بسم الله الرحمن الرحيم

Sudan University of Science and Technology

College of Computer Science and Information Technology

Title:

Building Models for the prediction of Leukemia in Children Using Decision Tree and Neural Networks Techniques

بناء نموذج للتنبؤ بسرطان الدم الأبيض لدى الأطفال بإستخدام خوارزميتي شجرة القرار و الشبكات العصبية

NOV 2018

PREPARED BY:

Asma Abbker Ishag

SUPERVISED BY:

Dr.Hwaida Ali Abdalgadir

THIS IS SUBMITTED AS A PARTIAL REQUIREMENTS OF MASTER DEGREE IN COMPUTER SCIENCE


بسم الله
قل هو الله أحد
الله الصمد
لم يلد ولم يولد
ولم يكن له كفوا أحد
سورة الإخلاص

Dedication

I present this research to a hospital National Center for Radiotherapy and Nuclear Medicine Khartoum and Blood Bank (Radiation& Isotopes Center –Khartoum) for their great efforts in treating children with cancer and wishing them full recovery.

I thank all my teachers who accompanied me in all the different stages of education and at their head

Dr. Hwaida Ali Abdalgadir.

Acknowledgements

I thank God for blessing the completion of this research.

I thank my mother and father and my brothers and sisters for encourage me all the times.

and also I would like thanks all my colleagues who stood by my side and accompanied me in all my steps.

List of Acronyms

- NN= Neural Network
- DT= Decision trees
- KDD= knowledge discovery in databases
- CBC =Count blood components
- AML = Acute Myeloid Leukemia
- ALL =Acute Lymphoblastic Myeloid

TABLE OF CONTENTS

List of Figures

List of Tables

Abstract:

The most common type of blood cancer is leukemia and prevalent among children and addition to the lack of medical staff against the increasing number of people with leukemia led to the need to create a model helps in the diagnosis process and facilitate the play by those present in the provision of efficient therapeutic services high and more accurate.

In this study, samples of the results of blood tests of the children, which aims to detect any types of leukemia, have been distributed among children in Sudan. samples were taken for a complete blood analysis from a hospital National Center for Radiotherapy and Nuclear Medicine Khartoum and blood Bank (Radiation & Isotopes Center –Khartoum) the tests included many types of cancers affecting children.

Two types of leukemia were concentrated:

Acute Myeloid Leukemia (AML).

Acute Lymphoblastic Leukemia (ALL).

Ranging from children on one to fifteen years of gender .the results were compared with the results of the annual analytical reports of the hospital's statistics office. the results were quite consistent in determining the increase of acute myeloid leukemia. as this type of leukemia affects cells established in the bone marrow which will later be granular blood cells (Granulocytes) and red blood cells (Erythrocyte).It contains all kinds of cells in the blood.

Data mining techniques using the decision tree and neural network algorithms helped to obtain the highest accuracy of 99.43% algorithms by dividing data to 70% for training dataset and 30% for testing dataset.

المستخلص :

يوجد نوعين اكثر انتشارا من غيرهما من امراض سرطان الدم الذي يصيب الاطفال وهو ما يعرف بابيضاض الدم النقوي او اللوكيميا بالاضافه الي قله الاطباء المتخصصين في هذا المجال وذلك بسب الهجره خارج البلاد مقابل الازدياد المستمر في عدد الاطفال المصابين بهذا النوع من سرطان الدم لذلك كان لابد لنا من القيام بهذه الدراسة العملية التي من خلالها يتم التنبؤ بايهما اكثر انتشارا وخطورة علي حياة الاطفال مستخدمين في ذلك التقنيات الحديثة في التنقيب عن البيانات.

تم في هذا البحث اخذ عينات لنتائج فحوصات الدم للاطفال والتي تهدف الي اكتشاف أي انواع سرطان الدم انتشارا بين الاطفال في السودان.

اخذت العينات لتحليل الدم الكامل من مستشفى الذرة (مركز الطب النووي والعلاج بالاشعة بالخرطوم) شملت الفحوصات العديد من انواع السرطانات التي تصيب الاطفال تم التركيز علي نوعين من امراض سرطان الدم وهما :-

سرطان الدم النخاعي الحاد (Acute Myeloid Leukemia) (AML) .

وسرطان الدم الليمفاوي الحاد (Acute Lymphoblastic Leukemia) (ALL).

تراوحت اعمار الاطفال بين عمر يوم الي خمسة عشر عاما من الجنسين وتمت مقارنة النتائج مع نتائج التقارير التحليلية السنويه لمكتب الاحصاء في المستشفى فكانت النتائج مطابقة تماما في تحديد تزايد سرطان الدم النخاعي الحاد (AML) . حيث ان هذا النوع من السرطان يصيب خلايا المنشأ في نخاع العظام التي ستكون فيما بعد خلايا الدم الحبيبية وهي Granulocytes والحمراء Erythrocyte وتضم كل انواع الخلايا الموجودة في الدم .

لقد ساعدت تقنية التنقيب عن البيانات باستخدام خوارزميات التصنيف شجرة القرار والشبكات العصبية بالحصول على اعلي دقة للخوارزميتين 99.43% بتقسيم البيانات الى 70% للتدريب و30% للاختبار .

Chapter One

Introduction

1-1 Background:

Despite the significant progress in various aspects of medicine during the twentieth century, it still holds cancer sensitive locations in the consciousness of the public people .in terms of raising the feelings of fear and anxiety to many. in spite of the great progress in the treatment of cancers of the blood and lymph tumors in particular, even more than the progress that has happened in other tumors, in spite of this substantial progress tumors of the blood and lymph nodes still cause panic and anxiety to many people and this is due to several reasons including the fact these diseases affect by the largest segment of young people, whether children or youth.
Cancer is a disease that affects cells in the body .there are many types of cancers, which arise from different types of cells.the capacity in which the participation of all types of cancer is that cancer cells are abnormal cells and do not respond to the natural control mechanism in the body . cancer generates large numbers of cancer cells as a result of replication on-controlled or because they live longer period of time than normal cells or perhaps combination of both leukemia is a type of cancers that affects the bone marrow cells (cells that grow and be blood cells) .

1-2 Problem Statement:

Cancer has become the leading cause of death worldwide. the most effective way to reduce cancer deaths is to detect it earlier. though cancer research is generally clinical and biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where data mining techniques have to be applied [1].
In this research preprocess a data set which contain twelve attributes and approximately 3100 cases of cancer recorded in each of the following years, 2013, 2014, 2015 and 2016 and classify them with data mining techniques . also implement accuracy in this classification by finding out the cancer blood which areas record highest cancer rate to this classification append the results of using the method of least .the most common type of blood cancer is leukemia and prevalent among children and addition to the lack of medical staff against the increasing number of people with leukemia led to the need to create a model helps in the diagnosis process and facilitate the play by those present in the provision of efficient therapeutic services high and more accurate .
Cancer is generally regarded as a disease of adults. But there being a higher proportion of childhood cancer (ALL-Acute Lymphoblastic Leukemia and AML-Acute Myeloid Leukemia) in Sudan .childhood cancer has increased over the last

years, but the increase is much larger the most effective way to reduce cancer deaths is to detect it earlier.

1-3 Research importance:

This research may provide valuable information to enhance the childhood cancer blood control program in the ministry of health and how It is prediction and determined by the type of cancer in the blood precise and clear manner. Building models help in the prediction of cancer in children.
Also the difficult diagnosis of blood cancer that faces doctors using mammogram .then most previous studies compare between several classification algorithms to define the differences between the algorithms.

1-4 Research Questions:

I. With Continued migration of medical staff and increased the number of cancer patients this led to the need for a computer systems help doctors who are to work efficiently and high accuracy. and also need computer systems to predict diseases and how to develop solutions to them ?
II. How to predict the types of blood cancers in children using modern techniques?
III. The most previous studies compare between several classification algorithms to define the differences between the algorithms, and define which one is better performance and accuracy to help doctors to diagnosis blood cancer perfectly .this research is using decision tree algorithm ,and Neural Network algorithm to prediction cancer blood.

1-5 Research Objectives:

I. Build models using the technique of data mining techniques classification and prediction using algorithms ((Decision Tree and Neural Network Technique)) for help doctors in the proper predict, and accurate operations for patients with leukemia in children.
II. This research may provide valuable information to enhance the childhood cancer blood control program in ministry of health.
III. It also stresses the point that child cancer should be controlled and the needed steps should be taken immediately.

1-6 Research Methodology:

Machine learning is a fast growing trend in the health care industry and helps medical experts to analyze data and identify trends. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks. the iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. they learn from previous computations to produce reliable, repeatable decisions and results[2].
Classification algorithms of data mining often used in the prediction of medical data analysis. classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. using tow algorithm (Decision tree and Backpropagation(a Neural Network technique)) and using Clementine tool preprocessing and classification methods .

1-7 Scope and Limitations:

The qualitative methodology used in this research does not claim to give a completely implementation for All classification algorithms.blood cancer just leukemia and implementation classification algorithms. It just uses neural network and decision tree algorithms. just focuses on (AML) and (ALL) blood cancer, data set from the National Center for Radiotherapy and Nuclear Medicine Khartoum and Blood Bank (Radiation & Isotopes Center –Khartoum).

1-8 Layout of the Thesis:

In chapter one, background of objective of this study ,a general overview of blood cancer problem, leukemia and related research works that has been done on the classification of leukemia cancer, chapter two the basic theory and concepts of data mining ,data mining techniques are briefly introduced.
The main types of leukemia and classification algorithms applied to datasets.
Chapter three gives a detail description data set dictionary and dataset stages of preprocessing, Clementine [10.1] desktop briefly introduced. chapter four presents the implementation of the classification model and experiment.

Chapter Two

Literature Review

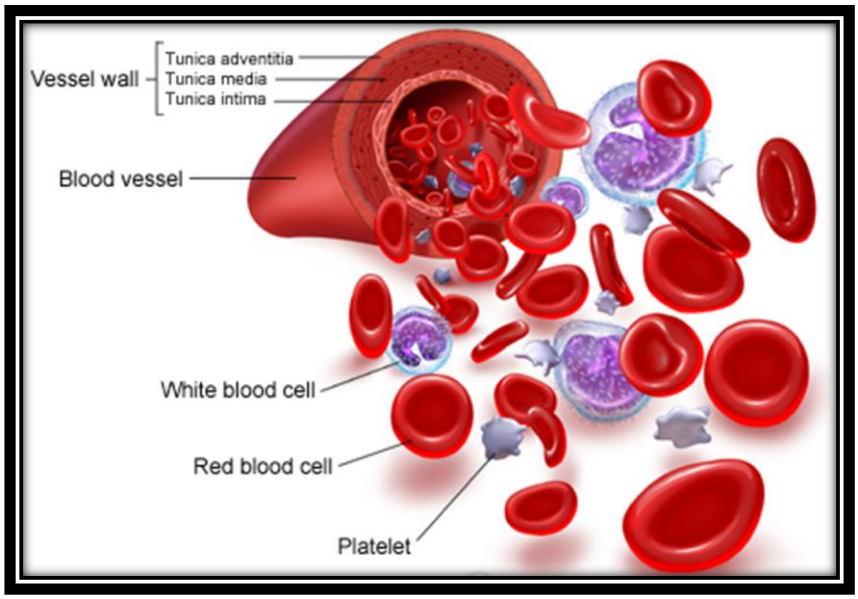
2-1 Introduction:

This chapter explains Leukemia and the technique of data mining using classification algorithms ,medical data classification and related work.

