A Data Mining Approach for Predicting Causes of Accident Using Emergency Medical Data: The Case of St. Paul Hospital

A Thesis Presented

by

Abdurazak kamil

To

The Faculty of Informatics

of

St. Mary's university

In partial Fulfillment of the requirements

For the Degree of Master of Science

In

Computer Science

January 2020

ACCEPTANCE

A Data Mining Approach for Predicting Causes of Accident Using Emergency Medical Data: The Case of St. Paul Hospital

By

Abdurazak kamil

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

Internal Examiner

Full Name _____ sign and Date_____

External Examiner

Full Name _____ sign and Date_____

Dean, Faculty of Informatics

Full Name _____ sign and Date_____

Date of defense _____

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

Abdurezak Kamil

_____

Signature

Addis Ababa

Ethiopia

This thesis has been submitted for examination with my approval as advisor.

Getahun semeon (PhD)
Name of Advisor

_____

Signature

Addis Ababa

Ethiopia

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor Getahun semeon (PhD) for his constant support and his helpful suggestions and comments he has been giving me, above all for his stimulating suggestions and encouragement that helped me in all the time of research .

Next, my deep gratitude is to St. paul Hospital Laboratory department w/t Meryem Sherif and Ato Eyob for his valuable explanation and advice on Emergency departments and they have kindly helped me to get the data I needed for this study.

Finally, I share the credit of my work for the rest of all my family, friends and University of St. Mary's students that I have not mentioned their name here.

TABLE OF CONTENTS

Contents

LIST OF ABBREVIATIONS AND ACRONYMS

DSS- Decision support system

ED-Emergency Department

NGO's - Non Governmental Organizations

KDP- Knowledge Discovery Process

DM-Data Mining

KDD- Knowledge Discovery in Databases

KDPM-Knowledge Discovery Process Models

CRISP-DM-Cross Industry Standard Process for Data Mining

SEMMA -Sample Explore Modify Model Assess

CSV -Comma Separated Values

ARFF- Attribute Relation File Format

IC- Information Center

ANN- Artificial Neural Network

MLP- Multi-Layer Perceptron

NN- Neural Network

ROC- Receiver Operating Characteristics

WEKA- Waikato Environment for Knowledge Analysis

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Emergency medical health problem is one of the critical challenges that affect many people in the world, especially in developing countries. As statistics shows, in African region in general and Ethiopia in particular, the emergency medical case situation is still worse and needs special attention particularly Road Traffic Accident (RTA).

A lot of demographic and clinical (related) data is recorded about patients who come and receive treatment in the emergency medical service unit of the hospital. As these data are getting larger and larger, there can be a probability in which hidden, implicit and non-trivial knowledge exist within these data. So far it is recognized among scientific scholars that traditional statistical methods might not be good enough to discover such hidden knowledge from large and complex volume of data. This is where data mining becomes very important to mine such hidden, complex, necessary data to generate vital knowledge.

The problem is to be able to handle this huge amount of data and information in such a way that they can identify what is important and be able to extract it from the accumulated data. Now a days, data mining technology is being used as a tool that provides the techniques to transform these mounds of data into useful information which in turn enables to derive knowledge for decision making. A number of data mining techniques and tools are available to perform this task.

Thus, the purpose of this study is to explore the potential applicability of data mining techniques in predicting the cause of accidents based on emergency medical data by taking St. Paul Hospitals as a case. Some machine learning algorithms from WEKA software.

Keywords: Emergency medical data, *WEKA*, *Predictive Model*

CHAPTER ONE

INTRODUCTION

1.1 Background

Nowadays, with the growing demand for emergency medical care, the management of hospital emergency department (EDs) has becoming increasingly important. The management of patient influx is one of the most crucial problem in EDs throughout the world.

EMERGENCY Department is the frontline for hospital to face patients in emergencies. The members in Emergency Department include doctors, nursing staff, technicians, social workers, emergency medical technicians, administrative staff, fellow workers, and volunteers. This department runs 24 hours a day, first aids, examination and surgical operations can all be conducted here. It is like a small hospital in the hospital. Points out in his research that Emergency Department service is the hospital's first line of the medical care service, In the emergency department "triage" refers to the methods used to assess patients' severity of injury or illness within a short time after their arrival, assign priorities, and transfer each patient to the appropriate place for treatment .

In recent years, hospital's emergency service unit has been frustrated with a lot of problems, which include, large number of patients'-(overcrowding), delayed patient treatment, long processing time, and high costs. There are a lot of internal and external causes for these problems such as, characteristics of patients', shortage of professionals in emergency unit, lack of access to health care providers, and qualification of man power, patient arrival time, management practices and treatment strategies selected by emergency unit. Understanding these causes properly is the first step to solve them and hence it helps to improve the efficiency of patient care in emergency departments.

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. It has been estimated that an acute care hospital may generate five terabytes of data a year [1]. The ability to use these data to extract useful information for quality healthcare is crucial.

Computer assisted information retrieval may help support quality decision making and to avoid human error. Although human decision-making is often optimal, it is poor when there are huge

amounts of data to be classified. Also efficiency and accuracy of decisions will decrease when humans are put into stress and immense work. Imagine a doctor who has to examine 5 patient records; he or she will go through them with ease. But if the number of records increases from 5 to 100 with a time constraint, it is almost certain that the accuracy with which the doctor delivers the results will not be as high as the ones obtained when he had only five records to be analyzed. In order to solve this and many other problems in the health sector related to disease diagnosis, one has to come up with a way to extract hidden information from enormous datasets that are collected in the past. Data mining can be a solution by generating rules from those enormous datasets.

Predicting causes of Accident are one of the most interesting and challenging tasks in which to develop data mining applications. In recent years new research avenues such as knowledge discovery in databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers who seek to identify and exploit patterns and relationships among large number of variables, and be able to predict the outcome of a disease using the historical cases stored within datasets.

Predictive data mining in clinical medicine enables to derive models that use patient information to support specific clinical decisions. Data mining models can be applied for building decision-making procedures such as diagnosis, and treatment planning, which once evaluated and verified. Even though the application of data mining to medical data analysis has been relatively limited until recently, the term data mining has been increasingly used in the medical literature over the past few years. [2]

1.2 Statement of the Problem

Emergency department are confronted to exceptional events due to seasonal epidemics (in winter: influenza, colds ,bronchiolitis, etc…in summer: trauma) health crisis … these event can cause strain situation when disequilibrium between the care load flow(demand) and care production capacity(supply) is observed. The management of patient flow in such situation is one of the most important problem managed by ED managers. To handle this influx of patients,EDs require significant human and material resources, but this are limited .

Hospitals'-especially those of the industrialized countries, have employed different types of hospital information systems to manage their healthcare or patient data. These systems typically generate vast amounts of data in the form of number, text, chart, and image. Due to the vast amount, fast growth of data content, size, and diversity, researchers have focused on techniques to find useful information from collections of data .[2]

A Specialized Hospital has great number of patients who visit it with various emergency medical cases such as Road Traffic Accident (RTA), sudden collapse, poisoning, fall down accident ,diarrhea, chest pain, suicide, foreign body swallow and so on. These varieties of cases and the existence of greater amount of emergency case patients from different areas of the country resulted in an increase in the amount of emergency medical data. Hence these lots of patients' emergency medical data make difficult on finding useful knowledge.

Now a days, hospitals Emergency department is not supported by tools and techniques that can extract patterns or useful information from hospitals database for competitive advantage. The hospitals database contain enough data and it is used to infer missing information using statistical method which are not supported by modern technology. The demand of patients emergency cause of visit increases from time to time based on good quality service in the hospital. However, EDs is still unable to succeed this domain because of various reasons, this drawback could consists of problems related to patients satisfaction by his/her treatment and predict future work based on demand.

Generally there are a lot of problems and challenges regarding emergency medical service in Ethiopia which is due to many factors such as shortage of health professionals, facilities, resources and discovery of non-trivial knowledge from the collected data. On the other hand, there is a gap in knowledge generation from the massive data which is kept in the emergency unit, and which can result in poor handling or planning on emergency medical services.

This study is different in three ways, the first one is the data that is used for this study is collected from St. Paul Hospital Emergency department, the second one is the study specifically concentrated on predicting the status of Emergency cause of accidents in addition to identifying the determinant factor that affect the EDs, the third one is this study uses data mining technology. The problems of previous research were not only related to the small proportion of

data set used but also the data analysis was conducted simple statistical techniques. To this end, this work attempts to explore and answer the following research questions:

Therefore; this thesis addresses the following research questions:

Which types of emergency cause of visit are most significant?

Which type of emergency case is at greatest risk of mortality?

What is the potential of data mining technique to predict major cause of attending emergency medical services?

Therefore, to improve the current high pressure on emergency medical unit based on better planning and decision making at national level, understanding the case of medical health care is very crucial. Identifying factors that contribute to high accident rate and increasing need for emergency health care can be the major focus of this study. The result of this study will initiate the respective health authorities and policy makers to take action in improving the emergency health care services and minimizing causes of accidents.

1.3 Objective of the study

1.3.1 General Objective

- ✓ The general objective of this study is to explore the potential of data mining technique for predict causes of Accidents based on emergency medical data.

1.3.2 Specific Objectives

- ✓ To investigate and transform the dataset in to a form, that can be used by data mining tools for additional research conduction.
- ✓ To recognize potentially appropriate attributes of emergency cause of visit and build models by conducting experiments.
- ✓ To extract rules and patterns of medical visit in addition to their relationships those can potentially influence accidents.
- ✓ To test and evaluate the results
- ✓ To construct a DSS using the model selected as a working model.

1.4 Scope and Limitations of the Study

The scope of this research is limited to assessing the possible application of data mining technology for St. Paul hospital data sets and identifying determinant factors that affect the emergency cause of visit. Different data mining techniques can be applied to the problem domain. Due to time constraint only classification mining techniques are applied.

During the study, there were also challenges that the researcher came across to develop the predictive model of emergency medical data. These challenges limited the model to perform more than this performance in one or the other way. In order to assess level of impact of each factor it is better to use different data source which holds attributes in parallel with influencing factors. Here, it is limited to develop the model using only St. Paul hospital in the emergency department and doesn't include other services in the hospital.

1.5 Significance of the Study

The research was targeted to explore the benefit of data mining technology for exploring significant knowledge from the emergency medical dataset which can be used to generate new or unseen knowledge that can improve the current status of the emergency medical health care by building applicable model and identifying dominant factors related to the cause of visit.

This research will give benefit in predicting and identify which type of emergency medical care to reduce emergency death and promote health care by improving pre and post health care services. It will support the National Health Care Policy in revising the existing rules, policies, and suggest new rules and policies. More specifically this research work can be used to revise and strengthen the already implemented emergency medical Care Process because it focuses on developing the medical health care profile and providing relevant emergency health care information for the specific emergency case victims. This reduces the emergency health bulky task i.e., predicting emergency medical risks. The result of the study can also be used the predictive model, as a frame work for keeping emergency cause of visit in medical health care.

In general the health center will get benefit from the research result to keep and improve emergency medical care in Ethiopia.

Moreover, the output of the study might be used as a benchmark for other researchers who have interest in such related issue or area

1.6 Ethical Considerations

The research has nothing to do with the personal identifiers (like name and address) of individual about whom the data is collected and therefore there is no problem of privacy and the confidentiality of individuals

The research is purely dedicated to academic purpose (Master's Thesis for the Partial Fulfillment of M.Sc. Degree in Computer Science)

The research is purely for public benefit in general and hospitals Emergency Departments under study to improve their cause of visit in particular.

1.7 Thesis Organization

This research report is arranged into five chapters. The first chapter briefly discusses background to the problem area and data mining technology, and states the problem, Objective of the study, significance of the study, and scope and limitation of the study. Chapter two reviews background materials necessary to understand the basic concepts and results of this research. Chapter three explains the methodology of the study. Chapter four presents the experimentation phase of the study. Results of training and testing of the models were also discussed here, and finally, describes a prototype that demonstrates how data mining in general, and decision tree in particular significantly improves emergency cause of visit. Finally, Chapter five provides conclusion, and offer recommendations for future work.

CHAPTER TWO

LITERATURE REVIEW

In this study, an attempted has been made to survey and incorporate a range of standards and steps used in data mining methodologies, various concepts, theories and practices of data mining in relation with practical emergency medical cause of visit problems and concepts. Related research works, which reveal the applicability of data mining technology in different domains, are also considered.

Using data mining with the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to apply powerful tool for analysis and interpretation of such data there by for the extraction of interesting knowledge that could help in decision-making.

2.1 Data Mining and Knowledge Discovery

The knowledge discovery process(KDP), also called knowledge discovery in databases, seeks new knowledge in some application domain. It is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [5]. The KDP has many steps (one of them is DM), each step has its own a particular completed discovery task and each accomplished by the application of a discovery method. Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain. Data mining is one step in the KDD process. It is the most researched part of the process [6]. Data mining can be used to verify a hypothesis or use a discovery method to find new patterns in the data that can make predictions with new, unseen data. Data mining has many definitions in many literatures; David defines as “data mining is the analysis of large observational data sets with the goal of finding unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner” [5]. A data set can be “large”, in the sense that it contains a large number of records or that a large number of variables is measured on each record.

Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases [7]. Data mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data; the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases.

2.2 Knowledge Discovery Process Models (KDPM)

The three models that are commonly used for data mining process models in this research are; Knowledge Discovery in Database (KDD), Cross-Industry Standard Process for Data Mining (CRISP-DM), Sample, Explore, Modify, Model, Assess (SEMMA). Furthermore, the integration of academic research (KDD) and industrial models (CRISP-DM) has introducing to the development of a hybrid model [6]. Model illustrates processes in each of KDPM steps that are carried out. While implementing a knowledge discovery project the KDPM Include a set of procedural steps to be followed by practitioners. Standardizing knowledge discovery projects within a general structure is the key motivation for introducing process models. As a result, the following aims will result cost and time saving, advance in understanding, success rates, and acceptance of such projects.

2.2.1 The KDD Process

It is the procedure of using DM techniques to mine what is considered knowledge based on the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database [7]. There are five phases in KDD As the figure below shows:

Selection- The selection phase includes emphasizing on a variable subset, or sample data's or building a data set target, on which finding is to be implemented.

Pre-processing- This phase comprises on the data target, preprocessing and cleaning in order to achieve consistent data. On this stage we also attempt to eradicate noise that existed in the data. Noise indicates that some form of error within the data. For filling missing values and elimination of duplicates in the database some of the tools can be used here.

