International Journal of Latest Trends in Engineering and Technology
Vol.(7)Issue(4), pp.352-359
DOI: http://dx.doi.org/10.21172/1.74.048
e-ISSN:2278-621X

A DATA MINING APPROACH FOR PRECISE DIAGNOSIS OF DENGUE FEVER

M.Bhavani[1] and S.Vinod kumar[2]

Abstract- Dengue is a eviscerate disease common in tropical countries. It is also known as break-bone fever. Dataset for dengue gives information about the patient suffering with the dengue disease. The Dataset consist of attribute like fever, bleeding, metallic taste, Fatigue. The main objective of this study is to calculate the performance of various classification Techniques and compare their performance. The classification techniques used in this study are REP Tree, J48, SMO, ZeroR and Random Tree. The performance of classification techniques were compared by plotting graphs and table. Weka the data mining tool is used for the classification.
Keywords— Data mining, Dataset, Classification, break bone fever, Weka.

I. INTRODUCTION

Data mining is the process of analyzing data from different aspect and converting it into useful information. Data mining is a well- known technique used by health organizations for classification of diseases such as dengue, diabetes and cancer in bioinformatics research. Dengue fever is a painful, debilitating mosquito-borne disease caused by dengue viruses. Nearly 390 million dengue infections occur worldwide each year, with about 96 million resulting in illness. Dengue is classified into two types namely type1 and type2.First one is called classical dengue and the other one is called dengue hemorrhagic fever(DHF).Dengue hemorrhage Fever is further classified into DHF I,DHF II,DHF III,DHF IV. Dengue Fever transmitted by the bite of an Aedes mosquito infected with a dengue virus. The mosquito becomes infected when it bites a person with dengue virus in their blood and it can be transmitted to other healthy person. It can't be spread directly from one person to another person. Symptoms always starts after four to six days from the day of infection and it will last for 10 days. The symptoms such as sudden fever, severe headaches, pain behind the eyes, Severe joint and muscle pain, Fatigue, nausea, vomiting, skin rash, mild bleeding are common among the patients. Sometimes, symptoms will be mild and it can be predicted as other viral infection such as flu, hence the accurate prediction of the fever is needed. At the beginning stages it is difficult to differentiate the dengue fever and dengue hemorraghic fever. Several Data Mining Techniques are used for the prediction of Dengue Fever.

[1] *Department of Computer Science and Engineering Rajalakshmi Engineering College, Chennai*
[2] *Department of Computer Science and Engineering Rajalakshmi Engineering College, Chennai.*

II. METHODOLOGY

Weka (Waikato Environment for Knowledge Analysis) is a data mining tool written in java developed at Waikato. WEKA is a very good data mining tool for the users to classify the accuracy on the basis of datasets by applying different algorithmic approaches and compared in the field of bioinformatics [6].It is also well-suited for developing new machine learning schemes. Our main objective is to identify that whether the patient is affected by Dengue or not. Some of the parameter are used for predicting the fever and compare the performance of the various classification techniques. The Various result obtained from the dataset using Weka tool are given below.

Figure-1 Weka.

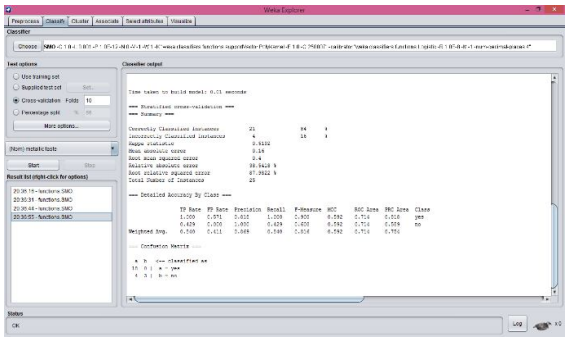

Figure-2.SMO classifier.

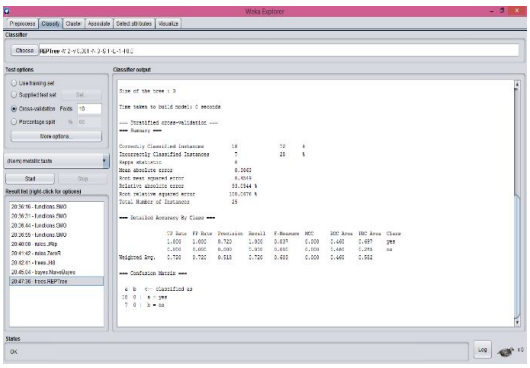

Figure-3: REP Tree classifier.