2-2 Definition Leukemia:

Leukemia's, which are cancers of the bone marrow and blood, are the most common childhood cancers. they account for about 30% of all cancers in children. the most common types in children are acute lymphocytic leukemia (ALL) and acute myelogenous leukemia (AML).
These leukemias can cause bone and joint pain, fatigue, weakness, pale skin, bleeding or bruising, fever, weight loss, and other symptoms. Acute leukemia's can grow quickly, so they need to be treated (typically with chemotherapy) as soon as they are found.
Causing widespread disease of cancer cells to the blood there are many types of leukemia and most of these type saris from cells that grown naturally made up of white blood cells. ((The word leukemia comes from the Greek language, which means white blood)).
normal blood consists Blood cells and see which with a microscope .and be 40% of the blood volume see Figure [2-1] Blood Components.
Blood cells (erythrocytes) they you earn that blood its red color one point of blood contains about five million red blood cell red cells contain a chemical called hemoglobin .and which combine oxygen . and oxygen from the lungs to take the rest of the body parts.
white blood cells (Leukocytes) there are different types of white blood cells which are called moderate cells (Neutrophils) are polymorphic and also there lymphocytes(lymphocytes) as well as the acidic cells
(Eosinophils) and single cells (Monocytes) the basic cells (Basophils) and all of them are considered part of the immune system the main role of defending the body against infections.
Platelets and be infinitesimal and help blood to clot if we cut ourselves. plasma it is the liquid part of blood and be 60% of the blood volume the water is the principal component of the plasma ,But they also contain many different types of proteins and other chemicals such as hormones and antibodies and enzymes and glucose and salts and fatty parts etc.
Leukemia is a cancer that starts in blood stem cells. Stem cells are basic cells that develop into different types of cells that have different jobs. Blood stem cells develop into either lymphoid stem cells or myeloid stem cells [3].

Leukemia develops when blood stem cells in the bone marrow change and no longer grow or behave normally. The normal cells are called leukemia cells.
Over the time, leukemia cells crowd out normal blood cells and prevent from doing their jobs [3].



Figure[2-1] Blood Components or Cells Blood

The most common types of childhood cancers that occur most often in children are different from those seen in adults. the most common cancers of children are:

- Leukemia
- Brain and other central nervous system tumors
- Neuro blastoma
- Wilms tumor
- Lymphoma (including both Hodgkin and non-Hodgkin)
- Rhabdo myosarcoma
- Retinoblastoma
- Bone cancer (including Osteosarcoma and Ewing sarcoma)

Other types of cancers are rare in children, but they do happen sometimes.
In very rare cases, children may even develop cancers that are much more common in adults.

2-2-1 The Main Types Of Leukemia:

- (ALL) (Acute Lymphoblastic Leukemia).
- (AML) (Acute Myeloid Leukemia).
- (CLL) (Chronic Lymphocytic Leukemia).

- (CML) (Chronic Myeloid Leukemia).

There are many sub-types falling under these types.

2-2-2 Acute Lymphoblastic Leukemia:

It is possible to appear in all the different age groups and all but 6 in 10 cases appear in children. It is therefore considered the most common types of blood cancers that affect children. (Despite being a non-common disease).

2-2-3 Acute Myeloid Leukemia:

Acute myeloid leukemia develops quickly to a certain extent and very quickly up to a serious degree. if you do not receive proper treatment. the acute myeloid leukemia disease is uncommon.

2-2-4 Chronic Lymphocytic Leukemia:

Grows and develops very slowly development may take months and maybe even years if treatment is not taken .the chronic lymphocytic leukemia, the most common types of blood cancers.

2-2-5 Chronic Myeloid Leukemia:

Grows and develops very slowly development may take months or even years Even if the treatment is not taken .the common chronic myeloid leukemia four of the rarest species in the blood cancers and it appears often in adults and becomes more common with age. some terms needs define to be clear**:**

- "Chronic" means a continuous or long and when talking singled leukemia the word chronic means that the disease grows and spreads slowly (even if not taken his treatment).
- " Acute " means develops quickly to a certain extent and very quickly up to a serious degree .
- "Lymphatic" means that cancer cells are abnormal lymph originated from stem cells.

2-3 Data Mining:

Data mining refers to extracting useful information from vast amounts of data. Many other terms are being used to interpret data mining, such as knowledge mining from databases, knowledge extraction, data analysis, and data archaeology. Nowadays, it

is commonly agreed that data mining is an essential step in the process of knowledge discovery in databases, or KDD. based on a broad view of data mining functionality, data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories[4].

Data mining is an iterative process. the iterative process consists of following steps like Data cleaning, Data Integration, Data Selection, Data transformation, Data Mining, Pattern Evaluation.

Data mining algorithms are divided into two types supervised algorithm and unsupervised algorithm. Supervised algorithm requires

Training and testing dataset and is capable of handle data set only.

Unsupervised algorithm does not require any training or testing data set and are capable of handling unlabeled data [5].

Analysis of cancer datasets is one of the important researches in data mining techniques data mining is the process of analyzing data from different perspectives and summarizing it into important information so as to identify hidden patterns from a large data set ,analysis of cancer datasets is one of the important research in data mining techniques [2].

data mining is one of the provoking and significant area of research. Data mining is implicit and non-trivial task of identifying the viable, novel, inherently efficient and perspicuous patterns of data.

Figure [2-2] represents the data mining as part of KDD [knowledge discovery from data] process.

The hidden relationships and trends are not precisely distinct from reviewing the data. data mining is a multi-level process involves extracting the data by retrieving and assembling them, data mining algorithms, evaluate the results and capture them.

Data mining is also revealed as necessary process where bright methods are used to extract the data patterns by passing through miscellaneous data mining processes[16].

Classification of datasets based on a predefined knowledge of the objects is a data mining .knowledge management technique is used in grouping the same data objects together. the ultimate goal of a supervised learning algorithm is to build a classifier that can be used to classify unlabeled instances a accurately.

Data classification contains supervised learning algorithms as it assigns class labels to data objects based on the relationship between the data sets with a pre-defined class label.

Classification algorithms have a very wide range of applications like fraud detection, churn prediction, artificial intelligence, neural networks and the credit card rating [2].

Data mining on medical data can help in simple classification to highly accurate

Predictions . The advantage over using classification on medical data would be to get over all idea of the data based on various attributes, so that the complexity can be reduced and detection of anomalies becomes easier [6].
Data Mining is used to discover knowledge out of data and presenting it in a form that is easily under-stood to humans. It is a process to examine large amounts of data routinely collected [14].

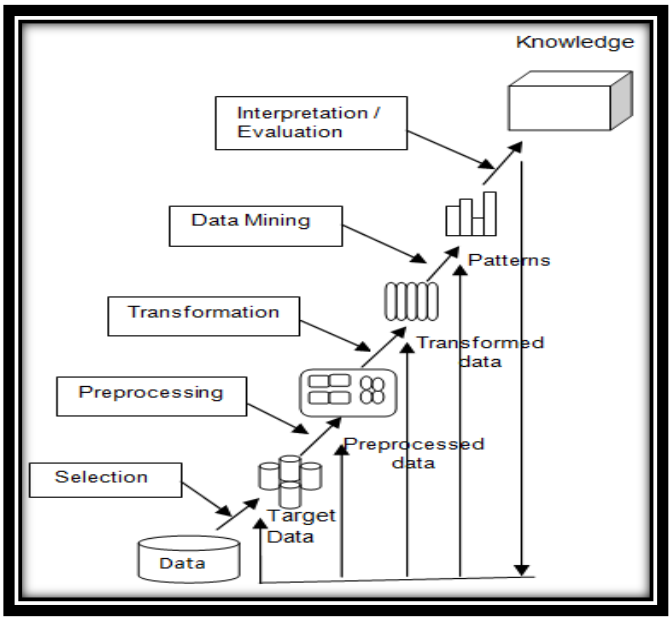


Figure [2-2]: Data Mining Processes

The cause of data mining effort is usually either to design a descriptive model or a predictive model. descriptive mining functions specify the common properties of the data in the database. Predictive mining functions perform the reasoning on the present data in order to form predictions. From a general view, there is a robust consent between both researchers and executives about the standard that all data mining techniques must meet.
Classification and prediction are two designs of data analysis that can be used to extract models illustrating the essential data classification or to predict the future data mode. such analysis can assist to provide with a better knowledge of the data tremendously. knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery.

2-3-1 Knowledge Discovery Process:

1. Data cleaning (to remove noise and inconsistent data).
2. Data integration (where multiple data sources may be combined) .

3. Data selection (where data relevant to the analysis task are retrieved from the database) .
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations) .
5. Data mining (an essential process where intelligent methods are applied to extract data patterns).
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures).
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users).

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining .the data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

2-4 Machine Learning:

There are some who regard data mining as synonymous with machine learning. There is no question that some data mining appropriately uses algorithms from machine learning. Machine-learning practitioners use the data as a training set, to train an algorithm of one of the many types used by machine-learning practitioners, such as Bayes nets, support-vector machines, decision trees, hidden Markov models, and many others.

There are situations where using data in this way makes sense. The typical case where machine learning is a good approach is when we have little idea of what we are looking for in the data. For example, it is rather unclear what it is about movies that makes certain movie-goers like or dislike it. Thus, in answering the "Netflix challenge" to devise an algorithm that predicts the ratings of movies by users, based on a sample of their responses, machine learning algorithms have proved quite successful [17].

2-5 Data Mining and Technologies Used:

Data mining has incorporated many techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, high performance computing, and many application domains. the interdisciplinary nature of data mining research and development contributes significantly to the success of data mining and its extensive applications .

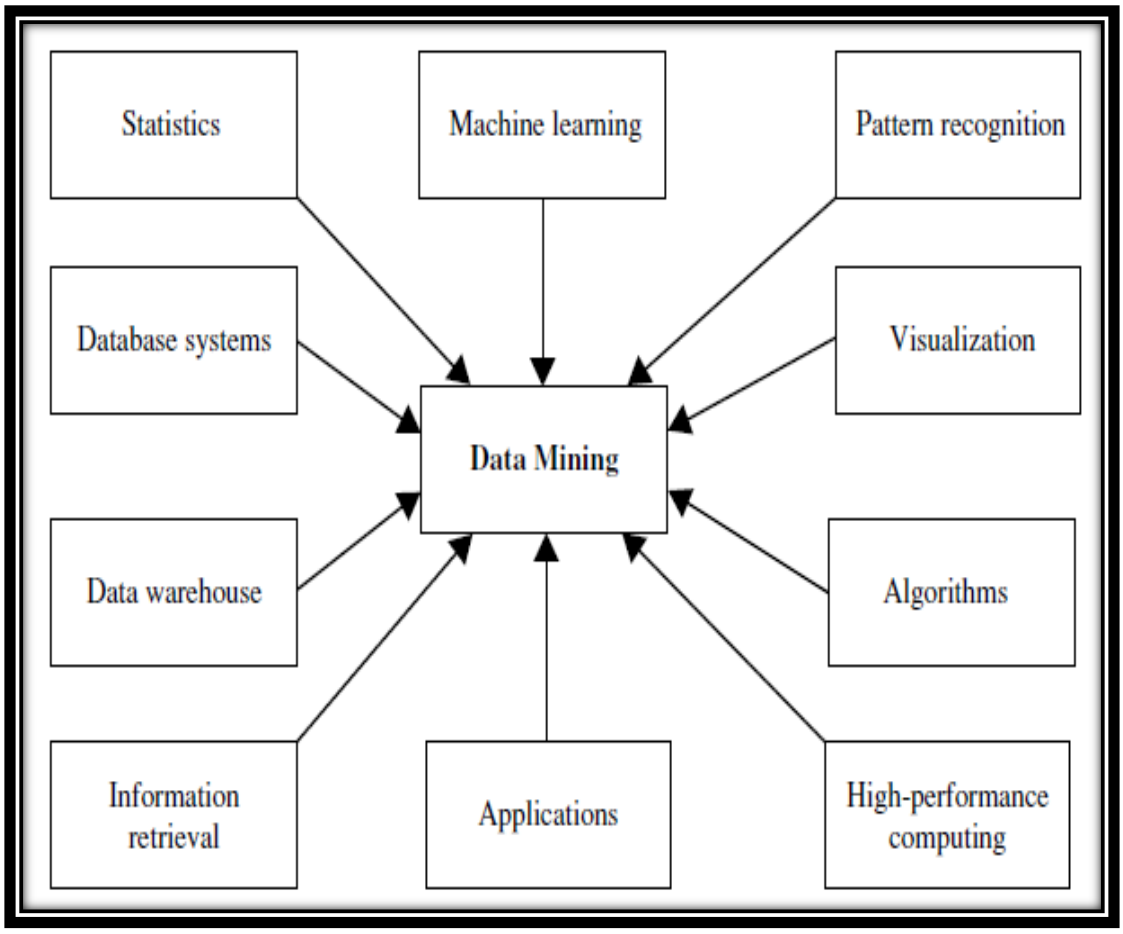


Figure [2-2]: Data Mining adopts techniques from many domains

2-6 Data Preprocessing:

Data have quality if they satisfy the requirements of the intended use. there are many factors comprising data quality, including accuracy, completeness, consistency, time lines ,believability, and interpretability.
the data collection instruments used may be faulty. there may have been human or computer errors occurring at data entry.
Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information.
There may be technology limitations such as limited buffer size for coordinating synchronized data transfer and consumption.
Incorrect data may also result from inconsistencies in naming conventions or data codes, or inconsistent formats for input fields. duplicate tuples also require data cleaning.