Data Transformation- This phase consists of discovering important attributes to represent the data, based on the aim of the task, and using transformation methods or dimensionality reduction to decrease the efficient number of variables under consideration.

Data Mining- this is the most important phase of KDD. It consists of three subcomponents: selecting data-mining task (e.g., Classification, clustering, summarization), selecting algorithms to be used in investigating for patterns, and selecting the correct algorithms for patterns.

Interpretation/Evaluation- This phase includes the interpretation of extracted patterns and combining discovered knowledge.

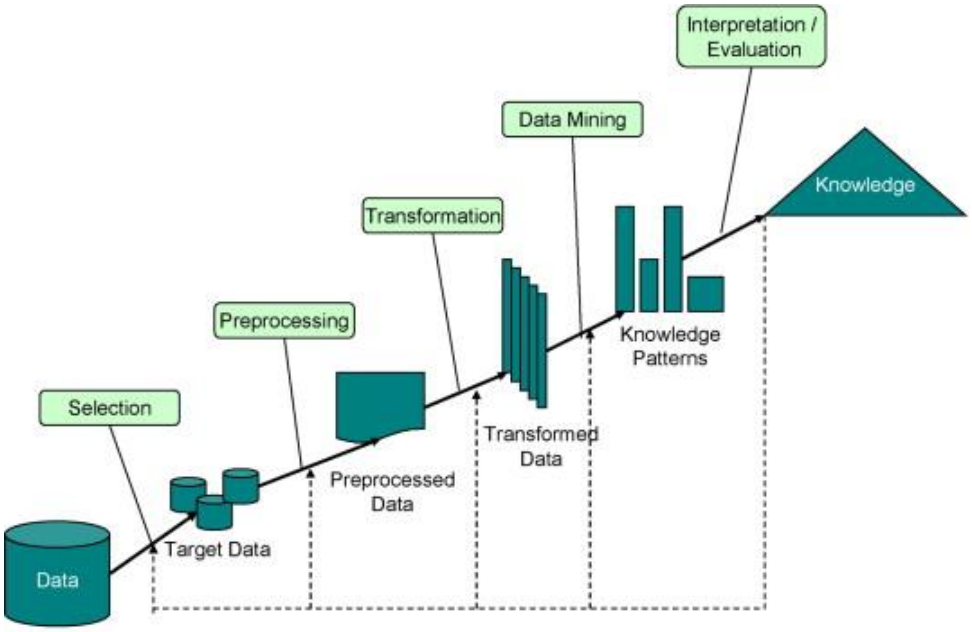

Figure 2.1 KDD Process

2.2.2 The Crisp-DM Process

The CRISP-DM process was initially composed by the means of the effort of a consortium with DaimlerChrysler, Statistical Package for Social Sciences (SPSS) and Numeric Character reference (NCR) then In the late 1990s four corporations namely: Integral Solutions Ltd. (A provider of commercial data mining solutions), NCR (a database provider), DaimlerChrysler (an automobile manufacturer), and OHRA (an insurance company) was first developed the CRISP-DM (CRoss-Industry Standard Process for Data Mining) [8]. The development of this process model has the benefit of strong industrial support. It has also been supported by the European

Strategic Program on Research in Information Technology **(**ESPRIT) program funded by the European Commission. As indicated in the figure below this process model has cyclic that encompasses six steps.

Figure 2.2 CRISP-DM process

Business understanding. This phase emphasizes on the understanding of aims and requirements from a business point of view. In addition it translates these into a DM problem definition, and proposes a preliminary project plan to obtain the objectives.

Data understanding: It's most important activities are collecting initial data, describing data, exploring data and verifying data quality [10]. This stage begins with an initial data collection and continues with activities such as identifying data quality problems, finding out first insights into the data or discovering interesting subsets to form hypotheses for hidden information in order to be came familiar with the data.

Data Preparation: The data preparation phase consists of all the activities necessary to build the final dataset from the initial raw data. After the data resources available are identified, they need to be selected, cleaned, built into the form desired, and formatted [9]. Data cleaning and data transformation in preparation of data modeling need to happen in this phase. Data exploration in-depth can be applied during this phase, and additional models utilized, again providing the opportunity to see patterns based on business understanding.

Modeling*:* Data mining software tools such as visualization (plotting data and establishing relationships) and cluster analysis (recognizing which variables go well together) are valuable for initial analysis. Tools such as generalized rule induction can develop initial association rules. Once greater data understanding is gained (often through pattern recognition triggered by viewing model output), more detailed models appropriate to the data type can be applied. The division of the data into training and test sets is also needed for modeling [8]. Generally, in this phase, various modeling techniques are selected and applied and their parameters are adjusted to optimal values.

Evaluation: After one or more models have been built that have high quality from a data analysis perspective, the model is evaluated from a business objective point of view. A review of the steps executed to construct the model is also performed. The key sub-steps in this step include evaluation of the results, process review, and determination of the next step [8]. Its key objective is to determine whether any important business issues have not been sufficiently considered.

2.2.3 The SEMMA Process

Data mining can be viewed as a process rather than a set of tools, and the acronym SEMMA stands for (Sample, Explore, Modify, Model, and Assess) refers to a methodology that clarifies this process. As explained in [11], SEMMA focuses on the model development aspects of data mining and involves the following logical steps (see figure 2.3):

1. **Sampling:**This stage extracts a portion of large dataset such that the sample taken is big enough to contain the significant information, or else small enough to manipulate quickly.

2. **Exploring:**The step illustratessearching for unanticipated trends and anomalies in order to gain understanding and ideas. Exploration helps refine the discovery process. If visual exploration does not reveal clear trends, it is possible to explore the data through statistical techniques.

3. **Modifying:** This stage includes creating, selecting, and transforming the variables to focus the model selection process. Based on the discoveries in the exploration phase,

one may need to manipulate the data to include information such as the grouping of clients and significant subgroups, or to introduce new variables. One may also need to look for outliers and reduce the number of variables, to narrow them down to the most significant ones.

One may also need to modify data when the mined data change. Because data mining is a dynamic, iterative process, one can update data mining methods or models when new information is available.

4. **Modeling:**This stage allows the software to search automatically for a combination of data that reliably predicts a desired outcome.

5. **Assessing:** This step involves data evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs. A common means of assessing a model is to apply it to a portion of data set aside during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, one can test the model against known data.

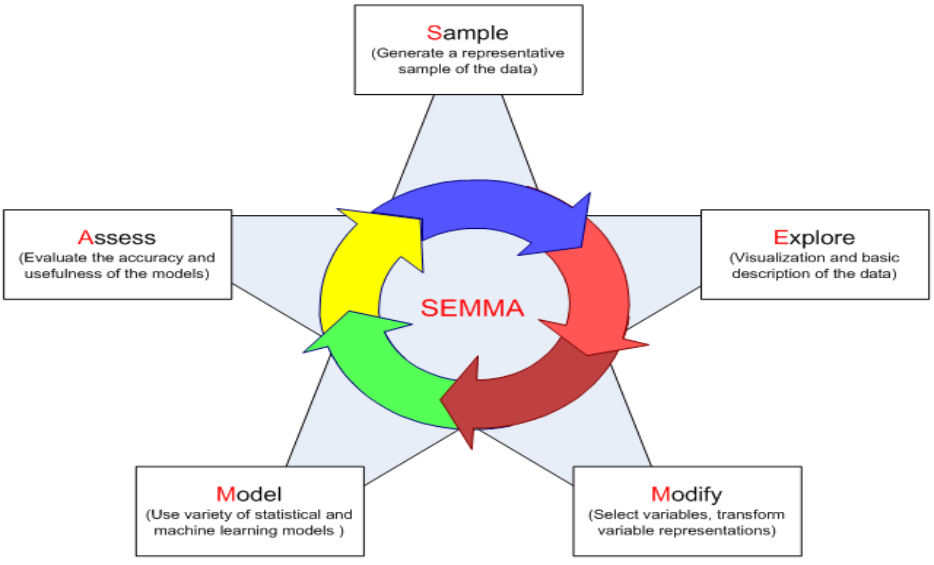

Figure 2.3: SEMMA process model

By assessing the results gained from each stage of the SEMMA process, one can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data.

The SEMMA process offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for its conception, creation and evolution, helping to present solutions to business problems as well as to find the DM business goals [11].

2.2.4 Hybrid Process Model

The development of both academic mainly the KDD and industrial oriented (CRISP-DM and other) data mining models has led to the growth of hybrid models that combines the features and job of both models. Hybrid model is a six-step KDP model (see Figure 3) developed by Cios et al. (2007) [12]. It was developed mainly based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include introducing several new explicit feedback mechanisms and in last steps the knowledge discovered for a particular domain may be applied in other domains.

According to [12] the description of the six steps are explained below

Understanding the Problem Domain: To define the problem and determine the research goals, identifying key people, and learning about current solutions to the problem this stage includes working closely with domain experts. It involves learning domain-specific terminology. A description of the problem and its restrictions is done. Research goals are translated into the DM goals, and include initial selection of DM tools to be used.

Understanding of the Data: Involves collection of sample data and deciding which data will be needed including its format and size. Background knowledge is used to guide these efforts. Data is checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Includes verification of the usefulness of the data in respect to the DM goals.

Preparation of the Data: This stage is one of the most important activities as the presence of irrelevant data during the analysis phase may yield wrong results. This is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about

half of the entire research effort. In this step, which data will be used as input for data mining tools of step 4, is decided. It may involve sampling of data, data cleaning like checking completeness of data records, removing or correcting for noise, etc. the cleaned data can be, further processed by feature selection and extraction algorithms(to reduce dimensionality), and by derivation of new attributes(also named as discretization). The result would be new data records, meeting specific input requirements for the planned to be used data mining tools.

Data Mining: This is another key step in the knowledge discovery process. Although it is the DM tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned DM tools and selection of the new ones. DM tools include many types of algorithms, such as classification, clustering, preprocessing techniques, machine learning, etc. This step involves the use of several DM tools on data prepared in step 3. First, the training and testing procedures are designed and the data model is constructed using one of the chosen DM tools; the generated data model is verified by using testing procedures.

Evaluation of the Discovered Knowledge: This step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only the approved models are retained. The entire data mining process may be revisited to identify which alternative actions could have been taken to improve the results.

Use of the Discovered Knowledge: This step is entirely in the hands of the owner of the database. It consists of planning where & how the discovered knowledge will be used. The application area in the current domain should be extended to other domains

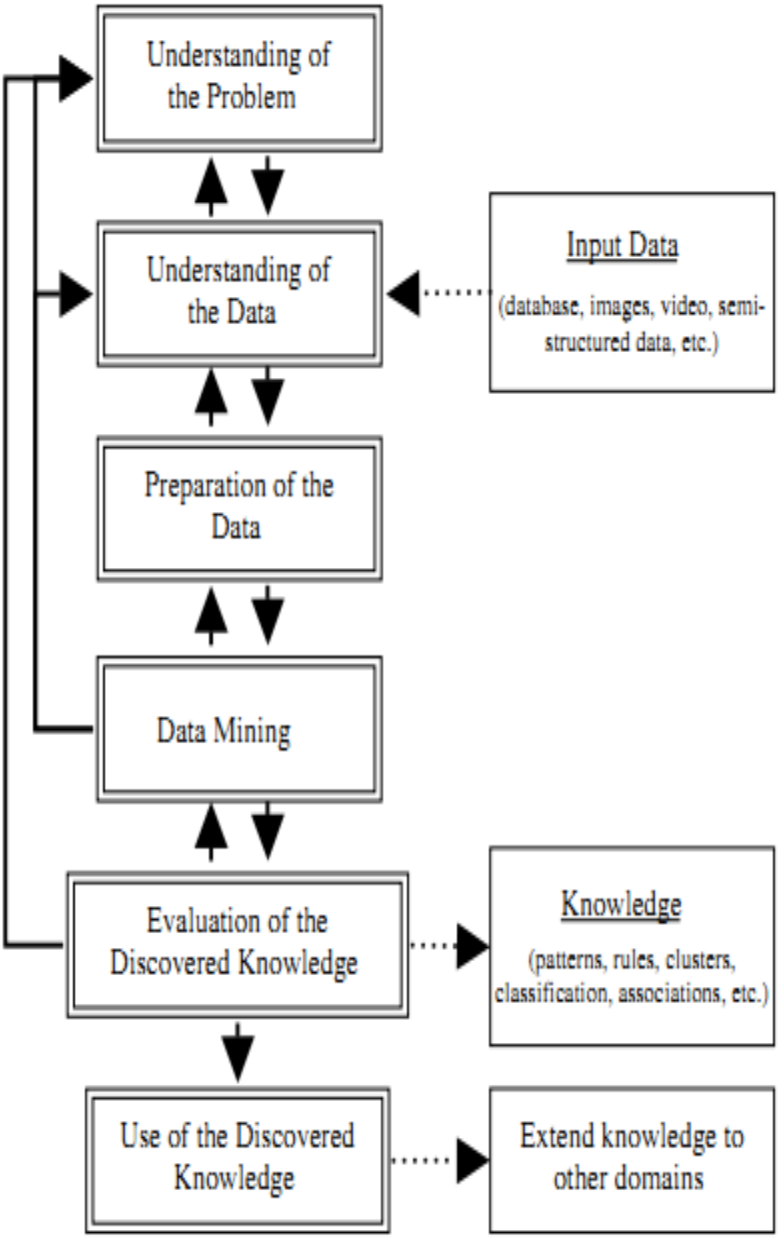

Figure 2.4.Cios et al. model

2.2.5 Summary of Data Mining Models

Nowadays, research efforts have been focused on proposing new models, rather than improving the design of a single model or proposing a generic unifying model. Despite the fact that most models have been developed in isolation, a significant progress has been made. The subsequent models provide more generic and appropriate descriptions. Most of them are not tied

specifically to academic or industrial needs, but rather provide a model that is independent of a particular tool, vendor, or application [12].

Data Mining Model	**KDD**	**SEMMA**	**CRISP-DM**	**HYBRID**
Number of steps	Pre KDD	----------	Business understanding	Understanding of the problem domain
	Selection	Sample	Data understanding	Understanding of the data
	Preprocessing	Explore		
	Model	Modify	Data preparation	Preparation of the data
	Data mining	Model	Modeling	Data mining
	Interpretation/evaluation	Assessment	Evaluation	Evaluation of the discovered knowledge
	Post KDD	----------	Deployment	Use of the discovered knowledge

Table 2.1: Comparison of Data mining process model

As shown in the table 2.1 most of the steps to be followed in all methods look like similar. Nevertheless, KDD and SEMMA do not have a step "understanding the problem" before selection and sample steps respectively. Definitely, understanding the domain problem is often a prerequisite step but missed by the two methodologies.

