Journal of
Health & Medical Informatics

Kale., J Health Med Informat 2015, 6:1
DOI: 10.4172/2157-7420.1000178

Research Article Open Access

Application of Data Mining Techniques to Discover Cause of Under-Five Children Admission to Pediatric Ward: The Case of Nigist Eleni Mohammed Memorial Zonal Hospital

Temesgen Dileba Kale*

Hossana College of Health Sciences, Ethiopia

Abstract

Background: Data mining has ability to extract useful knowledge that is hidden in huge data. Health care system is potential area to apply and take the advantage of data mining. The causes of child illnesses and admissions to hospitals that utilize the scarce resource in sub-Saharan region were easily preventable. Higher priority was given for prevention and control of these diseases at community level. However, for seriously ill children admissions should be facilitated in order to save the life of the child. Therefore, the purpose of this study was to apply data mining techniques on under-five children dataset in developing a model that support the discovery of the causes for under-five children admission to pediatric ward.

Methodology: The six-step cross industry standard process for data mining model was applied. Decision tree and artificial neural network algorithms were tested for classification. Exploratory data analysis techniques, graphs and tabular formats for visualization and accuracy, true positive rate, false positive rate, Receiver Operating Characteristic curve and the idea of experts were used for evaluation of the model.

Result: A total of 11,774 instances were used to construct the decision tree and artificial neural network. The decision tree algorithm J48 has higher accuracy (94.77%), weighted true positive rate (94.7%), weighted false positive rate (5.3%), weighted Receiver Operating Characteristic curve (0.99) and performs much faster than multilayer perceptron. According to the interesting rules in J48 presenting complaint of not taking any food, fluid or breast feeding (98.32%), was the top cause of under-five children admission to pediatric ward without any consideration of health information management system admission criteria.

Conclusion: In conclusion, encouraging results were obtained in classification tasks, data mining technique was applicable on pediatric dataset in developing a model that support the discovery of the causes of under-five children admission to pediatric ward.

Keywords: Under-five children; Cause of admission; Data mining; Under-five data; Pediatric- ward

Introduction

Background

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data [1]. It is an iterative process and search for new, valuable, and nontrivial information in large volumes of data and needs a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers [2]. It has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration [3].

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human endeavor, from this world (such as supermarket transaction data, government statistics, etc.) to the more exotic (such as molecular databases, and medical records). Interest has grown in the possibility of tapping these data, of extracting from them information that might be of value to the owner of the database [4].

When we see the evolution of data mining, it does not mean that data mining is different from other disciplines such as statistics. In statistics, researchers frequently deal with the problem of finding the smallest data size that gives sufficiently confident estimates. Data mining deals with the opposite problem, that, data size is large and we are interested in building a data model that is small but still describes the data well [5,6].

According to UNICEF report on issues of child mortality, a child born in developing country is over 13 times more likely die within the first five years of life as compared to economically advanced countries. Sub-Saharan Africa accounted for about half of these deaths in the developing world. Surprisingly the causes of illnesses and admissions to hospitals that utilized the scarce resource in the region were diseases that can be easily prevented. Child illnesses and deaths were higher for children from rural and poor families and whose mother lack basic education. An Ethiopian child is 30 times more likely die by his or her fifth birth day than a child in Western Europe [7,8].

***Corresponding author:** Temesgen Dileba Kale, Hossana College of Health Sciences, Ethiopia, Tel: +2510932583137; E-mail: temedile@yahoo.com

Received November 13, 2014; **Accepted** December 08, 2014; **Published** December 15, 2014

Citation: Kale TD (2015) Application of Data Mining Techniques to Discover Cause of Under-Five Children Admission to Pediatric Ward: The Case of Nigist Eleni Mohammed Memorial Zonal Hospital. J Health Med Informat 6: 178. doi:10.4172/2157-7420.1000178

When the issue of growth and development of a country is raised, the indicators of child illness and deaths have a prominent place. Ethiopia is among the highest country in child illnesses and deaths in the world. As indicated in health and health related indicators of 2011, neonatal mortality rate was 36/1000 live birth and infant mortality rate was 67/1000 live births [9]. Pneumonia (19%), diarrhea (18%), malaria (8%) and measles (4%) that easily prevented through simple improvements in basic health services and interventions, were the leading cause of child illnesses and deaths [10].

Hospitals and health centers were primarily medical care centers where huge data is collected regarding the patients and clients in daily bases. Unless this data has been stored and processed to provide appropriate information, it has no value. Appropriately processed information is a good source for health managers at different level of organizations in order to make informed decision. Huge data is stored either electronically or manually, however, lack of capacity in turning these data into information and then to use it as an input for decision making was in question. Therefore, the ability to use these data to extract useful information for quality health care is a crucial issue. Data mining is one of the solutions to analyze large amount of data and thus into information and knowledge.

Hospital admissions provided some measures of prevalence and severity of childhood illness. Admission rates and their cause reflect socioeconomic circumstances, the level of utilization of primary health care services and the health care seeking behavior of community. However, because of delayed health seeking behavior of parents and other care providers, easily preventable diseases may be complicated and even may lead to death of children.

Higher priority should be given for the prevention and control of preventable diseases at community level. However, for seriously ill children admissions should be facilitated in order to save the life of the child. This assists in managing the cases and facilitates the conditions for the next admission in order to utilize resources appropriately.

There is a rapidly widening gap between data-collection and data-organization capabilities and the ability to analyze the data. Whether the context is business, medicine, science, or government, the datasets themselves, in their raw form, are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use. The root of the problem is that the data size and dimensionality are too large for manual analysis and interpretation, or even for some semiautomatic computer-based analyses [1].

The underlying research problem that necessitated this research was the fact that, although large amount of data is available, they were not using it in a way that supports their objectives. Decision making bodies were not using this data for making informed decision. Thus, the data remains unutilized to the problems faced by the society due to lack of research in deploying appropriate data analysis and mining tools.

Therefore, this research was applied to discover the cause of under-five children admission by using data mining techniques on large data in pediatric dataset of Nigist Elleni Mohammed Memorial hospital in order to build predictive model for outcomes of under-five outpatient department visit. To this extent, the study provided answer for the research question in that:

- Preparing the data set for mining under-five OPD visit data
- Extracting hidden knowledge from under- five children admission

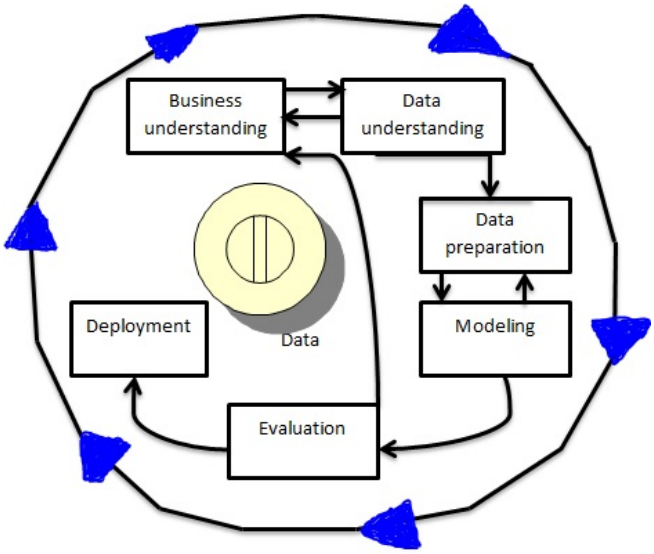

Figure 1: CRISP-DM 1.0 step-by-step data mining process model.

- Exploring the common signs and symptoms that were contributing to under-five children admission

Data Mining Models and Methods

The crisp-DM process model

CRISP-DM model was selected for this research in order to discover knowledge. It is the most popular knowledge discovery process model [5]. In CRISP-DM the numbers of steps followed are six. Each step has good documentation and divided into sub steps which help easily to identify all necessary details in the knowledge discovery process. The six step CRISP-DM process model is depicted in the figure below (Figure 1)

A model is an abstract representation of a real-world process. For example, $Y = 3X + 2$ is a very simple model of how the variable Y might relate to the variable X [3]. Two primary goals of data mining in practice are prediction and description in order to develop a model that predict or describe the variables [2,11].

