conceptual analysis (15.13%), Conceptual analysis/mathematical (74.13%) and concept implementation (proof of concept) (2.87%).

In building on this work, (Holz et al., 2006) mention concept implementation (also proof of concept or proof of principle) as “a claim about the value of a system design (or the design of a part of a system) is validated by building a system based on that design. Typically, the system that is built is not fully featured, but has enough functionality to convince the readers that the design can be effective. The proof-of-concept system is usually measured for performance or usability, to show that the new design is not so bad as to be unworkable”. It is on this method that this research wants to align its work, by concept implementation and present a prototype that uses a combination of association rule mining and C4.5 algorithms on a standardized EMR implementation.
3.1 Research Design

The research is structured as follows:
1. Problem Identification and Selection.
2. Literature Review and Concept Development.
3. Data Collection, Preparation and Processing.
4. Prototype Development.
5. Prototype Testing and Implementation.
6. Analyses and Presentation of Results.

3.2 Problem Identification and Selection

In this stage, an attempt is made to select and explain the problem that the research intends to solve which in our case is taking advantage of standardized EMRs and advances in the data-mining field. It is placed in perspective of the more general problem and there is an explanation of why it is a problem in the context of the research. There is also a brief description of how the researcher intends to approach the same and expected benefits that would accrue in tackling the issue at hand.

3.3 Literature Review and Concept Development

Here we attempt to understand the advances of data mining technology, in particular the use of ARM and the C4.5. We explore the purposes that ARM has been traditionally used and originally conceived. We also go ahead to look at the specific algorithms that are to be used for this research and we look at a few of the closest related work around the healthcare industry in general and in health prediction systems in particular. We develop the concept of ARM and C4.5 use in EMR systems and the several requirements in standards like ICD coding for diagnosis and HL7 for information exchange across the EMR systems that would be necessary to make the collection of data and analysis for this research possible.
3.4 Data Collection, Preparation and Processing

Here, we obtain data that is necessary for this research. The data should have met the standards defined earlier in order for it to be usable. For this research, we intend to obtain our data from Practice Fusion research data that is one of the leading EMR implementers in the United States of America. The dataset availed contains over 10,000 identified patient records (“Analyze This! | Research Division,” 2012) that contain over 98,000 individual contact points from over 150,000 medical practitioners across the country, and Practice Fusion’s Research Division is partnering with leading academic institutions and public health agencies to pursue ambitious new health studies (“Big Data Gets Put to Work for Public Health,” 2012).

We were unable to obtain Kenyan data from Kenyan EMR’s as attempts through the International Training and Education Center for Health (I-TECH) failed.

![Figure 2: section of the raw dataset](image-url)
The data obtained was then prepared and processed through the following steps:

i. Extracting the major diagnostic groups for each ICD-9 diagnosis for every patient record, and stored this to a relational database table.

ii. Designed as SQL procedure that could be used to sub-class the stored patient’s data.

3.5 Prototype Development

We have developed a prototype that implements a combination of C4.5 and association rule algorithm. The prototype feeds the c4.5 clustered output to Apriori algorithm we borrowed some implementations that have been used in market basket analysis. The prototype is capable of taking the data and finding associations based on the user defined values for the minimum support and confidence level. The prototype is implemented in Java through the Oracle ADF framework, and has a backend database of MYSQL server.

3.6 Prototype Testing and Implementation

The prototype was developed and tested using the identified patient records. First C4.5 was applied to the data to cluster and give us the overall association that existed on the dataset, then Apriori algorithm was applied to the dataset as is and finally a hybrid algorithm was run on the data and association rules generated was observed. Using varying minimum support and confidence values generated a number of rules. The top rules were those with the strongest confidence level above a support threshold. Since there is no globally accepted minimum support (as this is a custom user generated variable that depends on what they want to achieve, and how far deep they want to dig into the associations), we varied these values to observe the results and recorded each observation. Just as in other works using the values of support and confidence in this mining for strong associations like in (Kim et al., 2012) and that of (Tai and Chiu, 2009), we vary the same measures and indicate the values of support and confidence for each rule.
3.7 Analysis and Presentation of Results

Based on the rules observed, we compare this with the demographic data and select the demographic distribution of the top associations. These are mainly age groups and gender. We then use measures of central tendency (as appropriate for the nominal and ordinal variables) and classify the data into the different categories that they fall in. We also used a panel of experts drawn from the medical field who gave their opinion over the results. The survey was done using the questionnaire attached as Appendix A. The confidence levels and support for each question was left out deliberately so as to avoid user bias while answering the questions. We used the Likert scale to gather expert opinion and listened to their overall advice while noting explanations to some of their responses.
CHAPTER IV

4.0 SYSTEM ANALYSIS AND DESIGN

This chapter describes the system user interface design architecture and the back-end database design that have been used to structure the entire system prototype. Since this is a prototype and not a fully functioning system, areas such as system security (logging in and out) and complex input validation have not been implemented.

We hope to illustrate the implementation of a hybrid of C4.5 and Apriori algorithms in achieving our objectives. For an illustration of the research context see figure 3 below.

Figure 3: illustration of research context
4.1 Key system prototype elements

The system consists of these elements in order to accomplish its functions:

4.1.1 Input
This is the starting point of the system and is provided by the users. In this case it consists of the input variable of minimum support, minimum confidence threshold and the variable record number to be mined.

The minimum support determines the number of candidate item sets from which the rules will be generated and ordered by the confidence variable.

4.1.2 Transactions Data
This is the actual data that the system will process. For C4.5, we classify the data into classes of YES and NO, YES meaning the data contains the presence of HEART DISEASE while NO is the contra. For Apriori we apply the algorithm to the raw/unprocessed data as is without any prior processing, and for the Hybrid algorithm we pass the c4.5 classified data as its input. The c4.5 classification stage transforms multiple patients’ clinic visits into individual records for each; this information is passed to a MAP<key, Value> where patient ID is the key and the results of the clinic visits as the list of set values.

4.1.3 Output
The system generates three output results, that is, Apriori output, c4.5 output and the hybrid algorithm output. For C4.5 it’s a classification and grouping of data into various classes and number of occurrence. The Apriori will try to generate Strong association/rules from the unprocessed data based on minimum confidence and support, while for Hybrid algorithm the output will be the strongest rules that are based on the most frequent item sets and that satisfy the minimum confidence level.

System Design and Architecture
The prototype consists of three major components:
i. Database design
ii. Logic
iii. User interface
4.1.4 Database Design

Database consists of the following five tables:

i. Patients demographic records table(TRAINING_SYNCDIAGNOSTIC)

ii. Clinic visits per patient’s table(PROCESSDICDPCODEPERPATIENT)

iii. Rules table(Transactional) (Apriori Transactional table)

iv. RulesModified (Hybrid’s Transactional table)

v. C4.5 Transactional table (decisiontree_c45transdata)

The design of each and their relationship are as shown below.