In general, According to [14] all methodologies contain and followed same data mining tasks. Hence this research be conducted based on an understanding of the domain problem and business context, consequently, since it completes the limitations of the others, the optimal methodologies that fit to conduct this research is hybrid model.

2.3 Data Mining Tasks

Based on the kind of patterns to be mined from database tasks of data mining are very diverse and distinct. To discover different types of patterns and/or knowledge Data mining applications employ different kinds of data mining methods and techniques. According to [15], the different data mining tasks are Predictive modeling; descriptive modeling, exploratory data analysis, patterns and rules discovery, and retrieval by content.

2.3.1 Predictive Modeling

Predictive modeling allows the value of one variable to be predicted from the known Values of other variables. Some examples of predictive modeling are Classification, Regression, Time series analysis and Prediction. Many of the data mining applications are aimed to predict the future state of the data According to [16]. It does this by building a model of the real world based on data collected from a variety of sources which may include corporate transactions, customer histories and demographic information, process control data, and relevant external databases such as credit bureau information or weather data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefine groups or classes. It is a supervised learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that maps data item to real valued prediction variable.

2.3.2 Classification

Classification analysis is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown [17]. It is possible to organize data in a given class using classification. The classification uses given class labels to organize the objects in the data collection in an orderly manner. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern [17]. Classification model is one of the most commonly used supervised modeling techniques. In classification, a user needs to divide data into segments and then make distinct non-overlapping groups. For dividing data into groups, a user needs to have certain information about the data to be divided into segments. Classification problems aim to identify the characteristics that indicate the group

to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave.

According to [18] Classification approaches normally use a training set where all objects are already associated with known class labels. Then the classification algorithm learns from the training set and builds a model. The model is used to classify new objects. In other words, classification is a two-step process, first a classification model is built based on training data set and then the model is applied to new data for classification. There are different algorithms that are used for classification purpose such as, decision tree, neural network, genetic algorithm, naïve bayes, etc.

Decision Tree

A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions [17].In data mining, a decision tree is a predictive model which can be used to represent both classifiers and regression models. Decision tree algorithm is a data mining induction technique that recursively partitions a dataset of records using either depth-first greedy approach or breadth-first approach until all the data items belong to a particular class [17]. Decision trees constitute a way of representing a series of rules that lead to a class or value [19] and it is a powerful and popular tool for classification and prediction.

Decision trees are also useful for exploring data to gain insight into the relationships of a large number of candidate input variables to the target variable. Since decision trees combine both data exploration and modeling, they are a powerful first step in modeling process even when building the final model using some other techniques [18]. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable [21]. The target variable is usually categorical and the decision tree model is used either to calculate the probability that a given record belongs to each of the categories, or to classify the record by assigning it to the most likely class. Decision tree can also be used to estimate the value of continuous variable.

Decision trees are part of the Induction class of DM techniques. An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. Rules are applied successively one

after another, resulting in a hierarchy of segments within segments. In predictive modeling, the decision is simply the predicted value. When a decision tree is used for classification tasks, it is more appropriately referred to as a classification tree and when it is used for regression tasks, it is called the regression tree. The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent rules [21]. Rules can readily be expressed so that humans can understand them directly.

Decision tree models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make prediction [19]. Various decision tree algorithms such as CHAID (Chi-squared Automatic Interaction Detection), C4.5/5.0, CART (Classification and Regression Trees), J48 and any with less familiar acronyms, produce trees that differ from one another in the number of splits allowed at each level of the tree, how those splits are chosen when the tree is built, and how the tree growth is limited to prevent over-fitting [3].

Rule-Based

Rule induction is one of the techniques most used to extract knowledge from data, since the representation of knowledge as if/then rules is very intuitive and easily understandable by problem-domain experts [20]. Even though the pruned trees are more compact than the originals, they can still be very complex. Hence, to make a decision tree model more readable, it can be transformed into an IF-THEN decision rule [22]. As stated by Larose [23] decision rules can be generated from a decision tree by traversing any given path from the root node to any leaf. The complete set of decision rules generated by a decision tree is equivalent to the decision tree itself. A rule-based or decision rule classifier such as C4.5 uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form IF condition THEN conclusion. If the condition in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied and that the rule covers the tuples [20].

Neural Network

Neural network is a complex modeling technique based on a model of a human neuron. A neural net is given a set of inputs and weights to predict one or more outputs depending on the selected architecture [24]. One of the most common neural network architectures has three layers. The first layer is the input layer, the layer exposed to external signals. The input layer transmits signals to the neurons in the next layer, which is called a hidden layer. The hidden layer extracts

relevant features or patterns from the received signals. Those features or patterns that are considered important are then directed to the output layer.

Neural networks are powerful mathematical models suitable for almost all data mining tasks, with special emphasis on classification and estimation problems. Larose [23] stated that, neural networks are quite robust with respect to noisy data [25]. Han & Kamber [20] also stated the success of neural networks on a wide area of application including handwritten character recognition and medicine due to their high tolerance of noisy data and their ability to classify patterns on which they have not been trained. However, Han & Kamber [20] argue that, neural networks require long training times and have been criticized for their poor interpretability. Therefore, neural networks are more suitable for applications where this is feasible.

2.4 Related Works

Numerous works on different domains related to using data mining techniques for discovering and using new knowledge extracted from huge amount of datasets have motivated this study.

2.4.1 Application of Data Mining in Different Domains

A work done by Tsegayeet. al (2010) conducted a statistical research with a title "Pattern of Fatal Injuries in Addis Ababa, Ethiopia: A One-year Audit." in Addis Ababa, Ethiopia.

In their introduction they explained about injury which is one of the leading causes of death and disability in the developing countries and it is common but largely neglected health problem in developing countries including Ethiopia .It is the primary reason for an emergency hospital visit in Addis Ababa hospitals such as TikurAnbessa Specialized Hospital.

Their audit was devised mainly to assess the burden of fatal injuries together with identifying common causes of fatal injuries in Ethiopia. It was specifically designed to determine the profile or pattern of commonly occurring fatal injuries, and medical attention received before death.

Tariku Debela [26] an attempt was made on Developing a Predictive model for fertility preference of women of reproductive age using data mining techniques. 16515 dataset were used for training and testing within 15 attributes using EDHS 2011 (Ethiopia Demographic Health Survey) women's survey dataset collect by CSA. To develop a model the researchers followed Hybrid methodology. Three data mining classification algorithms, J48, Naïve Byes and neural

Network (Multilayer Perceptron), are experimented to build and evaluate the models. The research output indicated that J48 classifier performed well with the accuracy of 78.03% in predicting the fertility preference of women.

Other work done by Tesfahun Hailemariam [27] tries to construct adult mortality predictive model using data mining techniques so as to identify and improve adult health status using BRHP open cohort database. Under this research the methodology employed was hybrid knowledge discovery process model and a total data set of 62,869 was used. Decision tree and Bayes classifiers applied to build the predictive model. Finally the result indicated that J48 decision tree classifier using pruned technique was selected as optimal model. Finally, the study suggests that education plays a considerable role as a root cause of adult death, followed by outmigration.

Daniel Mamo [28] conducted to explore the potential applicability of the data mining technology in developing models that can detect and predict fraud suspicious in tax claims with a particular emphasis to Ethiopian Revenue and Custom Authority. The methodology applied in this research was the six-step Cios et al. (2000) KDD process model and a total of 13280 records were used. J48 decision tree classifier and PART rule induction algorithms were selected for experiments. The study implements clustering algorithm followed by classification techniques for developing the predictive model. The classification task of this study is carried out using the J48 decision tree , PART and neural network algorithms in order to create model that best predict cause of accidents..

Most of the research gaps regarding health care are classification for prediction is varying in processing method, machine learning algorithm, and datasets to get more accurate result. Different departments have their own describing attributes. EDs are inter connect for different departments and with different referral hospitals. Researches, conducted regarding this area is more of directly pertinent to specific scope. This paper mainly focuses on classification of EDs. There is no prior research conducted on classification of predicting Emergency cause of visit. During the review of prior research on the related area there is a gap between data types used and currently available data of the emergency medical departments in the hospital. In addition to this, most of the researches were conducted using either one or two machine learning algorithms. Here, it is proposed to apply several machine learning algorithms to show their effects on

classification of predicting emergency cause of visit in the EDs, and to add theoretical contribution to the area.

CHAPTER THREE

RESEARCH METHODOLOGY

The general purpose of this research is to build a model using Data Mining Approach for Predicting Cause of visit using Emergency medical data based on the datasets collected from the hospital emergency department. In this chapter the study describes the data mining research approach, source of data, methods, tools, techniques and algorithms that is used in building the model addressing the research problem.

3.1 Research Design

The overall research design follows hybrid model of discovering hidden knowledge from the St. Paul hospital emergency department through data mining technique. As it is discussed in chapter two data mining process models are developed for purely academic purposes. By combining the two a hybrid model is also developed. This hybrid model is organized by taking the cross Industry Standard Process for Data Mining (CRISP-DM) model by adopting it to academic researches. The process model adopted to undertake this research is the hybrid one due to the reason that this model describes each of the knowledge discovery process steps in a better way and it is flexible since it has a feedback mechanism in more steps than the CRISP-DM.

The hybrid model has six steps that are : understanding of the problem, understanding of the data , preparation of the data , data mining , Evaluation of the discovered knowledge and use of the discovered knowledge. The research is designed based on steps of this process model.

.

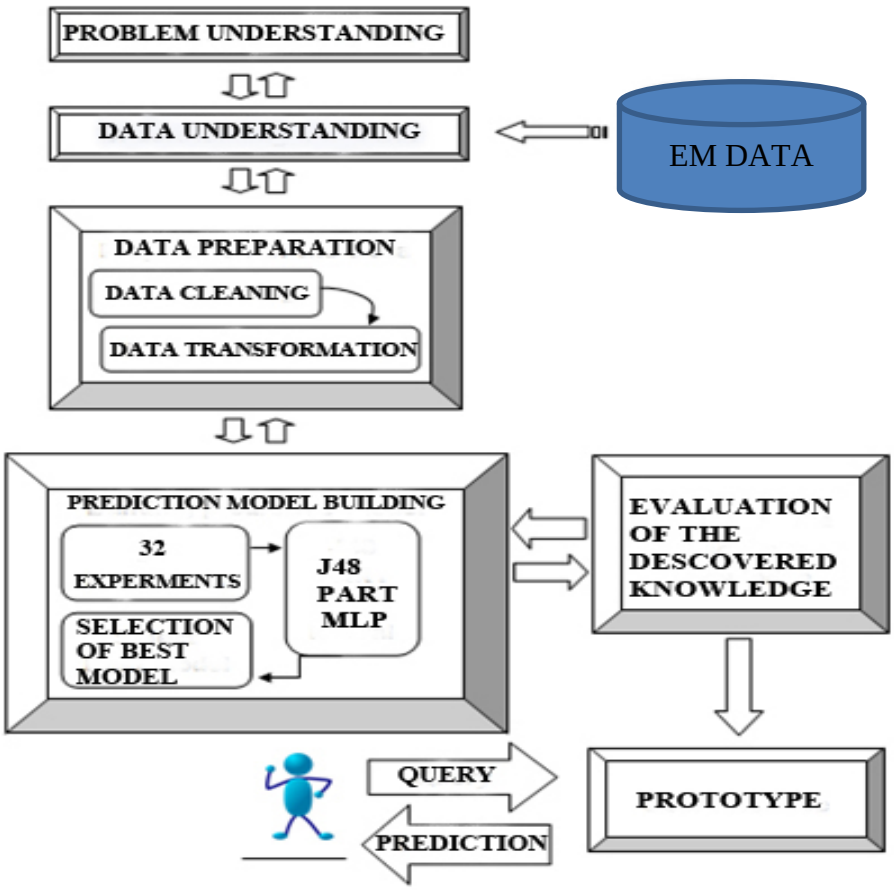

Figure 3.1: The overall process of the study

As described in figure 3.1, the Primary task is problem understanding during which the overall activity in the problem domain is analyzed and important attributes are selected. Then the next step is understanding of the data during which the information about the selected attributes are discussed in detailed and before the dataset passes into the data mining algorithm, preprocessing technique such as fill in missing values, smooth noisy data, identify or remove outliers, resolve inconsistencies, data reduction and data transformation are applied on the dataset. The preprocessing phase makes the input data suitable for the data mining algorithm. Once the preprocessing has been done, the preprocessed data are used to create the classification model by training the data mining algorithms. After having the classification model, the next very important step is the evaluation of the model with the test data set. By doing this we measure the performance of the model created. And finally we test our model with a real data.

3.2 Business Understanding

Business understanding is one of the essential phases in data mining process. Hence, in this research there were some measures taken to understand the business by using different methods such as reviewing documents, observation and discussion with domain experts were carried out to build business understanding.

A Couple of discussion questions were provided for understanding the problem domain. Based on the answer of domain experts the answers were organized as follows:-Now a days emergency medical data is the hottest research area in developed and developing countries especially in developing countries. Emergency medical data is a wide spread hottest research problem domain in the world. The primary goal of Health care is Prevention mechanism because it highly contributes in reducing the percentage of death rate.

3.2.1 Work Flow at Emergency medical service unit of St. Paul Hospital

The entire treatment process in emergency medical unit of the hospital may take short or long time depending on the type of emergency case, number of patients, availability of resources and other factors. In general, the overall processes followed to get service at emergency unit is presented below:

Step 1: Separation and registration in triage section: The patient whose case is suspected as emergency first comes to the Hospital emergency medical service unit Triage area. Here the nurse differentiates whether the patient is emergency case or not by observing the case, checking vital signs and asking the patient or his/her contact person. If the nurse believes the case is emergency, he or she provides give card to patient. Then, the patient takes out medical card from the hospital card room. In the card room, various socio- demographic information such as (name, address, age, sex, marital status and so on) is recorded on the medical card of the patient.

After separation, triage nurse fills various information such as socio demographic information, event information (i.e, activity, place of injury and intent) and some clinical information (i.e vital sign data and history of past medical illness) on the triage form by observation, asking the patient or his/her contact person, observing medical card information and values obtained from vital sign medical instruments.

Finally the nurse indicate or shows the patient where to go for treatment.