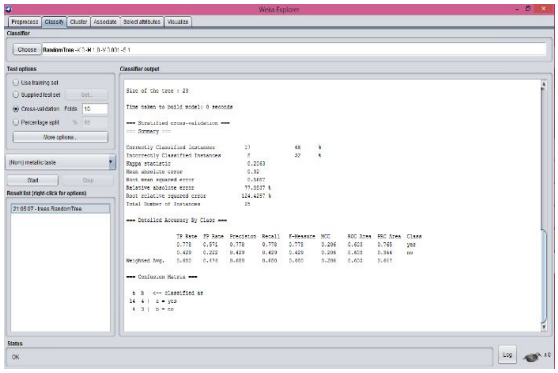

Figure-4.Random Tree classifier.

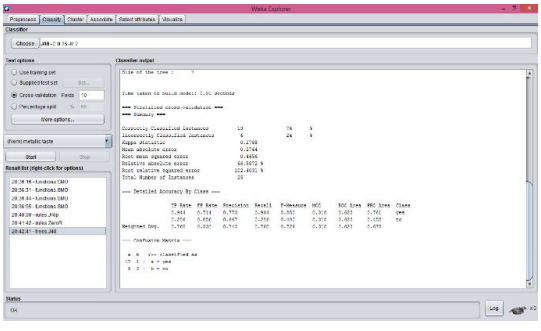

Figure-5 J48 classifier.

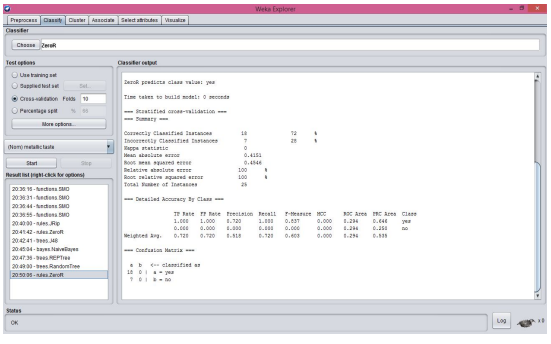

Figure-6 ZeroR classifier.

The main objective of this study is dengue disease prediction using data mining tool. The main task carried out in this study are:

1. Various Data mining classification techniques are used for the Prediction the dengue fever.
2. Comparing different classification techniques.
3. Finding best algorithm for the disease prediction.

A. Classification

It is the organization of data in given classes. Classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

Five techniques have been used in this study. They are REP Tree, RT, J48, ZeroR and SMO. Their performance was analyzed using some measures such as Accuracy, TP Rate, FP Rate, and ROC Area.

B. Dataset

Dataset is a collection of data. The dataset is contained in database table. The dataset is created in the CSV format. The dataset contains the attribute such as fever, bleeding, myalgia, flu, fatigue, pain, metallic taste. The symptoms are used as Attribute.

The dataset loaded in the Weka tool in ARRF format and the performance of the classification techniques are identified and tabulated.

Table -1 ATTRIBUTE DESCRIPTION

Attribute	possible values
EPID	any alpha-numeric values
Fever	Yes or No
Bleeding	Yes or No
Myalgia	Yes or No
Flu	Yes or No
Pain	Yes or No
Joint/Muscle pain	Yes or No
Metallic Taste	Yes or No
Result	Positive or Negative
Fatigue	Yes or No

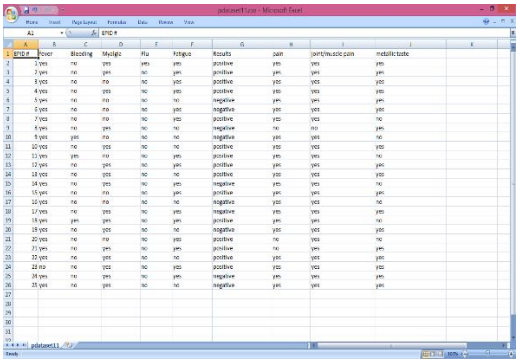

Figure-7 DATASET

III. EXPERIMENTS AND RESULT

There are numerous techniques available for the prediction of the dengue fever. The classification technique used in this study are following:

- REP Tree.
- J48.
- SMO.
- ZeroR.
- Random Tree.

