International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:5, 2016

Predication Model for Leukemia Diseases Based on Data Mining Classification Algorithms with Best Accuracy

Fahd Sabry Esmail, M. Badr Senousy, Mohamed Ragaie

***Abstract*—**In recent years, there has been an explosion in the rate of using technology that help discovering the diseases. For example, DNA microarrays allow us for the first time to obtain a "global" view of the cell. It has great potential to provide accurate medical diagnosis, to help in finding the right treatment and cure for many diseases. Various classification algorithms can be applied on such micro-array datasets to devise methods that can predict the occurrence of Leukemia disease. In this study, we compared the classification accuracy and response time among eleven decision tree methods and six rule classifier methods using five performance criteria. The experiment results show that the performance of Random Tree is producing better result. Also it takes lowest time to build model in tree classifier. The classification rules algorithms such as nearest- neighbor-like algorithm (NNge) is the best algorithm due to the high accuracy and it takes lowest time to build model in classification.

***Keywords*—**Data mining, classification techniques, decision tree, classification rule, leukemia diseases, microarray data.

I. Introduction

DATA mining plays an important role for predicting diseases. Recent advances in microarray technology offer the ability to measure expression levels of thousands of genes simultaneously. Analysis of such data helps us identifying different clinical outcomes that are caused by expression of a few predictive genes. The feature extraction and classification are carried out with combination of the high accuracy of ensemble based algorithms, and comprehensibility of a single decision tree. These allow deriving exact rules by describing gene expression differences among significantly expressed genes in leukemia. It is evident from our results that it is possible to achieve better accuracy in classifying Leukemia without sacrificing the level of comprehensibility. Some of the most important and popular data mining techniques are association rules, classification, clustering, prediction and sequential patterns [1].

Leukemia disease is a type of cancer that affects the blood and the bone marrow it is characterized by an abnormal proliferation of blood cells. Acute Myelogenous Leukemia (AML), Acute Lymphoblastic Leukemia (ALL), Chronic Myeloid Leukemia (CML) and Chronic Lymphocytic Leukemia (CLL) are categorized as leukemia diseases [2]. In general, leukemia is grouped by how fast it gets worse and what kind of white blood cells it affects [3].

Microarray is one such technology which enables the researchers to investigate and address issues which were once thought to be non-traceable by facilitating the simultaneous measurement of the expression levels of thousands of genes [4]. Microarray data sets are commonly very large, and analytical precision is influenced by a number of variables. So it is extremely useful to reduce the dataset to those genes that are best distinguished between the two cases or classes (e.g. normal vs. diseased). There are two common methods for in depth microarray data analysis such as clustering and classification [5]. Clustering is one of the unsupervised approaches to classify data into groups of genes or samples with similar patterns that are characteristic to the group. Classification is supervised learning and also known as class prediction or discriminate analysis. Generally, classification is a process of learning-from-examples. Given a set of pre-classified examples, the classifier learns to assign an unseen test case to one of the classes.

A DNA microarray technique allows to simultaneously observing the expression levels of thousands of genes during significant biological processes and across collections of related samples [6]. The rest of this paper is organized as the follows. In Section II, we discuss related works in this domain. In Section III, we explore the methodologies used in this work. In Section IV, we present experimental results and analysis. In Section V, we conclude the paper.

II. Literature Review

There are several gene selection methods for cancer classification using microarray datasets. However, most of them did not concern on identifying minimum number of informative genes with high classification accuracy [7].

Sivaraman et al. proposed a blood cancer prediagnosis system with the aid of Statistical Approach with Fuzzy Inference System and Feed Forward Back Propagation Neural Network. Their system was implemented on a huge set of test data. It was utilized to analyze of the outcomes. Thus the proposed Blood cancer pre diagnosis system offers a significant of accuracy, sensitivity and specificity. That used the method more precisely diagnosis the Blood cancer from the given test data by seeing the elevated rate of measurements

Fahd Sabry Esmail is demonstrator in Management Information Systems department in Modern Academy, Egypt (e-mail: fahdsabry985@gmail.com).
M.Badr Senousy is Professor of Computer Sciences and Information Systems in Sadat Academy for Management Sciences, Egypt (e-mail: badr_senousy_arcoit@yahoo.com).
Mohamed Ragaie is Professor of Computer Sciences and Information Systems in Arab Academy for Science Technology and Maritime Transport, Egypt (e-mail: ragaie2@mcit.gov.eg).

[8].

Priyanga et al. developed a system called data mining based cancer prediction system. The main aim of this model was to provide the earlier warning to the users, and it was also cost and time benefit to the user. It predicts three specific cancer risks. Specifically, cancer prediction system estimates the risk of the breast, skin, and lung cancers by examining a number of user-provided genetic and non-genetic factors. This system is validated by comparing its predicted results with the patient's prior medical record, and also this system analyzed using Weka. This prediction system is available in online [9].

Suji et al. get the oral datasets form the various diagnostic centers which contained both cancer and non-cancer patients' information and collected data was pre-processed for duplicate and missing information. Then they applied many classification algorithms on NMDS dataset. The performance of those algorithms had been analyzed. A classification rate of 100% was obtained for C4.5 algorithm and classification rate of 98.7% was obtained for Random Tree Algorithm. Classification rate of 99.5% was obtained for MPNN [10].