Major tasks in data Preprocessing ,the major steps involved in data preprocessing, namely, data cleaning, data integration, data reduction, and data transformation.

2-7 Data Mining Techniques:

Classification algorithms of data mining often used in the prediction of medical data analysis [3].
Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends .used tow algorithm ((Decision tree and Backpropagation (a Neural Network technique)) and comparison between them using clementine tools preprocessing and classification methods.
and compare the higher accuracy and time performance in medical field.
General approach to classification as a two-step process. In the first step, we build a classification model based on previous data. In the second step, we determine if the model's accuracy is acceptable, and if so, we use the model to classify new data.

2-8 Classification:

Classification is the process of using a model to predict unknown values (output variables), using a number of known values (input variables). The classification process is performed on data set D which holds following objects:

- Set size → $A = \{A1, A2, \ldots, A|A|\}$, where $|A|$ denotes the number of attributes or the size of the set A.
- Class label→ C: Target attribute; $C = \{c1, c2, \ldots, c|C|\}$, where $|C|$ is the number of classes and $|C| \geq 2$.

Classification technique can solve several problems in different fields like medicine, industry, business, science. basically it involves finding rules that categorize the data into disjoint groups. there are several classification discovery models and these are: the decision tree, neural networks, genetic algorithms and some statistical models [13].
Predictive modeling approaches predictive data mining is becoming an essential instrument for researchers and clinical practitioners in medicine. understanding the main issues underlying these methods and the application of agreed and uniform procedures is mandatory for their deployment and the broadcasting of results. Predictive models can be used to conjecture explicit values, based on patterns determined from known results.

this technique is becoming an essential appliance for researchers and clinical practitioners in medical . these methods may be applied to the edifice of decision models for procedures such as prognosis, diagnosis and treatment planning, which – once evaluated and verified –may be embedded within clinical information systems. Classification and prediction are major predictive data mining task [17].

2-9 Medical Data Classification:

As medicine plays a great role in human life, automated knowledge extraction from medical data sets has become an immense issue. Research on knowledge extraction from medical data is growing fast [16].
Medical and personal data from government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups.
Governments, companies and organizations such as hospitals, are stockpiling very important quantities of personal data to help them manage human resources, better understand a market, or simply assist clientele.
Regardless of the privacy issues this type of data often reveals, this information is collected, used and even shared. When correlated with other data this information can shed light on customer behavior and the like [18].
Using in the medical field two types of cancer AML and ALL with a machine learning with two algorithms neural network and decision tree, using Clementine tool.

2-10 Neural Networks (Back Propagation):

Classification is one of the most dynamic research and significant area of neural networks. In medical field, the neural network manipulates the predictive decision making by the described set of rules.
Neural network provide powerful mechanism to help the physicians to review, model and make sense of complex clinical data across medical applications. Backpropagation is a neural network learning algorithm.
Neural Network classifier especially, the neural network approach has been widely adopted in recent years. the neural network has several advantages, including its nonparametric nature, arbitrary decision boundary capability, easy adaptation to different types of data and input structures, fuzzy output values, and generalization for use with multiple images. Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions.

(Actual biological neural networks are incomparably more complex) Neural nets may use in classification problems (where the output is a categorical variable) or for regressions (where the output variable is continuous).
The architecture of the neural network consists of three layers such as input layer, hidden layer and output layer.
The nodes in the input layer linked with a number of nodes in the hidden layer. each input node joined to each node in the hidden layer.
The nodes in the hidden layer may connect to nodes in another hidden layer, or to an output layer. the output layer consists of one or more response variables[16].
Steps Performed in Neural Network Classifier:

I. Create feed-forward back propagation network.
II. Train neural network with the training samples and the group defined for it.
III. From the results of network and the samples trained in network classification rate is calculated using some mathematical formulas [16].

Neural Network (NN) is a prevailing AI technique that has the capability to learn a set of data and constructs weight matrixes to represent the learning patterns.
Neural networks are a parallel distributed processing models, are computer based,
Self-adaptive models that were first developed in the 1960s, but they reached great esteem only in the mid-1980s after the development of the back propagation algorithm by Rumel hart et al. [1986][17].
A neural network is a computational representation that takes as input a sequence of numbers, for example, encoded patient features, and outputs another sequence of numbers that is interpreted as, for example, survival probability of that patient . Computational nodes are connected in several layers (input, hidden and output) via weights that are typically adapted during the training phase to achieve high performance.
Implementation using data instead of possibly ill defined rules.
Noise and novel situations are handled automatically via data generalization.
Predictability of future indicator values based on past data and trend recognition.
Automated real-time analysis and diagnosis.
Allows rapid identification and classification of input data.
Reduces error associated with human fatigue and habituation.
Neural networks are very powerful at learning complicated, non-linear patterns in data [17].
They have a tendency of adapting themselves too much to the data, resulting
a phenomenon known as over fitting.

This may give rise to the discovery of spurious patterns. They belong to back-box methods, i.e., the relation between a neural network and the problem it represents is not easy to understand [17].

2-11 Decision Tree:

Decision trees are popular methods for inductive inference. they are robust to noisy data and learn disjunctive expressions. a decision tree is a k-array tree in which each internal node specifies a test on some attributes from input feature set representing data.

Each branch from a node corresponds to possible feature values specified at that node and every test results in branches, representing varied test outcomes.

The decision tree induction basic algorithm is a greedy algorithm constructing decision trees in a top-down recursive divide-and-conquer manner [3].

Decision Trees (DT) are trees that classify instances by sorting them based on feature values. each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume.

Instances are classified starting at the root node and sorted based on their feature values [11].

Classification Decision trees are authoritative classification algorithms that are becoming very admired with the advancement of data mining in the field of information systems.

As it implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy.

A decision tree structure consists of root, internal and leaf nodes.

The tree structure is used in classifying unknown data records.

A number of algorithms have been anticipated for decision tree induction. Following are the decision tree methods commonly used in most practical applications.

Following are the advantages and disadvantages of this method.

In this method, learned trees can be represented as a set of if-then rules that improve human readability.

This offers an easy approach to understand representation of clinical knowledge.

Decision tree methods are simple to understand and interpret by domain experts.

Tree building is very task loaded and computationally intensive as the training data set is traversed repeatedly.

Over-fitting in decision tree algorithm results in misclassification error.

Decision Tree has been applied in medical for many purposes like diagnosis of various chronic diseases, Predicting Risk of Mortality, feature selection to improve Classification accuracy, for reduction of the diagnosis cost [17].

2-11-1 Commercial Version 5.0 (C5.0):

C5.0 is an improved version of C4.5 and ID3 algorithms. It is a commercial product designed by Rule Quest Research Ltd Pty to analyze huge datasets and is implemented in SPSS Clementine workbench data mining software.
C5.0 uses common splitting algorithms include Entropy based information gain. The gain ratio is a robust and consistently gives a better choice of tests than the gain criterion (ID3) for large datasets.
The model works by splitting the sample based on the attribute that provides the maximum information gain [22].
Each subsample defined by the first split is then split again, usually based on a different attribute, and the process repeats until the subsamples cannot be split any further. Finally, the low-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned.
C5.0 model is quite robust in the presence of problems such as missing data and large numbers of input fields [22].
It usually does not require long training times to estimate. In addition, C5.0 models tend to be easier to understand than some other model types since the rules derived from the model have a very straightforward interpretation.
C5.0 also offers the powerful boosting method to increase accuracy of classification.
C5.0 uses information gain as a measure of purity, which is based on the notion of entropy.

2-11-2 Advantages and Disadvantages:

Table [2-1]: Advantages and Disadvantages of logarithm [4]

Methods	Advantages	Disadvantages
Neural network	1. error prone 2. robust in noise environment 3. very efficient	1.high complexity model with long duration of training 2. local minima 3. over fitting
Decision Tree	1.easy to understand 2.easily incorporated into real time system 3.produce a set of rules that are transparent	1.Small variation in data can lead to different decision trees. 2.Does not work very well on a small training set

2-12 Related Work:

Table [2-2]: Summary of relevant scientific papers

NO	Paper Name	Paper publisher & Date	Techniques	The Result	Open Issue
1	Cancer Spread Pattern –an Analysis using Classification and Prediction Techniques	P.Ramachandran 1, Dr.N.Girija2, Dr.T.Bhuvaneswari June 2013	Classification + Prediction	- Results leukemia tends to occupy the first position followed by breast and cervical cancer . -Classification Results women are more prone to cancer than men in almost all age groups. -Regression provides approximate results and not accurate results	-Divide children's age into three categories - Clarify each age group which type of cancer is most prevalent
2	A Study on Cancer Perpetuation Using the Classification Algorithms	ANITA KUMAR Month: April 2015 – September 2015	various classifications of algorithms	-Relative absolute error of LMT is high for cancer survival dataset. - Value of absolute relative error is greater than 50% for almost all the	-If one of the neural network algorithms is used in medical data, it better results . -High accuracy.

				algorithms.	
3	Comparative Analysis of Classification Algorithms for the Prediction of Leukemia Cancer	1Durairaj M, 2Deepika R August 2015	classification	-The results it is identified that naïve bayesclassifier is able to build good prediction model with 91.17% with less time of 0.16 seconds. -As a future dimension of this work, the accuracy of the data mining classification algorithms will be compared with the statistical model in order to propose suitable model for effective cancer prediction	analyzed using two factors such as prediction accuracy and time. Also we can add another factors size of memory .
4	A Survey on Data Mining Techniques in the Medicative Field	Chinky Gera M.Tech(CSE) Student Kirti Joshi Assistant Professor March 2015	classification algorithms Neural network Decision Tree	-Other algorithms can be applied on built-in dataset and the algorithm which gives best result will be applied on the test dataset. -good idea is	This paper also reviews the various techniques along with their pros and cons .

				taking also other algorithms to the experiments and compares their performance in medical field.	
5	Early Detection and Prevention of Cancer using Data Mining Techniques	P.Ramachandran N.Girija, T.Bhuvaneswari July 2014	-classification. -The decision tree model is build using the classification	-The first is the frequent and significant pattern discovery. -The second is mapping the cancer to its cluster and the third is prediction by giving risk score as output.	Detection and Prevention of Cancer is very important things in medical field
6	Healthcare Service Sector: Classifying and Finding Cancer spread pattern in Southern India Using Data Mining Techniques	1P.Ramachandran, 2Dr.N.Girija, 3Dr.T.Bhuvaneswari May 2012	classification	-find pattern using classified data. -higher technology and greater potential data mining can be helping had only if it can help not only finding newer pattern but using predictive methodologies to predict the future work.	In this paper used only one algorithm if using different algorithms the accuracy and result become better.