Data mining algorithms to achieve the predictive goals of data mining

Classification Algorithm: Classification of a collection consists of dividing the items that make up the collection into categories or classes. Different classification algorithms use different techniques for finding relations between the predictor attributes' values and the target attribute's values in the build data. These relations are summarized in a model; the model can then be applied to new cases with unknown target values to predict target values. The comparison technique is called testing a model, which measures the model's predictive accuracy. The application of a classification model to new data is called applying the model, and the data is called apply data or scoring data [12].

Classifier accuracy measures: The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. Classifier accuracy measures using the same dataset to derive a classifier or predictor and then to estimate the accuracy of the resulting learned model results in misleading overoptimistic estimates due to over specialization of the learning algorithm to the data. Then,

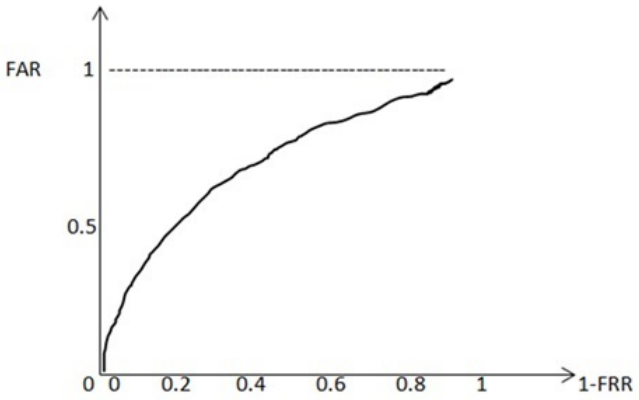

Figure 2: The ROC curve shows the trade- off between sensitivity and 1-specificity value.

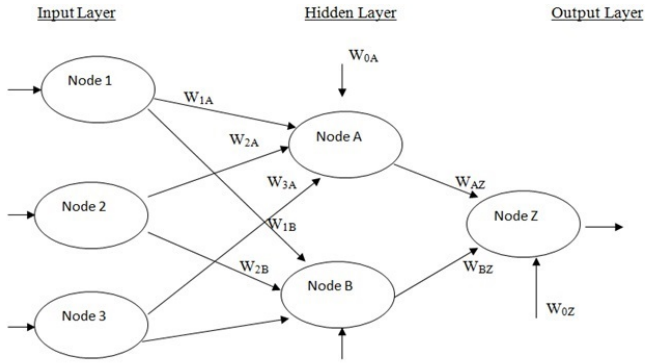

Figure 3: Simple neural networks. An introduction to data mining [15].

the classifier is applied on the test set and the number of instances that were assigned to actual classes and different class by the classifier is counted, a process whose result is effectively represented by confusion matrix [3].

Confusion matrix: The confusion matrix is a useful tool for analyzing how well classifier can recognized tuples of different classes. Given *m* classes, a confusion matrix is a table of at least size *m* by *m*. An entry, *CM i*, *j* in the first *m* rows and *m* columns indicates the number of tuples of class *i* that were labeled by the classifier as class *j*. For a classifier to have good accuracy, most of the tuples would be represented along the diagonal of the confusion matrix, from entry *CM*1, 1 to entry *CM m*, *m*, with the rest of the entries being close to zero [4].

Receiver operating characteristic curve: A classification model can be balanced by setting an appropriate threshold value to operate at a desired value of false acceptance rate (FAR). To analyze false acceptance rate and false reject rate (FRR) a parameter, the Receiver Operating Characteristic (ROC) Curve was developed. It is a plot of FAR versus FRR for different threshold values in the model. This curve permits to assess the performance of the model at various operating points (thresholds in a decision process using the available model) and the performance of the model as a whole (using as a parameter the area below the ROC curve). The ROC curve is especially useful for a comparison of the performances of two models obtained by using different data-mining methodologies [3,4] (Figure 2)

Decision Tree: Decision tree is a structure that can be used to divide a large collection of records into successively smaller sets of records by applying a sequence of decision rules [13]. It is a supervised learning method that constructs decision trees from a set of input-output samples. A typical decision-tree learning system adopts a top-down strategy that searches for solution in a part of the search space [2].

Decision tree consists of nodes and branches connecting the nodes. The nodes located at the bottom of the tree are called leaves and indicate classes. The top node in the tree, called the root, containing examples that are to be divided into classes. All nodes except the leaves are called decision nodes, since they specify decision to be performed at this node based on a single feature. Each decision node has a number of children nodes, equal to the number of values that a given feature assumes [5].

Tree induction: decision trees are used to predict and/or classify. There are two phases, the training and implementation. During the training phase, the data set is partitioned iteratively. During each pass (i.e. iteration), the data set is split on that feature (or attribute) that produces the most effective classification. Only those factors most significant to the partitioning are used. The implementation phase then produces decision rules which are equivalent to the partitions (or branching) created during the training phase. These rules are used to generate new information when presented with novel situations [1,5].

In pre-pruning, the growth of the tree stops when it is determined that no attribute will significantly increase the information gain in the process of classifying the data. While in post pruning, involves already-constructed trees. Complexity of the tree is resulted in observed loss in classification accuracy hence in order to make a good decision much of tree branches should be eliminated [2].

Artificial neural network (ANN): Artificial neural network is an abstract computational model of the human brain. It has the ability to learn from experiential knowledge expressed through inter unit connection strengths, and can make such knowledge available for use. It has the following capabilities; a typical neural network is composed of a potentially large number of neurons arranged in three different conceptual layers: an input layer representing the input variables, one or more hidden layers, and an output layer representing the output variables [2].

A neural network starts with an input layer, where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response nodes [14].

As indicated in Figure 3 below W_{1A}, W_{2B} ….W_{3B} represents weights of the nodes. Every node in a given layer is connected to every node in the next layer, although not to other nodes in the same layer. Each connection between nodes has a weight (e.g. W_{1A}) associated with it. At initialization, these weights are randomly assigned to values between zero and one. The number of input nodes usually depends on the number and type of attributes in the data set. The numbers of hidden layers, and the number of nodes in each hidden layer, are both configured by the user.

On computation of the weight value and methods used, the connection weights (W's) are the unknown parameters which are estimated by a training method (back propagation). First a combination function (usually summation) produces a linear combination of the inputs and the connection weights into a single scalar value. This will give the net for a given node. Once the net value for each node is

known; it will be used as an input to the activation function (usually sigmoid function) which is used to generate the output signal from the weighted average of inputs. For the total output node Z, net $_z$ can be calculated [14].

Artificial neural networks have memory that corresponds to the weights in the neurons. Neural networks can be trained offline and then transformed into a process where adaptive learning takes place. In the analysis of medical data, ANNs have become an alternative to classic statistical methods in recent years [15]. Several ANN techniques were applied in medicine, including methods for diagnosis and prognosis tasks, especially for survival analysis. Most applications of ANNs in medicine refer to classification tasks [16].

To this extent the decision tree and neural network classification algorithms were applied to achieve the predictive goals of data mining in pediatric dataset of Nigist Elleni Mohammed Memorial Zonal Hospital by using WEKA (Waikato Environment for Knowledge Analysis) as the classification tool.

Ethical considerations

The dataset contained the child's identifying information and services provided in the hospital while managing the condition of the child. The greatest contribution of this research is just assisting overall process of child admission to pediatric ward. There was no attempt that will be made to disclose participants' information. Therefore, to handle such an ethical issue to be raised in the study, the researchers obtained ethical clearance from the research and ethics committee of the School of Public Health of Addis Ababa University. In line with the ethical clearance, permission letter was also obtained from the medical director of the hospital. Furthermore, in order to manage confidentiality and privacy issues of the data, the identification information of the children was removed from the dataset.

Experimentation and Discussion of Result

Business understanding

Hadiya zone is one of the largest Zones in south nation nationalities peoples region (SNNPR), Ethiopia. It is located 232kms south west of Addis Ababa, the capital city of Ethiopia. According to the national census of 2007, the total population of the Zone estimated around 1,371,625 from these infants and under- five children comprised of 48,007 and 213'971 respectively. Hossana hospital recently called Nigist Elleni Mohammed Memorial (NEMM) Hospital; was the only Public Hospital in the Zone and governed by Zonal administration which is located in Hossana town. It provides service for the Zonal population and also people nearby the zone.