Shajahaan et al. compared various supervised learning algorithms to predict the best classifier. Experimental results showed that the effectiveness of the proposed method. Model was also evaluated using precision and recall. It was found that among various classification techniques random tree outperforms of all other algorithms with highest accuracy rate. Therefore, an efficient classifier was identified to determine the nature of the disease which was highly essential in a clinical investigation of life threatening disease like breast cancer [11].

Dash et al. provided a comparison between dimension reduction technique, namely Partial Least Squares (PLS) method and a hybrid feature selection scheme They evaluated the relative performance of four different supervised classification procedures such as Radial Basis Function Network (RBFN). Experimental results showed that the Partial Least-Squares (PLS) regression method was an appropriate feature selection method and a combined use of different classification and feature selection approaches made it possible to construct high performance classification models for microarray data [12].

Chandrasekar et al. presented effective classification techniques. After investigation of different classification algorithms, they chosen 6 classifier based on simulation performance and they used Tree Random classifier achieved overall classification accuracy 98%, which was significant [13].

Pujari et al. presented an ensemble model which was constructed to improve classification accuracy by combining the prediction of multiple classifiers. The performance measured gain, accuracy, specificity and sensitivity which were analyzed to handle ionosphere data using CART, CHAID and QUEST classification algorithms. From the experimental results, they concluded that the ensemble model with feature selection achieved highest accuracy of 93.84% on test data [14].

III. Methodology Research

This research uses data mining techniques for analysis and evaluation of classification algorithms of leukemia disease dataset. Through open source WEKA data mining techniques, we can generate predictive model for classification of leukemia disease, evaluate accuracies, and performance of several techniques.

A. Dataset Description

To compare these data mining classification techniques and comparison analysis, we need the datasets. This research chooses Leukemia data sets. Directly we can apply this data in the data mining tools (Weka) and predict the results. The chosen dataset "Testing data" on year 2010 contains 72 leukemia samples (47 ALL and 25 AML). Table I shows leukemia data sets description.

TABLE I
LEUKEMIA DATA SETS DESCRIPTION

Owner	Classes	Attribute Type	# Attributes genes	# Instances
BioInformatics_Seville [15]	ALL AML	Numeric	7129	72

B. Classification Algorithms

Classification divides data samples into target classes. The classification technique predicts the target class for each data points. Data classification approach is a supervised learning approach having known class categories [36]. Data set is partitioned as training and testing datasets. Using training dataset, we trained the classifier. Correctness of the classifier could be tested using test dataset. Classification is one of the most widely used methods of Data Mining in Healthcare organization [16]. However, the accuracy of such methods different according to the classification algorithm used. Identifying the best classification algorithm among all available is a challenging task. The present research proposes a comprehensive analysis of different classification algorithms, and performance of evaluate by applying leukemia micro-array data set. Hu et al. [16] used different classification method such as decision tree, SVM and ensemble approach for analyzing microarray data [16].

1. Utilization of Decision Tree Algorithms

Decision tree is one of the most popular and efficient technique in data mining. This technique has been established and well explored by many researchers. However, some decision tree algorithms may produce a large structure of tree size and it is difficult to understand [21]. Furthermore, misclassification of data often occurs in learning process. Therefore, a decision tree algorithm that can produce a simple tree structure with high accuracy in term of classification rate is a need to work with huge volume of data. Pruning methods have been introduced to reduce the complexity of tree structure without decrease the accuracy of classification.

In this research, we choose WEKA (The Waikato Environment for Knowledge Analysis) for running several algorithms in decision tree. Each algorithm was explained in subsections from A to K.

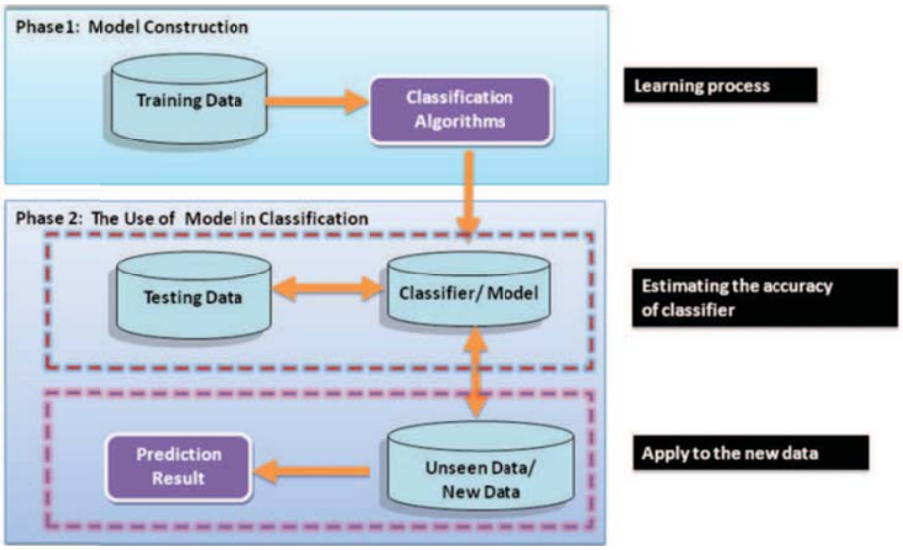

Fig. 1 Classification and Prediction in Data Mining [16]

a) J48

J48 classifier is a simple C4.5 decision tree for classification [17]. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to Zhao and Zhang [18], C4.5 algorithm produce decision tree classification for a given dataset by recursive division of the data and the decision tree is grown using Depth-first strategy. On data testing this algorithm will emphasize splitting dataset and by selecting a test that will give best result in information gain. In discrete attributes as well, these algorithms consider a test with a result of many as the number of different values and test binary attribute for each attribute will continue to grow in different values each attribute will be considered [18].