7	Childhood Cancer-a Hospital based study using Decision Tree Techniques	1K. Kalaivani and 2R. Shanmug alakshmi 2011	Decision Trees	-Female survivors showed greater functional disability in comparison to male survivors-demonstrated by poorer overall health status. -Family stress results from a perceived imbalance between the demands on the family and the resources available to meet such demands	In this paper they used data mining Techniques Decision Trees. If use anther tools like clementine or WEKA outcomes or result become perfect. .
8	Knowledge discovery from database Using an integration of clustering and classificatio n	Varun Kumar Nisha Rathee March 2011	Classification	-integration of clustering and classification technique gives a promising result with utmost accuracy rate and robustness among the classification and clustering algorithms -An experiment measuring the	I acceptant that integration of clustering and classification technique gives more accurate results .

				accuracy of binary classifier based on true positives, false negatives, and true negatives decision trees and decision tree. rules	
9	USING DATA MINING TECHNIQUES FOR DIAGNOSIS AND PROGNOSIS OF CANCER DISEASE	Shweta Kharya April 2012	Classification, Neural Network	Decision tree is found to be best predictor with 93.62% Accuracy on benchmark dataset	If implement Tools like WEKA or Clementine The Accuracy Become better
10	PREDICTION OF ACUTE MYELOID LEUKEMIA CANCER USING DATAMINING-A SURVEY	Durairaj M Assistant Professor Deepika R Research Scholar February 2015	Classification	That the correct accuracy of the classification algorithms are not stable and they differ from one another.	-If classification algorithms are not stable and they differ from one another. That means must use anther techniques and

					tools of data mining.
11	Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification	- Gopala Krishna Murthy Nookala - Bharath Kumar Pottumuthu - NagarajuOrsu -Suresh B. Mudunuri 2013	Classification	the results indicate that the performance of a classifier depends on the data set, especially on the number of attributes used in the data set and one should not rely completely on a particular algorithm for their study	-I accept that the results indicate that the performance of aclassifier depends on the data set. Andthe number of attributes used in the data set.

2-13 Discussion of Papers:

In this literature Survey we present an overview of the current research being carried out using the data mining techniques for the diagnosis and prognosis of various diseases. The following algorithms have been identified: Decision Trees, Artificial neural networks and Naïve Bayes.

All of the previous papers dealt with the treatment of medical data using the science of data mining and classification of data used different algorithms in comparison to high accuracy and performance best in terms of time.

Also addressed many cancers using algorithms including the decision tree algorithm and the network algorithm.

There are no available scientific papers specialized in leukemia in children. All available scientific papers include cancers of various kinds and how to benefit from medical data in the science of data exploration using different algorithms to know the speed of performance for the supplies and time of implementation.

Chapter Three

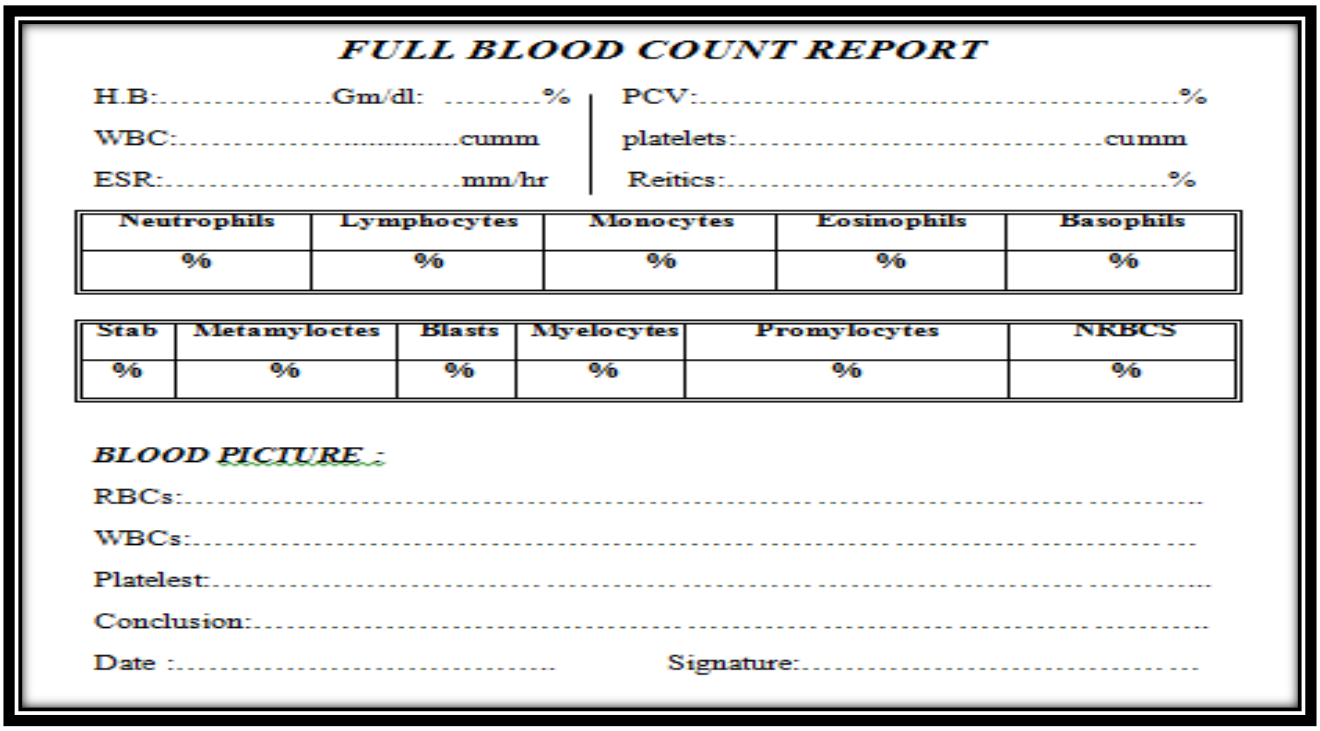
Methods and Techniques

3-1 Introduction:

This chapter content dataset description, dataset preprocessing, workflow diagram represents blood cancer and clementine10.1 desktop tools.

3-2 Data Set Description:

These data were collected from the National Center for Radiotherapy and Nuclear Medicine Khartoum and Blood Bank (Radiation & Isotopes Center –Khartoum).
the methodology for blood cancer classification is beginning with data collection. and cleaning the data from missing values. finally two classification algorithms applied to datasets accuracy.
in this research preprocess a data set which contain eleven attributes and approximately three a thousand and one hundred cases of cancer recorded in each of the following years [2013, 2014, 2015 and 2016] and classify them with data mining techniques it is a private data to children with cancer from day to fifteen years old. it includes a public examination of blood (CBC) the components in the blood .which are being tested to see if this child has cancer or not . these components are explained in detail as follows.



FULL BLOOD COUNT REPORT

H.B:.....Gm/dl: .....% | PCV:.....%

WBC:.....cumm | platelets:.....cumm

ESR:.....mm/hr | Reitics:.....%

Neutrophils	Lymphocytes	Monocytes	Eosinophils	Basophils
%	%	%	%	%

Stab	Metamyloctes	Blasts	Myelocytes	Promylocytes	NRBCS
%	%	%	%	%	%

BLOOD PICTURE :

RBCs:.....

WBCs:.....

Platelest:.....

Conclusion:.....

Date :..... Signature:.....

Figure [3-1]: Full Blood Count Report

Figure [3-1] represents the report that records the biocytes of the blood cells and the marrow cells HB the amount of the hemoglobin in the blood,Gm/dl the percentage of hemoglobin in the blood ,WBC the amount of white blood cells , PCV: Packed Cell Volume test is used to measure the amount of cells in the blood , Platelets the amount of blood platelets help the blood clotting process Reitics determine the number and/or percentage of reticulocytes in the blood .
are newly produced, relatively immature red blood cells , R : ESR Erythrocyte Sedimentation Rate is the rate at which red blood cells sediment in a period of one hour N : Neutrophils Is the rate at which red blood cells sediment in a period of one hour. L **:** Lymphocytes is one of the subtypes of white blood cell in a vertebrate's immune system M: Monocytes , Monocytes are a type of white blood cell, E: Eosinophils are a variety of white blood cell and one of the immune system components responsible for combating multi cellular parasites and certain infections in vertebrates.
B: Basophils Is a type of white blood cell, META: Metamyloctes Are a type of white blood cell, or leukocyte. They are the largest type of leukocyte and can differentiate into macrophages or dendritic cells, MYLO : Myelocytes Is a young cell of the granulocytic series, occurring normally in bone marrow BRO: Promylocytes is a granulocyte precursor, developing from the myeloblast and developing into the myelocyte NRBCS : Enucleated Red Blood Cells Enucleated red blood cell, or NRBC, is a red blood cell (RBC) that retains nucleus .Gender is type baby is masculine or feminine .

3-3 Dataset Preprocessing:

As mentioned earlier these data were collected from the National Center for Radiotherapy and Nuclear Medicine Khartoum and Blood Bank (Radiation & Isotopes Center –Khartoum) .
It is a private data to children with cancer from day old to fifteen years. it includes a public examination of blood (CBC) the components in the blood .which are being tested to see if this child has cancer or not. other tests, such as liver examination, kidney examination and bone marrow examination, are also available are also shown to confirm that the child has leukemia.
The bone marrow is examined every six months. what was mentioned earlier was to clarify the initial tests of knowledge of leukemia in children. in the being was converted data from manually to computerized database using an application Excel sheet file .the data entry from files Previous years starting from the year 2013 to 2016 .the number of records that have been entered three thousand and one hundred record[3100] for the pathological cases .

The process of converting data from a paper-based situation to a computational situation took about three months. This period is long but due to the following:
Firstly the laboratory does not download children's data separately.
Second, the children's section is allocated to them only on Tuesdays of the week in the lab.
Thirdly, the children are allocated to them in the field that is called the cost character **[F**], meaning free.
The number of attributes with fifteen, records three thousand and one hundred, they represent the different types of cancers affecting children.

3-3-1First Stage:

The data entry in the whole blood diseases as well as various other types of cancers for four years ago 2013 , 2014 , 2015 and 2016 for different ages .this process took three months to convert it from notebooks to Excel sheet file .we encountered several problems this is due to the following reasons:

- Not categorized in detail because there are no specialized computer system and the data is still recorded manually on the notebooks.
- Lack of interest in downloading the data full by age .the age of child determines in the growth stage that the cells are component of the body.
- Cancer cells are effective in every stage of cell formation**.**
- Few workers in the laboratory with increase in number of infected in addition to being the only center for the treatment of leukemia children in Khartoum .as well as neighboring countries such as Ethiopia, Somalia and other neighboring countries therefore, there are large numbers of patients in the hospital.
- Lack of a data reference archive in modern way.
- The lack of development of government hospitals in the use of advanced technology in improving the working environment and improve the efficiency of performance.
- Economic conditions adversely affect the provision of full health care from fully equipped hospitals and treatments available permanently.
- The lack of effective management in government hospitals also has a negative impact on the process of progress and development within hospitals.