The Hospital has four major inpatient departments (medical, surgical, pediatric, Gynecology/obstetrics), outpatient department and operation room. Under- five children were treated at under-five outpatient department (OPD) and pediatric ward. Children older than five-years and surgical cases were managed in other OPDs and surgical ward according to the cases and the hospital's management protocol. According to the statistical report of the hospital, the major causes of under-five OPD visits were pneumonia, malaria, diarrhea, malnutrition, and other febrile illnesses.

The information of a child in under-five OPD visit was recorded in daily bases in integrated under-five registration log book as per the visit. The log book captured records like age, sex, address, HMIS disease classification, HIV test, outcome of admission, danger signs, other unspecified signs and symptoms, presenting complaints (fever, cough, diarrhea, etc.), immunization status, vitamin 'A' supplementation , weight, weight for age, and length of stay. Based on the Hospital's working procedure outcome of under-five visit (admitted, not admitted) was taken as a class label and other attributes were checked for better result yield in data mining. There were 15,824 records of children in under- five OPD since June 7, 2007.

Under-five data understanding

A/ Data selection process: The dataset used for this study was the records kept from June 7, 2007 to February 7, 2012 in integrated under-five registration log book. A total of 15,824 records were obtained from the log books. Therefore the data collection for mining begun by coding these data into MS excel format.

B/ Basic data distribution: The dataset were organized in columns and rows where the columns represent an attribute and the rows represent single records of under-five OPD visit cases. In general there were 64 attributes (columns) and 15,824 records (rows) in under-five OPD dataset. Out of these under-five visits 1,264 records were stored in integrated under-five registration log book of birth to 2 months and the rest 14,560 records were stored in integrated under-five registration log book of 2 months to 5 year.

The records of some attributes in the log books of birth to two months and two months to five years were recorded differently. To select best attributes from those records, opinion of experts was taken. Finally, those attributes that were similar and mostly supportive for the diagnosis of the diseases were selected. Therefore, final data preprocessing was done on 28 attributes and 15,824 children records of under-five outpatient department (OPD) visits. The outcome of under-five OPD visit was either treated at OPD (not admitted) or treated in pediatric ward (admitted). Thus, some selected attributes, data types, and number of unique value the attributes taken with the number of missing values among 28 attributes were showed in Table 1 below.

Data preprocessing

As shown in Table 1, many selected attributes comprised of missed values in the records, incompleteness and noise. There were also fields that are redundant and containing derived values. Therefore,

S. NO	Name of Attribute	Data Type	Description of attributes	missing values
1	Age	Numeric	The age of under-five child	4 (0.03%)
2	Sex	Categorical	The sex of under-five child	8 (0.05%)
3	Weight	Numeric	The weight of under-five child	7 (0.05%)
4	Cough	Categorical	complaint of under-five child to seek medical care	14(0.09%)
5	Diarrhea	Numeric	complaint of under-five child to seek medical care	29(0.18%)
6	OPD visit	Categorical	Determining OPD visit as admitted or not admitted	0(0%)
7	HMIS**	Categorical	Diagnosis of under-five child made at OPD*	0(0%)

*-Outpatient department ** Health Management |Information system

Table 1: List of some selected attributes, descriptions and missed values in records of under-five outpatient department in NEMM Hospital from June, 2007- February, 2012.

such unnecessary attributes were removed from the records and the remaining attributes were considered for further preprocessing. As the dataset that is selected and prepared for the mining purposes should have initial features (relevant attributes); the selection of relevant attribute was done by working with experts with special care. Therefore, handling of missing values, incorrect values, noises and other irrelevant values for the attribute/fields selected was applied in the next section.

Missing values were observed in all selected attributes except the outcome of under-five visit and HMIS disease classification. Usually in data mining missing values are replaced either by mean value or modal value for numeric and nominal attributes. For numerical variables like age and weight the missing value was replaced by the mean value of age and weight respectively. Interestingly the tool WEKA replaces missing value by mean value. For all nominal variables like sex, cough, immunization status, oral trash, etc. the modal value was filled manually.

Regarding to outliers and noise field values, correction was done manually. Data was entered incorrectly for some attributes which created noise and contained outlier. For some other attributes its value was unique categorical, but incorrectly values other than the mentioned unique values were obtained. For example the sex of the child may be male (M) or female (F) and that is unique categorical value. But some of it was recorded as mm, ff, fm, & mf and these all create noise. The weight of the child was recorded as 85kg. This was a typical outlier and should be corrected. Such problems were handled and corrected values were substituted. The detail of how each attributes were handled in removing outliers and noise field values was described in the following section.

Age: - An attribute age was recorded in days, weeks, months and years. This makes it difficult for analysis. Based on the opinion of the experts, some disease classifications and management of child illness relied on the age classification of the children. For example neonatal sepsis, neonatal tetanus etc. were illnesses that were occurred at neonatal period. Child immunization took more attention at infantile period. Integrated management of neonatal and childhood illnesses (IMNCI) management in under-five OPD also relied on managing children less than two months in one category and children older than two months to five years in another category. Based on that, age data was transformed in that, 0-28 days to neonate, 29 day to 1year to infant and greater than 1year to five years to child. This transformed developmental age period also helps to identify the frequency of visits to OPD taken by those developmental age groups. Therefore, taking the name of child's development periods is relatively appropriate in classifying the age of children.

Address: - The address of child in the record describes the geographical area distribution of the children visited the OPD. As the Hospital is located in Hadiya Zone, the children who visited the hospital were primarily from districts of Hadiya Zone and other neighbor zones. The records of address were accordingly transformed as, those children visited from Hadiya zone were recorded as their respective districts such as Hossana, Lemo, Analemo, Misha, Gombora, Soro, Duna, Shashogo, and Gibe. Whereas, those children visited from Lera, Silte, Worabe, Wulbareg, Sankura, etc were recorded as Silte Zone, from Azernet, Endagagn, Wolkite, etc were recorded as Gurage Zone and Angecha, Hadaro, Doyogena etc were recorded as KT Zone.

Sex: - The sex of children who visit under-five OPD was also an important attribute selected for this study. It was recorded as either "m" or "f". The letter "m" is transformed to male and "f" to female, but its missing values was replaced by WEKA tool for its modal value.

Visit: - Another attribute selected was the frequency of visit that the children made to under-five OPD. Those children visited for the first time were recorded as "initial" visit and children visited more than one times were recorded as "follow up" visit, whereas the missing values were replaced by WEKA tool for its modal value.

Weight: - Weight is also an important attribute selected for this study, because it is direct reflection of the child's health and nutritional status. In relation to age, it is a normal child growth indicator because as the age of child increases the weight also increases at a proportional rate termed as appropriate weight for specified age. To the contrary, the child's weight which is not proportional to the age of child is termed as large weight for age, low weight for age or very low weight for age. As it is continuous numeric variable, it automatically descritized into five bins by using WEKA tool for analysis and respective model building. The summary of some selected attributes is shown on Table 2.

Presenting complaints of the child

The next attribute selected for this study was the presenting complaint of the child (symptoms) and physical findings observed by health care providers (signs). Under this attribute the symptoms presented on the child (subjective complaint of the patients) that urged the child to visit under-five OPD were recorded. Group of signs and symptoms presented on children are important clinical findings to know the specific cause of illnesses that resulted in under-five OPD visit. So that, which clinical findings are related with the specific diagnosis was an important area of this study. Therefore, in the following section clinical signs and symptoms, those are selected as best attribute for this study were briefly described.