b) REPTree

Basically Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. The basic of pruning of this algorithm is it used REP with back over fitting. It kindly sorts values for numerical attribute once and it handling the missing values with embedded method by C4.5 in fractional instances. In this algorithm we can see it used the method from C4.5 and the basic REP also count in it process [18].

c) LADTree

LAD Tree Logical Analysis of Data is the method for classification proposed in optimization literature. It builds a classifier for binary target variable based on learning a logical expression that can distinguish between positive and negative samples in a data set. The basic assumption of LAD model is that a binary point covered by some positive patterns, but not covered by any negative pattern is positive, and similarly, a binary point covered by some negative patterns, but not covered by positive pattern is negative. The construction of Lad model for a given data set typically involves the generation of large set patterns and the selection of a subset of them that satisfies the basic assumption of LAD model such that each pattern in the model satisfies certain requirements in terms of prevalence and homogeneity [19].

d) Random Forest

Random forest is an ensemble classifier which consists of many decision tree and gives class as outputs i.e., the mode of the class's output by individual trees. Random Forests gives many classification trees without pruning [20].

e) CART

CART stands for Classification and Regression Trees [37]. It is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. The splits are selected using the twoing criteria and the obtained tree is pruned by cost–complexity Pruning. When provided, CART can consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution. An important feature of CART is its ability to generate regression trees. Regression trees are trees where their leaves predict a real number and not a class. In case of regression, CART looks for splits that minimize the prediction squared error (the least–squared deviation). The prediction in each leaf is based on the weighted mean for node [21].

f) FT Tree

FT combines a standard univariate DT, such as C4.5, with linear functions of the attributes by means of linear regressions. While a univariate DT uses simple value tests on single attributes in a node, FT can use linear combinations of different attributes in a node or in a leaf. In the constructive phase a function is built and mapped to new attributes. A model is built using the constructor function. This is done using only the examples that fall at this node. Later, the model

is mapped to new attributes [19].

g) BFTree

BFTree is a classification algorithm that builds a decision tree using a best-first expansion of nodes rather than the depth-first expansion used by standard decision tree learners (such as C4.5). Pre- and postpruning options are available that are based on finding the best number of expansions to use via cross-validation on the training data. While fully grown trees are the same for best-first and depth-first algorithms, the pruning mechanism used by BFTree will yield a different pruned tree structure than that produced by depth-first methods [38]. Another tree base classification algorithm is FT that builds a functional tree with oblique splits and linear functions at the leaves [22].

h) Decision Stumps (DS)

Decision stumps (DS) are one level decision trees [23]. We can find the best stump just as we would learn a node in a decision tree: we search over all possible features to split on, and for each one, we search over all possible thresholds induced by sorting the observed values. In classification problems, each node in a decision stump represents a feature in an instance to be classified, and each branch represents a value that the node can take. Instances are classified starting at the root node and sorting them based on their feature values. In regression problems, DS (or regression stumps) do regression based on mean-squared error where each node in a decision stump represents a feature in an instance to be predicted, and each branch represents a value that the node can take. At worst a decision stump will reproduce the most common sense baseline, and may do better if the selected feature is particularly informative [24].

i) Logistic Model Tree

Logistic Model Tree (LMT) [25] algorithm makes a tree with binary and multiclass target variables, numeric and missing values. So this technique uses logistic regression tree. LMT produces a single outcome in the form of tree containing binary splits on numeric attributes.

j) NBTree

A Naive Bayes Tree (NBTree) Classifier Although the attribute independence assumption of naive Bayes is always violated on the whole training data; it could be expected that the dependencies within the local training data is weaker than that on the whole training data. Thus, NBTree [26] builds a naive Bayes classifier on each leaf node of the built decision tree, which just integrate the advantages of the decision tree classifiers and the naive Bayes classifiers. Simply speaking, it firstly uses decision tree to segment the training data, in which each segment of the training data is represented by a leaf node of tree, and then builds a naive Bayes classifier on each segment. A fundamental issue in building decision trees is the attribute selection measure at each non-terminal node of the tree.

k) RandomTree

A random tree is a tree drawn at random from a set of possible trees. In this context “at random” means that each tree in the set of trees has an equal chance of being sampled. Another way of saying this is that the distribution of trees is “uniform”. Random trees can be generated efficiently and the combination of large sets of random trees generally leads to accurate models. Random tree models have been extensively developed in the field of Machine Learning in the recent years [18].

2.Utilization of Rule Classifier Algorithms

Rule based classification algorithm also known as separate-and-conquer method. This method is an iterative process consisting in first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover [27]. Rule discovery or rule extraction from data is data mining techniques aimed at understanding data structures, providing comprehensible description instead of only black box prediction.