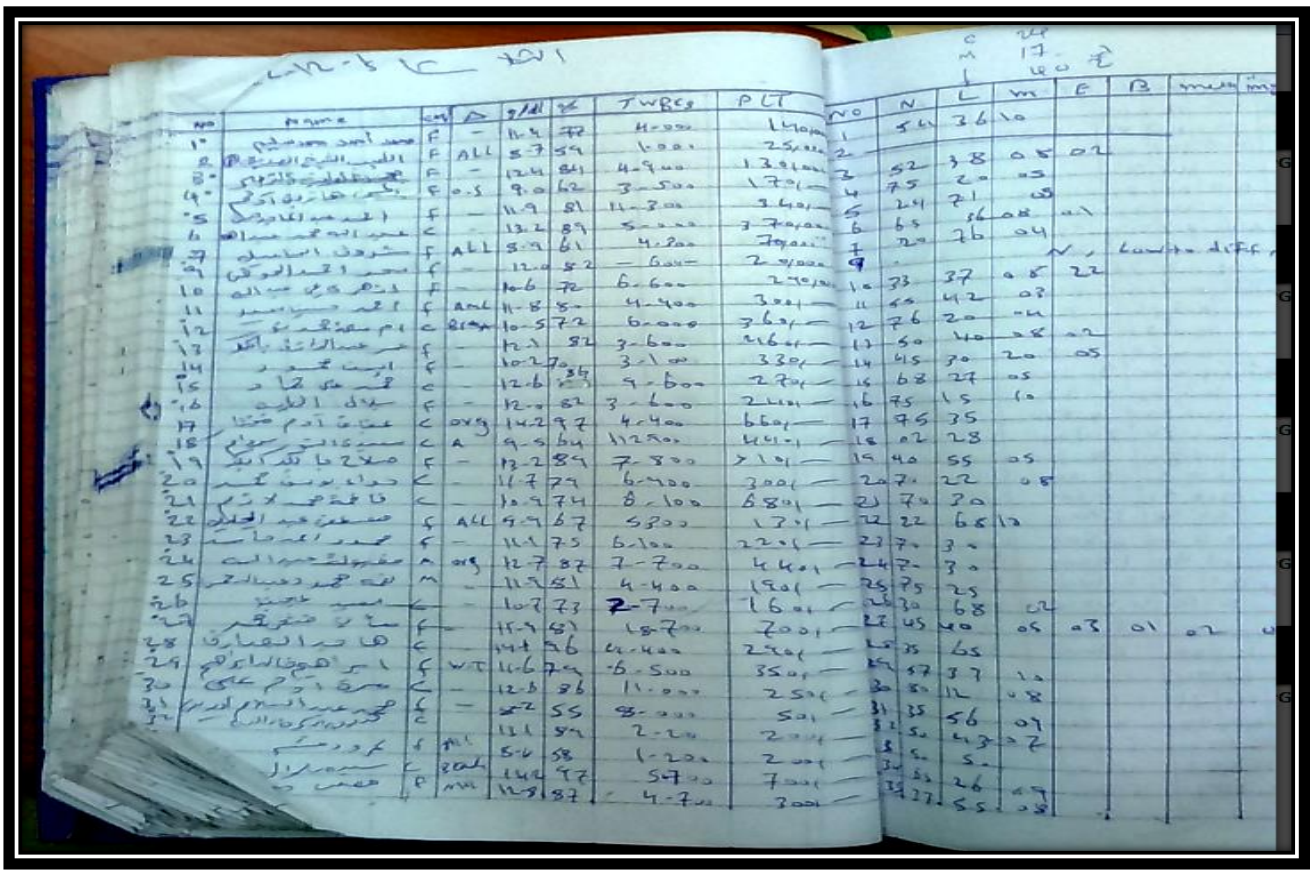


Figure [3-2]: Data Registry

Figure [3-2]: Data Registry represents the state of the data or the way in which the data is downloaded into the registry located in the laboratory. it contents many attributes gender , Diagnoses ,WBC Wight blood cell , PLT platelets , Neutrophils , Lymphocytes , Moncytes , Eosinphils ,Metamylocytes ,Blasts,Myelocytes , promyloctes, NRBCS all this blood contents.

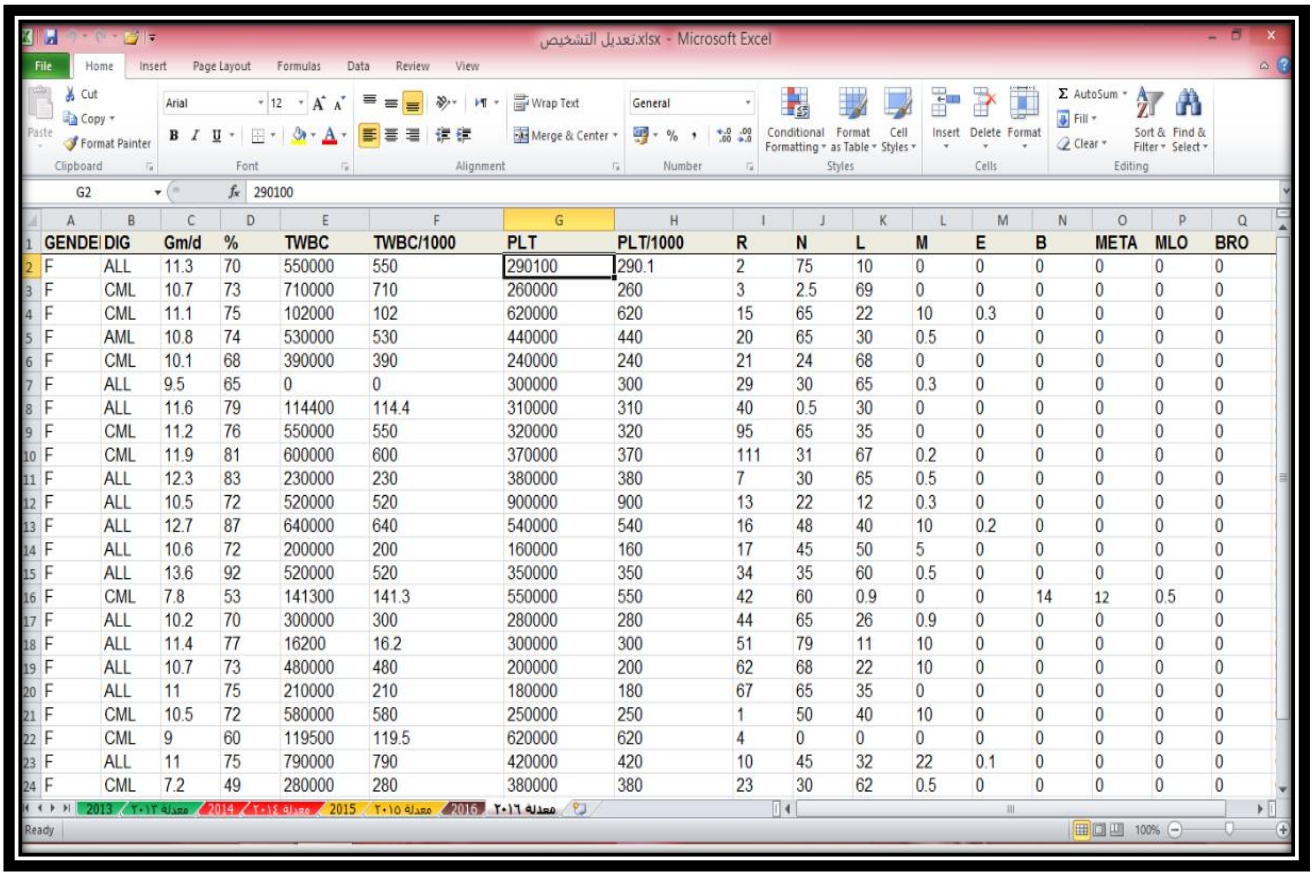
3-3-2 Second Stage:

After entering the data are aggregated the data were separated cancers of the blood and other type of cancers that affect organically for four years ago 2013 , 2014 , 2015 and of 2016 for different ages .

There are many types of leukemia threatens the lives of children .focused on leukemia because it is the field of study that we have done .and use it in the process of building the model that clearer and more accuracy.

- This phase was undertaken because the data on children were not available separately.

- The focus was on only two types of cancers that affect the blood (AML: Acute Myeloid Leukemia and ALL: Acute lymphoblastic Leukemia).
- The study included kinds of children under the age of fifteen years.
- The data is incomplete due to several reasons: the large number of patients in the hospital, few doctors in the laboratory, few devices used in the screening process.
- Children are screened for one day of the week.
- The only government hospital specializing in pediatric cancer.
- All the reasons in the first stage and the second stage led to an increase in data collection time.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	GENDE	DIG	Gm/d	%	TWBC	TWBC/1000	PLT	PLT/1000	R	N	L	M	E	B	META	MLO	BRO
2	F	ALL	11.3	70	550000	550	290100	290.1	2	75	10	0	0	0	0	0	0
3	F	CML	10.7	73	710000	710	260000	260	3	2.5	69	0	0	0	0	0	0
4	F	CML	11.1	75	102000	102	620000	620	15	65	22	10	0.3	0	0	0	0
5	F	AML	10.8	74	530000	530	440000	440	20	65	30	0.5	0	0	0	0	0
6	F	CML	10.1	68	390000	390	240000	240	21	24	68	0	0	0	0	0	0
7	F	ALL	9.5	65	0	0	300000	300	29	30	65	0.3	0	0	0	0	0
8	F	ALL	11.6	79	114400	114.4	310000	310	40	0.5	30	0	0	0	0	0	0
9	F	CML	11.2	76	550000	550	320000	320	95	65	35	0	0	0	0	0	0
10	F	CML	11.9	81	600000	600	370000	370	111	31	67	0.2	0	0	0	0	0
11	F	ALL	12.3	83	230000	230	380000	380	7	30	65	0.5	0	0	0	0	0
12	F	ALL	10.5	72	520000	520	900000	900	13	22	12	0.3	0	0	0	0	0
13	F	ALL	12.7	87	640000	640	540000	540	16	48	40	10	0.2	0	0	0	0
14	F	ALL	10.6	72	200000	200	160000	160	17	45	50	5	0	0	0	0	0
15	F	ALL	13.6	92	520000	520	350000	350	34	35	60	0.5	0	0	0	0	0
16	F	CML	7.8	53	141300	141.3	550000	550	42	60	0.9	0	0	14	12	0.5	0
17	F	ALL	10.2	70	300000	300	280000	280	44	65	26	0.9	0	0	0	0	0
18	F	ALL	11.4	77	16200	16.2	300000	300	51	79	11	10	0	0	0	0	0
19	F	ALL	10.7	73	480000	480	200000	200	62	68	22	10	0	0	0	0	0
20	F	ALL	11	75	210000	210	180000	180	67	65	35	0	0	0	0	0	0
21	F	CML	10.5	72	580000	580	250000	250	1	50	40	10	0	0	0	0	0
22	F	CML	9	60	119500	119.5	620000	620	4	0	0	0	0	0	0	0	0
23	F	ALL	11	75	790000	790	420000	420	10	45	32	22	0.1	0	0	0	0
24	F	CML	7.2	49	280000	280	380000	380	23	30	62	0.5	0	0	0	0	0