Before the selection of best signs and symptoms, knowing what clinical signs and symptoms were recorded in two integrated under-five registration log books (birth to 2 month and 2 month to 5yrs) in under-five OPD is a crucial issue. Clinical signs and symptoms such as cough, fever, diarrhea, vomiting, abdominal pain/cramp, trauma/injury/accident/poisoning, unable to drink/breast feed, convulsions, lethargic/unconscious, fast breathing, stridor, bulged fontanelle, irritable/restless, sunken eyes, oral trush, blood in stool, low weight for age,

Attributes	Old values recorded		New value it takes	Remark
Age	**Age in days, weeks, months, and years**		**Neonate, infant, child**	**Nominal**
Sex	**m ,f**		**Male, Female**	**Nominal**
Visit	**initial, follow up**		**Initial, Followup**	**Nominal**
Address	Hadiya zone	hossana, lemu, analemu, duna, gombora, misha , gibe, soro, shashogo, badewacho	Hossana, Lemu, Analemu, Duna, Gombora, Misha, Gibe, Soro, Shashogo, Badewacho	Nominal
	Neighbor Zones	lera, silte, worabe, wulbareg, sankura	Siltezone	
		doyogena, angecha, hadaro, obichaka	KTzone	
		azernet, endagagn,wolkite	Guragezone	
weight		Weight of children in kg	Weight of children in kg	Numeric continuous

Table 2: Summary of some selected attributes and transformed value in records of under-five outpatient department in NEMM Hospital from June, 2007- February, 2012.

S. no	Attributes	Description	Value it takes	Remark
1	Cough	Presenting complaint of the child	Yes or no	2 categorical
2	Fever	Presenting complaint of the child	Yes or no	2 categorical
3	Diarrhea	Presenting complaint of the child	Yes or no	2 categorical
4	Vomiting	Presenting complaint of the child	Yes or no	2 categorical
5	Abdominal pain/cramp	Presenting complaint of the child	Yes or no	2 categorical
6	Trauma/injury/accident/poisoning	Presenting complaint of the child	Yes or no	2 categorical
7	Unable to drink/breast feed	Presenting complaint of the child considered as danger sign	Yes or no	2 categorical
8	Convulsion history/now	Presenting complaint of the child considered as danger sign	Yes or no	2 categorical
9	Lethargic/unconscious	Presenting complaint of the child considered as danger sign	Yes or no	2 categorical
10	Fast breathing	Presenting complaint of the child	Yes or no	2 categorical
11	Stridor	Presenting complaint of the child	Yes or no	2 categorical
12	Chest indrawing	Presenting complaint of the child	Yes or no	2 categorical
13	Bulge fontanelle	Presenting complaint of the child	Yes or no	2 categorical
14	Restless/irritable	Presenting complaint of the child	Yes or no	2 categorical
15	Blood in stool	Presenting complaint of the child	Yes or no	2 categorical
16	Sunken eyes	Presenting complaint of the child	Yes or no	2 categorical
17	Skin pinch	Presenting complaint of the child	Normal, low, veryslow	3 categorical
18	Weight for age	Presenting complaint of the child	AWFA*,VLWA**, LWFA***	3 categorical
19	Oral trash	Presenting complaint of the child	Yes or no	2 categorical

*Appropriate weight for age, **Very low weight for age, ***Low weight for age

Table 3: Some selected clinical signs and symptoms as an attribute and descriptions in records of under-five outpatient department in NEMM Hospital from June, 2007-February, 2012.

Attribute	Description	Value it takes	Remark
Immunization status	The child's immune status against targeted diseases	Completed, uptodate, defaulted, notstarted	4 categorical

Table 4: Immunization status, descriptions, transformed values in records of under-five outpatient department in NEMM Hospital from June, 2007- February, 2012.

chest in drawing were recorded in both birth to 2 month and 2 month to 5yrs integrated under-five registration log books. Clinical signs and symptoms such as nasal flaring, grunting, gestational age, movement less than normal, jaundice, breast feeding, supplementary feeding were recorded in birth to 2 month integrated under-five registration log book only whereas stiff neck, red eye, mouth ulcer, ear pain/discharge, foot edema, Mid-upper arm circumference (MUAC) <10 cm, pallor, appetite were recorded in 2 month to five year integrated under-five registration log book only.

For the selection of attributes, the opinion of the experts was taken as an important input. Children may complain for cough, fever, diarrhea, vomiting, abdominal pain/cramp, trauma/injury/accident/poisoning, unable to drink/breast feed, convulsions, lethargic/unconscious, fast breathing, stridor, bulged fontanelle, irritable/restless, sunken eyes, oral trush, blood in stool, low weight for age, chest indrawing and others. All these, clinical signs and symptoms are grouped into subjective complaints or objective findings that may results in under-five OPD visit and/or pediatric ward admission. By interpreting all these findings and professionals judgment HMIS disease classifications was termed for each under-five child visit or admission. Hence, all mentioned clinical findings that are recorded in both log books were taken as a best attribute for this study as presented in Table 4.

Immunization status

Immunization status of the child is another important risk factor in assessing the health status of the children. Children may be vaccinated against targeted diseases such as tuberculosis, polio, diphtheria, tetanus, whooping cough, hepatitis B virus, hemophylus influenza type B, measles etc. in governmental health organizations and non-governmental organizations. Therefore, children immunized against targeted diseases are assumed to be protected from those diseases. In this sense knowing the immune status of the child in relation to child's disease condition is one of the crucial issues. Accordingly, immunization status of the child was also selected as attribute for this study.

Immunization status was recorded as completed, up to date, defaulted or not started. Immunization status is completed to mean the child is completely vaccinated against targeted diseases, up to date means the child has started vaccination but not completed, defaulted means the child has started vaccination but not continued the next schedules to complete it and not started means the child yet not started immunization. A detail of transformed values on immunization status is shown in Table 4.

HMIS disease classification

Under-five children visit under-five OPD for their illnesses. Ministry of health classifies these illnesses as HMIS diseases. HMIS diseases are termed as HMIS admission disease classification based on the child's admission diagnosis made in under-five OPD. It describes for what illness the child was admitted to the pediatric ward. For example, HMIS admission disease classification is malaria means the child is admitted to pediatric ward for the illness malaria and takes medical care in the ward. Therefore, HMIS admission disease classification is the diagnosis made in under-five OPD for the illness that child to visit under-five OPD. Hence, it is one of the attributes selected for this study.

HMIS disease classification

Under-five children visited under-five OPD for their illnesses. HMIS diseases are termed as HMIS admission disease classification based on the child's admission diagnosis made in under-five OPD.

A total of 116 HMIS admission disease classifications were recorded with the highest frequency of pneumonia (4064), 25.68% and the other 2 diagnosis each consists of (3) 0.01%. This shows that, there was high difference between common cases and rare illnesses. From these points of view it needs some adjustments and the experts' involvement was

S. No	Attribute		
	Old value	**New value**	**Value it takes**
1	Malaria	malaria	Malaria
2	TAPF	Trauma, accident, poisoning,fructure	TAPF
3	Helminthiasis	helminthiasis	Helminthiasis
4	Burn	burn	Burn
5	Severe pneumonia	severepneumonia	Severepneumonia
6	UTI	UTI	UTI
7	Skin infection	skininfection	skininfection
8	Meningitis	meningitis	Meningitis
9	Tonsillitis	tonsilitis	tonsilitis
10	Diarrhea	diarrhea	diarrhea
11	Foreign body	foreignbody	foreignbody
12	Constipation	constipation	constipation
13	Severe acute malnutrition	SAM	SAM
14	Typhoid fever	Typhoid fever	Typhoidfever
15	Nephritis	Nephritis	Nephritis
16	Conjunctivitis	Conjunctivitis	Conjunctivitis
17	Hernia	Hernia	Hernia
18	Genitalia problems	Genitalia problems	Genitalia problems
19	Infected wound	Infected wound	Infectedpneumonia
20	Cellulitis	Cellulitis	Cellulitis
21	Liver and GI problems	Liver and GI problems	LGIP
22	Pneumonia	Pneumonia	Pneumonia
23	Dysentery	Dysentery	Dysentery
24	Tumor	Tumor	Tumor
25	Neonatal sepsis	Neonatal sepsis	Neonatalsepsis
26	Intestinal obstruction	Intestinal obstruction	Intestinalobstruction
27	Epilepsy	Epilepsy	Epilepsy
28	Rectal prolapsed	Rectal prolapsed	Rectalprolapse
29	Measles	measles	Measles
30	AFI	Acute febrile illness	AFI
31	Seizure	Seizure	Seizure
32	Tuberculosis	Tuberculosis	Tuberculosis
33	Prematurity	Prematurity	Prematurity

Table 5: Some of HMIS disease classifications and transformed values in records of under-five outpatient department in NEMM Hospital from June, 2007- February, 2012.

needed. So that, by giving primary consideration for the objective of the study and the impact of using 116 values for one attribute on algorithms to be used, the following work was done by involving experts.