Classification algorithms are widely used in various applications. Data classification is a two steps process in which first step is the training phase where the classifier algorithm builds classifier with the training set of tuples and the second phase is classification phase where the model is used for classification and its performance is analyzed with the testing set of tuples [28]. There are various classification rule algorithms such as NNge, JRip, Ridor, DTNB, PART, OneR, ZeroR and so on. In this research, we have analyzed classification rule algorithms namely OneR, JRip, NNge, PART, Ridor and ZeroR.

a) OneR

OneR, short for “One Rule”, is a simple classification algorithm that generates a one-level decision tree. OneR is able to infer typically simple, yet accurate, classification rules from a set of instances. Comprehensive studies of OneR’s performance have shown it produces rules only slightly less accurate than state-of-the-art learning schemes while producing rules that are simple for humans to interpret. OneR is also able to handle missing values and numeric attributes showing adaptability despite simplicity. The OneR algorithm creates one rule for each attribute in the training data, and then selects the rule with the smallest error rate as its ‘one rule’. To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to their majority class; one such binding for each attribute value of the attribute the rule is based on [29].

b) JRip

In 1995 JRip was implemented by Cohen, W. W, in this algorithm were implemented a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction

(RIPPER). By the way, Cohen implementing RIPPER [30] in order to increase the accuracy of rules by replacing or revising individual rules. Reduce Error Pruning was used where it isolates some data for training and decided when stop from adding more condition to a rule. By using the heuristic based on minimum description length as stopping criterion. Post-processing steps followed in the induction rule revising the regulations in the estimates obtained by global pruning strategy and it improves the accuracy.

c) NNge

Nearest-neighbor-like algorithm (NNge) is a nearest neighbor method with generalization. Instance-based learners are "lazy" in the sense that they perform little work when learning from the data set, but expend more effort classifying new examples. The simplest method, nearest neighbor, performs no work at all when learning. NNge does not attempt to out-perform all other machine learning classifiers. Rather, it examines generalized exemplars as a method of improving the classification performance of instance-based learners [31].

d) PART

PART algorithm [32] is a relatively simple algorithm who does not execute global optimization to generate accurate rules, but it is practiced separately and-conquer strategy, for example it builds a rule, removes the instances it covers, and continues to create a recursive rule for instances rest until there is no longer the instances is left. Furthermore, Eibe and Witten [32] said that the algorithm producing sets of rules called 'decision lists' which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in every iterative and makes the "best" leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning.

e) Ridor

Brian R. Gaines and Paul Compton [33] has develop Ridor or RIpple-DOwn Rule learner. This algorithm generates default rule first and after that it generate the exceptions for default rule along with the least error rate. Then it generates the "best" exceptions for each exception and iterates until pure. Thus it performs a tree-like expansion of exceptions. The exceptions are a set of rules that predict classes other than the default. IREP is used to generate the exceptions.

f)ZeroR

ZeroR is the simplest classification method which depends on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods [34].

g)Performance Factors Evaluation

Accuracy is the proportion of the total number of predictions that were correct. It is determined using:

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp} [35]$$

where, TP rate = positives correctly classified / total positives, FP rate = negatives incorrectly classified / total negatives.

Precision is the proportion of the predicted positive cases that were correct, as calculated using:

$$\text{Precision} = \frac{tp}{tp+fp}$$

Recall or Sensitivity or True Positive Rate (TPR): It is the proportion of positive cases that were correctly identified, as calculated using:

$$\text{Recall} = \frac{tp}{tp+fn}$$

F-measure: The F-Measure computes some average of the information retrieval precision and recall metrics.

$$F_{ij} = \frac{2 * \text{Recall}(i,j) * \text{Precision}(i,j)}{\text{Precision}(i,j) + \text{Recall}(i,j)}$$

Receiver Operating Characteristic (ROC) Curve: It is a graphical approach for displaying the tradeoff between true positive rate (TPR) and false positive rate (FPR) of a classifier.

IV. Experimental Results

In this section, we conducted an experiment using Weka application. Weka is a comprehensive suite of Java class libraries that perform many advanced machine learning and data mining algorithms [25]. We analyze and compare the performance of decision tree algorithms namely Decicion Stump, FT, J48(C4.5), LADTree, REPTree, LMT, NBTree, CART, Random Forest and RandomTree, and compare the performance of Rule classifier algorithms namely JRip, NNge, OneR, PART, Ridor, ZeroR.

A. Accuracy Measures

This approach has been implemented on two different machines (M1 and M2) as shown in Table II. The simulation results are partitioned into several sub items for easier analysis and evaluation. Different performance matrix like accuracy, Time Taken to Build Model (Seconds), True Positive rate, False Positive rate, Precision, Recall, F Measure, and Receiver Operating Characteristics (ROC) Area are presented in numeric value during training and testing phase. The summary of those results by running the techniques in WEKA is reported in Tables III-VI.

Figs. 2 and 3 show the comparison based about the accuracy by each learning algorithm. Based on Figs. 2 and 3, we can clearly see that the highest accuracy is 100% and the lowest is 65.27%. In fact, the highest accuracy belongs to the NNge from Rule Classifier and FT, LAD tree, LMT, NBtree, Random forest and random tree from tree classifier. The total time required to build the model is also a crucial parameter in comparing the classification algorithm.

In this simple experiment, from Tables II-V, we can say that a Zero R from rule classifier requires the shortest time which is around 0 seconds consecutive with compared to random tree from tree classifier which requires the longest model building time which is around 0.02 seconds. as shown in Figs. 7 and 8.

B.Response Time

The total time required to build the model is also a crucial parameter in comparing the classification algorithm. In Tables VII-X, we have summarized two main measures of evaluation for each algorithm such as time taken to build the model and accuracy.