Figure 4: Database Schema
Figure 5: Patient’s Demographics records table (Training syncDiagnosis).

Figure 6: Clinic visits per patients table (Processed ICD9 Codes per Patient)

Figure 7: Rules Table
4.1.5 Logic

This is the most important component as it implements the algorithm, using the input entered by the user, we follow these steps to obtain the classification, get frequent items, and association rules for the C4.5, Apriori and the hybrid algorithm:

4.1.6 C4.5 Algorithm

The row data is organized into heart disease and non-heart disease attributes, this is achieved through an SQL procedure, which groups data into a yes/no class for all the diagnosis ICD9Code attributes. The algorithm (C4.5 implemented in java) then traverses the data provided and groups individual ICD9codes into the two classes each with an individual count for the number of occurrence and the information gain for each. E.g.

```plaintext
value: 401.9,
classes: YES,
counts: 12,
value: 812.02,
classes: NO,
counts: 1,
```
For ICD9Code 401.9 which is “Unspecified Hypertension” occurred 12 times on the supplied data and is of class yes meaning, the attribute is categorized under heart Disease. The resultant attribute is displayed on a webpage for further interpretation; sample extract is as shown above.

### 4.1.7 Apriori Algorithm

Find all sets of items (item sets) that have transaction support above minimum support. The support for an item set is the number of transactions that contains the item set. Item set with minimum support are called large item sets, and all other are small item sets. E.g. if we define a support of 40% means we only selects those diagnoses that meets this threshold. E.g. [401.1], [401.9]

This would mean that these item sets meet the defined minimum support and are here in called frequent item-set. In the first iteration of algorithm each item is a member of the set of candidate. The set of frequent 1-itemsets, L1, consists of candidates’ 1-itemsets satisfying minimum support.

We keep joining the newly formed itemsets with themselves and use Apriori property or downward–closure property according to (Agrawal et al., 1993), that all subset of frequent itemsets must also be frequent thus obtaining itemsets such as [[401.9],[401.1]]. We iterate until there is no itemsets meeting our minimum threshold. Then we generate the basic association out of the current item-set, a rule is basic if its consequence contains only one item, and finally we try to mine the association rule. See the figure below for more information.

Figure 10: Resultant association

```
[HEART DISEASE] -> [ 493, 401.1, 401.9]: 1.0
[401.1, 401.9] -> [ 493, HEART DISEASE]: 1.0
```

### 4.1.8 The Hybrid algorithm

The algorithm works in similar way as the Apriori, the only difference is the input data, which comes from the classified C4.5 output. That is, it feeds on the C4.5 output and uses the data to mine association, generate basic rules and finally the association rules as shown in fig 11. below.
4.1.9 User Interface

This is where the user interacts with the system, they are able to provide the values of minimum support and minimum confidence and with this several associations can be obtained. The user is then presented with a data grid with the results along with their confident values. As shown below in the sample screen.
Figure 12: Login screen.

Figure 13: Landing page
Figure 14: Resultant page before user input.

Figure 15: Hybrid Screen, After user input
CHAPTER V

5.0 Results and Discussion

Here we are able to present results of the research within the context of the study. After execution of the runs and aggregation of the same, we were able to come up with a number of rules based on a support factor of 3.8% and minimum confidence of 90% for the first instance to see if we could obtain fewer rules from which we expect to see rules that are known by medics, also known as trivial rules (Barry and Linnoff, 2004, pp. 296-298). The research expanded to all the dataset as there were fewer rules generated when restriction was made to only heart disease related datasets, this was done with the aim of testing accuracy of the hybrid algorithm as larger dataset would provide us with the enabling environment.

5.0 Potentially Trivial association

Sample results of the run are shown below. In it are the top association and the associated support levels.

<table>
<thead>
<tr>
<th>Rule No</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[691.2] -&gt; [677, 731.4]</td>
</tr>
<tr>
<td>2</td>
<td>[924.2] -&gt; [752.3, 463]</td>
</tr>
<tr>
<td>4</td>
<td>[307.4] -&gt; [366, 401]</td>
</tr>
<tr>
<td>6</td>
<td>[723.5] -&gt; [495, 430.0]</td>
</tr>
<tr>
<td>7</td>
<td>[252.5] -&gt; [495, 389.8]</td>
</tr>
<tr>
<td>8</td>
<td>[455, 380.4] -&gt; [722.5, 472]</td>
</tr>
<tr>
<td>9</td>
<td>[307.4] -&gt; [396, 401]</td>
</tr>
<tr>
<td>10</td>
<td>[401] -&gt; [396, 204]</td>
</tr>
<tr>
<td>11</td>
<td>[292.6] -&gt; [302.8, 490]</td>
</tr>
<tr>
<td>12</td>
<td>[302.8, 477.4] -&gt; [451, 722.5]</td>
</tr>
<tr>
<td>13</td>
<td>[455, 722.5] -&gt; [320.8, 477.4]</td>
</tr>
<tr>
<td>14</td>
<td>[455, 722.5] -&gt; [353.6, 303.5]</td>
</tr>
<tr>
<td>15</td>
<td>[455, 722.5] -&gt; [486, 466]</td>
</tr>
<tr>
<td>16</td>
<td>[616.1] -&gt; [491, 386.9]</td>
</tr>
<tr>
<td>17</td>
<td>[616.1] -&gt; [491, 386.9]</td>
</tr>
<tr>
<td>18</td>
<td>[401] -&gt; [491, 386.9]</td>
</tr>
<tr>
<td>19</td>
<td>[753.0] -&gt; [491.3, 469]</td>
</tr>
<tr>
<td>20</td>
<td>[491] -&gt; [391, 207]</td>
</tr>
<tr>
<td>21</td>
<td>[491] -&gt; [391, 207]</td>
</tr>
<tr>
<td>22</td>
<td>[401] -&gt; [391, 207]</td>
</tr>
</tbody>
</table>
Figure 16: Sample association rules