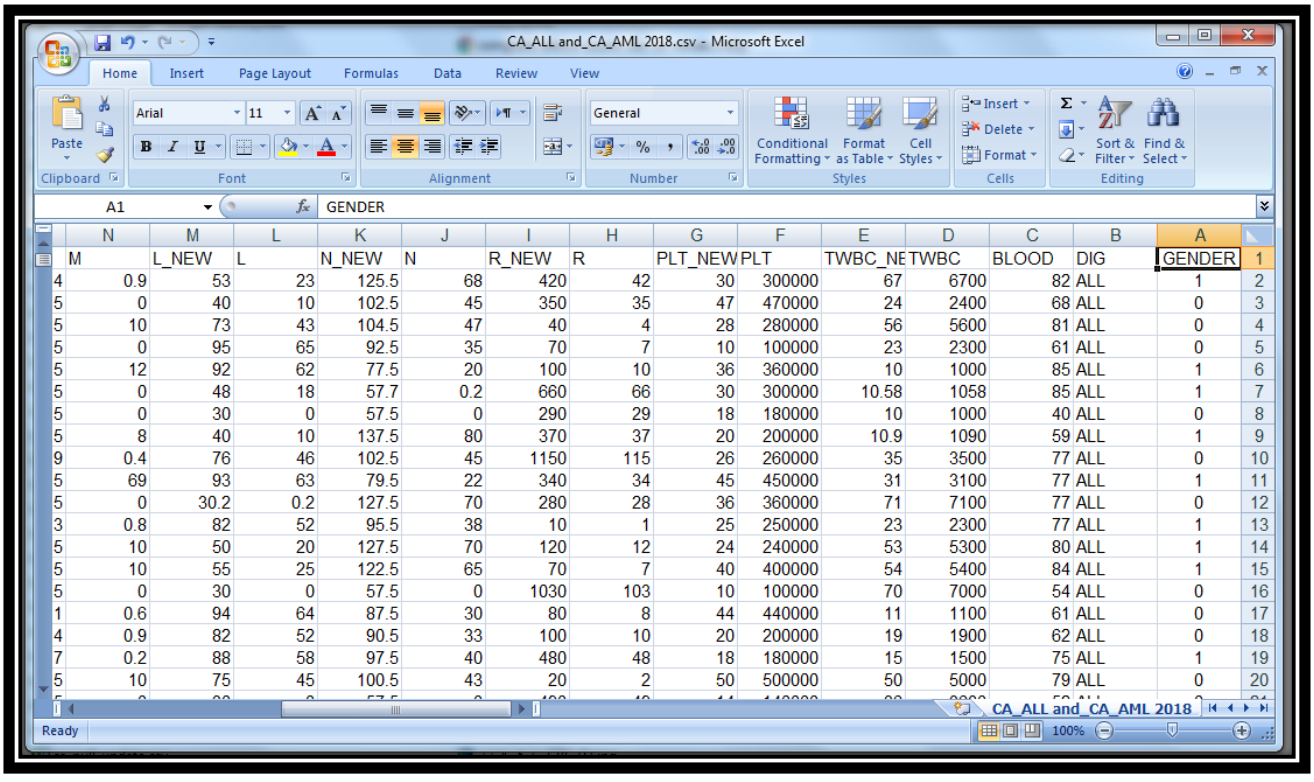
Figure [3-3]: Data In Excel Sheet

Figure [3-3] represent all data of cancer used excel sheet they are many type of cancer that affective in blood. this column also blood contents all cells. in excel sheet . attributes selected for the experiments that were the most important.

3-3-3 Third Stage:

It has been compiled Hematology for the three previous years. It was chosen over the spread of blood diseases among children . the most types spread in the Sudan is, Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML).

- This is the final stage in data collection, with a record number of children 580 record and field is 12.data cleaning, and fill incomplete, noisy and inconsistent prepared it.
- data cleaning routines attempt to fill on missing values, smooth out noise while identifying outliers and correct inconsistencies in the data.
- Data has been prepared and processed in various statistical methods until it is entered into a Clementine tools .these tools can include statistical models, mathematical algorithm.
- After this stage the data has been processed to be ready to apply the algorithms.



N	M	L_NEW	L	N_NEW	N	R_NEW	R	PLT_NEW	PLT	TWBC_NE	TWBC	BLOOD	DIG	GENDER	
4	0.9	53	23	125.5	68	420	42	30	300000	67	6700	82	ALL	1	2
5	0	40	10	102.5	45	350	35	47	470000	24	2400	68	ALL	0	3
5	10	73	43	104.5	47	40	4	28	280000	56	5600	81	ALL	0	4
5	0	95	65	92.5	35	70	7	10	100000	23	2300	61	ALL	0	5
5	12	92	62	77.5	20	100	10	36	360000	10	1000	85	ALL	1	6
5	0	48	18	57.7	0.2	660	66	30	300000	10.58	1058	85	ALL	1	7
5	0	30	0	57.5	0	290	29	18	180000	10	1000	40	ALL	0	8
5	8	40	10	137.5	80	370	37	20	200000	10.9	1090	59	ALL	1	9
9	0.4	76	46	102.5	45	1150	115	26	260000	35	3500	77	ALL	0	10
5	69	93	63	79.5	22	340	34	45	450000	31	3100	77	ALL	1	11
5	0	30.2	0.2	127.5	70	280	28	36	360000	71	7100	77	ALL	0	12
3	0.8	82	52	95.5	38	10	1	25	250000	23	2300	77	ALL	1	13
5	10	50	20	127.5	70	120	12	24	240000	53	5300	80	ALL	1	14
5	10	55	25	122.5	65	70	7	40	400000	54	5400	84	ALL	1	15
5	0	30	0	57.5	0	1030	103	10	100000	70	7000	54	ALL	0	16
1	0.6	94	64	87.5	30	80	8	44	440000	11	1100	61	ALL	0	17
4	0.9	82	52	90.5	33	100	10	20	200000	19	1900	62	ALL	0	18
7	0.2	88	58	97.5	40	480	48	18	180000	15	1500	75	ALL	1	19
5	10	75	45	100.5	43	20	2	50	500000	50	5000	79	ALL	0	20

Figure [3-4]: ALL cancer Excel Sheet

Figure [3-4] represent one type of cancer that [ALL] acute lymphoblastic leukemia and all cells contents blood . convert from manual register to excel sheet.

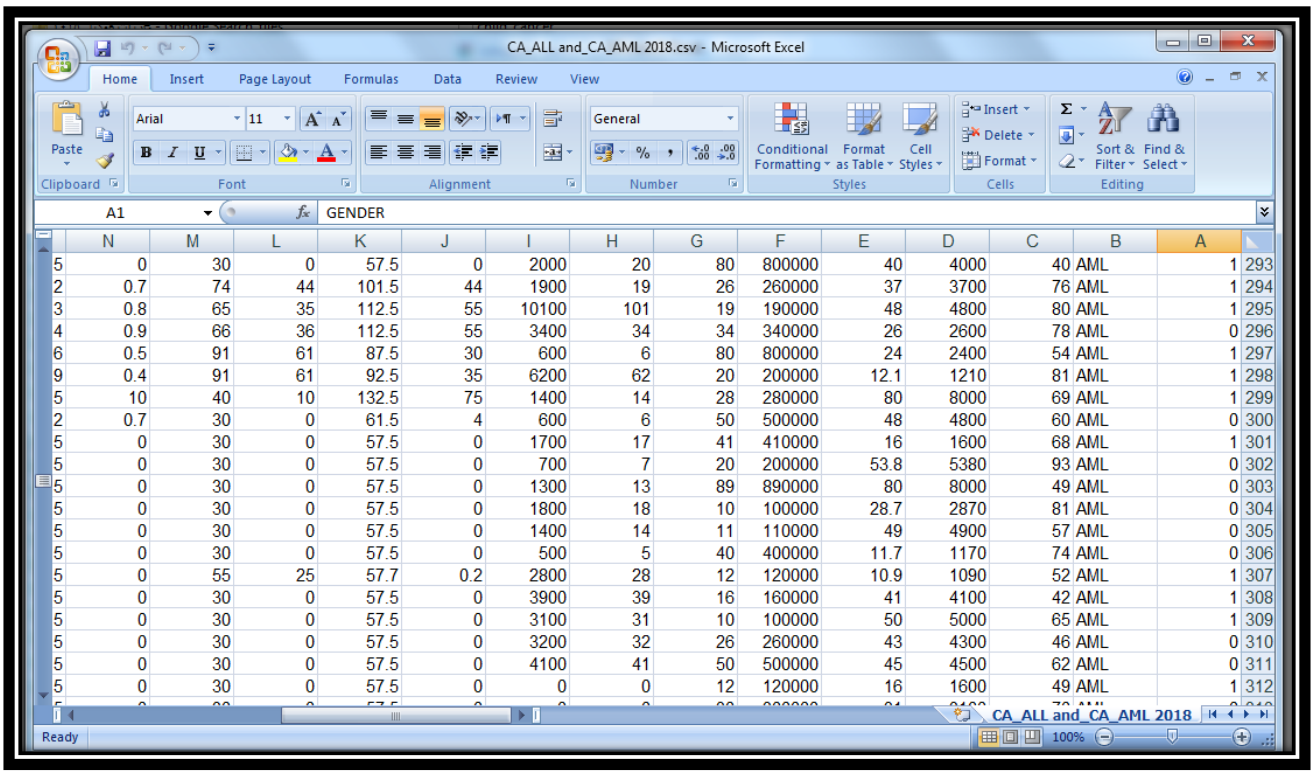


Figure [3-5]: AML Cancer Excel Sheet

Figure [3-5] represent other type of blood cancer AML acute myeloid leukemia and all cells contents blood. convert from manual register to excel sheet .

3-4 Statistical Analysis:

the data were analyzed statistically to find .At this stage of the preparation of the data some missing values of data. the data that were used in this research type numeric and character . It found an average of some of the fields that the missing values. and some of the big values Fields
It has been reduced values dividing by a thousand. And others have been converted to decimal values .We used the following statistical equations :
Sum: Total of numbers in field.
Account: The number of digits in field.
Mean: Total of numbers in field divided in to the number of digits in field.
Equations Arithmetic mean:
This equation was applied to the field of blood platelets and white blood cells so that you can read well data and finding a high accuracy of the algorithms that used in the analysis process .some attributes are of a class nature with the upper limit and the minimum. Therefore we used the central distributions to find the center of the class used equations of Mid Point.

As a data mining application, Clementine offers a strategic approach to find useful relationships in large data sets. In contrast to more traditional statistical methods, you do not necessarily need to know what you are looking for when you start. You can explore your data, fitting different models and investigating different relationships, until you find useful information .Working in Clementine is working with data. in its simplest form, working with Clementine is a three-step process. First, you read data into Clementine, then run the data through a series of manipulations and finally send the data to a destination.

This sequence of operations is known as a data stream because the data flows record by record from the source through each manipulation and, finally, to the destination- either a model or type of data output.

Most of your work in Clementine will involve creating and modifying data streams.

At each point in the data mining process, Clementine's Visual interface invites your specific business expertise.

Modeling algorithms, such as prediction, classification, segmentation and association detection, ensures powerful and accurate models .model results can easily be deployed and read into databases, SPSS and a wide variety of other applications.

You can also use the add-on component, Clementine Solution Publisher, to deploy entire data streams that read data into a model and deploy result swith out a full version of Clementine.

This brings important data closer to decision makers who need it.the numerous features of clementine's data mining workbench are integrated by a visual programming interface. You can use this interface to draw diagrams of data operations relevant to your business.

Each operation is represented by an icon or node and the nodes are linked together in a streamer presenting the flow of data through each operation [7].

In some fields, the category center for the value in the field was added .

(Eosinophils , Basophils ,Metamyloctes , Myelocytes , Promylocytes).

Blood components are values with a range of between two peaks with low values, high values and a mean value as follows:

RBC 6…….4.7milion/mm3 male and 5.4……4.2 female.

Hemoglobin 18……13.5mg% male and 16……12.5 female mg%.