TABLE II
DESCRIPTION OF MACHINES

Machine Name	Specification
M1	Intel Core 2 Due 2.13 GHz Processor 4 GB RAM
M2	Intel 3.00 GHz Processor 2 GB RAM

TABLE III
ACCURACY MEASURE FOR CLASSIFICATION RULE ALGORITHMS (M1)

Methods	Time Taken to Build Model (Seconds)	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Accuracy %
JRIP	1.09	0.958	0.041	0.959	0.958	0.959	0.959	95.83
NNge	1.39	1	0	1	1	1	1	100
One R	0.28	0.958	0.041	0.959	0.958	0.959	0.959	95.83
PART	0.68	0.986	0.026	0.986	0.986	0.986	0.98	98.61
Ridor	0.55	0.944	0.086	0.945	0.944	0.944	0.929	94.44
Zero R	0	0.653	0.653	0.426	0.653	0.516	0.5	65.27

TABLE IV
ACCURACY MEASURE FOR CLASSIFICATION TREE ALGORITHMS (M1)

Methods	Time Taken to Build Model (Seconds)	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Accuracy %
BF Tree	4.87	0.958	0.041	0.959	0.958	0.959	0.959	95.83
Decision Stump	0.26	0.944	0.03	0.952	0.944	0.945	0.957	94.44
FT	3.75	1	0	1	1	1	1	100
J48(C4.5)	0.56	0.986	0.026	0.986	0.986	0.986	0.98	98.61
LADTree	9.52	1	0	1	1	1	1	100
REP Tree	0.3	0.931	0.037	0.942	0.931	0.932	0.947	93.05
LMT	5.25	1	0	1	1	1	1	100
NBTree	3.62	1	0	1	1	1	1	100
CART	4.48	0.958	0.041	0.959	0.958	0.959	0.959	95.83
Random Forest	0.91	1	0	1	1	1	1	100
Random Tree	0.02	1	0	1	1	1	1	100

TABLE V
ACCURACY MEASURE FOR CLASSIFICATION RULE ALGORITHMS (M2)

Methods	Time Taken to Build Model (Seconds)	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Accuracy %
JRIP	2.17	0.958	0.041	0.959	0.958	0.959	0.959	95.83
NNge	2.59	1	0	1	1	1	1	100
One R	0.36	0.958	0.041	0.959	0.958	0.959	0.959	95.83
PART	1.49	0.986	0.026	0.986	0.986	0.986	0.98	98.61
Ridor	1.8	0.944	0.086	0.945	0.944	0.944	0.929	94.44
Zero R	0	0.653	0.653	0.426	0.653	0.516	0.5	65.27

TABLE VI
ACCURACY MEASURE FOR CLASSIFICATION TREE ALGORITHMS (M2)

Methods	Time Taken to Build Model (Seconds)	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Accuracy %
BF Tree	8.38	0.958	0.041	0.959	0.958	0.959	0.959	95.83
Decision Stump	0.59	0.944	0.03	0.952	0.944	0.945	0.957	94.44
FT	7.17	1	0	1	1	1	1	100
J48(C4.5)	1.45	0.986	0.026	0.986	0.986	0.986	0.98	98.61
LADTree	10.92	1	0	1	1	1	1	100
REP Tree	0.41	0.931	0.037	0.942	0.931	0.932	0.947	93.05
LMT	11.03	1	0	1	1	1	1	100
NBTree	4.59	1	0	1	1	1	1	100
CART	8.98	0.958	0.041	0.959	0.958	0.959	0.959	95.83
Random Forest	0.33	1	0	1	1	1	1	100
Random Tree	0.02	1	0	1	1	1	1	100

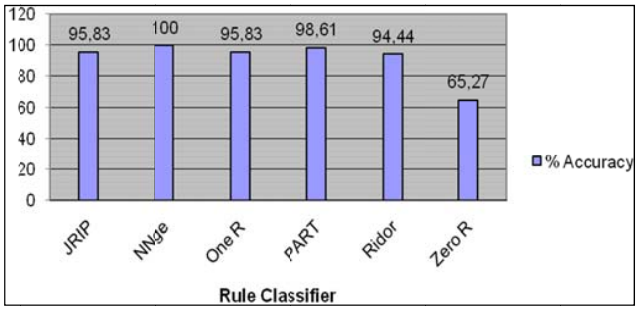

Fig. 2 Accuracy % of Rule Classifiers

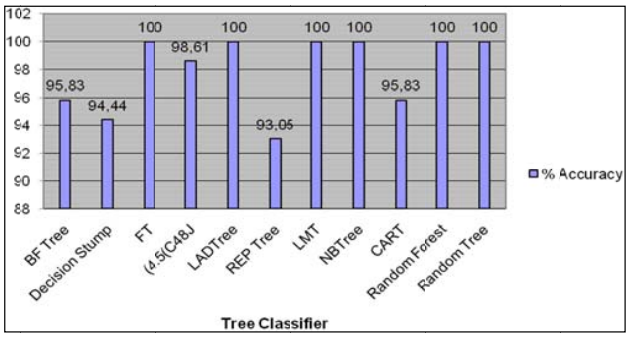

Fig. 3 Accuracy % of Tree Classifiers

TABLE VII
COMPARISON OF RULE CLASSIFIER METHODS (M1)

Methods	Time Taken to Build Model (Seconds)	Accuracy %
JRIP	1.09	95.83
NNge	1.39	100
One R	0.28	95.83
PART	0.68	98.61
Ridor	0.55	94.44
Zero R	0	65.27