These Translates to ICD description as shown in (figure 5.2) below

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>401.169-348.5,207.41</td>
<td>Essential Hypertension, Hearing Loss</td>
</tr>
<tr>
<td>107.41</td>
<td>986.407.389</td>
</tr>
<tr>
<td>107.41</td>
<td>401.377.41</td>
</tr>
<tr>
<td>722.51,477.9-940.5,80.8</td>
<td>Degeneration of Lumbar or Lumbarosacral Intervertebral Disc</td>
</tr>
<tr>
<td>722.51,435.6,280.8,477.9</td>
<td>Degeneration of Lumbar or Lumbarosacral Intervertebral Disc</td>
</tr>
<tr>
<td>940.5,80.8</td>
<td>940.5,277.3</td>
</tr>
<tr>
<td>940.5,80.8</td>
<td>722.51,477.9</td>
</tr>
<tr>
<td>107.41</td>
<td>986.407.389</td>
</tr>
<tr>
<td>107.41</td>
<td>401.377.41</td>
</tr>
<tr>
<td>401.169-348.5,207.41</td>
<td>Essential Hypertension, Hearing Loss</td>
</tr>
<tr>
<td>107.41</td>
<td>986.407.389</td>
</tr>
<tr>
<td>401.169-348.5,207.41</td>
<td>Essential Hypertension, Hearing Loss</td>
</tr>
<tr>
<td>107.41</td>
<td>986.407.389</td>
</tr>
<tr>
<td>401.169-348.5,207.41</td>
<td>Essential Hypertension, Hearing Loss</td>
</tr>
<tr>
<td>107.41</td>
<td>986.407.389</td>
</tr>
</tbody>
</table>

Figure 17: ICD9 Descriptions

5.1 Demographic Comparison

We compared the diagnoses with the demographic prevalence where they were comorbid and the result is as indicated in appendix B. It is worth noting that the average age of the population was age 52.
5.2 Validation of Results

In this stage we are able to compare the results of our prototype and the opinion of experts regarding whether the associations obtained here are known to them or not, and if not whether they agree that they could be linked (probably indirectly) and by how much (strongly or otherwise). This is done through a questionnaire survey (see appendix A). Each of the questions can be scored as follows:

![Figure 18: Likert scale scores](image)

Each of the question has a score associated that is calculated from the median of responses from all experts (since the scale (Likert) consist of ordinal values).

5.3 Discussion of results

After running through the dataset, we were able to generate several associations that differed based on what we set as the minimum support and confidence level. We did not find a universally applicable or acceptable threshold for minimum support and confidence, as this seems to be applicable in different ways to different domains, depending on what patterns the end user intends to accept or reject and the available data. As earlier discussed, as in other works using the values of support and confidence in this mining for strong associations like in (Kim et al., 2012) and that of (Tai and Chiu, 2009), we vary the same measures and indicate the values of support and confidence for each rule.

It is possible to obtain a very large number of rules since these increases as the values of minimum support and minimum confidence decreases towards zero. The outliers in the data in this case will be the rules that may not necessarily meet the selected user-specific threshold for minimum support and confidence. It is therefore up to the user to decide what the most acceptable values for minimum support and confidence are, and what criteria to use to discard or accept the generated associations.
We observed that some rules were generated which happened to be consistent with common knowledge amongst the members of the medical fraternity, for example the link between Essential Hypertension and Transient Disorders of Initiating or maintaining sleep, or Cataract (as shown in the first three rules of Figure 5.2 and subsequent description in Figure 5.2). The panelists agreed with this concurrence of 4.5/5 translating to a 90% nod. These known associations also had all high measures of confidence (between 60-90% from our system) as had earlier been discussed.

Figure 5.2 some diagnosis were also consistent with some of previous specific research like that of (Kim et al., 2012) that indicate the strongest link between Type 2 Diabetes mellitus and Essential Hypertension with a confidence of 34.86%. This is captured as rule 97 in our results with a confidence of 90%.

There are other rules which most of the panelists chose to neither agree nor disagree. They attributed this to the fact that some of the associations may be incidental to some specific patients and it may be observed in a few cases but not necessarily a majority of the cases. The presence of one qualifying diagnosis from amongst the set on the left being linked to that on the right also caused a mixed reaction in most of the practitioners, an example being that of: Internal hemorrhoids without mention of complication, Fracture of unspecified part of neck of femur, closed → Degeneration of lumbar or lumbosacral intervertebral disc.

In such a case, the panelists argued that it is the link of Degeneration of lumbar or lumbosacral intervertebral to internal hemorrhoid and not the other way round that would trigger the association. We also observe that some associations were out rightly rejected by the same panel of experts as expected (e.g. the association between esophageal diseases→Disorders of Lipoid metabolism).

Some experts indicated that some of the associations could be comorbid but not necessarily linked, that is without a cause-effect relation and that some conditions coexist but are not very frequent.
Another panelist was also keen to indicate that the associations that we seek to investigate can only be investigated as comorbidity patterns and causal relations may not necessarily be possible to state comprehensively at this level. This is what the research emphasizes as the output of its findings.

Findings to mining medical datasets require a lot of domain expertise to interpret the rules as was reiterated by (Roddick et al., 2003). Most of them will be known but others may be less known while those that seem unusual may be discarded at a first look. However, output to this research may prove to be of utmost importance to curious specialists since some of the rules generated, however few, could be used as a starting point for future research by the domain experts. Of great interest would be to attempt to establish the reasons for comorbidity amongst our associations that seem unusual or unknown. These reasons could be causal links or outright co-existence due to the condition of the patient. As one panelist explained, a patient diagnosed separately with allergic rhinitis, bronchitis and eczema (dermatitis) will have allergic tendencies that make such conditions, whereas unrelated, to be present in the same patient over time. When this happens frequently in the sampled population, some associations like these will certainly emerge from our system, and only further investigation by domain experts will show that.

Comparisons with demographics showed some expected patterns like some disease prevalence being higher in older patients e.g. the combination of Hypertension and Diabetes Mellitus being found in patients with an average age of 63.5, presenting a distance of 11.5 years above our average age. This is true for the most common diagnosis associations from our results. Further demographic analysis could be done on individual sets of associations as far as one would desire to find more relevant demographic patterns and compare them with the expected patterns.
CHAPTER VI SUMMERY CONCLUSION AND RECOMMENDATIONS

6.0 Summery

This research set out to design and develop a hybrid of C4.5 and Apriori algorithms to efficiently and accurately identify any hidden diagnosis patterns within the big EMR data. We intended to find out the current use of association rule mining and C4.5 algorithms in both the commercial world in general and the health sector in particular. Of great interest, was the adaption of the hybrid algorithm by combining C4.5 and Apriori algorithms on strength basis to mine the associations by prototype implementation. We also intended to investigate the applicability of the developed algorithm in the context of electronic medical record systems that adhere to certain standards, and as well as to show that we can generate and discover strong rules (relationships) that indicate multi-morbidity trends from the EMR data with varying confidence and support levels.