Haematocrit 52……..42 % male and 47……37 female mg%.

WBC 10.500 ……..4000 mm3

Differential (count) :

Neutrophiles 70……..…..45%.

Lymphocytes 40……..…..20% .

Monocytes 6…….….1%.

Eosinophiles 4……..…..1%.

Basophiles 1.....…….……0%.
Platelets 450.000……150.000/mm3
Reticulocyte 2……. 0.2%
ESR 13……..0.0 mm% male and 20……..0.0mm female.
Bleeding Time 5……..1 min.
Coagulation Time 10 …… 5 min.
Prothrombin Time 100…….80%.
INR 1……...1.2.

3-5 Work flow Diagram Represents Blood Cancer:

Figure [3-1] workflow diagram represents blood cancer cell steps as follow .from data set collected, data cleaning, feature selection and classification logarithm used.

3-5-1 Feature selection:

Data mining problems may involve hundreds, or even thousands, of fields that potentially may be used as predictors. As a result, a great deal of time and effort may be spent examining which fields or variables to include in the model. To narrow down the choices, the Feature Selection algorithm can be used to identify the fields that are most important for a given analysis.

The Feature Selection node screens predictor fields for removal based on a set of criteria (such as the percentage of missing values) then it ranks the importance of remaining predictors relative to a specified target.

Feature selection consists of three steps:

- Screening. Removes unimportant and problematic predictors and records or cases, such as predictors with too many missing values or predictors with too much or too little variation to be useful.
- Ranking. Sorts remaining predictors and assigns ranks based on importance.
- Selecting. Identifies the subset of features to use in subsequent models for example, by preserving only the most important predictors and filtering or excluding all others [24].

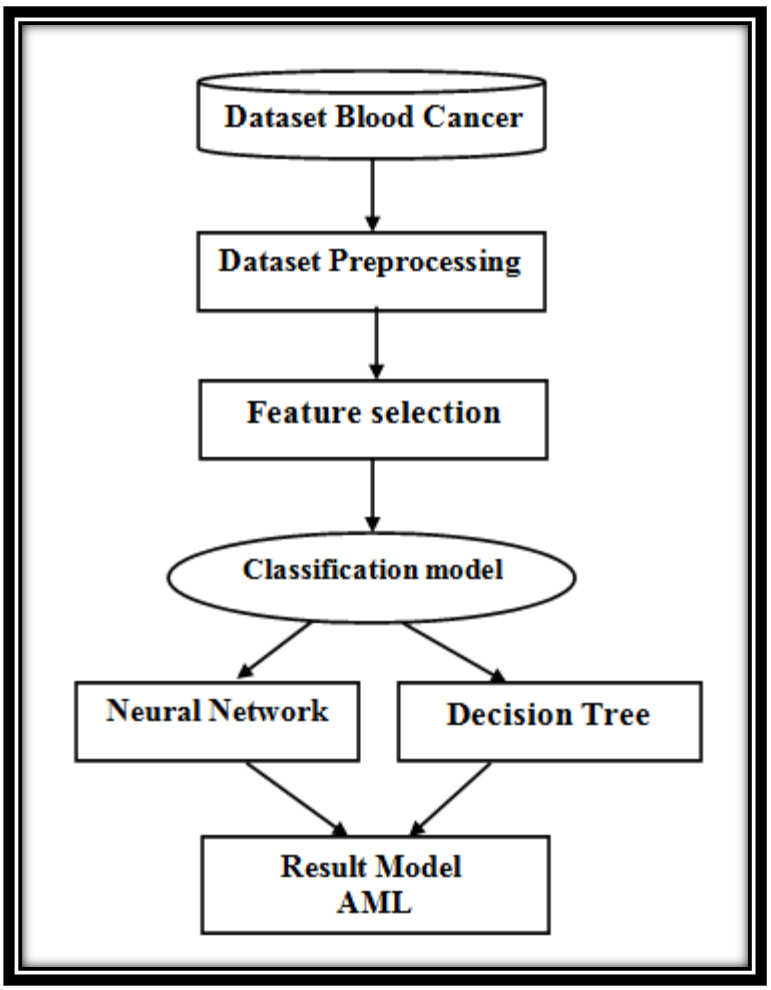


Figure [3-6]: Workflow Diagram for Blood Cancer

3-6 Clementine10.1 Desktop Tools:

Clementine is the SPSS enterprise-strength data mining workbench. Clementine helps organizations to improve customer and citizen relationships through an in-depth understanding of data. organizations use the insight gained from Clementine to retain profitable customers, identify cross-selling opportunities, attract new customers, detect fraud, reduce risk, and improve government service delivery. Clementine's visual interface invites users to apply their specific business expertise, which leads to more powerful predictive models and shortens time-to-solution. Clementine offers many modeling techniques, such as prediction, classification, segmentation, and association detection algorithms.
once models are created, Clementine Solution Publisher enables their delivery enterprise-wide to decision makers or to a database.

Clementine includes a number of machine-learning and modeling technologies, which can be roughly grouped according to the types of problems they are intended to solve: prediction, clustering, and association.

- Predictive modeling methods include decision trees, neural networks, and Statistical models.
- Clustering models focus on identifying groups of similar records and labeling the records according to the group to which they belong. clustering methods include kohonen, K-Means, and two step.
- Association rules associate a particular conclusion (such as the purchase of a particular product) with a set of conditions (the purchase of several other products).
- Screening models can be used to screen data to locate fields and records that are most likely to be of interest in modeling, and identify outliers that may not fit known patterns. Available methods include feature selection and anomaly detection.

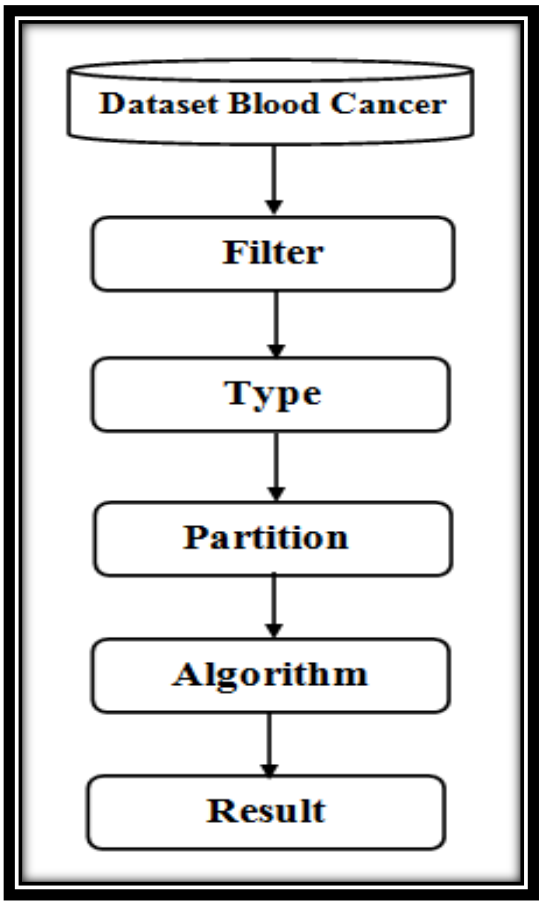


Figure [3-7] Steps of Execution Clementine

Figure [3-7] Steps Execution Clementine tool explain steps from implementation dataset in Clementine tool .data is filtered according to attributes required. and select type or target then partitioning data into tow part testing and training after that chose algorithm in this search tow algorithms neural network and decision tree .we execute model and we get result .

Chapter Four

Experimental Results

4-1 Introduction:

In this chapter explain modeling of algorithms

4-2 Modeling of Logarithms:

To evaluate the effectiveness of our methods, experiments on blood cancer dataset is conducted by Implementation tools .Clementine as a data mining tool that combines advanced modeling technology with ease of use, Clementine helps you discover and predict interesting and valuable relationships within your data. you can use Clementine for decision-support activities.

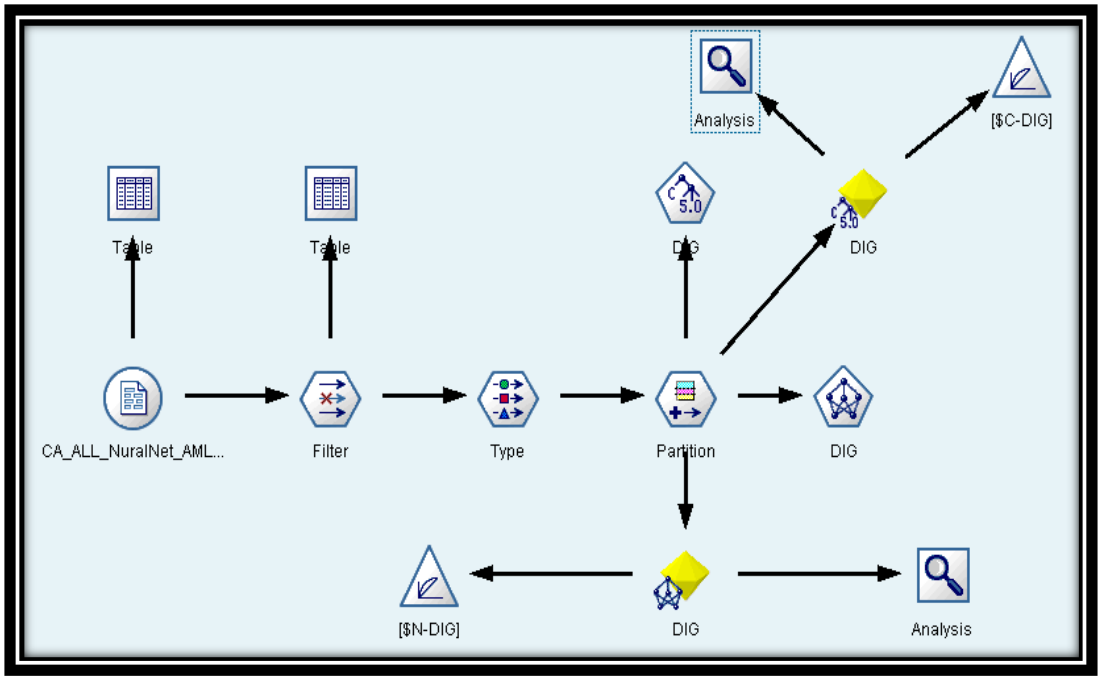


Figure [4-1] Modeling of Logarithms

Figure [4-1] explains the steps to build the model using clementine10.1.it's one of most powerful tools and programs used in data mining.
Clementine is visual programming interface is so easy to learn each node has a clearly defined function.

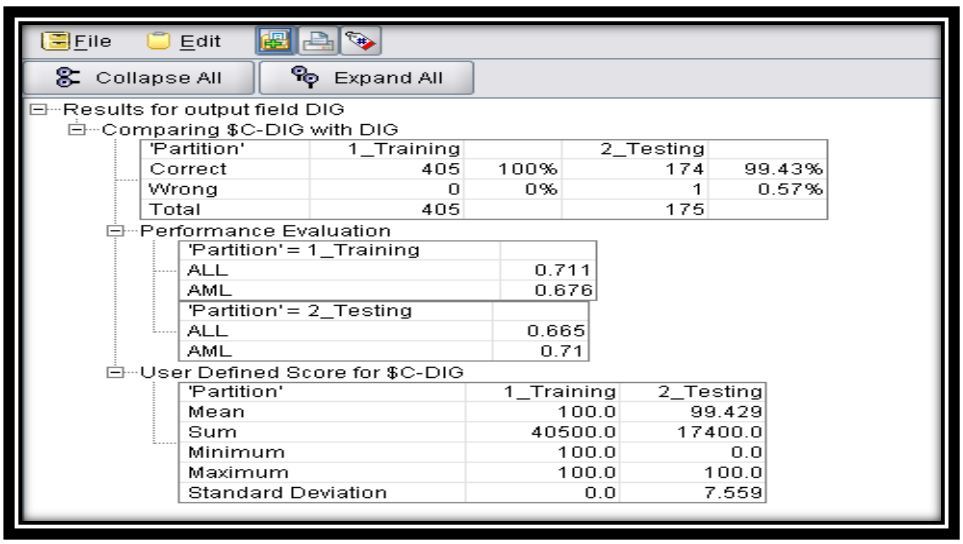


Figure [4-2] Decision Trees Analyses

Figure [4-2] represented result of using algorithms Decision Trees and Analyses it explain partitioning the data 70% training dataset and 30% testing dataset the correct or is 100% and wrong is 0% . The testing data set wrong is one and correct is 99.43%.