TABLE VIII
COMPARISON OF TREE CLASSIFIER METHODS (M1)

Methods	Time Taken to Build Model(Seconds)	Accuracy %
BF Tree	8.38	95.83
Decision Stump	0.59	94.44
FT	7.17	100
J48(C4.5)	1.45	98.61
LADTree	10.92	100
REP Tree	0.41	93.05
LMT	11.03	100
NBTree	4.59	100
CART	8.98	95.83
Random Forest	0.33	100
Random Tree	0.02	100

Tables VII-X show that NNge from rule classifier take maximum amount of time to build the model i.e. is around 1.39-1.59 seconds. Next highest LMT is around 11.03-5.25 and LADtree 10.92-9.52 seconds to build the model from tree classifier. In terms of second measure of evaluation, Random tree has the highest percentage of accuracy is 100% and has the longest model building time which is around 0.02 seconds and the best Measure among all and the Next highest accuracy is 100% belongs to NNge and take time to build model is around 1.39-1.59 seconds that also lowest time compare to others. Hence, we conclude that Random tree has performed better than all the other classifiers in the analysis by two machines of our dataset.

TABLE IX
COMPARISON OF RULE CLASSIFIER METHODS (M2)

Methods	Time Taken to Build Model (Seconds)	Accuracy %
JRIP	2.17	95.83
NNge	2.59	100
One R	0.36	95.83
PART	1.49	98.61
Ridor	1.8	94.44
Zero R	0	65.27

Figs. 4-9 show the comparison of all the algorithms of two machines with respect to the time taken to build the model.

TABLE X
COMPARISON OF TREE CLASSIFIER METHODS (M2)

Methods	Time Taken to Build Model (Seconds)	Accuracy %
BF Tree	4.87	95.83
Decision Stump	0.26	94.44
FT	3.75	100
J48(C4.5)	0.56	98.61
LADTree	9.52	100
REP Tree	0.3	93.05
LMT	5.25	100
NBTree	3.62	100
CART	4.48	95.83
Random Forest	0.91	100
Random Tree	0.02	100

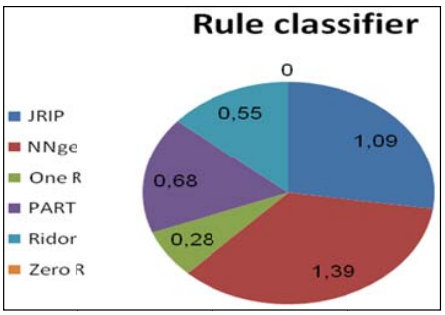

Fig. 4 Time Taken to Build Model of M1 (Seconds)

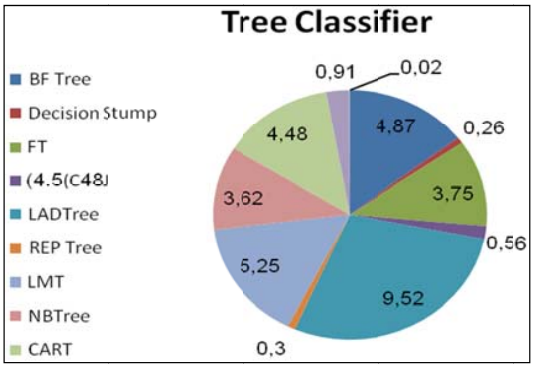

Fig. 5 Time Taken to Build Model of M1 (Seconds)

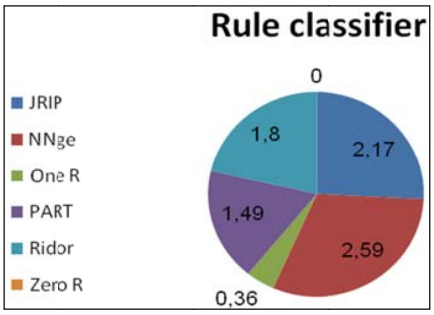

Fig. 6 Time Taken to Build Model of M2 (Seconds)

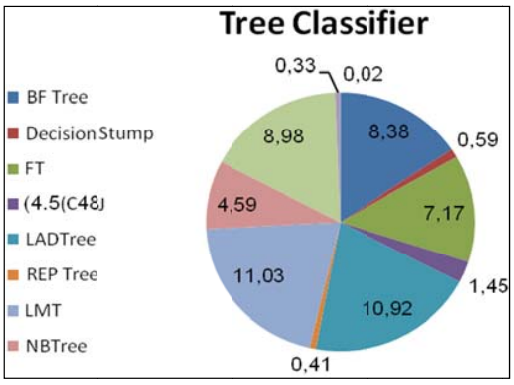

Fig. 7 Time Taken to Build Model of M2 (Seconds)

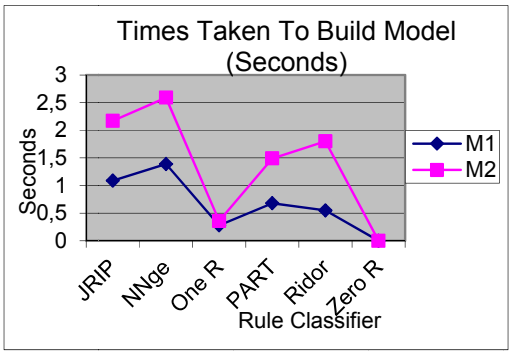

Fig. 8 Time Taken to Build Model of M1 and M2 (Seconds)