6.1 Conclusion

Using this prototype, we are now able to mine data from EMR systems that implement any standardized diagnosis coding guideline. In our case, it is the WHO recommended standard of ICD-CM coding. Multiple systems can exchange their data and we are therefore able to take advantage of big data and generate patterns from it based on user defined measures of interestingness on what suits one as the minimum support and confidence levels. It is also of importance to note that the data used for mining the associations was primarily intended for other clinical purposes. In this research, we were able to take advantage and build our system to find interesting patterns that could arise from this kind of well-organized big data. This goes to demonstrate the power of having standardized clinical data across multiple implementations of electronic medical records systems. We were able to see that although the medical practitioners agreed on some already known associations, it would not be prudent to expect them to agree on all previously unknown associations. This research would therefore prove to be key as input to another research on causation, and would be a good starting point for any medical researcher seeking to unlock multi-morbidity trends amongst patients in any given patient population.
We are particularly encouraged by previous studies that seemed to suggest that Vitamin D deficiency is associated with Hypertension but the causal relationship is not known (Vimaleswaran et al., 2014). This is the same way in which there could be a (perhaps less prevalent but nonetheless unknown and important) relationship between Vitamin D deficiency and disorders of lipoid metabolism mostly hyperlipidemia (mixed and unspecified type). This would ideally then be used as input to another study that seeks to concentrate on the specific association and finding if there is any causal association. The demographic prevalence of our associations showed no much difference with the expected outcome as discussed in the previous chapter.

6.2 Limitation and challenges

There were several limitations and challenges faced while carrying out this research. Of most importance was obtaining the data used for this research, as getting EMR data is an issue that carries a lot of confidentiality and medical-legal challenges that goes with it. Most hospitals would therefore be unwilling to release such data. We also required data that adhered to a specified standard in order to make data mining possible and this is currently a challenge in Kenya since the standards and guidelines for this have only been recently established. Not all major CDO’s have been using EMR systems. This reduces even further the pool from which to obtain the expected data. We were thus unable to obtain Kenyan data within the duration of this research but should this become available in future, the same system could be applied to the data and hopefully interesting patterns will be obtained. Interpreting results obtained from mining medical data also requires familiarity with diagnosis coding and expertise in the medical field.
6.3 Future work

As stated in the previous chapter, we recommend that the output of this research particularly with results from the rules that had higher confidence but lower support levels be investigated further by another domain specific study to explain the comorbidity trends to those that are unknown to the medical fraternity. In order to better performance while handling large data, we recommend that another enhancement be done to the algorithm. If this works then it would speed up and even encourage the use of the same prototype implementation on larger datasets that are more likely to yield further and more usable associations. Adoption and use of the Kenyan data with the same implementation would also be encouraged as soon as there is enough data adhering to the specified standard.
REFERENCES


Chitra R and Seenivasagam V. (2013, JULY). REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES. *ICTACT JOURNAL ON SOFT COMPUTING, 03(04, 2013.).


Data Mining. *International J. of Healthcare & Biomedical Research.*


Ghaziabad, India.


---

**APPENDIX**

Appendix A

<table>
<thead>
<tr>
<th>Main Association</th>
<th>Strongly agree</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Essential Hypertension, Hearing loss, Cataract, Transient Disorder of Initiating or Maintaining sleep</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Transient Disorder of Initiating or Maintaining sleep, Cataract, Essential Hypertension, Hearing loss</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Cataract, Transient Disorder of Initiating or Maintaining sleep, Essential Hypertension, Hearing loss</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Degeneration of lumbar or lumbosacral intervertebral disc, Allergic rhinitis, cause unspecified, Internal hemorrhoids without mention of complication, Fracture of unspecified part of femur, closed</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Degeneration of lumbar or lumbosacral intervertebral disc, Internal hemorrhoids without mention of complication, Fracture of unspecified part of femur, closed, Allergic rhinitis, cause unspecified</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Internal hemorrhoids without mention of complication, Fracture of unspecified part of femur, closed, Cataract, Essential Hypertension, Hearing loss</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Transient disorder of initiating or maintaining sleep, Cataract, Essential Hypertension, Hearing loss</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Essential Hypertension, Cataract, Transient disorder of Initiating or Maintaining sleep</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Appendix B

<table>
<thead>
<tr>
<th>Code</th>
<th>Male</th>
<th>Female</th>
<th>Male%</th>
<th>Female%</th>
<th>AVG Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>401,389→366,307.41</td>
<td>98</td>
<td>94</td>
<td>51</td>
<td>49</td>
<td>64</td>
</tr>
<tr>
<td>307.41→366,401,389</td>
<td>98</td>
<td>94</td>
<td>51</td>
<td>49</td>
<td>64</td>
</tr>
<tr>
<td>366,307.41→401,389</td>
<td>98</td>
<td>94</td>
<td>51</td>
<td>49</td>
<td>64</td>
</tr>
<tr>
<td>722.52,477.9→455,820.8</td>
<td>98</td>
<td>123</td>
<td>44</td>
<td>56</td>
<td>59.5</td>
</tr>
<tr>
<td>722.52→455,820.8,477.9</td>
<td>98</td>
<td>123</td>
<td>44</td>
<td>56</td>
<td>59.5</td>
</tr>
<tr>
<td>455,820.8→722.52,477.9</td>
<td>98</td>
<td>123</td>
<td>44</td>
<td>56</td>
<td>59.5</td>
</tr>
<tr>
<td>307.41→366,401</td>
<td>98</td>
<td>94</td>
<td>51</td>
<td>49</td>
<td>64</td>
</tr>
<tr>
<td>401→366,307.41</td>
<td>98</td>
<td>94</td>
<td>51</td>
<td>49</td>
<td>64</td>
</tr>
<tr>
<td>369→401,307.41</td>
<td>98</td>
<td>94</td>
<td>51</td>
<td>49</td>
<td>64</td>
</tr>
<tr>
<td>457.1→481,786.05</td>
<td>98</td>
<td>94</td>
<td>51</td>
<td>49</td>
<td>64</td>
</tr>
<tr>
<td>366→401,385,307.41</td>
<td>98</td>
<td>94</td>
<td>51</td>
<td>49</td>
<td>64</td>
</tr>
<tr>
<td>401→389,307.41</td>
<td>98</td>
<td>94</td>
<td>51</td>
<td>49</td>
<td>64</td>
</tr>
</tbody>
</table>