Accordingly, representing some of the values by some other higher concepts was done by considering the relative prevalence of the illnesses, illnesses affecting the organ system of the body, those illnesses that affect the same organ but having different diagnosis, diagnosis that were given different names based on its severity and illnesses of very rare diagnosis. Hence, a total of 53 categories were selected for the attribute HMIS admission disease classification. HMIS admission disease classifications and transformed value are shown in Table 5.

Outcome of under-five OPD visit: - Dependent attribute (outcome variable) for this study was the outcome of under-five OPD visit. Based on the findings assessed at under-five OPD, the Doctors reached at the diagnosis of the child's illness. These illnesses are assigned into HMIS disease classification, where it needs immediate treatment under direct observation of physician and nursing care (admitted) or provided appropriate treatment at OPD level or else referred to other health facility for further management(not admitted). Therefore, prediction of cause of under-five children admission is possible at this point by taking different attributes as risk factors and taking admitted or not admitted as an outcome variable.

Data reduction and feature selection: Inputs from domain expert are essential in data reduction and feature selection process. Because expertise in assessing child health condition and management is not trivial, their involvement in the preprocessing step is more essential. Thus, candidate variables were selected primarily by taking the research objectives and bases of experts influence in assessing child health condition and management. The reduction process applied on attributes was described as follows.

1. Attributes like serial number, medical registration number, date of visit, date of admission, data of discharge, etc. have no value in predicting admission. In addition attributes like amount of charge, amount paid, voucher and treatment contained too many missing values (more than 80%).
2. Attributes such as HIV test offered, HIV test performed, temperature in degree centigrade, vomits everything, breathing problem and birth weight, were removed because all these attributes are redundant.
3. Attribute like duration of cough, respiration rate in minutes, risk area of malaria, check Vitamin A supplementation in last six months, etc. were removed because it contained too many missing values (more than 80%).
4. Attributes like drinking poorly, history of measles last six

S.No	Attributes	Value it takes	Composition in number	Composition in %
1	Age category	Child	8165	51.6
		Infant	6144	38.8
		Neonate	1515	9.6
2	Sex	Male	9065	57.29
		Female	6759	42.81
3	Address	Hossana	9643	60.94
		Lemo	2923	18.47
		Misha	775	4.90
		Shashogo	318	2.01
		Analemo	348	2.20
		Soro	295	1.86
		Gibe	209	1.32
		KTzone	281	1.78
		Gombora	278	1.76
		Siltezone	217	1.37
		Duna	338	2.14
		Badewacho	26	0.16
		Guragezone	162	1.02
4	Weight in kg	Less than 1.7	24	0.15
		1.7-2.7	56	0.35
		2.7-5.95	1583	10.00
		5.95-10.25	7311	46.21
		Greater than 10.25	6850	43.29
5	Visit	Initial	15257	96.42
		Follow up	567	3.58

Table 6: Selected socio-demographic attributes, transformed value, number and percent compositions in records of under-five outpatient department in NEMM Hospital from June, 2007- February, 2012.

S. No	Attributes	Value it takes	Composition in number	Composition in %
1	Cough	Yes	7639	48.27
		No	8185	51.73
2	Fever	Yes	4807	30.38
		No	11017	69.62
3	Diarrhea	Yes	3056	19.31
		No	12768	80.69
4	Vomiting	Yes	3147	19.89
		No	12677	80.11
5	Abdominal cramp/ pain	Yes	569	3.60
		No	15255	96.40
6	Trauma/injury/ accident	Yes	86	0.54
		No	15738	99.46
7	Others	Yes	2832	17.90
		No	12992	82.10
8	Unable to breast feed/drink	Yes	1079	6.82
		No	14738	93.18
9	Convulsion history/ convulsing now	Yes	238	1.50
		No	15585	98.50
10	Lethargic or unconscious	Yes	292	1.85
		No	15532	98.15
11	Fast breathing	Yes	7076	44.72
		No	8748	55.28
12	Stridor	Yes	587	3.71
		No	15237	96.29

Table 7: Selected signs and symptoms as attribute, transformed value, number and percent compositions in records of under-five outpatient department in NEMM Hospital from June, 2007- February, 2012.

months, generalized skin rash, deep mouth ulcer, cough/ runny nose/red eyes, tender swelling behind the ear, etc. were recorded in two months to five year integrated registration log book only. Whereas gestational age, nasal flaring, grunting, condition of umbilicus, skin pustules, etc. were recorded in birth to two months integrated registration log book only. Thus, all were omitted because of lack of uniformity in records and the values they took.

Finally, 28 attributes and 11,774 records were selected for experimentation tasks. Table 6-9 describe the selected attributes, transformed values, as well as number and percent composition.

AS indicated in Tables 6-9 among under-five children who visited under-five OPD, nearly over half (51.6%) of children were at the age category of child, 57.29% were male and 60.94% were visiting from Hossana town followed by Lemu woreda (18.47%). More than 95% of admissions to the pediatric ward were initial visits.

Regarding to the presenting complaint of children, 48.27% of children were presented with cough and 44.72% of children had fast breathing. 30.38% of children visited to under-five OPD were presented with fever, 19.89% with vomiting, 19.31with diarrhea and 14.74% had chest indrawing whereas 17.9% of complaints were other unspecified complaints. 96.85% of children who visited under-five OPD had almost appropriate weight for their age.

Pneumonia (25.68%) and severe pneumonia (15.8%) only accounts for 41.48% of under-five OPD visits followed by diarrhea which comprised of 17.43% of cases. 10.62% of under-five OPD visits were due to tonsillitis, 3.14% were due to other respiratory infections such as croup, asthma, bronchitis whereas neonatal sepsis accounted for 2.94% and 2.92% were due to malaria. Severe acute malnutrition comprised of 2.42% of under-five OPD visits, whereas 2.07%, 1.18% and 1.04% of under-five OPD visits were due to helminthiasis, tuberculosis and meningitis respectively. Only 14.78% of under-five OPD visits were due to other illness other than pneumonia, severe pneumonia, diarrhea,

S.No	Attributes	Value it takes	Composition in number	Composition in %
1	Chest indrawing	Yes	2332	14.74
		No	13492	85.26
2	Bulged fontanelle	Yes	29	0.18
		No	15795	99.82
3	Restless/irritable	Yes	157	0.99
		No	15667	99.01
4	Blood in stool	Yes	91	0.58
		No	15733	99.42
6	Sunken eyes	Yes	667	4.22
		No	15157	95.78
7	Skin pinch	Normal	15087	95.34
		Slow	473	2.99
		Very slow	264	1.67
8	Weight for age	Appropriate weight for age	15319	96.85
		Low weight for age	443	2.80
		Very low weight for age	54	0.35
9	Oral trash	Yes	7	0.05
		No	15817	99.95
10	Immunization status	Completed	9211	58.21
		Up to date	5392	34.08
		Not started	1214	7.67
		Defaulted	7	0.04

Table 8: Selected physical examination findings as attribute, transformed value, number and percent compositions in records of under-five outpatient department in NEMM Hospital from June, 2007- February, 2012.