Step2: Treatment: In the treatment section of the emergency unit responsible doctors diagnose the patient for further investigation. Depending on the case, the doctor may order laboratory,

X-ray, surgery ,medicine, refer for admission in emergency department, to other units even to other Hospitals .In this section, the doctor records the details of the investigation on the patient's medical card.

Step3: Entering patient information on the emergency medical registration data base:

In this section, the data clerk (health informatics in profession) enters or fills the details of the patient socio- demographic, event and triage data from the medical card and triage form to the database.

After the clerk finishes entering all the necessary data, the medical card together with triage form is returned and stored to the card room.

Step4:Follow up: If the patient is admitted, then, normal follow up process is continued .Here, nurses and doctors carry out the process.

As statistics shows emergency medical cases are among the main causes of disability and mortality in Ethiopia. There are large number of patients who visit St.Paul Hospital. Since it is one of the specialized Hospital in the country, it is also a referral Hospital that can accept patients all over the country and its location also make it the center for many of the emergency medical cases. These and other reasons make the hospital very crowded with patients in general and emergency case patients in particular. As a result, there is a huge volume of emergency medical patients' data which is kept unprocessed with regard to knowledge generation.

Therefore, the main reason that necessitated this research is to explore these huge volumes of data that can be used to discover knowledge and patterns that can play a role in awareness, planning and decision-making in emergency care environment. The data can be used beyond simple statistical analysis such as data mining. It uncovers important data patterns that contribute greatly to business strategies in providing a novel knowledge that can be used as a base for guidance and decision making.

Identifying the important factors or variables among the emergency patient variables have significant impact for the problem domain. Moreover, predicting or classifying which age group, sex, place, activity and other independent variables are more likely to be associated with the cause of visit dependent variable which is important for the situation. Hence, data mining technology can offer enormous potential in predicting and mining the hidden characteristics and patterns that exist within the dataset.

3.3 problem understanding

As describe on appendix 1, we have asked a couple of discussion questions to domain experts for understanding the problem domain. Based on the answer of domain experts the answers are organized as follows:- Is emergency cause of visit in the hospital is a hot issue now a day? What techniques do you currently use to identify emergency causes? What are the existing solutions for reducing the emergency cases in the hospitals?

3.4 Understanding the Data

After understanding the problem to be addressed clearly in this study, the next step is analyzing and understanding the data available. The outcome of DM and knowledge discovery heavily depends on the quality and quantity of available data [22]. The data that is used in this research was initially collected. The study was conducted using secondary data obtained from St. Paul hospital Emergency department. Originally the data is collected in paper format and from its database and then converted it into excel format.

At this stage, the data that was used in this research was described briefly. The description includes listing out attributes, their respective values, data types, and evaluation of their importance etc… as well as visualization of the data to see data distribution department wise. Careful analysis of the data and its structure was done together with domain experts by evaluating the relationships of the data with the problem at hand and the particular DM tasks to be performed.

The researcher consulted domain experts to understand the situation related to emergency cause of visit in the hospital and to have an overview of the problem domain. The domain experts that were communicated include doctors, nurses, data clerks and database administrator from emergency unit of the Hospital. Among these experts, doctors and nurses were concerned with

the process of patient treatment and the consultation with them provide what exactly the business process is and what kind of data are being captured during emergency medical care.

The Data clerks are actually concerned with entering the data from the medical card and triage form to the database through user interfaces. The consultation of them provides the overall picture of data encoding to the system such as the flow or logic, problems encountered and so on.

The Database administrator is concerned with the database design and management basically taking backup, maintaining the system and training users. The consultation with the administrator provide the overall picture of emergency medical registration system such as the database tables, relationships among tables, attributes and interfaces .The source data was also obtained by consulting the administrator. Data quality problems were identified with the discussion with database administrator, triage nurses and data clerks. As a result the following factors can be raised; inconsistent data encoding, incorrect source data value, missed or unknown source data value, wrongly recorded data value on the patient card or database and so on.

At this stage, the data that is used in this research is described briefly. The description includes listing out attributes, their respective values, data types, and evaluation of their importance etc… as well as visualization of the data to see data distribution department wise. Careful analysis of the data and its structure is done together with domain experts by evaluating the relationships of the data with the problem at hand and the particular DM tasks to be performed.

3.5 Data collection

The source data employed in this research was collected from St. Paul Hospital emergency medical registration database. A backup of the database tables which were exported in to excel format was taken from emergency unit.

The Hospital emergency medical registration database contains around 10200 total records during the time of data collection (until January 2019 G.C) but 10080 records were used for the research experimentation since 120 records were removed due to inconveniences because some were duplicate and did not have medical card number

3.5.1 Data source description

The emergency medical database is designed by MySQL Database Management System (DBMS). The data clerks (who are nurses by profession) have the interface (designed by web-

based system) and through which they can enter patient demographic, Event, Triage, Trauma and Non-trauma related data. The data clerk first enters his or her user name and password for authentication purpose. Once the user is authorized the system displays personal information or socio demographic page and the user fills the necessary data from the card and triage form. The database contains more than four tables or files, but for the research purpose four tables were selected based on significance to the problem domain. The tables were Personal information or socio demographic, Event, Trauma triage and Non-trauma triage.

Each table has more than five attributes (variables) but based on the appropriateness to the problem domain the following fourteen were selected from all tables. The selected attributes from the Personal information table include Medical Record Number,Sex ,Age ,Marital status ,Subcity and Region more description for all selected attributes is given in this section on Table 3.1.The selected attributes from the Event table include Medical Record Number, cause of visit, intent, activity, place of injury and referred from. The selected attributes from each of the two tables (Trauma and Non-trauma triage tables) include Medical Record Number, past medical illness, Triage assessment and Transferred to. These four tables were joined by Medical Record Number (i.e it is primary key which uniquely identifies records)to get the full integrated attributes for each individual patient record.

Full description for the above selected attributes is given in the table 3.1.

S.No	**Attributes**	**Description**	**Values**	**Data Type**
1	MRNo	Medical Record Number	Numeric value	numeric
2	Gender	Gender of the patient	Female and Male	Nominal
3	Age	Numeric age value	Age of the patient	Numeric
4	Marital Status	Marital Status of the patient	Married, Divorced, Single,999	Nominal
5	Subcity	The sub city of the patient	Addis Ketema, Bole, AkakiKaliti, Gulele,Yeka,Kirkos,Ledeta, Arada NefasSelkLafto and KolfeKeranio	Nominal

6	Triage Assessment	The level of emergency. Values are arranged from the most critical (Red) to lowest(Green) Black value represents the patient dead on arrival	Red Green Black	Nominal
7	Intent	Describes how the injury or illness occurred	Accident, Violence and Other	Nominal
8	Activity of the patient	Describes the activity of the patient during injury or illness	Traveling, Working, Studying, Playing and Other	Nominal
9	Referred From	Describes the referral of the patient	Governmental Health Facility , Non Governmental Health Facility and Self	Nominal
10	place of injury	The place where the injury or illness occurred	Home, Street, Work, School, Recreational Place and Other	Nominal
11	Transferred to	The place or Special unit that the patient was transferred	EmergencyOPD,Stabilization room, ,Resuscitation room, Home, Regular OPD, and Referred to other Hospital	Nominal
12	Past medical illness	Describe whether the patient has any previously (Past) known medical illness	Yes, No and Unknown	Nominal
13	Cause of visit	The cause or event for patient's Hospital visit	Trauma Non-Trauma	Nominal

Table 3.1: The original attributes and their description

3.6 Data Quality and Management

3.6.1 Data Preparation and Preprocessing

Data preprocessing is one of the most important activities as the presence of irrelevant data during the analysis phase may yield wrong results. Data preprocessing steps are concerned about sampling, running correlation and significance tests, and data cleaning. It includes checking the completeness of data records, removing or correcting for noise and missing values. The final aim of data preprocessing is creating a data that meet the specific input requirements for the DM tools.

Data preprocessing transforms data into a format that was more easily and efficiently processed for the purpose of the user. The main task of data preprocessing is to select standardized data from the original data.

Data preparation is the fundamental stage of data analysis which involves the construction of the final data set. Before data is fed into a DM algorithm, it must be collected, inspected, cleaned and selected. Since the optimal predictor will fail on bad data, data quality and preparation is crucial.

Data Cleaning

Data cleaning is a process which fills in missing values, removes noise and corrects data inconsistencies. Usually, a real world data contains incomplete, noisy and inconsistent data and such unclean data may cause confusion in the data mining process [12].

Missing Value Handling

Missing values refer to the values for one or more attributes in a data that do not exist. In real world application data are rarely complete. Many real-life data sets are incomplete. The problem with missing attribute values is a very important issue in data mining. In data mining the problem with the missing values has become a challenging issue [11].

Moreover, the fact that a value is missing may be significant in itself. A widely applied approach is used to calculate a substitute value for missing fields, for example, the median, mode or mean of available. Accordingly, we have analyzed the dataset and identified missing values and take measure to solve the problem as follows. As we depicted above, missing values are clearly

identified and calculated as the percentage of the value against total records. Therefore, based on the above principles the missing values were handled.

Summary of handled missing value is illustrated in the table 3.2 below.

No	Attribute	% Of missing value	Replace with	Reason/Technique applied
1	Transferred to	0.1	Home	Mode
2	Past medical illness	0.1	Unknown	Mode
3	Triage Assessment	0.2	Green	Mode

Table 3.2 Summary of handled missing value

Detecting noisy data and outliers:

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. In this research, the data objects do not have any outlier since the values of each column fits the database's general behavior.

Data Transformation and Reduction

Data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations. The data transformation process include (1) discretization, where methods are used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals, Interval labels can then be used to replace actual data values, data cube aggregation, dimension reduction (irrelevant or redundant attributes are removed), and data compression (data is encoded to reduce the size, numerously reduction (models or samples are used instead of the actual data)). The data should be reduced in order to make the analysis process manageable and cost-efficient. (2)normalization, where the attribute data are scaled so as to fall within a small specified range; (3) attribute construction, where new attributes are constructed and added from the given set of attributes to help the mining process, and aggregation, where a summary or aggregation operations are applied to the data [11,14].

Attribute Selection

One of the factors that affect the success of data mining algorithms on a given task is the quality of the data. If information is irrelevant or redundant, or the data is noisy and unreliable, then knowledge discovery during training is more difficult. Attribute subset selection is the process of identifying and removing as much of the irrelevant and redundant information as possible for achieving the success of DM goals .

Attribute selection is the process of selecting a subset of relevant features for use in model construction. There are several reasons for attribute selection like creating simple and easier transparent model, fast model induction etc. The data set contains 12 attributes and to decide on the relevant attributes for this study, the original table was contained 26 attributes. From this a total of 12 attributes were selected for the research based on their relevance and pre-processing activities of the problem. There were 12 attributes which were excluded in the preliminary data observation like 'Medical record number', 'name of patient', 'address of patient', 'name of contact person', 'address of contact person', 'laboratory number', 'drug', 'month', 'days of month', since they are no more important for the mining purpose. 'Medical record number' and 'laboratory number' have no important to know the research output. The other four attributes namely 'name of patient', 'address of patient', 'name of contact person' and 'address of contact person' have values such as name, phone-number, and so on. Therefore, those values can make the model complicated and not understandable. Accordingly, the 'drug' attribute has only one value and is not necessary for mining activity. Hence, the decision was made to remove those columns from the data set used for model building. Table 3.1 shows final attributes used for model building and their description.

Descriptive Statistical Summary of the Selected Attributes

The dataset has been described and visualized using dataset relative to the whole records. Simple statistical analysis has been performed to verify the quality of the dataset such as missing values, error values and to obtain high level information regarding the data mining questions. Hence, the selected attributes used for model building are statistically described in details below. This is helpful for understanding of the dataset for experimentation and increasing the accuracy of the model.

Attribute name	Attribute Values	Frequency	Percent
	Male	6755	67.03
	Female	3324	32.97

SEX	Total	10080	100
AGE	< 20 Year	1467	14.5
	21- 50 years	4302	42.7
	51 and Above	4310	42.8
	Total	10080	100
Marital Status	SINGLE	4010	39.8
	MARRIED	5971	59.3
	DIVORCE	98	0.97
	Total	10080	
Cause of visit	Non-Trauma	4120	40.87
	Trauma	5960	59.129
	Total	10080	100
Triage Assessment	GREEN	6163	61.14
	RED	3720	36.9
	BLACK	197	1.95
	Total	10080	100
Intent	ACCIDENT	4892	48.53
	VIOLENCE	2052	20.36
	OTHER	3136	31.11
	Total	10080	100
Activity			
	Traveling	2742	27.20
	Working	3131	31.06
	Playing & Recreating	3030	30.05
	OTHERS	1177	11.67
	Total	10080	100
Referred From	GOV	5676	56.30
	NGHS	2347	23.28
	SELF	2057	20.40
	TOTAL	10080	100
place of injury	HOME	4991	49.51
	STRAIGHT	1564	15.51
	WORK	2153	21.35
	PLAYING	1372	13.61
	TOTAL	10080	100
Transferred to	Emergency OPD	7536	74.76
	Resuscitation room	1760	17.46

	Stabilization room	391	3.87
	Regular OPD	196	1.94
	Referred to other	197	1.95
	Total	10080	100
Past medical illness	Yes	2347	23.28
	No	6164	61.15
	unknown	1569	15.56
	Total	10080	10099

Table 3.3 Descriptive statistical summary of the selected attributes

AGE: - it is the age of patients having 3 Categories 0-20 years, 21-50 years and above 50 years old. It is Nominal value

As the tables above shows 14.5 % of the emergency patients have 0-20 years category, 42.7 % have 21-50 years old and 42.8 % of the emergency patients have 50 years and above.

MARITAL STATUS: - This attribute shows marital status of the emergency patients. It has three values single , married and divorce.

As shown in the above table 39.8 % of the patients marital status has single, 59.3% have married and 0.97 % of the patients marital status has divorce.

CAUSE OF VISIT: - This attribute shows the number of patients cause of visit in the emergency department. It has two values Non-Trauma and Trauma.

As the table above shows 59.13 % of the patients cause of visits are Trauma and 40.87 % of the patients cause of visits are Non-Trauma.

TRIAGE ASSESSMENT: - This attribute shows the numbers of patients are curable, worst case or passed on. It has three values Green, Red and Black.

As the table above shows 61.14 % of the Triage Assessment have cow Green or curable, 36.9 % of the Triage Assessments have Red or worst cases and rest 1.95 % of the Triage Assessment have Black or Dead.