Appendix C : Sample codes

Modified.jspx

```xml
<?xml version='1.0' encoding='windows-1252'?>
<jsp:root xmlns:jsp="http://java.sun.com/JSP/Page" version="2.1"
    xmlns:f="http://java.sun.com/jsf/core"
    xmlns:h="http://java.sun.com/jsf/html"
    xmlns:af="http://xmlns.oracle.com/adf/faces/rich"
    xmlns:dvt="http://xmlns.oracle.com/dss/adf/faces">

    <jsp:directive.page contentType="text/html;charset=UTF-8"/>
    <f:view>
        <af:document id="d1" title="Hybrid Algorithm">
            <af:messages id="m1"/>
            <h:outputText escape="false"
                value="&lt;link rel="icon" type="image/png"
                href="/images/favicon.png"'/
                id="fav"/>
    </f:view>
</jsp:root>
```
<af:pageTemplate viewId="#{Rendering.template}" id="fms">
  <f:facet name="center">
    <af:group id="g1">
      <af:panelBox text="PanelBox1" id="pb4" showHeader="never" inlineStyle="width:99%;">
        <table cellspacing="0" cellpadding="0" border="0" width="100%">
          <tr>
            <td width="50%" align="left">
              <af:inputText id="it1" helpTopicId="FMSbankAccounts" columns="13" readOnly="true" shortDesc="Help"/>
            </td>
            <td width="100%" align="right">
              <af:statusIndicator id="si1"/>
            </td>
          </tr>
        </table>
        <h:panelGrid columns="1" id="pg1">
          <af:outputLabel value="The Algorithm Details" id="ol1" inlineStyle="font-weight:bold;"/>
          <af:separator id="s1"/>
        </h:panelGrid>
        <af:panelGroupLayout id="pgl2">
          <af:panelBox text="PanelBox2" id="pb1" showHeader="never" inlineStyle="width:99%;">
            <af:panelTabbed id="pt1" inlineStyle="height:450.0px;width:1000.0px;">
              <af:showDetailItem text="Hybrid Algorithm" id="sdi2"/>
            </af:panelTabbed>
          </af:panelBox>
        </af:panelGroupLayout>
      </af:panelBox>
    </af:group>
  </f:facet>
</af:pageTemplate>
<af:outputLabel value="Apriori with C45" id="ol5"
inlineStyle="font-weight:bold;"/>
<af:separator id="s4"/>
</h:panelGrid>

<af:panelGroupLayout id="pgl112">
<af:panelBox text="PanelBox6" id="pb6"
showHeader="never"
inlineStyle="width:99%;"
rendered="true">
<f:facet name="toolbar"/>
<h:panelGrid columns="3" id="pg10">
<af:commandButton text="Add" id="cb7"
icon="/images/create.gif"
partialSubmit="true"
action="#{ModifiedController.launchSetUp}"/>
<af:popup id="AprioriModified"
contentDelivery="lazyUncached">
<af:dialog id="d3" type="none"
title="Generate Rules For Hybrid Algorithm"/>
<af:panelBox text="PanelBox3" id="pb2"
showHeader="never"
inlineStyle="width:99%;">"/>
<f:facet name="toolbar"/>
<h:panelGrid columns="2" id="pg11">
<af:outputLabel value="Minimum Confidence" id="ol9"/>
```xml
<af:table value="#{bindings.ModifiedManip.collectionModel}" var="row"
        rows="#{bindings.ModifiedManip.rangeSize}"
        emptyText="#{bindings.ModifiedManip.viewable ? 'No data to display.' : 'Access Denied.'}"
        rowBandInterval="0"
        selectedRowKeys="#{bindings.ModifiedManip.collectionModel.selectedRow}"
        selectionListener="#{bindings.ModifiedManip.collectionModel.makeCurrent}"
        rowSelection="single"
        fetchSize="#{bindings.ModifiedManip.rangeSize}"
        filterModel="#{bindings.ModifiedManipQuery.queryDescriptor}"
        filterVisible="true"
        queryListener="#{bindings.ModifiedManipQuery.processQuery}"
        varStatus="vs"
        id="t2" binding="#{ModifiedController.datatable}">
    <af:column sortProperty="#{bindings.ModifiedManip.hints.itemNo.name}"
              filterable="true"
              sortable="true" headerText="Rule No" id="c1">
        <af:outputText value="#{row.itemNo}"
                       shortDesc="#{bindings.ModifiedManip.hints.itemNo.tooltip}"
                       id="ot2"/>
    </af:column>
    <af:column sortProperty="#{bindings.ModifiedManip.hints.rule.name}"
              filterable="true"
              sortable="true" headerText="Rule" id="c2" width="700">
        <af:outputText value="#{row.rule}"
                       shortDesc="#{bindings.ModifiedManip.hints.rule.tooltip}"
                       id="ot3"/>
    </af:column>
</af:table>
```
ModifiedController.java
package hybridAlgorithView.modifiedTree;

import Sacco.view.base.ADFUtils;
import Sacco.view.base.GlobalCC;

import hybridAlgorithView.DBConn;
import hybridAlgorithView.apriori.AprioriMain;
import java.sql.Connection;
import java.sql.SQLException;
import java.sql.Statement;
import javax.faces.context.FacesContext;
import oracle.adf.view.rich.component.rich.RichDialog;
import oracle.adf.view.rich.component.rich.data.RichTable;
import oracle.adf.view.rich.component.rich.input.RichInputText;
import oracle.adf.view.rich.context.AdfFacesContext;
import org.apache.myfaces.trinidad.render.ExtendedRenderKitService;
import org.apache.myfaces.trinidad.util.Service;
public class ModifiedController {

    private RichTable datatable;
    private RichInputText minimumSupport;
    private RichInputText minimumConfidence;

    private Connection conn=null;
    private Statement state=null;
    private RichInputText limit;

    public ModifiedController() {
        super();
    }

    public void setDatatable(RichTable datatable) {
        this.datatable = datatable;
    }

    public RichTable getDatatable() {
        return datatable;
    }

    public String generate() {
        AprioriModified.demo(minimumSupport.getValue().toString(),minimumConfidence.getValue().toString(),limit.getValue().toString());

        ExtendedRenderKitService erkService1 = Service.getService(FacesContext.getCurrentInstance().getRenderKit(), ExtendedRenderKitService.class);
    }
}
erkService1.addScript(FacesContext.getCurrentInstance(),
   "var hints = {autodismissNever:false}; " +
   "AdfPage.PAGE.findComponent(" +
   "fms:AprioriModified" + ").hide(hints);";)
ADFUtils.findIterator("findAccountGroupIterator").executeQuery();