SMO:

Sequential Minimal Optimization (SMO) is one way to solve the SVM training problem that is more efficient than standard QP solvers. The Sequential Minimal Optimization is commonly called as Platt's SMO algorithm and it is well organized with good computational efficiency.

SMO uses heuristics to partition the training problem into smaller problems that can be solved analytically. Whether or not it works well depends largely on the assumptions behind the heuristics (working set selection). We got the outcomes mentioned in the (Table 4).various measures obtained as a result was plotted as a graph (Figure 8).

J48:

C4.5 is the technique used to create a decision ID3 developed by Ross Quinlan. C4.5 is an addition of Quinlan's earlier ID3 Technique. The decision trees created by C4.5 can also be used for classification, and for this purpose, C4.5 is often stated to as an arithmetical classifier [6].

The commonly used method is information gain or entropy measure in which the measure that corresponds to level uncertainty in the information. Thus it is like tree structure with root node, intermediate and leaf nodes. Node holds the decision and helps to achieve our result. The dataset is evaluated in Weka Data Mining tool with J48 technique; we got the outcomes

mentioned in the Table (Table 3).various measures obtained as a result was plotted as a graph (Figure 12).

REP Tree:

REP Tree is a fast decision tree learner. Builds a decision/regression tree using entropy as impurity measure and prunes it using reduced-error pruning. It only sorts values for numeric attributes once. The dataset is evaluated in Weka Data Mining tool with REP Tree technique; we got the outcomes mentioned in the Table (Table 2).various measures obtained as a result was plotted as a graph (Figure 8).

Random Tree:

Random Tree is an ensemble learning algorithm that generates many individual learners. It is an algorithm for constructing a tree that considers K random features at each node. It involves a bagging idea to produce a random set of data for constructing a decision tree. In standard tree each node is split using the best split among all variables. The dataset is evaluated in Weka Data Mining tool with Random Tree technique; we got the outcomes mentioned in the Table (Table 2).various measures obtained as a result was plotted as a graph (Figure 11).

ZeroR:

ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). The dataset is evaluated in Weka Data Mining tool with ZeroR technique; we got the outcomes mentioned in the Table (Table 5).various measures obtained as a result was plotted as a graph (Figure 10).