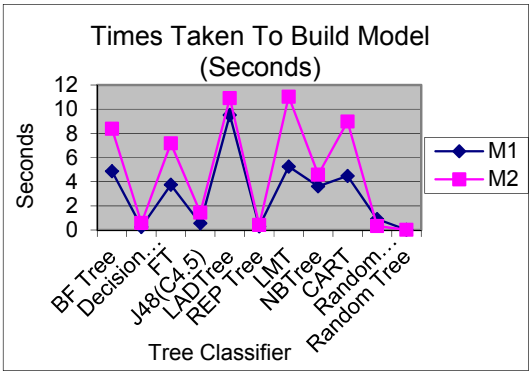

Fig. 9 Time Taken to Build Model of M1 and M2 (Seconds)

C.Analysis of Classification Algorithms Results

This study, has been examine the performance of different classification methods that could generate accuracy and predict best model to disease diagnosis the data set. According to Figs. 2 and 3; Tables VII-X, we can clearly see the highest accuracy is 100% belongs to NNge and lowest accuracy is 65.27% that belongs to Zero R. The total time required to build the model is also a crucial parameter in comparing the classification algorithm. Based on Figs. 8 and 9; Tables VII-X, we can compare time taken to build model among different classifiers in WEKA. We clearly find out that Zero R is the best, second best is the Random tree. An algorithm which has highest accuracy and lowest time to build model will be preferred as it has more powerful classification capability and ability in terms of medical and bioinformatics fields.

Based on Figs. 10 and 11, we can clearly see that the NNge methods is best comparatively other classifiers cause 100% accuracy achieved by NNge and take time to build model is around 1.39-2.59 seconds that also lowest time compare to others. In fact, the highest accuracy belongs to the decision tree classifier by Random tree has the highest percentage of accuracy is 100% and has the longest model building time which is around 0.02 seconds and the best measure among all.

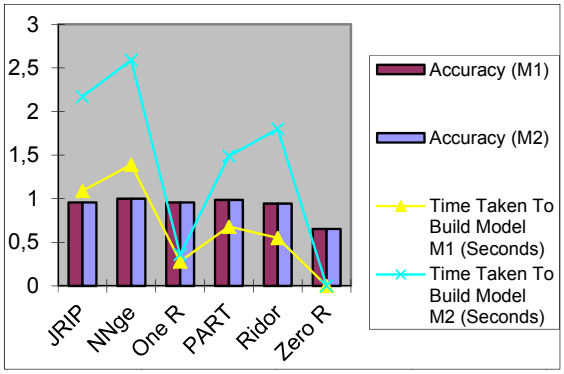

Fig. 10 Accuracy, T1 and T2 of Rule Classifiers

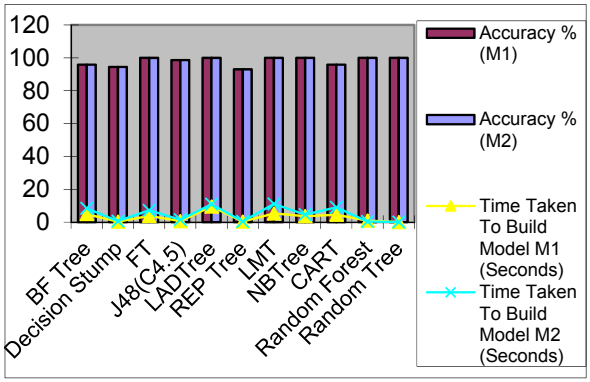

Fig. 11 Accuracy, T1 and T2 of Tree Classifiers

V. Conclusion

In this work, focuses on finding the right algorithm for classification of data that works better on diverse data sets, we have met our objective which is used to evaluate and investigate seventeen selected classification algorithms based on Weka tool to predict of best model of leukemia diseases. The best algorithm based on the Leukemia data is Random Tree classifier with an accuracy of 100% and the total time taken to build the model is at 0.02 seconds. These results suggest that among the machine learning algorithm tested because it has the potential to significantly improve the conventional classification methods to be used in medical field or in general, bioinformatics field.