AdfFacesContext.getCurrentInstance().addPartialTarget datatable;
GlobalCC.saveSuccessfull();

return null;
}

public void setMinimumSupport(RichInputText minimumSupport) {
   this.minimumSupport = minimumSupport;
}

public RichInputText getMinimumSupport() {
   return minimumSupport;
}

public void setMinimumConfidence(RichInputText minimumConfidence) {
   this.minimumConfidence = minimumConfidence;
}
public RichInputText getMinimumConfidence() {
    return minimumConfidence;
}

public String launchSetUp() {
    Deleted();
    ExtendedRenderKitService erkService =
        Service.getService(FacesContext.getCurrentInstance().getRenderKit(),
        ExtendedRenderKitService.class);
    erkService.addScript(FacesContext.getCurrentInstance(),
        "var hints = {autodismissNever:false}; " +
        "AdfPage.PAGE.findComponent(" + "fms:AprioriModified" +
        ");
    return null;
}

public void Deleted() {
    String query = "DELETE FROM  rulesmodified";

    try{
        conn=DBConn.getDBCponnection();
        state=conn.createStatement();
        state.executeUpdate(query);
        ADFUtils.findIterator("findAccountGroupIterator").executeQuery();
        AdfFacesContext.getCurrentInstance().addPartialTarget(datatable);
    }
}
catch(Exception e){
    e.printStackTrace();
}

finally {
    try {
        if (conn != null) {
            conn.close();
        }
        if (state != null) {
            state.close();
        }
    } catch (SQLException sqlee) {
        sqlee.printStackTrace();
    }
}

public void setLimit(RichInputText limit) {
    this.limit = limit;
}

public RichInputText getLimit() {
    return limit;
}

/*public static void main(String [] args){
    ModifiedController mc=new ModifiedController();
*/
mc. generate();
 */
}

protected static void demo(String minConfidence, String minSupport, String noOfRecords) {
    double confidence = Double.parseDouble(minConfidence) / 10; // 0.4
    double support = Double.parseDouble(minSupport) / 100; // 0.06
    int limit = Integer.parseInt(noOfRecords);
    AprioriFrequentItemsetGenerator<String> generator =
        new AprioriFrequentItemsetGenerator<>();

    // Further code
List<Set<String>> itemsetList = new ArrayList<>();
for (Set<String> rowDat : ModifiedTree.getValueTree(limit)) {
    itemsetList.add(rowDat);
}

long startTime = System.nanoTime();
FrequentItemsetData<String> data = generator.generate(itemsetList, support);
long endTime = System.nanoTime();

int i = 1;

System.out.println("--- Frequent itemsets ---");

for (Set<String> itemset : data.getFrequentItemsetList()) {
    System.out.printf("%2d: %9s, support: %1.1fn", i++, itemset, data.getSupport(itemset));
}

System.out.printf("Mined frequent itemset in %d milliseconds.
", (endTime - startTime) / 1_000_000);

startTime = System.nanoTime();
List<AssociationRule<String>> associationRuleList =
    new AssociationRuleGenerator<String>()
        .mineAssociationRules(data, confidence);
endTime = System.nanoTime();
System.out.printf("Mined association rules in %d milliseconds.
", endTime - startTime);
System.out.println();
System.out.println("-------------- Association rules ---------------------------------------");
startTime = System.nanoTime();
for (AssociationRule<String> rule : associationRuleList) {
    System.out.printf("%2d: %sn", i++, rule);
    RecordData.insertRulesModified(rule.toString());
}
endTime = System.nanoTime();
System.out.printf("generated association rules in %d milliseconds.n",
    (endTime - startTime) / 1_000_000);
}

FrequentItemsetData.java

package hybridAlgorithView.apriori;
import java.util.List;
import java.util.Map;
import java.util.Set;
public class FrequentItemsetData<I>  {
    private final List<Set<I>> frequentItemsetList;
    private final Map<Set<I>, Integer> supportCountMap;
    private final double minimumSupport;
    private final int numberOfTransactions;

    FrequentItemsetData(List<Set<I>> frequentItemsetList,
            Map<Set<I>, Integer> supportCountMap,
            double minimumSupport,
            int transactionNumber) {
        }
this.frequentItemsetList = frequentItemsetList;
this.supportCountMap = supportCountMap;
this.minimumSupport = minimumSupport;
this.numberOfTransactions = transactionNumber;
}

public List<Set<I>> getFrequentItemsetList() {
    return frequentItemsetList;
}

public Map<Set<I>, Integer> getSupportCountMap() {
    return supportCountMap;
}

public double getMinimumSupport() {
    return minimumSupport;
}

public int getTransactionNumber() {
    return numberOfTransactions;
}

public double getSupport(Set<I> itemset) {
    return 1.0 * supportCountMap.get(itemset) / numberOfTransactions;
}

}

ModifiedTree.java

package hybridAlgorithmView.modifiedTree;

import java.util.stream.Collectors;
import hybridAlgorithmView.C45.*;
import java.io.IOException;
import java.util.ArrayList;
import java.util.HashSet;
import java.util.*;
import hybridAlgorithmView.frequentGenerator.RecordData;
import java.lang.Math;
import java.util.Arrays;
import java.util.HashMap;

public class ModifiedTree {
    /*  public static void main(String[] args) throws IOException {
        getValueTree(800);
    }*/

    public static double calcIofD(List<Integer> classesCount){
        double IofD = 0.0;
        double temp = 0.0;
        int totalNumClasses = 0;
        for(int i : classesCount){
            totalNumClasses += i;
        }
    }
}
for(double d : classesCount){
    temp = (-1 * (d/totalNumClasses)) * (Math.log((d/totalNumClasses)) / Math.log(2));
    IofD += temp;
}
return IofD;

public static List<Set<String>> getValueTree(int limit){
    List<Val> items = getDataOriginal(limit);
    Map<String, Set<String>> result =
        items.stream().collect(
            Collectors.groupingBy(Val::getValueName,
                Collectors.mapping(Val::getItClass , Collectors.toSet()))
        );
    List<Set<String>> dataAsSet = new ArrayList<>();
    String splitBy="","
    for (Map.Entry<String, Set<String>> entry : result.entrySet()) {
        String first=entry.getValue().toString().replaceAll("\[|\]", "");
        String second=first.replaceAll("\[|\]", "");
        String third=second.replaceAll("\[|\]", "");
        // System.out.println("Third===="+third);
        Set<String> itemset = new HashSet<>();
        // itemset.add((second).replaceAll("\[|\]", "")
        //     );
        String[] cols = third.split(splitBy);
        if(cols.length<5){
            

continue;
}
dataAsSet.add(new HashSet<>(Arrays.asList(cols[0], cols[1], cols[2], cols[3])));
dataAsSet.add(itemset);

// System.out.println(dataAsSet);
return dataAsSet;

public static List<Val> getDataOriginal(int limit) {
    List<Val> itemsetList = new ArrayList<Val>();
    for(C45Model a: RecordData.getProcessedDataList(limit)) {
        Val inV = new Val(a.getAttribute(), a.getItsClass());
        itemsetList.add(inV);
    }
    return itemsetList;
}
import java.sql.ResultSet;
import java.sql.SQLException;
import java.sql.Statement;
import java.util.ArrayList;
import java.util.Arrays;
import java.util.HashSet;
import java.util.List;
import java.util.Set;

public class RecordData {
    public RecordData() {
        super();
    }
    static Connection conn=null;
    public static List<HashSet<String>> getTestData(int limit,int end){