INTENT: - It is nominal type attribute with 3 values. Accident , Violence and Others.

As the table above shows 48.53 % of the patients have Accident, 20.36 % of the emergency patients have Violence and the rest 31.11 % of emergency patients have Others.

ACTIVITY: - refers to the patients activities during accidents. It is nominal values and it has 4 values.

REFERRED FROM: - this attribute shows the number of patients comes from. Its attribute is nominal type. It has 3 categories GOV, NGHS and SELF.

As the table above shows 56.30 % of the Patients come from governmental institute, 23.28 % of the Patients come from non-governmental health institute and the rest , 20.40 % of the Patients from self-come.

PLACE OF INJURY: - refers to the patients place of injury. It is nominal values and it has 4 values Home, straight, work and playing.

TRANSFERRED TO: - This attribute is nominal type. The modal value is emergency OPD (74.76%) and the least frequent value is regular ODP with 1.94 %. Statistical summary is presented in table 3.5

PAST MEDICAL ILLNESS: - It is nominal type attribute with three values. Yes, no, and unknown. The modal value is No(61.15%) ,23.28% of the patients are past medical illness and the least frequent value is unknown with 15.56 %.

3.7 Data Mining Tool Selection

For conducting this study the WEKA (Waikato Environment for Knowledge Analysis) data mining software is chosen. WEKA is chosen because of its widespread application in different data mining researches and familiarity of the researcher with the software.

WEKA, a machine-learning algorithm written in Java, is adopted for undertaking the experiment. WEKA constitutes several machine learning algorithms for solving real-world data mining problems [33]. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from one's own Java code. WEKA is open source software. The WEKA data mining software included classification, clustering, association rule learner, numeric prediction and several other schemes. In addition to the learning schemes, WEKA also comprises several tools that can be used for datasets pre-processing.

3.8 Data Preparation for Weka Software

Weka needs the data set to be prepared in some Weka understandable formats. First, we have entered the original paper format data into Microsoft Excel. And then Microsoft Excel was used for performing pre-processing activities and then the file is saved into Weka acceptable comma separated values (CSV) or comma delimited file format. The Weka native data format is known as the ARFF (Attribute Relation File Format). It is basically a CSV (comma separated value) format with some extra headers to specify what type each attribute is (numerical, binary, nominal). Finally, we have used 12 attributes with 10,080 instances that are ready for experimentation process.

3.9 Data Mining

In this phase, appropriate techniques of data mining were applied to the data set available. Typically, there are several techniques for the same data mining problem. Data mining techniques: clustering and classification were applied to the dataset available.

After the data was cleaned and prepared, it was analyzed using a data mining tool. There are varieties of tools available for data mining such as the knowledge Studio, Weka, xlminer, and others. Among those tools, Weka is selected since the whole suite of Weka is written in java, so it can be run on any platform. In addition to this, the package has three different interfaces: a command line interface, an Explorer GUI interface which allows for preparation, transformation and modeling algorithms on a dataset, and an Experimenter GUI interface which allows to run different algorithms in batch and to compare the results.

Although the choice of data mining techniques for classification tasks seems to be strongly dependent on the application, the data mining techniques that are frequently employed for classification tasks are neural networks, decision trees and rule induction rule induction. As it is indicated previously, for the purpose of this research work the researcher experimented the potential applicability of data mining technology in developing a model that predicts the occurrence of patient cause of accidents.

This software partitions the dataset prepared for analysis into training and test facts where training facts are used to train and build the models and test facts are used to test the performance of the model. By default the software automatically sets aside 10% of the prepared dataset for testing purposes. Besides for validation purpose the researcher splits the dataset 75% for training

and the remaining 25% for testing. 75% of the original data was selected for training purpose since the classifier learns more from large amount of data and increases its performance. The test data (25%) was selected from the original data using Simple Random Sampling technique.

In this research, numerous models were built by using the Weka3.7.7 software, and the proposed models was tested using test sets of data. Besides, the validity and performance of the model was tested to check its efficiency and effectiveness. Finally, the confusion matrix was used to evaluate the accuracy and performance of the model built with the decision tree algorithm.

3.10 Methods of Classification

In this research an attempt was made to build a model for the purpose of classifying the status of EM data using the three well known supervised machine learning algorithms called decision tree, PART rule indication, and neural network. Basically supervised machine learning algorithms automatically build a classifier by learning the characteristics of the category from a set of classified data, and then use the classifier to classify documents into predefined categories.

3.11 Evaluation of the Discovered Knowledge

Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge.

The different classification models in this research were evaluated using a test dataset based on their classification accuracy. For this purpose, the classifier was evaluated using different confusion matrices (True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR), and Relative Operating Characteristics (ROC)), the number of correctly classified instances, and number of leaves and the size of the trees, execution time.

Moreover, models were compared with respect to their speed, robustness, scalability, and interpretability. The confusion matrix of the classifier model was analyzed in terms of the following variables depicted in Table 3.6.

		Predicted Cause of Accident	
		Positive (YES)	**Negative (NO)**
Actual Cause of Accident	**Positive (YES)**	TP	FN
	Negative (NO)	FP	TN

Key: TN = True Negative FP= False Positive FN= False Positive TP= True Positive

Table 3.4 confusion matrixes

The representation of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are defined as follows:

- ✓ True Positive (TP): The number of food security records that are correctly identified as Trauma Cause of Accident .
- ✓ False Negative (FN): The number of records that are incorrectly classified as they are Non-Trauma Cause of Accident however in fact they are Trauma Cause of Accident.
- ✓ False Positive (FP): The number of records that are incorrectly identified as Trauma Predicted Cause of Accident however in fact they are not.
- ✓ True Negative (TN): The number of Non-Trauma Cause of Accident record that correctly classified as Non-Trauma Cause of Accident.

The performance of predictive/ classification was measured by sensitivity, specificity, false negative rate, correctly classified instance and incorrectly classified. These performance measurements are described as follows.

Sensitivity (TPR): indicated by the proportion of in Non-Trauma Cause of Accident records that are correctly classified as positive, TPR= TP/ TP + FN

Specificity (TNR): The proportion of not having the in Non-Trauma Cause of Accident that are correctly classified as negative, TNR= TN/ TN+FN

False Negative Rate (FNR): The proportion of not having the Non-Trauma Cause of Accident that are incorrectly classified as negative i.e. FNR= 1 – TPR or FNR = FN/ TP + FN hence, TPR + FPR =1.

Correctly Classified Instances (Accuracy): To compute the proportion of correctly classified among a total classifications, Accuracy = (TP + TN) / (TN + FP + FN + TP)

Incorrectly Classified Instances (Error Rate): The proportion of incorrectly classified among total classifications, Error Rate = (FP + FN) / (TN + FP + FN + TP).

Recall and Precision are two widely used metrics to compute the effectiveness and efficiency of the model. As a result, recall was computed for both positive and negative classes. So the formula of the recall for negative class is TN/ (TN + FP) and for positive class TP/ (TP+FN). Precision is the number of class members classified correctly over the total number of instances classified as class members. The formula of precision is (TP*100%) / (TP+FP).

Receiver Operating Characteristic (ROC) Curve

This procedure is a useful way to evaluate the performance of classification schemes in which there is one variable with two categories by which subjects are classified [35].

To assess the accuracy of a model, one can measure the area under the curve which is a portion of the area of the unit square and its value is ranged from 0-1. It is assumed that increasing numbers on the scale represents that the subject belongs to one category while decreasing numbers on the scale represent the increasing belief that the subject belongs to the other category [3]. Thus, from the ROC curve, the closer the ROC curve of a model is to the diagonal line, the less accurate the model is closer to the area of 0.5.

ROC Area	Performance
0.9-1.0	Excellent(A)
0.8-0.9	Good (B)
0.7-0.8	Fair (C)
0.6-0.7	Poor (D)
0.5-0.6	Fail

Table 3.6: Performance Measures of ROC Area

The model with perfect accuracy will have an area of 1.0 i.e. the larger the area, the better performance of the model or the larger values of the test result variable indicate the stronger

evidence for a positive actual state (1.00) [20] [40]. For example, in Table 3.6, classifier A performs better than B.

By using ROC analysis one can identify predictors in order to find the one with optimal characteristics and their associated cut-points. Therefore, sensitivity, specificity, precision, F-measure, and ROC area were taken in to account when the classifier performance is evaluated.

In this study we used the above preprocessing, classification and evaluation techniques to apply data mining. Basically after the document preprocessing the final list of attributes are provided to the algorithm as a training data. Once the classifier trained with the preprocessed training data, then, the next very important step is to evaluate the classifier using the test data in the same domain. Using the test data the classifier automatically predicts the given data as Trauma or Non-Trauma and from the produced result; the classifier performance is measured using accuracy, recall and precisions.

3.12 use of the Discovered Knowledge

This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge was created and the entire project was documented.

Finally, the discovered knowledge was used by constructing/ designing user interface for showing the application of data mining techniques to the environment. In addition to this, other researchers can use this study as reference as well as getting initial information for doing researches related to cause of accidents.

CHAPTER FOUR

EXPERIMENTATION AND DISCSSION

This describes about techniques that have been used in developing a model to predict Emergency cause of accidents in St. Paul Hospital. The study has been conducted according to a hybrid model, which was discussed in the previous chapter.

4.1 Model Building

Modeling phase of data mining is the process of providing the processed data to the selected classification algorithm and selecting model that shows better performance. There are a number of tasks involved during model building like selecting the modeling technique, experimental setup or design, building the model and evaluating the model.

4.1.1 Selecting Modeling Technique

The first step in model building is selecting an appropriate modeling technique based on the objectives of the study. Three classification algorithms namely J48, PART induction rule and Neural Network algorithms were selected and deployed through Weka machine learning software. J48 and PART were selected, because they are easy for model building, interpretation and understanding. Neural Network was selected because of its high learning capability and has high tolerance for noisy data even though it is difficult to interpret.

J48 is one of the most common decision tree algorithms that are used today to implement classification technique using WEKA. A PART algorithm is common types of rule induction technique which generate a model as a set of rules. A Neural Network is also common types of Function algorithm which generate a model as a Network topology. In the meantime, the J48 algorithms of decision tree generate a model by constructing a decision tree where each internal node is a feature or attribute. The leaf nodes are class output.

4.1.2 Experimental Setup

In any data mining process before building a model, there is a need to generate a procedure or mechanism to test the model's quality and validity. For instance, in supervised data mining tasks such as classification, it is common to use classification accuracy measure or error rates as quality measures for data mining models.

Weka 3.7.7 software was used to measure the quality, validity and test of the selected model. For purposes of this study k-fold (10-folds) cross validation and percentage split test options are used because of their relatively low bias and variations. In 10 fold cross validation; the data were divided in to 10 folds where 9 folds were used as training data whereas the remaining one folds as test data. In the percentage split method, where 75 % of the data was used as training and the remaining 25 % was used as test data. In supplied test set the researcher specified 70 % for training and the remaining 30% for testing data. Accuracy, Precision, Specificity, ROC curve, Recall and confusion matrix standard metrics were also used for evaluation of the results.

4.1.2.1 Experimental Set up for J48 and PART

Establishing scenario for a model to be developed is very important to see the analysis of each result; by comparing the evaluation criteria, the better the performance model was selected. Consequently for both of the methods the following three scenarios has been done with different parameter values of WEKA 3.7.7 software.

Scenario 1: Decision tree without pruning and with pruning.

Scenario 2: Rule induction without pruning and with pruning.

Scenario 3: Neural Network

Once the modeling tool is chosen and performance evaluation criteria are established, building model with a number of parameters that govern the model generation process would be the next task.

Experiments	Parameters			
	Unpruned	**Confidence factor**	**(min Numobj)**	**Test option**
Experiment #1	True	0.25	2	10 fold cross validation
Experiment #2	True	0.25	5	10 fold cross validation
Experiment #3	True	0.50	5	80 % percentage split
Experiment #4	True	0.50	5	70 % percentage split
Experiment #5	False	0.50	2	10 fold cross validation
Experiment #6	False	0.25	2	10 fold cross validation
Experiment #7	False	0.25	5	10 fold cross validation
Experiment #8	False	0.25	2	70 % percentage split
Experiment #9	False	0.50	5	80 % percentage split

Table 4.1: Values of parameters used for 9 experiments

4.1.2.2 Experimental Set up for Neural Network

A neural network is an adaptable system that can learn relationships through repeated presentation of data, and is capable of generalizing to new, previously unseen data. They are a large set of interconnected neurons, which execute in parallel to perform the task of learning.

In the case of ANN, we need to select appropriate topology and learning algorithm that best fits to the proposed classification model. Hence, its basic components are described as follows

Architecture: It is the topology of the network that describes the pattern of connections between neurons. In our case, we will use feed forward multilayer perceptron (MLP) model. A feed-forward multilayer ANN is a topology in which neurons are arranged in layer and has the following properties.

The first layer gets the input data from the environment and is called input layer.

The last layer generates output class and is called output layer.

Layers other than input and output layers are called hidden layers.

The i^{th} layer gets input from each of the neurons in the i – 1 layer weighted by some weight factor for $1< I =N$.

Learning algorithm: Learning in ANN indicates the methods used to determine the adaptation of weights between the connections of two neurons. A MLP feed-forward network uses back

propagation learning algorithm for adapting its weights to a sequence of training samples during a learning phase. A back propagation algorithm is a supervised learning algorithm that propagates classification errors from the output layers back toward to the input layers and modify the weight to minimize the total error.

Activation function: It is the processing logic that computes the neuron's final output state. Back propagation requires continuous and differentiable activation function. Hence, sigmoid function was used in order to provide smooth control of the input-output relationship.

As described above we have used feed forward multilayer perceptron (MLP) model with back-propagation learning rule which is based on supervised learning and by default WEKA use one hidden layer with the number of hidden neurons = (# of input attributes + # of classes) / 2.)

Experiments	**Parameters**				
	Hidden Layers	**Learning Rate**	**Momentum**	**Learning Time**	**Test option**
Experiment #1	1	0.3	0.2	500	10 fold cross validation
Experiment #2	1	0.5	0.2	400	10 fold cross validation
Experiment #3	1	0.5	0.3	500	80 % percentage split
Experiment #4	1	0.5	0.2	400	70 % percentage split
Experiment #5	2	0.3	0.2	500	10 fold cross validation
Experiment #6	2	0.5	0.2	400	10 fold cross validation
Experiment #7	2	0.5	0.3	500	10 fold cross validation
Experiment #8	2	0.5	0.2	400	70 % percentage split
Experiment #9	2	0.5	0.2	400	80 % percentage split

Table 4.2: Values of parameters used for 9 experiments

4.2 Model Building and Evaluation Using J48 Decision Tree

4.2.1 Model Building

In the Hybrid methodology, Model building is an iterative process. Therefore, it is important to conduct different experiments to find the optimal model for solving the problem. In this study, different experiments were conducted altering parameters of the J48 decision tree and PART rule induction algorithm for building the optimal predictive model.