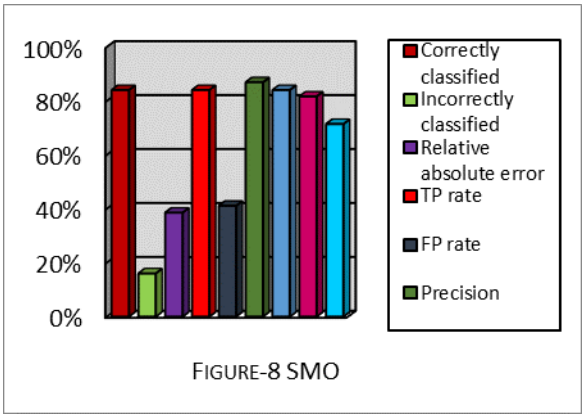

FIGURE-8 SMO

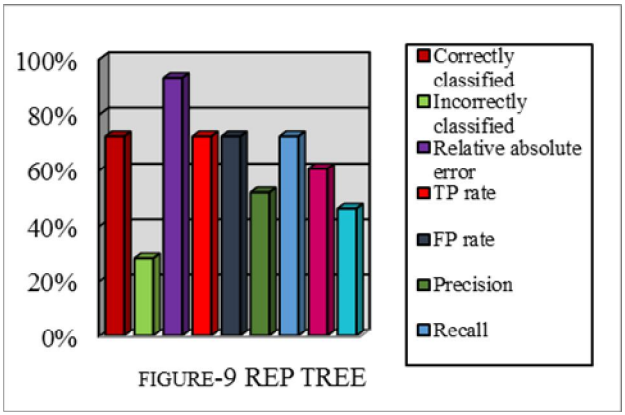

FIGURE-9 REP TREE

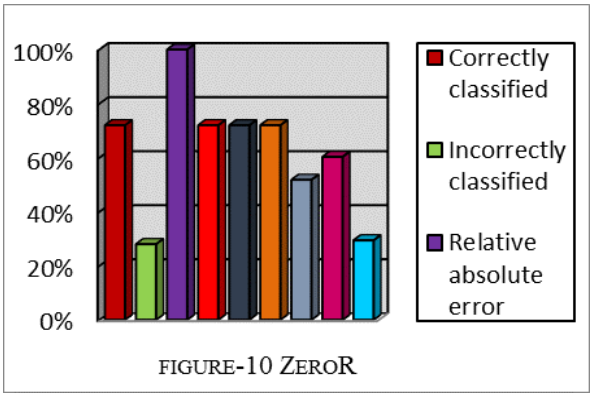

FIGURE-10 ZEROR

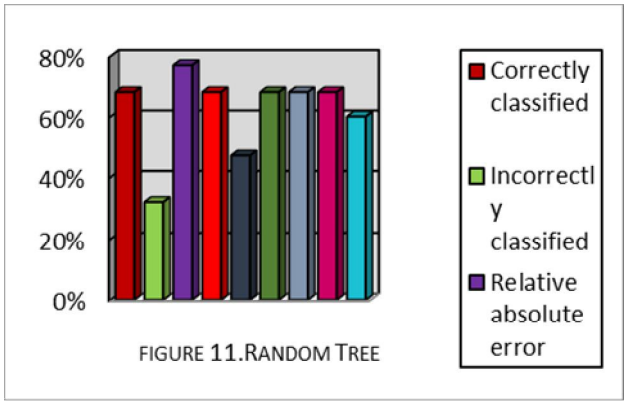

FIGURE 11.RANDOM TREE

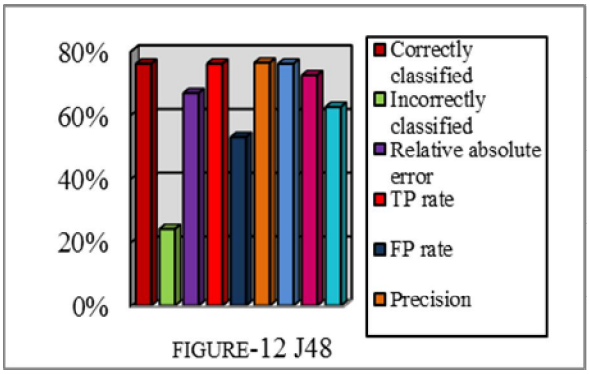

FIGURE-12 J48

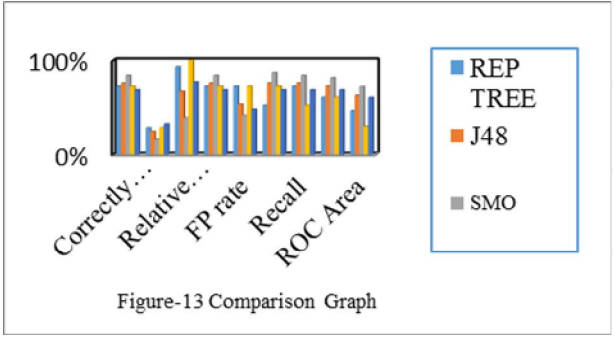

Figure-13 Comparison Graph

Table-2 REP TREE

Attributes name	measure
Correctly classified	72%
Incorrectly classified	28%
Relative absolute error	93.05%
TP rate	0.72
FP rate	0.72
Precision	0.518
Recall	0.72
F-measure	0.603
ROC Area	0.46

Table-3 J48

Attributes name	**measure**
Correctly classified	76%
Incorrectly classified	24%
Relative absolute error	66.59%
TP rate	0.76
FP rate	0.53
Precision	0.763
Recall	0.76
F-measure	0.724
ROC Area	0.623

Table-4 SMO

Attribute name	measure
Correctly classified	84%
Incorrectly classified	16%
Relative absolute error	38.54%
TP rate	0.84
FP rate	0.411
Precision	0.869
Recall	0.84
F-measure	0.816
ROC Area	0.714

Table-5 ZeroR

Attribute name	measure
Correctly classified	72%
Incorrectly classified	28%
Relative absolute error	100%
TP rate	0.72
FP rate	0.72
Precision	0.72
Recall	0.518
F-measure	0.603
ROC Area	0.294