        List<HashSet<String>> itemsetList = new ArrayList<>();
        PreparedStatement st=null;
        conn=null;
        ResultSet rs=null;
        try{
            conn=DBConn.getDBCconnection();
            st=conn.prepareStatement("SELECT
TRIM(training_SyncDiagnosis.PatientGuid),TRIM(ICD9CODE),TRIM(BMI),(SystolicBP/DiastolicBP)rate,PhysicianSpecialty FROM training_SyncDiagnosis inner join training_synctranscript on training_synctranscript.PatientGuid=training_SyncDiagnosis.PatientGuid where (SystolicBP/DiastolicBP)>0 and BMI!='NULL' AND id between "+limit+" and "+end+"");
            rs=st.executeQuery();
            while(rs.next()){
                itemsetList.add(new
HashSet<>(Arrays.asList(rs.getString(1), rs.getString(2), rs.getString(3), rs.getString(4), rs.getString(5)));

catch (Exception e) {
    e.printStackTrace();
}
finally {
    try {
        if (conn != null) {
            conn.close();
        }
        if (st != null) {
            st.close();
        }
    } catch (SQLException sqlee) {
        sqlee.printStackTrace();
    }
}
return itemsetList;

public static void insertRules(String rule) {
    PreparedStatement state = null;
    String query = "insert into rules(rule) values(?)";
    try {
        conn = DBConn.getConnection();
        state = conn.prepareStatement(query);
        state.setString(1, rule);
        state.executeUpdate();
    } catch (SQLException e) {
        e.printStackTrace();
    }
}
state.setString(1, rule);
state.executeUpdate();

catch(Exception e){
    e.printStackTrace();
}
finally {
    try {
        if (conn != null) {
            conn.close();
        }
        if (state != null) {
            state.close();
        }
    } catch (SQLException sqlee) {
        sqlee.printStackTrace();
    }
}

public static int getCount(){

    PreparedStatement st=null;
    conn=null;
    ResultSet rs=null;
    int total = 0;
    try{
        conn=DBConn.getDBCponnection();
        st=conn.prepareStatement("SELECT count(*) FROM training SyncDiagnosis");

    }
where DiagnosisGuid is not null ");
    rs=st.executeQuery();
    if(rs.next()){
        total=Integer.parseInt(rs.getString(1));
    }
}
}
catch(Exception e){
    e.printStackTrace();
}
finally {
    try {
        if (conn != null) {
            conn.close();
        }
        if (st != null) {
            st.close();
        }
    } catch (SQLException sqlee) {
        sqlee.printStackTrace();
    }
}
return total;
}
public static List<HashSet<String>> getData(){
    String splitBy="\,";
    List<HashSet<String>> itemsetList = new ArrayList<HashSet<String>>();
PreparedStatement st=null;
conn=null;
ResultSet rs=null;
try{
    conn=DBConn.getDBCpnnnection();
    st=conn.prepareStatement("select patientId,ICD9Codes,class from icd9perpatient limit 50 ");
    rs=st.executeQuery();
    while(rs.next()){
        String values=rs.getString(2);
        String[] cols = values.split(splitBy);
        if(cols.length<5){
            continue;
        }
        insertToProcessdICD(rs.getString(1),cols[0],cols[1],cols[2],cols[3],rs.getString(3));
        itemsetList.add(new HashSet<>(Arrays.asList(cols[0],cols[1],cols[2],rs.getString(3))));
    }
}
catch(Exception e){
    e.printStackTrace();
}
finally {
    try {
        if (conn != null) {
            conn.close();
        }
    } catch (Exception e) {
        e.printStackTrace();
    }
}
if (st != null) {
    st.close();
}
}
} catch (SQLException sqlee) {
    sqlee.printStackTrace();
}
}
return itemsetList;

public static List<String> getDataTree() {

    String splitBy="",";
    List<String> itemsetList = new ArrayList<String>();
    PreparedStatement st=null;
    conn=null;
    ResultSet rs=null;
    try{
        conn=DBConn.getDBCponnection();
        st=conn.prepareStatement("select patientId,ICD9Codes from icd9perpatient
limit 400");
        rs=st.executeQuery();
        while(rs.next()){
            String values=rs.getString(2);
            String[] cols = values.split(splitBy);
            if(cols.length<4){
                continue;
            }
            String val=cols+"","+rs.getString(1);
            itemsetList.add(val);
        }
    }
    return itemsetList;
}
catch(Exception e){
    e.printStackTrace();
}
finally {
try {
    if (conn != null) {
        conn.close();
    }
    if (st != null) {
        st.close();
    }
} catch (SQLException sqlee) {
    sqlee.printStackTrace();
}
}
return itemsetList;

public static void insertToProcessdICD(String patientguid,String fisrt,String second,String third,String fouth){
    PreparedStatement state=null;
    String query="insert into processdicdcodeperpatient(patientGuid,first_symptom,second_symptom,third_symptom,fouth_symptom) values(?,?,?,?);
    try{
        conn=DBConn.getDBCponnection();
    }
state=conn.prepareStatement(query);
state.setString(1, patientguid.replace(',', ''));
state.setString(2, fisrt.replace(',', ''));
state.setString(3, second.replace(',', ''));
state.setString(4, third.replace(',', ''));
state.setString(5, fouth.replace(',', ''));

;
state.executeUpdate();

} catch(Exception e){
    e.printStackTrace();
}
finally {
try {
    if (conn != null) {
        conn.close();
    }
    if (state != null) {
        state.close();
    }
} catch (SQLException sqlee) {
    sqlee.printStackTrace();
}
}

public static List<HashSet<String>> getProcessedData()

    String splitBy="",;
List<HashSet<String>> itemsetList = new ArrayList<HashSet<String>>();
PreparedStatement st=null;
conn=null;
ResultSet rs=null;
try{
    getData();
    conn=DBConn.getDBCponnection();
    st=conn.prepareStatement("select first_symptom,second_symptom,third_symptom,fouth_symptom,patientGuid,class from processdicdcodeperpatient ");
    rs=st.executeQuery();
    while(rs.next()){\n           itemsetList.add(new HashSet<>(Arrays.asList(rs.getString(1),rs.getString(2),rs.getString(3),rs.getString(4),rs.getString(5),rs.getString(6))));