S.No	Attributes	Value it takes	Composition in number	Composition in %
1	HMIS disease classification	Malaria	462	2.92
		Tonsillitis	1679	10.62
		Helminthiasis	327	2.07
		Severepneumonia	2500	15.80
		Pneumonia	4064	25.68
		Skininfection	484	3.07
		Meningitis	151	1.04
		diarrhea	2758	17.43
		SAM	383	2.42
		Othersrespiratoryinfections	496	3.14
		Conjunctivitis	82	0.52
		Genitalproblems	87	0.55
		Infectedwound	108	0.68
		Lymphadenitis	93	0.59
		intestinalobstruction	73	0.46
		AFI	146	0.92
		Tb	172	1.18
2	Outcome of under-five OPD visit	Admitted	5846	36.94
		Notadmitted	9978	63.06

Table 9: Some selected HMIS diagnosis and outcome of OPD visit as attribute, transformed value, number and percent compositions in records of under-five outpatient department in NEMM Hospital from June, 2007- February, 2012.

tonsillitis, croup, bronchitis, asthma, severe acute malnutrition, malaria, meningitis, helminthiasis tuberculosis and neonatal sepsis.

The researchers understood from these points that there were great variation of values among attributes and needs balancing for each attribute. The WEKA tool has Synthetic Minority Over-sampling TEchnique (SMOTE) where it automatically balances the dataset. Therefore, the SMOTE technique was applied up to 300 percent to balance the data before the experimentation was done.

Classification sub Phase: The researchers applied experimentation task to classify pediatric dataset item into one of predefined classes. The classification outputs obtained on different algorithms were compared. Finally, the best selected model was used to predict cause of under-five children admission to pediatric ward.

The outcome of under-five OPD visit as admitted and not admitted values were used as a decision variable. The decision variables were used for predicting cause of admission of under-five children to pediatric ward. The total dataset (28 attributes and 11,774 records) were used to construct the decision tree and artificial neural network (ANN). J48, Random tree and REP tree algorithms for decision tree and multilayer perceptron algorithm for ANN were implemented by using the WEKA tool, which was evaluated by 10 fold cross-validation test option.

Decision Tree Model Building Experiment: Decision tree model

Algorithm	No. of attributes	Tree size	Time elapsed in second	Correctly classified instances	Accuracy in %	WTPR	WFPR	WROC
J48	28	375	0.48	11291	95.90	96.5	3.5	0.998
	26	401	0.45	11262	95.86	96.4	3.7	0.996
	24	386	0.38	11281	95.87	96.3	3.7	0.989
	22	386	0.37	11281	95.87	96.3	3.7	0.989
	20	223	0.34	11230	95.80	96.3	3.7	0.989
	16	153	0.23	11111	**94.77**	**94.7**	**5.3**	**0.99**
Random tree	28	2160	0.08	11274	95.75	96.2	3.9	0.980
	26	2450	0.05	11291	95.90	96.3	3.8	0.982
	24	2475	0.09	11255	95.59	96.2	3.9	0.981
	22	2915	0.11	11280	95.80	96.2	3.8	0.980
	20	1368	0.13	11238	95.45	96.0	4.0	0.981
	16	909	0.05	11156	94.75	94.8	5.3	0.981
REP tree	28	373	0.27	11291	95.90	96.3	3.7	0.990
	26	354	0.42	11286	95.86	96.3	3.7	0.989
	24	352	0.22	11284	95.84	96.2	3.8	0.987
	22	352	0.22	11284	95.84	96.2	3.8	0.988
	20	178	0.36	11253	95.58	96.1	3.9	0.988
	16	**109**	**0.13**	11152	94.72	94.3	5.7	0.983

Table 10: Performances of J48, Random tree, REP tree classification algorithms on decision tree model building to predict under-five children admission to pediatric ward in NEMM Hospital, 2012.

building experiment was implemented, on 28 attributes and 11,774 records of children visited under-five OPD. In order to improve the accuracy and the performance of the algorithm, attributes were selected by using their information gain. The WEKA attributes selection information gain algorithm was applied on 10 fold cross-validation modes. The outputs of information gain result based on their rank were:

1. HMIS disease classification
2. Chest in-drawing,
3. Unable to feed or drink
4. Skin pinch
5. Stridor
6. Sunken eyes
7. Weight for age
8. Fast breathing
9. Lethargic or unconscious
10. Convulsion history/convulsing now
11. Weight
12. Vomiting
13. Cough
14. Diarrhea
15. Restless or irritable
16. Age category
17. Bulged fontanelle
18. Visit type
19. Fever
20. Blood in stool
21. Other unspecified sign and symptom
22. Trauma/injury/accident
23. Abdominal pain/cramp
24. Sex
25. Decision tree experimentation was conducted, where three different kinds of algorithms (J48, Random tree and REP tree) were tested in WEKA. The test was done continuously by using default parameter settings and also varying the parameters of the algorithm until the best predictive model result. First all 28 attributes of children who visited under-five OPD were tested on all algorithms which is followed by ignoring two attributes (sex and abdominal cramp/pain). Experiment three was also done in similar fashion, by changing the parameter settings and various attributes. Generally six experiments were applied by reducing the number of attributes and changing the parameter settings of the algorithms. The details of the experimentation outputs on decision tree model building experiments were summarized on Table 10.

As shown in table 10, for decision tree the models were built for each classifier. When models were compared, in terms of accuracy, time required for building the model, size of the tree, Weighted True Positive Rate (WTPR), Weighted False Positive Rate (WFPR) and Weighted Residual Outcome Curve (WROC), the model built by running J48 algorithm has relatively higher accuracy, WTPR, WFPR and WROC than REP tree and Random tree. REP tree has small number of tree sizes, time required to build the model was also shorter than J48, but the accuracy of the model was relatively lower. When compared to REP tree and J48 the model built by implementing random tree algorithm was faster but lower accuracy, WTPR, WFPR, WROC and also the size of the tree was not manageable. Therefore, J48 performance was better than REP tree and random tree.

Artificial neural network (ANN) model building experiment

Artificial neural network (ANN) was the second classification algorithm applied to build a model on this study. All 11,774 records with 30 attributes of children admitted to pediatric ward were tested in the experiment. The algorithm supported by WEKA for this experiment was Multilayer perceptron. ANN accepts inputs that were in binary

No of attributes	Time in (seconds)	No of (nodes) Hidden layers	instances classified	Accuracy in %	WTPR	WFPR	WROC
28	2567.88	53	11164	94.82	94.8	5.2	0.977
26	3797.87	51	11201	95.13	95.1	4.9	0.975
24	1418.18	50	11191	95.05	95.0	5.0	0.977
22	1342.15	49	11194	95.07	95.1	5.0	0.976
20	950.23	40	11184	94.99	95.0	5.0	0.980
16	858.6	38	11138	94.60	94.6	5.4	0.979
16	186.2	8	11135	94.57	94.6	5.5	0.981
16	188.14	8	11101	94.28	94.3	5.8	0.976

Table 11: Performance of multilayer perceptron classification algorithm at learning rate 0.2 and momentum 0.3 on ANN model building to predict under-five children admission to pediatric ward in NEMM Hospital, 2012.

Algorithms	No. of attributes	Time elapsed in second	Accuracy in %	WTPR	WFPR	WROC
J48 Decision tree	28	0.48	95.90	96.5	3.5	0.998
	26	0.45	95.86	96.4	3.7	0.996
	24	0.38	95.87	96.3	3.7	0.989
	22	0.37	95.87	96.3	3.7	0.989
	20	0.34	95.80	96.3	3.7	0.989
	16	**0.23**	**94.77**	**94.7**	**5.3**	**0.99**
Random tree decision tree	28	0.08	95.75	96.2	3.9	0.980
	26	0.05	95.90	96.3	3.8	0.982
	24	0.09	95.59	96.2	3.9	0.981
	22	0.11	95.80	96.2	3.8	0.980
	20	0.13	95.45	96.0	4.0	0.981
	16	**0.05**	**94.75**	**94.8**	**5.3**	**0.98|1**
REP tree decision tree	28	0.27	95.90	96.3	3.7	0.990
	26	0.42	95.86	96.3	3.7	0.989
	24	0.22	95.84	96.2	3.8	0.987
	22	0.22	95.84	96.2	3.8	0.988
	20	0.36	95.58	96.1	3.9	0.988
	16	**0.13**	**94.72**	**94.3**	**5.7**	**0.983**
Multilayer perceptron ANN	28	2567.88	94.82	94.8	5.2	0.977
	26	3797.87	95.13	95.1	4.9	0.975
	24	1418.18	95.05	95.0	5.0	0.977
	22	1342.15	95.07	95.1	5.0	0.976
	20	950.23	94.99	95.0	5.0	0.980
	16	**858.6**	**94.60**	**94.6**	**5.4**	**0.979**

Table 12: Comparison of performances on J48, Random tree, REP tree and multilayer perceptron algorithms to predict under-five children admission to pediatric ward in NEMM Hospital, 2012.

form. The normalization process was handled by using WEKA tool. All attributes except the target were normalized.