The J48 decision tree algorithm builds decision trees from a set of predefined training dataset using the concept of information entropy and attribute ordering. It uses the fact that each attribute of the data was used to make a decision by splitting the data into smaller subsets.

Performance Measurements	**Experiments**								
	Without pruning				**With pruning**				
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**
Accuracy (%)	0.969	0.969	0.966	0.965	0.956	0.969	**0.969**	0.965	0.966
Mean absolute error	0.085	0.085	0.086	0.085	0.085	0.086	**0.085**	0.085	0.086
Numbers of leaves	58	58	58	58	58	58	**58**	58	58
Size of tree	73	73	73	73	73	73	**73**	73	73
Time taken	0.13	0.06	0.05	0.03	0.06	0.06	**0.05**	0.05	0.05
AV.TP rate	0.959	0.956	0.953	0.953	0.956	0.956	**0.956**	0.953	0.953
AV.FP rate	0.05	0.05	0.051	0.052	0.05	0.05	**0.05**	0.052	0.051
AV. Precision	0.956	0.956	0.953	0.953	0.956	0.956	**0.956**	0.953	0.953
AV. Recall	0.956	0.956	0.953	0.953	0.955	0.956	**0.956**	0.953	0.953
AV.ROC area	0.951	0.951	0.949	0.949	0.951	0.951	**0.951**	0.949	0.949
CCI	9631	9631	1922	2881	9631	9631	**9631**	2881	1922
ICCI	448	448	94	143	448	448	**448**	143	94

Key: CCI: Correctly classified Instance, ICCI (Incorrectly classified Instance), Accuracy: Registered performance of model, AV: Average, TP: True Positive. FP: False Positive, ROC: Relative Optical character curve.

Table4.3: Experimentation result of J48 Algorithms

As we can see the above table the result of each experiment demonstrate that the model developed with pruning scored higher accuracy than the unpruned experiment. As a result Experiment # 7 (building decision tree pruned with 10-fold cross validation) scored the highest accuracy (96.9%) and better performance in terms of time taken .Experiment # 6 , Experiment # 2 and Experiment # 1 also showed highest performance next to experiment # 7. All these experiments Experiment #6 and Experiment #7 are pruned experiments. The unpruned experiment # 2 has also good performance next to the above experiments and better than all the other unpruned experiments. In general the pruned experiments have shown good performance than the unpruned experiments.

4.2.2 Model Evaluation

When evaluating a classifier, there are different ways of measuring its performance. The experiments conducted above have been analyzed and evaluated in terms of classifiers performance values, accuracy, confusion matrix values, TP and FP Rate, number of leaves, and size of the tree generated, ROC curves and execution time.

As shown in Table 4.3, performance of the classifier on the testing set increased as the MinNumObj increased up to about 5. Experiment #7 showed an accuracy of 96.9 %. At this accuracy correctly and incorrectly classified instance are 9631 and 448 respectively from 10079 instances. From nine different trials experiment #7 is the optimal model in terms of accuracy and minimized incorrectly classified instances. The Confusion Matrix of Experiment #7 below shows the number of instances of each class that are assigned to all possible classes according to the classifier's prediction. The columns represent the predictions, and the rows represent the actual class.

		Predicted Cause of Accident		
		Positive	**Negative**	**Total**
Actual Cause of Accident	**Positive (Trauma)**	5706	183	5889
	Negative (Non-Trauma)	265	3925	4190
	Total	5971	4108	10079

Table 4.4 summery confusion matrixes for J48

From the table 4.4 confusion matrix we can say the following:

True positives refer to the positive tuples that were correctly labeled by the classifier. True negatives are the negative tuples that were correctly labeled by the classifier. False positives are the negative tuples that were incorrectly labeled (tuples which are actually incorrect or false but the classifier predicted as correct). Similarly, false negatives are the positive tuples that were incorrectly labeled (tuples which are actually correct but the classifier predicted incorrectly).

The above confusion matrix shows that 5706 instances were correctly predicted as Trauma cause of accident (True positive). True positive of the actual class of the test instance Trauma and the classifier correctly predicts cause of accident in the hospital. The numbers of instance which were correctly predicted as Non-Trauma cause of accident are 3925 instances (True negative). In this case of true negative the actual class of the test instance is Non-Trauma cause of accident and the classifier correctly predicts the class as Non-Trauma . Therefore, correctly classified instances are the sum of diagonal values of the table, which are 5706 instances correctly classified from 10079 instances.

The following result has been extracted from Experiment #7 model. True Positive rate shows the percentage of low weight instances whose predicted values of the class attribute are identical with the actual values. FP rate shows the percentage of instances whose predicted values of the class attribute are not identical with the actual values.

	TP	**FP**	**Precision**	**Recall**	**F-Measure**	**ROC**	**Class**
	0.969	0.063	0.956	0.969	0.962	0.951	TRAUMAE
	0.937	0.031	0.955	0.937	0.946	0.95	NONE TRAUMA
Weighted Av	0.956	0.05	0.956	0.956	0.955	0.951	

Table 4.5 Detailed accuracy by class

If we take the first level where ‘Predict cause of accident = Trauma’ TP Rate is the ratio of Trauma cause of accident cases predicted correctly to the total of positive cases, there were 5706

instances correctly predicted as Trauma, and 3925 instances in all that were None-Trauma So the TP Rate (True Positive Rate) of Trauma cause of accident level= 3925/4190 = 93.67. The FP Rate is then the ratio of None Trauma cause of accident level of incorrectly predicted as Trauma cause of accident level to the total of in cause of accident level cases. 183 None Trauma cause of accident level instances were predicted as Trauma cause of accident level and there were 5889 None Trauma cause of accident level in all. So the FP Rate is 183/5889 =0.031. We can follow the same method to calculate for 'cause of accident = Trauma ' but as we can see from detailed accuracy by class TP Rate and FP Rate of None Trauma cause of accident class level are 0.969 and 0.063 respectively. The model performance is good quality because it has high true positive rates with low false positive rates.

As can be seen from the detailed accuracy by class output, the ROC (Receiver Operating Characteristics) area of this model is highest (0.951). The Area under the ROC curve in table 4.3 is higher. Higher numbers here indicate the model is the more accurate. The ROC curve is a plot of how the classifier is performed over the entire range of possible choices of cutoff values. Each point on the curve represents the True-Positive Rate plotted on the y-axis and the False-Positive Rate plotted on the x-axis that resulted from a particular cut-off value.

4.3 Generating Specific Rules from J48 Decision Tree Algorithm

In knowledge discovery, it is crucial to investigate interaction between attributes in order to induce interesting and useful prediction rules. Decision-tree algorithms generate a model by constructing a decision tree where each internal node is a feature or attribute.

Among all the experiments that was fitted using J48 classifier algorithm without pruning and with pruning, experiment #7 model with pruning had the greatest performance and selected as the optimal model for rule extraction. The model was fitted using (0.25 confidence Factor, min Numobj=5 and Unpruned=False). It is possible to extract important rules simply by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node. Based on the domain experts the following 15 were the most important rules/patterns extracted from the decision tree model.

> If Place of injury is Street and Intent is accidental then the Cause of visit is likely to be for 'Trauma' (the patient will come to the Hospital emergency unit due to Trauma case).

If Place of injury is Street and intent is violence and Activity is working and whose Region is Addis Ababa then the Cause of visit is likely to be for ‘Trauma’.

If Place of injury(illness) is Home and Intent is accidental and Transferred to the Emergency Out Patient Department (OPD) and Activity is Playing then the Cause of visit is likely to be for ‘Non-Trauma’.

If Place of injury is Home and Intent is violence then the Cause of visit is likely to be for ‘Non-Trauma’.

If Place of injury is Home and intent is accidental and transferred to the emergency OPD and activity is working then the Cause of visit is likely to be for ‘Non-Trauma’.

If Place of injury is working area (work) and intent is accidental and activity is working then the Cause of visit is likely to be for ‘Non-Trauma’.

If Place of injury is other and intent is accidental then the Cause of visit is likely to be for ‘Trauma’.

If Place of injury is other and Intent is violence and Past medical illness is No then the Cause of visit is likely to be for ‘Trauma’.

If Place of injury is recreational place then the cause of visit is highly likely to be predicted as ‘Trauma’.

If Place of injury is school and activity is studying then the Cause of visit is likely to be for ‘Non-Trauma’.

If Place of injury is working area and Activity is other and intent is accidental and transferred to emergency OPD and Past medical illness is no and whose Subcity is Arada then the Cause of visit is likely to be for ‘Non -Trauma’.

If Place of injury is work and activity is other and intent is accidental and transferred to Resuscitation(recovery) room then the Cause of visit is likely to be for 'Non -Trauma'.

If Place of injury is school and Activity is playing or recreating and marital status is married then the Cause of visit is likely to be for 'Non- Trauma'.

If Place of injury is Home and Intent is accidental and Transferred to the Emergency OPD and activity is other and Triage assessment is orange and whose Age is young, and whose Region is Addis Ababa and who are referred from Governmental Health Facility, then, Cause of visit is likely to be for 'Trauma' .

If Place of injury is Home and intent is accidental and transferred to the emergency OPD and activity is other and triage assessment is orange and whose Age is adalt and Past medical illness is no and whose sex is female then the Cause of visit is likely to be for 'Trauma'.

4.4 Model Building Using PART Rule Induction Algorithms

The second data mining technique used in this research was a PART Rule induction algorithm. Rule induction is one of the techniques most used to extract knowledge from data, since the representation of knowledge as if/then rules is very intuitive and easily understandable by problem-domain experts [13].

To build the Rule induction model, WEKA software package and the same number of datasets were used as an input respectively. The experiments were divided into two as without pruning and with pruning by including three test option that are 10-fold cross validation, percentage split evaluator and supplied test set .The following table shows the summary of the PART rule induction algorithm.

Performance Measurements	Experiments								
	Without pruning				With pruning				
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**
Accuracy (%)	0.969	0.969	0.966	0.965	0.969	**0.9691**	0.969	0.965	0.966
Mean absolute error	0.085	0.085	0.086	0.086	0.085	**0.084**	0.084	0.086	0.086
Numbers of Rules	62	62	62	62	62	**16**	16	16	16
Time taken	0.61	0.17	0.16	0.17	0.22	**0.09**	0.16	0.08	0.09
AV.TP rate	0.956	0.956	0.953	0.953	0.956	**0.956**	0.956	0.953	0.953
AV.FP rate	0.05	0.05	0.051	0.052	0.05	**0.05**	0.05	0.052	0.051
AV. Precision	0.956	0.956	0.953	0.953	0.956	**0.956**	0.956	0.953	0.953
AV. Recall	0.956	0.956	0.953	0.953	0.956	**0.956**	0.956	0.953	0.953
AV.ROC area	0.948	0.948	0.946	0.946	0.948	**0.954**	0.954	0.949	0.949
CCI	9631	9631	1922	2881	9631	**9631**	9631	2881	1922
ICCI	448	448	94	143	448	**448**	448	143	94

Table 4.6: Experimentation result of PART rule induction Algorithms.

As shown in the above table, the registered performance in case of induction rule learner, the pruned is better than the unpruned one. Among the 9 experiments Experiment #6 (10-fold cross validation) registered the optimal performance of 96.9 % and time taken 0.09 second. This shows that out of the total set of 10079 records, 9631 (95.55 %) of the records are correctly classified, while 448 (4.45 %) of the records are misclassified. Experiment #2 also registered better performance out of all the experiments using unpruned parameter with an accuracy of 96.6 and time taken 0.17 seconds. The following table shows the confusion matrix of the experiment #6.

		Predicted Cause of Accidents		
		Positive (Trauma)	**Negative** (None-Tauma)	**Total**
Actual Cause of Accidents	**Positive**	5706	183	5889
	Negative	265	3925	4190
	Total	5971	4108	10079

Table 4.7 summery confusion matrixes for PART

Furthermore, the resulting confusion matrix shown in Table 4.7 depicts that out of the total 5889 Trauma cause of accident instances 5706 (96.8%) of them are correctly classified in their respective class, while 183 (3.1%) of the records are incorrectly classified as N oneTrauma cause of accident level. On the other hand, out of the total Cause of accident instances 3925 (63.67 %) of them are correctly classified as Trauma cause of accident level and 265 (6.32 %) of the records are misclassified.

4.5 Generating Rules from PART Algorithm

The numbers of rules produced by PART algorithm are 171. The rules generated using PART algorithms are more clear and understandable. Domain experts selected interesting patterns or rules among 171 rules produced by the PART rule induction algorithm are presented as follows.

The following are some of the rules or patterns which were generated by PART algorithm.

Place of injury = Street and Intent = accidental: Trauma

Place of injury = school and Activity =studying: Non-Trauma

Place of injury = street and Intent = violence and Activity = recreating: Trauma

Place of injury = recreational place: Trauma

Place of injury = street and Intent = violence and Activity = travelling: Trauma

Most of the rules that were obtained from PART algorithm are similar with that of the rules obtained from J48 models. Such as Place of injury = Street and Intent = accidental: Trauma, and Place of injury = school and Activity =studying: Non-Trauma .

When we compare the accuracy measure and the results obtained from both classification algorithm(J48 and PART) models , we can say they are nearly similar. In terms of accuracy, execution time and ROC, J48 is slightly better than PART. The rules generated using both algorithms have similar patterns. We can select J48 model because it has slightly better performance and accuracy measure than PART classification model.

4.6 Model Building Using Neural Network Function Algorithms

The third data mining technique used in this research was a NEURAL NETWORK Function algorithm.

To build the Neural Network Function model, WEKA software package and the same number of datasets were used as an input respectively. The experiments were divided into two as without pruning and with pruning by including test option that are 10-fold cross validation and supplied test set. The following table shows the summary of the Neural Network Function algorithm.