Table 8 PERFORMANCE COMPARISON

Techniques	TP Rate	ROC Area	FP Rate	Accuracy
SMO	0.84	0.714	0.411	0.84
J48	0.72	0.623	0.53	0.76
REP Tree	0.72	0.46	0.72	0.72
ZeroR	0.72	0.294	0.72	0.72
Random Tree	0.68	0.6	0.47	0.68

Table-6 RANDOM TREE

Attribute name	measure
Correctly classified	68%
Incorrectly classified	32%
Relative absolute error	77.08%
TP rate	0.68
FP rate	0.474
Precision	0.68
Recall	0.68
F-measure	0.68
ROC Area	0.6

Table-7 PERFORMANCE MEASURES

Attribute Name	Description
Correctly classified	percentage of test instance correctly classified
Incorrectly classified	percentage of test instance incorrectly classified
TP Rate	true positives: number of examples predicted positive that are actually positive
FP Rate	false positives: number of examples predicted positive that are actually negative
ROC area	It is the useful technique for visualizing and selection of classifier based on their performance
Precision	It is the fraction of retrieved instances that are relavant.
Recall	It is the fraction of relavant instances that are retrieved
Accuracy	A measure of predictive model that reflects the proportionate number of time that model is correct when applied to dataset
F-measure	It is harmonic mean of precision and recall.

IV. COMPARISON

In this study, five techniques were used to classify the dengue dataset. After the analysis of each classification techniques, comparison was made between them. Some measures like TP Rate, FP Rate, Correctly classified, incorrectly classified and precision were used for comparison; their values are given in the table.(table 8) and the graph is plotted comparing the values obtained as a result (figure 13). From the comparison made, the SMO and J48 give more accuracy than other classifier.

V. CONCLUSION

The main objective of this study is to predict the dengue fever with high accuracy; five classifiers namely REP Tree, J48, SMO, ZeroR and Random Tree were used for this purpose. These techniques were applied to the dengue dataset using Weka Data mining tool. These techniques were evaluated and their performance was compared. From the analysis, SMO and J48 outperforms well than other algorithms, they achieved the accuracy of 84 % and 76 % respectively. The performance measure used for comparison are listed in the table (Table 7).

REFERENCES:

[1] Kamran Shaukat, et.al," Dengue Fever Prediction: A Data Mining Problem", Data Mining Genomics Proteomics, Vol 6, 2015.

[2] Farooqi W, Ali S, "A Critical Study of Selected Classification Algorithms for Dengue Fever and Dengue Hemorrhagic Fever". Frontiers of Information Technology (FIT), 11th International Conference on IEEE, 2013.

[3] Shameem Fathima, Nisar Hundewale, "Comparison of Classification Techniques-SVM and Naives Bayes to predict the Arboviral Disease-Dengue", *International Conference on Bioinformatics and Biomedicine Workshops on IEEE*, 2011.

[4] Hlaudi Daniel Masethe, Mosima Anna Masethe, "Prediction of Heart Disease using Classification Algorithms", *Proceedings of the World Congress on Engineering and Computer Science* ,Vol 2, 2014.

[5] Tina R. Patil, Mrs. S.S.Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", *international Journal of Computer Science and Applications*, Vol 6, 2013.

[6] Shakil KA et.al. (2015), "Dengue disease prediction using weka data mining tool", arXiv preprint arXiv: 1502.05167.

[7] Stany Leena Princy and A. Muruganandam ," An Implementation of Dengue Fever Disease Spread Using Informatica Tool with Special Reference to Dharmapuri District", *International Journal of Innovative Research in Computer and Communication Engineering*, 2016.

[8] Hlaudi Daniel Masethe, Mosima Anna Masethe , "Prediction of Heart Disease using Classification Algorithms" , *World Congress on Engineering and Computer Science*, Vol 2, 2014.

[9] Jyoti Rohilla, Preeti Gulia, "Analysis of Data Mining Techniques for Diagnosing Heart Disease", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 5, 2015.

[10] Dr.T.Christopher, et.al, "Study of Classification Algorithm for Lung Cancer Prediction", *International Journal of Innovative Science, Engineering & Technology*, volume 3, 2016.