The experimentation was conducted by using the outcomes of under-five OPD visit (admitted and not admitted) as target attribute. In similar to decision tree, the test was done continuously by using default parameter settings and also varying the parameters of the algorithm until the best predictive model result. First all 28 attributes of children who visited under-five OPD were tested on multilayer perceptron algorithm which is followed by ignoring two attributes (sex and abdominal cramp/pain). Experiment three was also done in similar fashion, by changing the parameter settings and various attributes. Generally eight experiments were applied by reducing the number of attributes and changing the parameter settings of the algorithms. The learning rate and number of hidden layers were modified and the result for which the performance is best was finally selected.

The result of multilayer perceptron algorithm for ANN classifier revealed that, when the numbers of attributes were excluded from the test, the time required for building the model and the number of nodes also decreased by assigning learning rate and momentum at constant. The accuracy (the instances of the children that were correctly classified) of the model was greater than 94.5% and WTPR, WFPR and WROC of the classifier were also higher. Meanwhile when 16 attributes were tested by assigning nodes 8 and learning rate and momentum at 0.3 & 0.2, the time elapsed to build the model was much lower than the previous tested value and also slight decrease on accuracy was observed. For the same attributes when the node was assigned at 8 and learning rate and momentum at 0.5 & 0.4, the accuracy still decreased slightly but the time required for building the model increased for about 1.94 seconds. The details of the experimentation outputs on ANN model building experiments were summarized on Table 11.

In general the number of nodes and the time required for building the model was very high. Moreover, the outputs of the algorithms were too difficult to interpret for domain experts.

Comparison of models

As observed in the previous sections, models were built for classification, different attributes and parameters were set in order to get the classifier that has good performance. Selecting the classifier that built the model with high performance has paramount importance in order to apply on actual working environment. The other important thing that was considered during comparison of models was the

simplicity of the model for users and time required for building the models. In addition to that sensitivity and specificity has greater importance in clinical and medical fields than using general accuracy of the classifiers rather models better compared based on WROC area. Therefore, for the comparison of the models primarily the accuracy, WTPR, WFPR, WROC of the classifier and the time required for building the models were taken. The details of comparison on models were summarized on Table 12.

As shown in Table 12, Models were built by applying tests on decision tree and ANN algorithms in order to predict admission of under-five children to pediatric ward. However, which of these models best perform the prediction of admission was the question to be answered by selecting best classifier. Therefore, researchers compared the performance (the predictive ability) of algorithms, in terms of accuracy, WFPR, WFNR, WROC and time required for building the models. Best performances of all algorithms were obtained by experimenting 16 attributes. Based on that, the model built by running J48 algorithm has relatively higher accuracy, WTPR, WFPR, and WROC than REP tree, Random tree and multilayer perceptron algorithms.

REP tree has small number of tree size than J48 and Random tree (Table 10) and also performs predictions very fast; however, it has relatively lower accuracy and WROC. J48 has also very much small number of tree size than Random tree as the result interpretation of the model could be easy for domain experts. Random tree algorithm required shorter time for building the models than all other algorithms but, the size of the tree was not manageable. Multilayer perceptron algorithm required the longest time period for building the models and also has lower accuracy, WTPR, WFPR and WROC than all other algorithms.

Generally, when the performances of all algorithms were compared, J48 performance was better than REP tree and random tree in decision tree classifiers. Similarly, J48 algorithm has relatively higher accuracy, WTPR, WFPR, WROC and performs much faster than multilayer perceptron. . Therefore, the performances of the decision tree classifier were higher than ANN. According to the findings, the researchers selected J48 algorithm for prediction of under-five children admission to pediatric ward.

Decision rules of J48 algorithm and its interpretations

From the outputs of J48, 35 interesting rules were obtained. According to that unable to feed or drink was the first attribute in determining the admission of children to pediatric ward followed by weight of the child. among all rules obtained, five of the first rules generated and their interpretations were discussed in the following section. Details of rules in J48 were stated in Table 13, Figure 4.

According to rule one, if the child not able to take any food or drink orally, the child must be admitted to pediatric ward because, parenteral or Naso-gastric feeding is the only option for the child. It was obviously known that in the working procedure at under-five OPD, to admit children primarily the HMIS admission disease classification has to be settled based on the child's presenting complaint. But for any child who visited under-five OPD with presenting complaint of not taking any food, fluid or breast feeding, without any consideration of HMIS admission disease classification the child should be admitted to pediatric ward for parenteral or naso-gastric tube feeding. Out of

Rule No	'If' J48	'Then' J48	Success Ratio	Percent
1	Child unable to drink or breast feed = Yes	Admitted	1051/18	98.32
2	Child unable to drink or breast feed = No and sunken eye = No and Weight for age = Low weight for age	Admitted	36/3	92.31
3	Child unable to drink or breast feed = No and weight for age = very low weight for age and restlessness or irritability = no	Admitted	354/6	98.33
4	Child unable to drink or breast feed = No and restlessness or irritability =yes	Not admitted	6/0	100
5	Child unable to drink or breast feed = no and weight for age = appropriate weight for age and the health management information system (HMIS) disease classification =			
	Malaria	Admitted	235/6	97.5
	Trauma, accident, poison and fracture	Admitted	130/39	76.9
	Burn	Admitted	135/17	88.8
	Severe pneumonia	Admitted	1964/16	99.2
	Meningitis	Admitted	30/0	100
	Kwashiorkor	Admitted	38/0	100
	Nephritis	Admitted	51/0	100
	Hernia	Admitted	23/9	71.9
	Cellulitis	Admitted	65/17	79.3
	Intestinal obstruction	Admitted	71/1	98.6
	Mastoditis	Admitted	7/0	100
	Rectal prolapse	Admitted	18/2	90
	Measles	Admitted	17/0	100
	Anemia	Admitted	20/0	100
	Pyomyocitis	Admitted	7/0	100
	Angina	Admitted	18/0	100
	Meconium aspiration	Admitted	9/0	100
	Tetanus	Admitted	6/0	100
	Abscess	Admitted	30/0	100
	Hypothermia	Admitted	12/0	100
	Tuberculosis	Admitted	171/1	99.4
	RVI	Admitted	21/0	100

Table 13: Interesting rules in J48 to predict admission of under-five children to pediatric ward, in NEMM Hospital, 2012.

children presented with this complaint 1051 (98.32%) were admitted to pediatric ward. This is an interesting and new knowledge that was hidden in under-five OPD data set and discovered.

Another interesting rule discovered in this study was for the child able to feed or drink, it first looks for the weight for age classification of the child. If it is low weight for age but not true for sunken eye then the outcome of under-five OPD visit was also admission. Therefore, without any consideration of HMIS admission disease classification, again for children who are able to feed or drink and whose weight is under low weight for age classification only by looking absence of sunken eyes the child could be admitted to pediatric ward. Out of children presented with this complaint 36 (92.31%) were admitted to pediatric ward.

Regarding to the third interesting rule obtained in this study, again first look weight for age classification of the child. If the weight for age classification of the child is under very low weight for age but no restlessness or irritability (i.e. the child's weight classification is not related with acute fluid loss from the body) then the outcome of under-five OPD visit was also admission. Therefore, without any consideration of HMIS admission disease classification, simply first weight the child and look for the child's weight for age classification then if it is classified at under very low weight for age but the child is not irritable or restless then the child could be admitted to pediatric ward only for its complaint of very low weight for age by excluding fluid loss. Out of children presented with this complaint (354/360) 98.33% were admitted to pediatric ward

The fourth rule also revealed that for the child presented with the complaint of irritability or restlessness admission was not recommended. The rule also supported that all children complained for restlessness or irritability but able to take food and drink were not admitted to pediatric ward 6 (100%) this complaint might be due to some dehydration as the result of mild fluid loss. It was also usual to treat those children at oral rehydration salt therapy corner than admission.

