Performance Evaluation of Classification Algorithms in Predicting Hepatitis Virus

Musa Wakil Bara[1]

Yusuf Abubakar[2]

Mohammed Abubakar[3]

[1,3] *Department of Computer Science, Mai Idris Alooma Polytechnic P.M.B. 1020 Geidam, Yobe State. +2348065845924,* asadiq4all@yahoo.com

[2]*Department of Computer Science, Nuhu Bamalli Polytechnic, Zaria Kaduna State. +2348036423364, yusufabukausar@gmail.com*

Abstract

Nowadays, the number of hepatitis victims is increasing daily and the chance of the survivability of the patients is becoming difficult to predict. Hepatitis is a disease that is caused by many factors such as alcohol consumption, eating polluted foods, drugs and so on. Due to the increase of the number of hepatitis patients and various causes of the disease, doctors find it difficult to accurately diagnose and predict the disease. Researchers have applied data mining methods to extract valuable information from hepatitis database to help in predicting the presence of the disease. In this paper, five (5) classification algorithms were selected to accurately predict the presence or absence of the disease, and to compare their performance against RandomForest which recorded the highest performance in the literature. The experiment was conducted on hepatitis dataset obtained from UCI machine learning repository using WEKA data mining tool. IBk classification algorithm outperformed the rests with accuracy of 88.80%, recall of 98.80% and precision of 83.20%. However, RandomForest still remains the favourite in terms of accuracy. It is recommended that RandomForest algorithm can be used in hepatitis disease prediction where accuracy is preferred. IBk algorithm is recommended in applications that prefer recall while MODLEM is recommended in applications that give preference to precision.

Keywords: *Hepatitis, WEKA, Data Mining, Classification*

1. INTRODUCTION

Hepatitis is a liver disease that is caused by many factors such as alcohol consumption, smile harm gases, eating polluted foods, vinegar and drugs. The disease kills many people every year [8]. Accurate diagnosis of hepatitis increases the chance of living of the patients. There are different laboratory factors (based on enzymes level in blood) used by doctors to diagnose hepatitis. But no better way offers to predict chance of hepatitis prediction. So providing suggestions to predict hepatitis patient's survivability is very important to doctors. Application of data mining techniques on medical issues has been a popular technique of late [6, 7]. Data mining algorithms play a significant role in prediction and diagnosis of the diseases [6]. The growth of medical databases is very high; this rapid growth is the main motivation for researchers to mine useful information from

these medical databases. As the volume of stored data increases, data mining techniques play an important role in finding patterns and extracting knowledge to provide better patient care and effective diagnostic capabilities. Researchers used different data mining techniques in the prediction of hepatitis disease in order to help the medical professionals to easily identify and give good diagnosis. Classification algorithms are mostly used in predicting hepatitis disease.

In the literature, it was noted that most of the researchers evaluate the performance of their work based on accuracy of the algorithm(s) used [10] while some consider accuracy and time [6]. However, accuracy alone or accuracy and time (in some cases) are not the only measure of the performance of classification algorithm.

In this research, the performance evaluation will be based on accuracy, precision and recall. The research is aimed to use some selected classification algorithms and compare their performance in predicting hepatitis.

2. RELATED WORKS

G. Sathya Devi [1] proposed the application of CART algorithm in Hepatitis Disease Diagnosis using decision trees C4.5 algorithm, ID3 algorithm and CART algorithms. It classifies the hepatitis diseases and compares the effectiveness, correction rate among them. From that the CART derived model showed the extended definition for identifying (diagnosing) hepatitis disease provided a good classification accuracy of 83.2%.

Fadl Mutaher et al. [2] presented the comparative analysis in the prognostic of hepatitis data using Rough set technique over Multi-Layer Neural Network using back-propagation algorithm. The prediction of the outcome is more specific and accurate using Rough set technique. Performance and time taken to run the hepatitis data is fast in Naive Bayes algorithm. The results obtained were compared with other algorithms like, Naive Bayes up-datable algorithm, FT Tree algorithm, Kstar algorithm, J48 algorithm, LMT algorithm and neural network. Attributes were fully classified and the result obtained was of 86.52%. Based on the experimental results the classification accuracy is found to be better using Naïve Bayes algorithm compared to other algorithms.