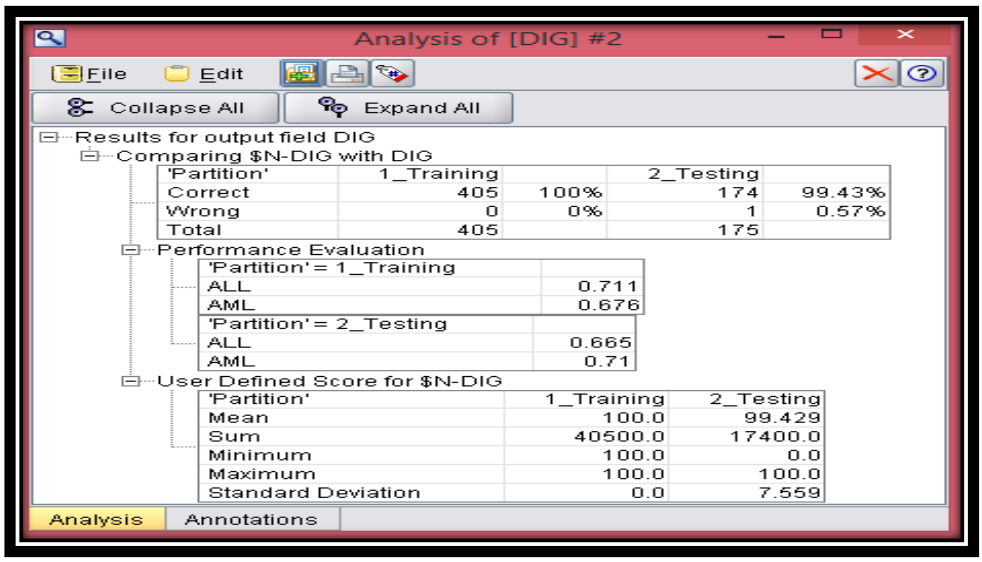


Figure [4-3] Neural Network Analyses

Figure [4-3] represented result of using algorithms Neural network analyses demonstrates the accuracy of algorithms in model construction in training dataset and testing data set .where the dataset portioning it 70% training and 30% testing. Correct is 100% in training data set and wrong is 0% , testing data set correct is 99.43% and wrong is one.

Table [4-2]: Classification Accuracy on CBC Dataset

No	Classification Algorithm	Algorithm(Accuracy %)
1.	Decision Tree	99.43%
2.	Neural Network	99.43%

These results were reached after many experiments, we observed from table [4-2] the same accuracy with decision tree and neural network obtained 99.43% using partition 70% training dataset and 30% testing dataset .
We found the result that was monitored by the bureau of statistics it is the most common type of leukemia spread among children is (AML) (Acute Myeloid Leukemia) .

4-3 Conclusions:

In this research we concluded that blood cancer (AML) Acute Myeloid Leukemia is a threat to the lives of children in Sudan. the ministry of health has be carry out on intensive program to educate citizens about the seriousness of the disease.in this research we used two algorithm to predict childhood leukemia ,we chose two type of leukemia (AML) Acute Myeloid Leukemia and (ALL) Acute Lymphoblastic Myeloid, we found (AML) more widespread (ALL).

4-4 Recommendations:

- Providing data in computerized manner in a database in modern way especially in the health system .because of their role in health care .and to make the right decision in advancing health awareness and continuous development .also provides researchers with saving time.

- Data should be documented in a clear and understandable manner.
- Use other classification algorithm to predict childhood leukemia.
- Research that includes a geographic allocation to reduce the incidence of the disease.
- Knowledge of the disease in geographical locations makes it easier to know the causes, through behavior of individuals and for early prevention.

References:

[1] P.Ramachandran1, Dr.N.Girija2, Dr.T.Bhuvaneswari3 2013, JuneCancer Spread Pattern –an Analysis using Classification and Prediction Techniques.(International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 6, June 2013).

[2]SAS.(2018). Evolution of machine learning. [Online] Available at: https://www.sas.com/en_us/insights/analytics/machine-learning/ [Accessed 21 Nov. 2018].

[2] P.Ramachandran1, Dr.N.Girija2, Dr.T.Bhuvaneswari3 2015, Month: April - 2015, September A Study on Cancer Perpetuation Using the Classification Algorithms.(International Journal of Recent Research in Mathematics Computer Science and Information Technology Vol. 2, Issue 1, pp: (96-99), Month: April 2015 – September 2015, Available at: www.paperpublications.org)

[3] 1Durairaj M, 2Deepika R 2015 August Comparative Analysis of Classification Algorithms for the Prediction of Leukemia Cancer.(Volume 5, Issue 8, August 2015 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www. ijarcsse.com).

[4] Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.

[5] P.Ramachandran- N.Girija, Ph.D - T.Bhuvaneswari, Ph.D.2014 July Early Detection and Prevention of Cancer using Data Mining Techniques.
(International Journal of Computer Applications (0975 – 8887) Volume 97– No.13, July 2014).

[6] 1P.Ramachandran, 2Dr.N.Girija, 3Dr.T.Bhuvaneswari2012,May.Healthcare Service Sector: Classifying and Finding Cancer spread pattern in Southern India Using Data Mining Techniques .(P.Ramachandran et al. / International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397Vol. 4 No. 05 May 2012) .

[7] 1K. Kalaivani and 2R.Shanmugalakshmi 2011, Childhood Cancer-a Hospital based study using Decision Tree Techniques.(Journal of Computer Science 7 (12):

1819-1823, 2011 ISSN 1549-3636© 2011 Science Publications J. Computer Sci., 7 (12): 1819-1823, 2011).

[8] Varun Kumar –Nisha Rathee 2011, March .Knowledge discovery from database Using an integration of clustering and classification .((IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011).

[9] Shweta Kharya Sr. Assistant Professor, Bhilai Institute of Technology, Durg-491 001, Chhattisgarh, India 2012, April .USING DATA MINING TECHNIQUES FOR DIAGNOSISAND PROGNOSIS OF CANCER DISEASE.
(International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012)

[10] Durairaj M Assistant Professor School of Comp. Sci., Engg. &Applications, BharathidasanUniversity,Trichy,TN,Indiadurairaj.bdu@gmail.comDeepika R Research Scholar School of Comp. Sci., Engg. &Applications, Bharathidasan university, Trichy,TN,Indiadeepikar9191@gmail.com 2015, February
PREDICTION OF ACUTE MYELOID LEUKEMIA CANCER USING DATAMINING-A SURVEY. (International Journal of Emerging Technology and Innovative Engineering Volume I, Issue 2, February 2015 ISSN: 2394 - 6598).

[11] Nagaraju Orsu Department of Computer Science Government Degree College Macherla, Andhra Pradesh, India Suresh B. Mudunuri* Centre for Bioinformatics Research & S/w Development Grandhi Varalakshmi Venkatarao Institute of Technology Bhimavaram, Andhra Pradesh, India 2013 Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification .
(IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5, 2013).

[12] Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, pp.3-24.

[13] Muhammad, I. and Yan, Z., 2015. SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. ICTACT Journal on Soft Computing, 5(3).

[14] Deulkar, M.D.S. and Deshmukh, R.R., 2016. Data Mining Classification. Imperial Journal of Interdisciplinary Research, 2(4).

[15] Jiawei Han University of Illinois at Urbana-Champaign Micheline Kamber Data Mining: Concepts an Techniques Second Edition.
[16] Esfandiari, N., Babavalian, M.R., Moghadam, A.M.E. and Tabar, V.K., 2014. Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications, 41(9), pp.4434-4463.

[16] Chinky GeraM .Tech(CSE) student Kirti Joshi Assistant Professor 2015 March A Survey on Data Mining Techniques in the Medicative Field.(International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 13, March 2015).

[17] http://infolab.stanford.edu/~ullman/mmds/ch1.pdf Access 04:08 12/5/2019

[18] https://www.exinfm.com/pdffiles/intro_dm.pdf Access 04:16 after morning Osmar R. Zaïane, 1999 CMPUT690 Principles of Knowledge Discovery in Databases.

[18] Pediatric AML: From Biology to Clinical Management
Jasmijn D. E. de Rooij, C. Michel Zwaan and Marry van den Heuvel-Eibrink *
Department of Pediatric Oncology, Erasmus MC-Sophia Children's Hospital, 3015CN Rotterdam,The Netherlands; E-Mails: j.d.e.derooij@erasmusmc.nl (J.D.E.R.);c.m.zwaan@erasmusmc.nl (C.M.Z.)* Author to whom correspondence should be addressed; E-Mail: m.vandenheuvel@erasmusmc.nl;
Tel.: +31-107-036-691.Academic Editor: Celalettin Ustun
Received: 17 October 2014 / Accepted: 28 November 2014 / Published: 9 January 2015.

[19] A Concise Review Jennifer N. Saultz 1 and Ramiro Garzon 2, 2518Academic Editor: Jeffrey E. Rubnitz Received: 11 December 2015; Accepted: 29 February 2016; Published: 5 March 2016 Correspondence: ramiro.garzon@osumc.edu; Tel.: +1-614-247- Acute Myeloid Leukemia

[20] Alaa M. Elsayad 2010 Predicting the Severity of Breast Masses with Ensemble of Bayesian Classifiers Department of Computers and Systems, Electronics Research Institute,12622 Bohoth St., Dokki, Geza, Egypt.

[21] Ashish Chhabbi,Lakhan Ahuja,Sahil Ahir and Prof Y.K.Sharma Vishwakarma 19 March 2016 Heart Disease Prediction Using Data Mining Techniques International Journal of Research in Advent Technology (E-ISSN:2321-9637) Special Issue

[24] Clementine Desktop10.1 user guide book.

Appendix:

There are many diseases of cancers including:
NHL :Non-Hodgkin's lymphoma
NPC :Nasopharyngeal Carcinoma
NPH : Neutral protamine Hagedorn
LHC :Lymphopietic Cancers
RBL : Rat basophilic Leukemia
HL : Hodgkin Lymphoma type of Lymphoma
HCC : Hepatocellular Carcinoma
HD :Hodgkin's Disease or Lymphoma
GCI :gastric cancer incidence
RECTUM : Rectum
FIBRO: Fibroblasts in cancer Fibro-Osseous,Fibro- sarcoma
Boon Cancer: Cancer in the bone marrow
Ostusacoma: Cancer in the bone marrow
Oesophaqus
Stomach
Oral Cancer
Bowel Cancer
Brain Tumer
WillimsTumour
Thymoma
Liver
Kidney& Renal
HCC : Liver Cancer
Lung & Bronchus
Thyroid
Lip sarcoma
Unknown Primary
Malignant Melanoma
Adenocarcinoma
Nasopharynx
Neurablastoma
Histosytoma
Retinoblastoma
Pharynx
Hypophargnx
Soft tussuse
Rhabdomysarcoma

Scalp
BCC
ARM
GCL