Performance Measurements	**Experiments**								
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**
Accuracy (%)	0.967	0.965	0.968	0.939	**0.969**	0.966	0.966	0.965	0.968
Mean absolute error	0.114	0.838	0.123	0.174	**0.084**	0.092	0.085	0.084	0.125
Time taken	8.44	119.57	11.02	6.56	**13.58**	9.99	12.42	16.53	16.14
AV.TP rate	0.938	0.953	0.934	0.905	**0.956**	0.95	0.953	0.953	0.934
AV.FP rate	0.074	0.052	0.079	0.109	**0.05**	0.057	0.053	0.052	0.079
AV. Precision	0.939	0.953	0.935	0.905	**0.956**	0.95	0.953	0.953	0.935
AV. Recall	0.938	0.953	0.934	0.905	**0.956**	0.95	0.953	0.953	0.934
AV.ROC area	0.938	0.946	0.915	0.867	**0.95**	0.949	0.951	0.949	0.913
CCI	9455	2881	1883	2736	**9631**	9574	9603	2881	1883
ICCI	624	143	133	288	**448**	505	476	143	133

Table 4.8: Experimentation result of Neural Network Function Algorithms.

As shown in the above table, the registered performance in case of Neural Network learner, the Experiment with two hidden layers is better than the Experiment with one hidden layer. Among the 9 experiments Experiment #5 (10 folds) registered highest performance of 96.9 %. This shows that out of the total set of 1079 records, 9631(95.55 %) of the records were correctly classified, while 448 (4.44 %) of the records are misclassified.

The following table shows the confusion matrix of the experiment #5.

		Predicted Cause of Accidents		
		Positive (Trauma)	**Negative (none Trauma)**	**Total**
Actual Cause of Accidents	**Positive (Trauma)**	5706	183	5889
	Negative (none trauma)	265	3925	4190
	Total	5971	4108	10079

Table 4.9 summery confusion matrixes for Neural Network

Furthermore, the resulting confusion matrix shown in Table 4.9 depicts that out of the total 5889 Cause of Accidents **as Trauma** 5706 (96.89 %) of them are correctly classified in their respective class, while 183 (3.11 %) of the records are incorrectly classified as Cause of Accidents as none Trauma. On the other hand, out of the total Cause of Accidents instances 3925 (93.67 %) of them are correctly classified as Cause of Accidents are non-Trauma and 265 (6.32 %) of the records are misclassified.

4.7 Model Comparison

For all experiments carried out in three algorithms, and then selecting a better classification technique for building a model to predict the factor of Cause of Accidents. Consequently, the three selected classification models J48, PART and NEURAL NETWORK with respective better performance accuracy, Precision and number of instances correctly classified and misclassified are listed in the table

Types of algorithms	**Accuracy (%)**	**Time taken (sec/)**	**Correctly classified**	**Misclassified**
J48	96.91	**0.05**	9631	448
PART	96.90	**0.09**	**9631**	448
NEURAL NETWORK	96.90	**13.58**	**9631**	448

Table 4.10 Performance comparison of selected optimal models

As shown in Table 4.10, J48 decision tree algorithm classifier outperforms of all the selected classifiers with an accuracy of 96.91 % and it was the optimal classifier in predicting Cause of Accidents. At the same time both PART and NEURAL NETWORK achieved 96.90% accuracy rate.

The model developed from this experiment was validated using the separated test set 30% and the performance of the model increases to 96.91 %. This registers high performance and is better when compared it with the previous experimentations.

4.8 Evaluation of Discovered Knowledge

Evaluation of Knowledge discovery task involves several steps to evaluate the performance of the system in terms of how the system correctly classifies the instances inside the data set into distinct values of the class attribute. About 25 rules/patterns were generated by the J48 algorithm from the experiment #6. Consequently, to evaluate the importance of the discovered knowledge/rules, whether they are acceptable/not and whether they go in line with what is already known in the real world practice.

Finally, the experts selected 15 rules generated by the J48 algorithm as optimal rules. Then Rule 1 – Rule 5 discussed below were selected as the most interesting and optimal rules or discovered knowledge.

Rule 1: If Place of injury is Street and Intent is accidental then the Cause of Accident is likely to be for 'Trauma' (the patient will come to the Hospital emergency unit due to Trauma case).

This rule described If Place of injury is Street and Intent is accidental the Cause of Accident have Trauma cases defiantly.

Rule 2: If Place of injury(illness) is Home and Intent is accidental and Transferred to the Emergency Out Patient Department (OPD) and Activity is Playing then the Cause of visit is likely to be for 'Non-Trauma'.

As the above rule presented If Place of injury(illness) is Home, have Transferred to the Emergency Out Patient Department (OPD and have Activity is Playing have high probability to became 'Non-Trauma' Cause of Accident.

Rule 3: If Place of injury is other and Intent is violence and Past medical illness is No then the Cause of visit is likely to be for 'Trauma'.

The rule showed If Place of injury is other , Intent is violence and Past medical illness is No Cause of Accident have became 'Trauma' .

Rule 4: If Place of injury is Home and Intent is accidental and Transferred to the Emergency OPD and activity is other and Triage assessment is orange and whose Age is young, and whose Region is Addis Ababa and who are referred from Governmental Health Facility, then, Cause of visit is likely to be for 'Trauma' .

Rule 5:- If Place of injury is Home and intent is accidental and transferred to the emergency OPD and activity is other and triage assessment is orange and whose Age is adalt and Past medical illness is no and whose sex is female then the Cause of visit is likely to be for 'Trauma'.

Finally, we agreed with the general rules that the model produced and findings of the current research.

4.9 Use of Discovered Knowledge

In order to show how to use the discovered knowledge for the domain expert, the researcher design user interface by using JAVA programming language. User Interface is the interaction point between the user and the system. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from Java code. WEKA is written in the Java language and contains a Graphical User Interface (GUI) for interacting with data files and producing visual results. It also has a general Application Programming Interface (API); WEKA can be embedded like any other library in applications. Hence, it is easy to call the instances, filter, algorithms and evaluation on Java codes.

Therefore, the java code written for the interface for this study is directly calling the selected model with all its parameter. The figures 4.2 and 4.3 below shows the predicting result based on the filled attribute values.

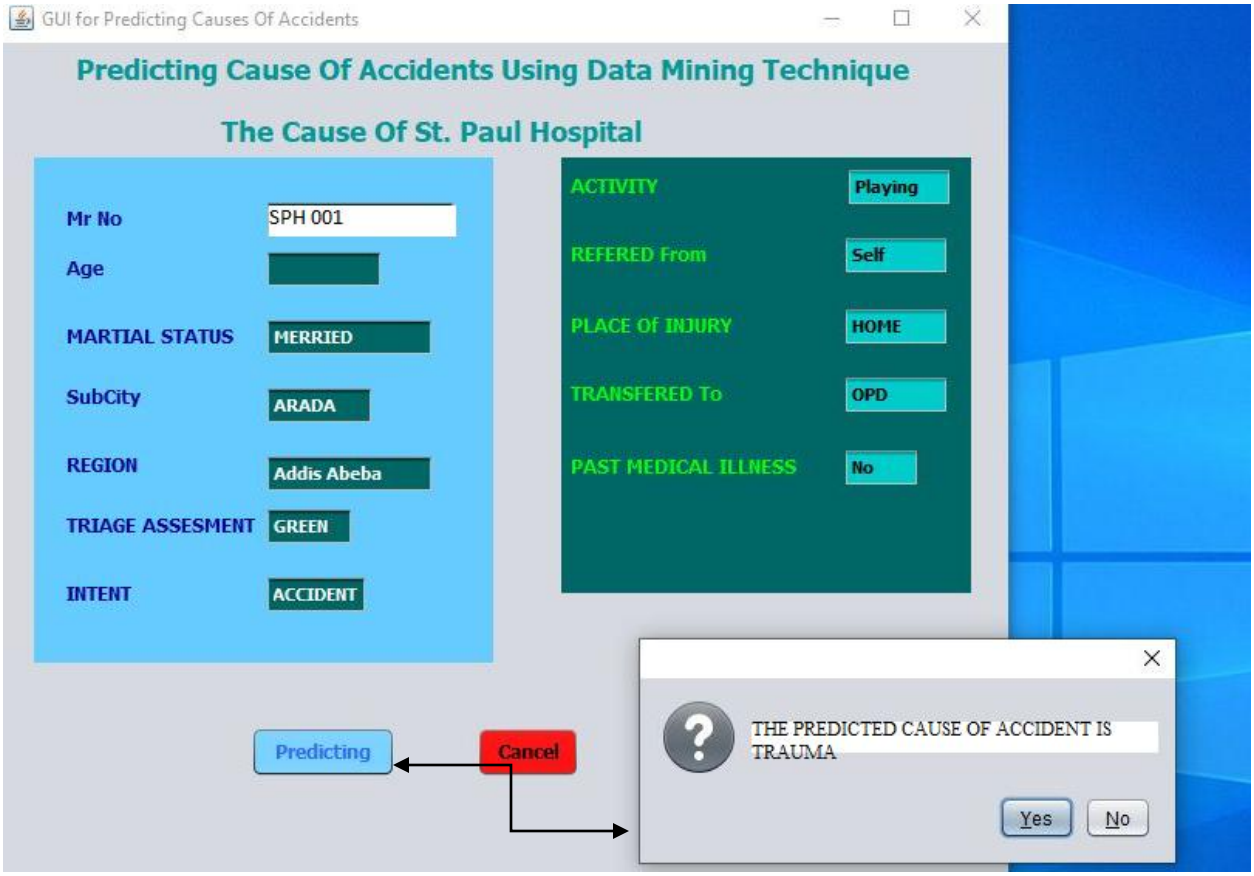

Figure 4.2: A model predicting Cause of Accident as TRAUMA

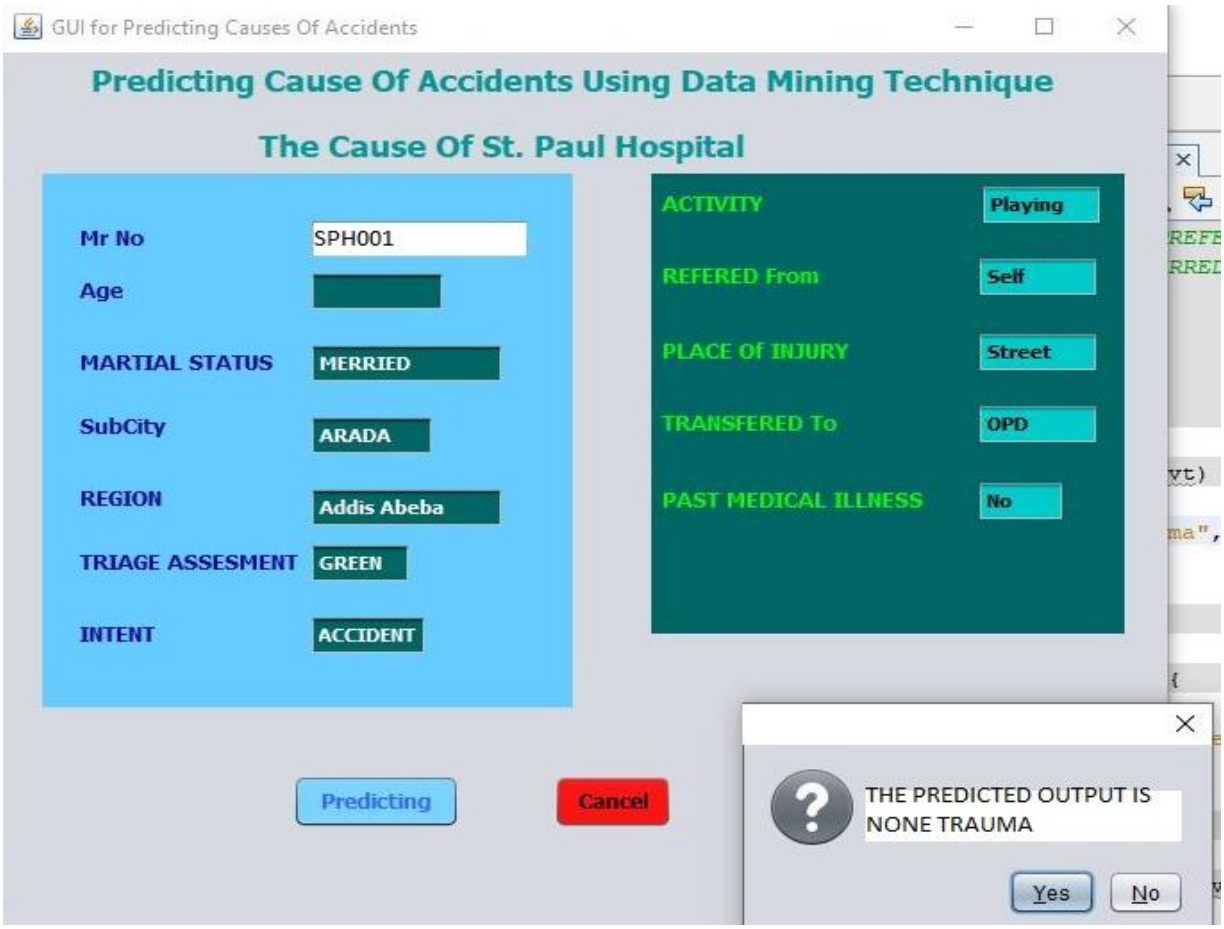

Figure 4.3: A model predicting Cause of accident as NONE TRAUMA

From this we can say that the users or domain experts can predicting cause of Accidents so that the system can give advice before emergency cause of visit in addition to recommend what measures to take in order to have a better cause of accidents .

4.10 Discussion

This research was designed to build a model that can predict and identify determinant factors of predicting cause of Accidents in St.Paul Hospital, Addis Ababa by using data mining techniques. The result from this study showed that the objective was met. `In this study three classification algorithms namely J48,PART and ANN algorithms were selected and deployed using Weka machine learning software to build a model that predicts whether cause of Accidents result in Trauma or None-Trauma from the St.Paul Hospital Emergency Department 2011 E.C dataset. different experiments were examined in each algorithm to build model and the optimal

experiment was selected from each algorithm for prediction. The performance of the classification algorithms (J48, PART and ANN) was evaluated based on their accuracy, sensitivity, specificity, recall, precision and area under the curve (ROC curve). J48 was selected as a working model because it achieves highest value on the performance evaluation measurements i.e. 96.91 % Accuracy, 0.85 Sensitivity, 0.956 Precision, 0.956 Recall and 0.951 Area under ROC curve. Finally, by considering the confidence level of the extracted rules and using the results of j48 (the working model) this study extracted most interesting rules.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusions

Health facilities, especially governmental ones such as Hospitals, deliver a lot of clinical services for many patients. They take or store tremendous demographic data as well as clinical history of the patients who are treated at these facilities. Hence, these data grows exponentially from time to time requiring powerful analysis tools for uncovering the hidden patterns and producing decision support information. Therefore, data mining techniques are solutions in discovering hidden and potentially useful information or knowledge that can be used for decision support out of large volume of data collected overtime from various sources.