        }
    }
}
catch(Exception e){
    e.printStackTrace();
}
finally {
try {
if (conn != null) {
    conn.close();
}
if (st != null) {
    st.close();
}
}
public static List<String> getDatafortest()
{
    String splitBy=",";
    List<String> itemsetList = new ArrayList<String>();
    PreparedStatement st=null;
    conn=null;
    ResultSet rs=null;
    try{
        conn=DBConn.getDBCconnection();
        st=conn.prepareStatement("select ICD9Code from training_syncdiagnosis where ICD9Code REGEXP '^\[0-9\]+$'=1 order by ICD9CODE limit 20");
        rs=st.executeQuery();
        while(rs.next()){
            itemsetList.add((rs.getString(1)));
        }
    }catch(Exception e){
        e.printStackTrace();
    }
    finally {
    }
try {
    if (conn != null) {
        conn.close();
    }
    if (st != null) {
        st.close();
    }
} catch (SQLException sqlee) {
    sqlee.printStackTrace();
}

return itemsetList;

}
if (state != null) {
    state.close();
}
}

public static void insertC45Data(double original, double predicted, String rule) {
    PreparedStatement state = null;
    String query = "insert into c45data(original, predicted, rule) values(?,?,?)";
    try {
        conn = DBConn.getDBCconnection();
        state = conn.prepareStatement(query);
        state.setDouble(1, original);
        state.setDouble(2, predicted);
        state.setString(3, rule);
        state.executeUpdate();
    } finally {
        try {
            if (conn != null) {
                conn.close();
            }
            if (state != null) {
                state.close();
            }
        } catch (SQLException sqlee) {
            sqlee.printStackTrace();
        }
    }
}
public static List<HashSet<String>> getDataApriori()
{
    String splitBy="",";
    List<HashSet<String>> itemsetList = new ArrayList<HashSet<String>>();
    PreparedStatement st=null;
    conn=null;
    ResultSet rs=null;
    try{
        String Y="YES";
        String N="NO";
        // String query="select ICD9CODE,PatientGuid,case when(ICD9CODE between 390 and 459) then "'+Y'+" ELSE "'+N'+" END AS CLASS from training_syncdiagnosis where ICD9CODE between 390 and 459  ORDER by ICD9CODE limit 40";
        conn=DBConn.getDBCponnection();
        st=conn.prepareStatement("select patientId,ICD9Codes,class from icd9perpatient limit 50 ");
        rs=st.executeQuery();
        while(rs.next()){
            String values=rs.getString(2);
            String[] cols = values.split(splitBy); //
            if(cols.length<5){
                continue;
            }
            //
            insertToProcessdICD(rs.getString(1),cols[0],cols[1],cols[2],cols[3],rs.getString(3));
        }
    }
} catch (SQLException sqlee) {
    sqlee.printStackTrace();
}
}
itemsetList.add(new
HashSet<>(Arrays.asList(cols[0],cols[1],cols[2],rs.getString(3))));

}

catch(Exception e){
    e.printStackTrace();
}
finally {
    try {
        if (conn != null) {
            conn.close();
        }
        if (st != null) {
            st.close();
        }
    } catch (SQLException sqlee) {
        sqlee.printStackTrace();
    }
}
return itemsetList;

}

public static List<C45Model> getProcessedDataList(int limit)
{

    List<C45Model> itemsetList = new ArrayList<C45Model>();
    PreparedStatement st=null;
    conn=null;
ResultSet rs=null;
try{
    getData();
    conn=DBConn.getDBCponnection();
    String y="YES";
    String n="NO";
    st=conn.prepareStatement("select ICD9CODE,PatientGuid,case when(ICD9CODE between 390 and 459) then "+y+" ELSE "+n+" END AS CLASS from training_syncdiagnosis
    order by PatientGuid,ICD9CODE limit "+limit+" ");
    rs=st.executeQuery();
    while(rs.next()){
        C45Model p=new C45Model();
        p.setDisease(rs.getString(2));
        p.setAttribute(rs.getString(2));
        p.setItsClass(rs.getString(1));
        itemsetList.add(p);
    }
}

} catch(Exception e){
    e.printStackTrace();
}
finally {
try {
    if (conn != null) {
        conn.close();
    }
    if (st != null) {
        st.close();
    }
}
public static List<C45Model> getProcessedDataListOld() {

    List<C45Model> itemsetList = new ArrayList<C45Model>();
    PreparedStatement st=null;
    conn=null;
    ResultSet rs=null;
    try{
        try{
            getData();
            conn=DBConn.getDBCponnection();
            String y="YES";
            String n="NO";
            st=conn.prepareStatement("select ICD9CODE,PatientGuid,case when(ICD9CODE between 390 and 459) then "+y+" ELSE "+n+" END AS CLASS from training_syncdiagnosis order by PatientGuid,ICD9CODE limit 500 ");
            rs=st.executeQuery();
            while(rs.next()){
                C45Model p=new C45Model();
                p.setDisease(rs.getString(1));
                // p.setAttribute(rs.getString(2));
                p.setItsClass(rs.getString(3));
                itemsetList.add(p);
            }
        }
    }

    return itemsetList;
}
catch(Exception e){
    e.printStackTrace();
}
finally {
    try {
        if (conn != null) {
            conn.close();
        }
        if (st != null) {
            st.close();
        }
    } catch (SQLException sqlee) {
        sqlee.printStackTrace();
    }
}
return itemsetList;

public static void insertC45(String value,String clases,int count){
    PreparedStatement state=null;
    String query="insert into decisiontree_c45transdata(value,class,count) values(?,?,?)";
    try{
        conn=DBConn.getDBCponnection();
        state=conn.prepareStatement(query);
        state.setString(1, value);
        state.setString(2, clases);
        state.setInt(3, count);
        state.executeUpdate();
    }
}
catch(Exception e)
    
    e.printStackTrace();
    
} finally {
    try {
        if (conn != null) {
            conn.close();
        }
        if (state != null) {
            state.close();
        }
    } catch (SQLException sqlee) {
        sqlee.printStackTrace();
    }
    
}

public static void insertRulesModified(String rule)
{
    PreparedStatement state=null;
    String query="insert into rulesmodified(ruleValue) values(?)";
    try{
        conn=DBConn.getDBConnection();
        state=conn.prepareStatement(query);
        state.setString(1, rule);
        state.executeUpdate();
    }
    catch(Exception e){
        e.printStackTrace();
    }
    finally {
        try {
            if (conn != null) {
                conn.close();
            }
            if (state != null) {
                state.close();
            }
        } catch (SQLException sqlee) {
            sqlee.printStackTrace();
        }
        
    }
}
try {
    if (conn != null) {
        conn.close();
    }
    if (state != null) {
        state.close();
    }
} catch (SQLException sqlee) {
    sqlee.printStackTrace();
}