Pushpa latha and Pandya reviewed different Data Mining techniques which are used to diagnosis Hepatitis disease and shows the performance of different Data Mining techniques which were implemented [3].

Ramasamy, M. et al. [4] Highlights the performance of seven decision tree classification algorithms (Decision Stump, Hoeffding Tree, J48, LMT, Random Forest, REP Tree and Random Tree) on hepatitis prognostic dataset aimed at enabling the classifiers to accurately carryout categorization of medical data. The classification accuracies are evaluated using 10-fold cross validation technique. The results show that Random Forest outperformed all other classifiers with an overall accuracy of 87.5%

Even though the researchers have tried in comparing the performance of classifiers, they however concentrated on accuracy only to evaluate the models. Performance evaluation of

classifiers does not only rely on accuracy, but there are other factors to be considered which include precision, recall, F-measure, Area Under Curve (AUC) among other. In [1], X-transformed mouse liver cancer using association rule mining, and the research accuracy was 78.9%.

The rest of the paper is organized as follows: Methodology, Experiment and Results, Conclusion and References.

3. METHODOLOGY

In this section, the dataset used, proposed system design and the selected algorithms to be used are discussed.

A. Hepatitis Dataset Description

This study conducts experiments with hepatitis dataset which is obtained from UCI machine learning repository. The dataset contains 155 instances, 19 independent attributes and the class attribute stating the life prognosis YES or NO. The main goal of the dataset is to predict the presence or absence of hepatitis virus and to train the classifiers to classify new dataset. The detailed description of the training dataset is listed in Table 1.

Table 1 Hepatitis Dataset Description

Attributes	Data Type	Values
Class	Categorical	Die, Live
Age	Numerical	Numerical Values
Gender	Categorical	Male, Female
Steroid	Categorical	Yes, No
Antivirus	Categorical	Yes, No
Fatigue	Categorical	Yes, No
Malaise	Categorical	Yes, No
Anorexia	Categorical	Yes, No
Liver_Big	Categorical	Yes, No
Liver_Firm	Categorical	Yes, No
Spleen_Palpable	Categorical	Yes, No
Spiders	Categorical	Yes, No
Ascites	Categorical	Yes, No
Varices	Categorical	Yes, No
Bilirubin	Numerical	0.39, 0.80, 1.20, 2.00, 3.00, 4.0033,
Alk_Phosphate	Numerical	80, 120, 160, 200, 250
SGOT	Numerical	13, 100, 200, 300, 400, 500
Albumin	Numerical	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Protime	Numerical	10, 20, 30, 40, 50, 60, 70, 80, 90
Histology	Categorical	Yes, No