Data mining, extracting meaningful patterns and rules from large quantities of data, is clearly useful in any field where there are large quantities of data and something worth learning. In this respect, cause of accident is a potential area for data mining. It is filled with lots of data and professionals who already make sense of all the data to manage Accidents.

In this research, an attempt was made to assess the potential applicability of DM technology in support of Predicting Accident assessment activity with the aim of identifying and predicting causes of accidents in the hospital. The hybrid process model designed by (Ciso.et al, 2007) has been followed during undertaking the experimentation and discussion. The data set used in this study has been taken from St. Paul Hospital Emergency unit department for the year 2011 E.C. After taking the data, it has been preprocessed and prepared in a format suitable for the DM tasks.

Accordingly, this experimental research, which employed the commonly used methodological approach in data mining researches, made use of three predictive modeling techniques, decision tree, rule induction and neural networks, to address the problem. For model building and experimentation J48, PART and MLP algorithms were used. Experimentation was conducted using three scenarios in three test options (10 fold cross validation, percentage split and supplied test set) for each algorithm. By changing the test options and the default parameter values of the algorithm these models were tested and evaluated. J48, PART and MLP algorithm performed 96.91 %, 96.90 % and 96.90 % accuracy respectively. As a result J48 decision tree algorithms

registered optimal performance with 96.91 % accuracy running by10-fold cross validation with parameters (confidence Factor=0. 25 and min Numobj=5).

In general, the results from this study can contribute towards encouraging and support the decision for Emergency cause of accidents. The extracted rules in both algorithms are very effective for predicting cause of accidents. But based on the accuracy of the algorithm J48 algorithm was used for predicting whether the patients' cause of visit will have TRAUMA or NONE TRAUMA level.

5.2 Recommendation

This research work is conducted mainly for academic purpose. However, it is the researcher's belief that the findings of the research will help initiate patients cause of visit in the hospital to work on the application of data mining technology to gain competitive advantage in their field. The results of this study have shown that the data mining technology particularly the decision tree technique is well applicable. Hence, based on the findings of this study, the following recommendations can be forwarded.

Database standardization is required for the Hospital emergency medicine registration database because, there were so many duplicate, invalid, missing and inconsistent values in many of the attributes and records which consume much of the researcher's time to clean or preprocess the data set.

There exists only very few health related data on softcopy form (computerized form) in the hospital. In addition, most hardcopy data are also not convenient (to mean the data recorded on the card is bulky, inconsistent, and it is mostly not understandable by non-professionals for the area) for data mining research because they are collected and stored for other purposes .Hence, as much as possible this situation need improvement such as using standardized data collection and storing system like using integrated hospitals software in the country.

The researcher beliefs the study can be used as a reference for the research works which will be done in this related area in the future, especially those research that apply data mining.

5.3 Future Work

This research has been attempted to identify the factor that affect patients cause of visit with limited data set and 13 attributes. Further researches can also be conducted by increasing the number of attributes and datasets.

This study was carried out using classification algorithms such as J48 decision tree, PART rule induction algorithms and Neural Networks. So further investigation needs to be done using other classification algorithms such as Naive Bayes and Support Vector Machine plus using a clustering technique and association rule discovery.

The predictive model, which is developed in this research, generated various patterns and rules. For the Emergency Department to use it effectively there is a need to design a knowledge base system, which can provide advice for the domain experts.

REFERENCES

[1]. Huang, H. et al. “Business rule extraction from legacy code”, Proceedings of 20th International Conference on Computer Software and Applications, IEEE COMPSAC’96, 1996, pp.162-167

[2]. Bellazzi, R. & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics, 77,81–97.

[3]. Cios Krzysztof J., Pedrycz Wiltod., Swiniarski Roman W. Kurgan Lukasz A. Data Mining: A knowledge Discovery approach. New York: Springer-Verlag Science Business Media LLC; 2007.

[4]. Witten I. H. and Frank E. 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2rd ed. Amsterdam: Morgan Kaufmann

[5]. K. MUSILEK, "A survey of Knowledge Discovery and Data Mining process models," *The Knowledge Engineering Review,* vol. Vol. 2 pp. PP. 1–24, 2006.

[6]. D. Hand, Mannila, H. and Smyth, P., *Principles of Data Mining*. Massachusetts Institute of Technology, Massachusetts: The MIT Press 2008

[7] K. MUSILEK, "A survey of Knowledge Discovery and Data Mining process models," *The Knowledge Engineering Review,* vol. Vol. 2 pp. PP. 1–24, 2006.

[8] G. J. ROGERS, E., "Mining Your Data for Healthcare Quality Improvement," *Journal of Healthcare Information Management,* vol. Vol 19(2): 65, 2005.

[9]. T. C. Corporation, *Introduction to Data Mining and Knowledge Discovery*. USA: Potomac, 2005.

[10]. J. H. a. Kamber, *Data mining concepts and techniques*, 2nd ed. San Francisco, USA: Morgan kaufman publishers, 2008.

[11]. E. Hajizadeh, D. Ardakani and Shahrabi, "Application of Data Mining Techniques," Journal of Economics and International Finance, vol. 2(7), pp. 109-118, 2010.

[12] M. Kantardzic, *Data mining concepts, methods and algorithms*, 2nd ed. Hoboken, New Jercy: Jone wily and Sons, Inc., 2011

[13]. M. L. Berry, G., *Data mining techniques for marketing, sales, and customer relationship management*, 2nd ed. Indiana: Wiley publishing, 2004

[14] K. CIOS, WITOLD, P, ROMAN, S AND KURGAN, A., *The Knowledge Discovery Process*. USA: Springer, Inc, 2007.

[15] D. Hand, Mannila, H. and Smyth, P., *Principles of Data Mining*. Massachusetts Institute of Technology, Massachusetts: The MIT Press 2008.

[16]. M. Bramer Principles of data mining. London: Springer, 2007.

[18] T. Larose, *Data mining methods and models*. New Jersey: John wily and Sons Inc., 2006.

[19]. M. Bramer Principles of data mining. London: Springer, 2007

[20] J. H. a. Kamber, *Data mining concepts and techniques*, 2nd ed. San Francisco, USA: Morgan kaufman publishers, 2008.

[21] E. Hajizadeh, D. Ardakani and Shahrabi, "Application of Data Mining Techniques," Journal of Economics and International Finance, vol. 2(7), pp. 109-118, 2010.

[22] M. Kantardzic, *Data mining concepts, methods and algorithms*, 2nd ed. Hoboken, New Jercy: Jone wily and Sons, Inc., 2011.

[23] T. Larose, *Data mining methods and models*. New Jersey: John wily and Sons Inc., 2006.

[24] M. Bramer *Principles of data mining*. London: Springer, 2007.

[25] D. T. Larose, *Discovering knowledge in data: an introduction to data mining*. New Jersey: John Wiley & Sons, 2005.

.[26] T. Debela." Developing a Predictive Model for Fertility Preference of Women of Reproductive Age Using Data Mining Techniques" Msc. Thesis, Addis Ababa University, Ethiopia, 2013

[27] T. Hailemariam." Application of Data Mining for Predicting Adult Mortality" Msc. Thesis, Addis Ababa University, Ethiopia, 2013

[28] D. Mamo. "Application of Data Mining Technology to Support Fraud Protection: The Case of Ethiopian Revenue and Custom Authority" Msc. Thesis, Addis Ababa University, Ethiopia, 2013

[29] A. Roberts, "guide to Weka.," 2005. Retrieved January 13, 2011, from http://www.comp.leeds.ac.uk/andyr

[30] I. W. a. E. Frank, *Data mining: practical machine learning tools and techniques with java implementations*, 2nd ed. San Francisco: Morgan kaufmann publishers, 2000.

[31] 41 .P. a. N. Clark, T. , "The CN2 induction algorithm," *Machine Learning*, vol. 3(4), pp. 261-283, 1989

[32] A. B. a. S. S. THERLING K. (2010, 20 September). *An Overview of Data Mining Techniques*. Available: http://www.thearling.com/text/dmtechniques/dmtechniques.htm

[33] H. M. MOSHKOVICH, MECHITOV,ALEXANDER I., AND OLSON,DAVID L., "Rule induction in data mining," *Ltd. Technical University of Crete*, vol. 22, pp. p.303-311, 2002.

[34] MDMW. (2008). *My Data Mining Weblog Rule learner (or Rule Induction)*. Available:

APPENDIX

Appendix 1: **Discussion questions forward to domain experts**

Is emergency cause of visit in the hospital is a hot issue now a day?

What techniques do you currently use to identify emergency causes?

What are the existing solutions for reducing the emergency cases in the hospitals?

Appendix 2:- Sample J48 Decision Tree With 10 Fold Cross Validation

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Transferred to = Emergency OPD

| Region = ADDIS ABABA

| | Age = adalt

| | | place of injury = PLAYING: Trauma (294.0/11.0)

| | | place of injury = HOME: None Trauma (585.0/25.0)

| | | place of injury = STREET: None Trauma (0.0)

| | | place of injury = WORK: None Trauma (390.0/13.0)

| | Age = child: None Trauma (588.0/23.0)

| | Age = old

| | | Subcity = Addis Ketema

| | | | Intent = Accident: None Trauma (196.0/7.0)

| | | | Intent = violence: None Trauma (0.0)

| | | | Intent = Other: Trauma (196.0/10.0)

| | | Subcity = Akaki Kaliti: None Trauma (196.0/11.0)

| | | Subcity = Arada

| | | | Sex = M: Trauma (196.0/4.0)

| | | | Sex = F: None Trauma (195.0/9.0)

| | | Subcity = Gulele: Trauma (195.0/8.0)

| | | Subcity = Kolfe Keranio: None Trauma (392.0/22.0)

| | | Subcity = Outside AA: Trauma (0.0)

| | | Subcity = Kirkos

| | | | Past medical illness = NO: Trauma (588.0/34.0)

| | | | Past medical illness = YES: None Trauma (195.0/7.0)

| | | | Past medical illness = UNKNOWN: Trauma (0.0)

| | | Subcity = Lideta

| | | | place of injury = PLAYING: Trauma (0.0)

| | | | place of injury = HOME: None Trauma (98.0/3.0)

| | | | place of injury = STREET: None Trauma (98.0/9.0)

| | | | place of injury = WORK: Trauma (196.0/8.0)

| | | Subcity = Bole: Trauma (196.0/8.0)

| | | Subcity = Nifas Silk

| | | | Sex = M: None Trauma (196.0/11.0)

| | | | Sex = F: Trauma (196.0/14.0)

| | | Subcity = Yeka: Trauma (392.0/14.0)

| Region = OROMIA

| | Triage Assessment = Green: Trauma (881.0/32.0)

| | Triage Assessment = Red

| | | Sex = M: None Trauma (196.0/10.0)

| | | Sex = F: Trauma (195.0/7.0)

| | Triage Assessment = Black: Trauma (0.0)

| Region = AMHARA

| | Sex = M: None Trauma (196.0/7.0)

| | Sex = F: Trauma (196.0/10.0)

| Region = SNNP: None Trauma (98.0/2.0)

| Region = afar: Trauma (98.0/4.0)

| Region = AFAR: None Trauma (98.0/3.0)

Transferred to = Resuscitation room

| Subcity = Addis Ketema: Trauma (0.0)

| Subcity = Akaki Kaliti: Trauma (293.0/11.0)

| Subcity = Arada: Trauma (488.0/20.0)

| Subcity = Gulele: Trauma (0.0)

| Subcity = Kolfe Keranio: None Trauma (196.0/8.0)

| Subcity = Outside AA: Trauma (195.0/9.0)

| Subcity = Kirkos: Trauma (0.0)

| Subcity = Lideta: Trauma (392.0/21.0)

| Subcity = Bole: Trauma (0.0)

| Subcity = Nifas Silk: Trauma (196.0/9.0)

| Subcity = Yeka: Trauma (0.0)

Transferred to = Stabilization room

| Subcity = Addis Ketema: Trauma (0.0)

| Subcity = Akaki Kaliti: Trauma (0.0)

| Subcity = Arada: Trauma (0.0)

| Subcity = Gulele: Trauma (0.0)

| Subcity = Kolfe Keranio: None Trauma (195.0/13.0)

| Subcity = Outside AA: Trauma (0.0)

| Subcity = Kirkos: Trauma (0.0)

| Subcity = Lideta: Trauma (0.0)

| Subcity = Bole: Trauma (0.0)

| Subcity = Nifas Silk: Trauma (0.0)

| Subcity = Yeka: Trauma (196.0/9.0)

Transferred to = Regular OPD : Trauma (196.0/10.0)

Transferred to = Referred to other Hospital: Trauma (196.0/12.0)

Appendix 3 : Sample PART Decision Rule List With 10-fold cross validatio

Transferred to = Resuscitation room AND

Subcity = Arada: Trauma (488.0/20.0)

Age = child: None Trauma (784.0/31.0)

Transferred to = Resuscitation room : Trauma (1076.0/50.0)

Subcity = Kolfe Keranio: None Trauma (782.0/42.0)

place of injury = STREET AND

Subcity = Outside AA: Trauma (684.0/24.0)

Transferred to = Emergency OPD AND

Age = adalt AND

place of injury = HOME: None Trauma (585.0/25.0)

Transferred to = Emergency OPD AND

Triage Assessment = Green AND

Referred From = GOV AND

Intent = Accident: Trauma (1175.0/55.0)

Transferred to = Emergency OPD AND

Activity = other AND

Triage Assessment = Red: None Trauma (784.0/33.0)

Subcity = Outside AA: Trauma (686.0/33.0)

Past medical illness = UNKNOWN: Trauma (784.0/35.0)

Referred From = NGHS: None Trauma (586.0/20.0)

Subcity = Kirkos: Trauma (294.0/18.0)

Referred From = GOV AND

Subcity = Arada AND

Sex = M: Trauma (196.0/4.0)

Subcity = Lideta AND

place of injury = WORK: Trauma (196.0/8.0)

place of injury = HOME AND

Activity = Traveling : Trauma (392.0/18.0)

: None Trauma (587.0/32.0)