References

[1] B. Rajeswari and Aruchamy Rajini, Survey On Data Mining Algorithms to Predict Leukemia Types, Ijrset Volume 2, Issue 5, (2010).
[2] Sujata Dash, Bichitrananda Patra, B.K. Tripathy. A Hybrid Data Mining Technique for Improving the Classification Accuracy of Microarray Data Set, I.J. Information Engineering and Electronic Business, (2012), 2, 43-50.
[3] Monica Madhukar, Sos Agaian, Deterministic Model for Acute Myelogenous Leukemia Classification, IEEE International Conference on Systems, Man, and Cybernetics (2012).
[4] Schena M, Shalon D, Davis RW, Brown PO, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science. 1995 Oct 20; 270(5235):467-70.
[5] Mutch DM, Berger A, Mansourian R, Rytz A, Roberts MA, Microarray data analysis: a practical approach for selecting differentially expressed genes,Genome Biol. (2001); 2(12):PREPRINT0009.
[6] Arma R, Marcos IL, Taboada V, Ucar E, Irantzu B, Fullaondo A, Pedro L, Zubiaga A. Microarray analysis of autoimmune diseases by machine learning procedures. IEEE Trans Inform Biomed (2009);13(3):341–50.
[7] CRISTINA OPREA, Performance evaluation of the data mining classification methods, Information society and sustainable development, (2014).
[8] Arunkumar Sivaraman, S. Arun Rajesh, Dr.M. Lakshmi, "Optimistic Diagnosis of Acute Leukemia Based On Human Blood Sample Using Feed Forward Back Propagation Neural Network diagnosis system", International Journal of Innovative Research in Science, Volume 3, Special Issue 3, March (2014).
[9] A. Priyanga, S. Prakasam, Effectiveness of Data Mining - based Cancer Prediction System (DMBCPS), International Journal of Computer Applications, Volume 83 – No 10, December (2013).
[10] Jaya Suji. R1, Dr. Rajagopalan S.P, An automatic Oral Cancer Classification using Data Mining Techniques, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10, October (2013).
[11] S. Syed Shajahaan, S. Shanthi, V. ManoChitra, Application of Data Mining Techniques to Model Breast Cancer Data, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 11, November (2013).
[12] Sujata Dash, Bichitrananda Patra, B.K. Tripathy. A Hybrid Data Mining Technique for Improving the Classification Accuracy of Microarray Data Set, I.J. Information Engineering and Electronic Business, (2012), 2, 43-50.
[13] R.M. Chandrasekar Ph.D, V. Palaniammal M.C.A., M. Phil, Performance and Evaluation of Data Mining Techniques in Cancer Diagnosis, IOSR Journal of Computer Engineering (IOSR-JCE),(2013).
[14] Pushpalatha Pujari and Jyoti Bala Gupta, "Improving Classification Accuracy by Using Feature Selection and Ensemble Model", International Journal of Soft Computing and Engineering, ISSN:2231-2307, Vol.2, Issue 2, May (2012).
[15] C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller and etl. "Leukemia dataset "http://archiveorg.com/page/4089630/05-06-2014/http://tunedit.org/repo/mad_ssk/Leukemia.arff (2010).
[16] H. Hu, J. Li, A. Plank, H. Wang and G. Daggard, "A Comparative Study of Classification Methods for Microarray Data Analysis", Proc. Fifth Australasian Data Mining Conference (AusDM2006), Sydney, Australia. CRPIT, ACS, vol. 61, (2006), pp. 33-37.
[17] I. H. Witten, and E. Frank, "Data Mining Practical Machine Learning Tools and Techniques," Second Edition, Morgan Kaufmann Publisher, United States of America, (2005).
[18] Y. Zhao and Y. Zhang, "Comparison of Decision Tree Methods for Finding Active Objects," National Astronomical Observatories, Advances of Space Research, (2007).
[19] Trilok Chand Sharma, Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April (2013)
[20] L. Breiman, "Random Forests" Machine Learning. 45(1):5-32, (2001).
[21] Lior Rokach, Oded Maimon, "Data Mining and Knowledge Discovery Handbook", Chapter 9: Decision Trees (2005).
[22] Kawsar Ahmed, Tasnuba Jesmin, "Comparative Analysis of Data Mining Classification Algorithms in Type-2 Diabetes Prediction Data Using WEKA Approach", Internat. J. Sci. Eng., Vol. 7(2)2014:155-160, October (2014).
[23] W. Iba, & P. Langley, Induction of one-level decision trees. Proc. of the Ninth Inter. Machine Learning Conference (1992). Scotland: Morgan Kaufmann.
[24] S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas,"Local Boosting of Decision Stumps for Regression and Classification Problems", Journal of Computers, vol. 1, no. 4, July (2006).
[25] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees". for Machine Learning, Vol. 59(1-2), pp.161-205, (2005).
[26] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid," ser. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press (1996), pp. 202–207.
[27] Ashish Kumar Dogra, Tanuj Wala, "A Review Paper on Data Mining Techniques and Algorithms", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 5, May (2015).
[28] Mirza Nazura Abdulkarim, "Classification and Retrieval of Research Papers: A Semantic Hierarchical Approach" (2010).
[29] Gaya Buddhinath and Damien Derry, "A Simple Enhancement to One Rule Classification", Department of Computer Science & Software Engineering. University of Melbourne, Australia, (2006)
[30] F. Leon, M. H. Zaharia and D. Galea, "Performance Analysis of Categorization Algorithms," International Symposium on Automatic Control and Computer Science, (2004).
[31] B. Martin. Instance - Based Learning: Nearest Neighbour with generalisation, Department of Computer Science, University of Waikato, Hamilton, New Zealand, (1995)
[32] E. Frank and I. H. Witten, "Generating Accurate Rule Sets Without Global Optimization," International Conference on Machine Learning, pages 144-151, (1998).
[33] B. R. Gaines and P. Compton, "Induction of Ripple-Down Rules Applied to Modeling Large Databases," J. Intell. Inf. System.5(3), pages 211-228, (1995).
[34] I.H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques" ISBN: 0-12-088407-0, (2005)
[35] Powers, David M W. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies 2(1): 37–63, (2011).

[36] Divya Tomar and Sonali Agarwal, A survey on Data Mining approaches for Healthcare, International Journal of Bio-Science and Bio-Technology Vol.5, No.5, pp. 241-266 (2013).
[37] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.I. (1984). Classification and regression trees. Belmont, Calif.: Wadsworth.
[38] Gama, J., Functional trees, Machine Learning, 2004, 55(3):219–250.