This study also showed in the fifth rule that, for the child able to drink or breast feed and whose weight for age classification was appropriate weight for age, then the health management information system (HMIS) lists of disease classification were the cause of admission. Among many HMIS disease classifications malaria constituted 235 (97.51%), trauma accident, poison and fracture together 130 (76.92%), burn 135 (88.82%), severe pneumonia 1964 (99.19%), meningitis 30 (100%), kwashiorkor 38 (100%), nephritis 51 (100%), hernia 23 (71.88%), cellulitis 65 (79.27%), intestinal obstruction 72 (98.67%), mastoditis 7, (100%), rectal prolapse 18 (90%), measles 17 (100%), anemia 20 (100%) , pyomyocitis 7(100%), angina 18 (100%), meconium aspiration 9 (100%), tetanus 6 (100%), abscess 30, (100%), hypothermia 12 (100%), tuberculosis 171 (99.45%) and RVI 21 (100%) under-five children admission to pediatric wards.

Generally 35 decision rules were discovered in this study; the researchers discussed and interpreted only five of these rules. Other interesting rules discovered were shown in Table 14.

Among the data mining classification techniques experimented, both decision tree and neural network showed comparative accuracy and performance for outcomes of under-five OPD visit. Models of decision tree and ANN were compared for the outcome of under-five OPD visit, in terms of accuracy, WTPR, WFPR, WROC and time required for building the models. The decision tree algorithm J48 has higher accuracy (94.77%), weighted true positive rate (94.7%), weighted false positive rate (5.3%), weighted receiver operating characteristics curve (0.99) and performs much faster than multilayer perceptron. In addition to that, models built by using neural network, were incomprehensible for a human and the extraction of business knowledge from it was found to be difficult. Hence, the decision tree approaches seems more appropriate to the users. Moreover, the decision tree algorithms have a simple feature which can be easily understandable by users.

According to interesting rules in J48, presenting complaint of not taking any food, fluid or breast feeding (98.32%), low weight for age without sunken eyes (92.31%) and very low weight for age but not in association with restless or irritable (98.33%) were among the cause of under-five children admission to pediatric ward without any consideration of health information management system admission disease classification criteria.

In summary, encouraging results were obtained in classification tasks. Therefore, data mining technique is applicable on pediatric dataset in developing a model that support the discovery of the causes of under-five children admission to pediatric ward.

Rule No	'If' J48	'Then' J48	Success Ratio	Percent
6	Child unable to drink or breast feed = no and sunken eye = no and weight for age = appropriate weight for age and HMIS disease classification = diarrhea and skin pinch = very slow or slow	Admitted	40/1	97.5
7	Child unable to drink or breast feed = no and sunken eye = no and weight for age = appropriate weight for age and HMIS disease classification = severe acute malnutrition	Admitted	38/1	97.4
8	Child unable to drink or breast feed = no and sunken eye = no and weight for age = appropriate weight for age and HMIS disease classification = other respiratory infections and fast breathing = yes and vomiting = yes	Admitted	6/0	100
9	Child unable to drink or breast feed = no and sunken eye = no and weight for age = appropriate weight for age and HMIS disease classification = typhoid fever	Not admitted	73/10	88
10	Child unable to drink or breast feed = no and sunken eye = no and weight for age = appropriate weight for age and HMIS disease classification = infected wound and fever = yes	Admitted	3/1	75
11	Child unable to drink or breast feed = no and sunken eye = no and weight for age = appropriate weight for age and HMIS disease classification = pneumonia and stridor no	Not admitted	2008/7	99.7
12	Child unable to drink or breast feed = no and sunken eye = yes and HMIS disease classification = severe pneumonia	Admitted	16/8	66.7
13	Child unable to drink or breast feed = no and sunken eye = yes and HMIS disease classification = severe acute malnutrition	Admitted	51/0	100
14	Child unable to drink or breast feed = no and sunken eye = no and weight for age = very low weight for age and restlessness = no	Not admitted	354/6	98.3
15	Child unable to drink or breast feed = no and sunken eye = no and weight for age = appropriate weight for age and HMIS disease classification = sepsis and fast breathing = yes	Admitted	18/7	72

Table 14: Interesting rules 6-15 in J48 to predict admission of under-five children to pediatric ward in NEMM Hospital, 2012.

```
=== Run information ===
Scheme:     weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:   16selectedOPD-weka.filters.supervised.attribute.Discretize-Rfirst-last
Instances:  11774
Attributes: 16
Size of the tree: 153
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       11111               94.3689 %
Incorrectly Classified Instances       663                5.6311 %
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure  ROC Area  Class
                 0.922    0.035      0.963      0.922      0.942      0.979    admitted
                 0.965    0.078      0.926      0.965      0.945      0.979    not-admitted
W/Avg.           0.944    0.057      0.944      0.944      0.944      0.979

=== Confusion Matrix ===
    a     b   <-- classified as
 5376   457 |     a = admitted
  206  5735 |     b = not-admitted
```

Figure 4: Summary statistics of J48-C0.25-M2.

Recommendation

This study showed the potential applicability of data mining techniques in pediatric dataset in developing a classification model. Based on the study, the following recommendations are put forwarded for health managers, Ministry of Health (MOH), Non-governmental organizations (NGOs), and other stakeholders:

- Use of integrated dataset with other routes of child admissions such as emergency OPD, weekends and holidays data is also recommended.
- More research and development efforts need to be conducted to enable and explore the variety of data mining techniques that can be applied in pediatric dataset.
- Integration of data mining techniques into existing system and computerizing manual recording systems in database is a priority issue.
- Training is highly recommended for data handlers. Therefore, immediate managers of the organization, MOH, NGO's and other stakeholders must facilitate conditions for the overall improvement of data handling and storing.
- Besides computerizing the data, consulting experts on recording formats and information to be registered is also a crucial issue in improving the quality of health services.
- Implementation of the findings primarily in Nigist Eleni Mohammed Memorial hospital and other similar settings.

The size of the dataset has an impact on data mining research. Especially proportional dataset will enhance the performance of the algorithms. Further researches can be conducted using large dataset.

References

1. Witten IH, Frank E (2005) Data mining practical machine learning Tools and techniques (2nd edn). Elsiever Inc, USA: Morgan Kaufman publisher.
2. Kantardzic M (2003) Data mining: concepts, models, methods and algorithms. IEEE press.
3. Han J, Kamber M (2006) Data mining concepts and techniques (2nd Edn) Morgan Kaufmann publisher, New York, USA.
4. Hand D, Mannila H, Smyth P (2001) Principles of data mining. The MIT press, London, UK.
5. Krzysztof JC, Witold P, Swiniarki RW, Kurgan LA (2007) Data Mining: A Knowledge Discovery Approach. Springer Science Business Media LLC, New York, USA.
6. Nisbet R, Elder J, Miner G (2009) Hand book of statistical analysis and data mining applications. Canada, Elsevier Inc.
7. UNICEF MDG, Reduce child mortality (2011).
8. Federal Ministry of Health, Maternal and child health package Addis Ababa (2003).
9. UNICEF, Statistical report on Ethiopia (2011).
10. United Nations, Millennium Development Goal report (2008) Newyork.
11. Fayyad U, Piatetsky-Shapiro G, Smith P (1996) From Data Mining to Knowledge Discovery in Databases.
12. Taft M, Krishnan R, Hornick M, Muhkin D, Tang G, Thomas S et.al (2005) Oracle data mining concepts.
13. Berry M, Linoff G (2004) Data mining techniques for marketing, sales and customer relationship management. 2nd ed. Indianapolis, Indiana: Wiley publishing, Inc.
14. Two Crows Corporation (2005) Introduction to data mining and Knowledge discovery.
15. Larose DT (2005) Discovering knowledge in data: An introduction to data mining. Canada ,A Jhon Wiley & Sons Inc.
16. Ripley BD, Ripley RM (1997) Neural networks as statistical methods in survival analysis, in Artificial Neural Networks: Prospects for Medicine.Texas: Landes Biosciences.