B. Proposed System Design

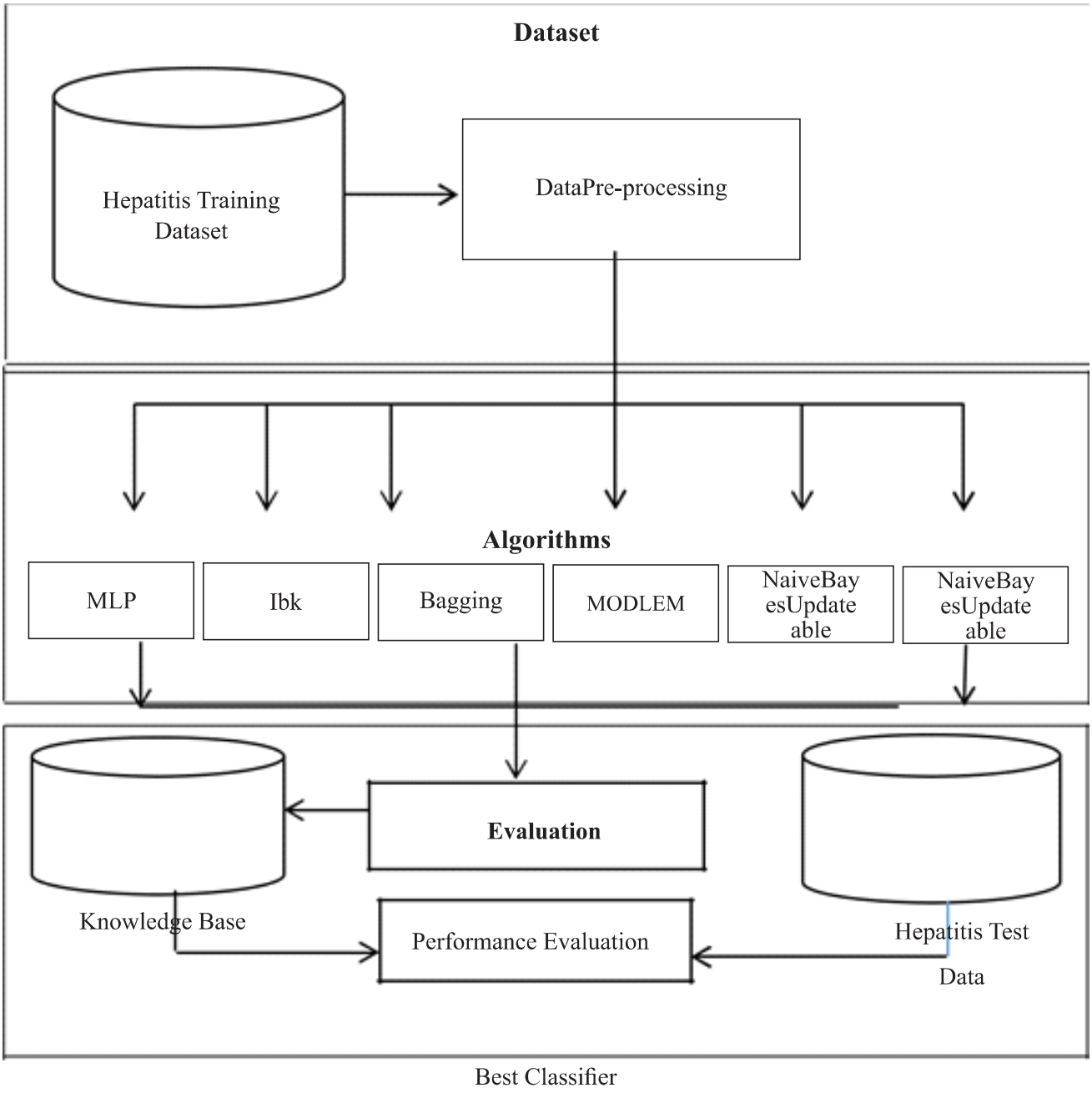


Figure 1 Proposed System Design

C. Classifiers

This section describes the classification algorithms used from each classifier to construct the models. A total of 6 classification algorithms have been used in this research. The classifiers in WEKA have been categorized into the following groups Bayes, Functions, Lazy, Rules, Meta, Misc and Tree classifiers etc. A good mix of algorithms has been chosen from these groups that include:

i. Function (MultilayerPerceptron)

ii. Lazy (IBk)

iii. Decision Tree (RandomForest)

iv. Naïve Bayes (NaïveBayesUpdatable)

v. Meta (Bagging)

vi. Rules (MODLEM)

Multilayer Perceptron

Multilayer Perceptron (MLP) is a non-linear classifier based on the perceptron. The learning rule for the Multi-Layer Perceptron is named as Back Propagation Rule (also known as Generalized Delta Rule). MLP is a back propagation neural network with one or more layers.

IBk

IBk is a k-nearest-neighbor classifier that uses the same distance metric. K-NN is a type of instance based learning or lazy learning where the function is only approximated locally and all computation is deferred until classification. The final classification in this algorithm is decided by a majority vote of its neighbors. IBk supports numeric class problems by calculating the average target value of the nearest problems.

This algorithm is one of the highly accurate machine learning algorithms that involves no learning cost and builds a new model for each test. The testing may become costly if the number of instances in the input data set increases.

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. Each classification tree gives a certain number of votes for each class. Among all the trees, the algorithm chooses the classification with the most number of votes.

Random forest runs efficiently on large datasets but is comparatively slower than other algorithms. It can effectively estimate missing values and hence is suitable for handling datasets with large number of missing values.

Naïve Bayes Updateable

The Bayesian classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a just way by determining probabilities of the outcomes. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms such as Naïve Bayes, naïve Bayes kernel and Naïve Bayes updateable.

The Naive update calculates the prior probabilities of a class, and the conditional prior probabilities for a set of features given a class.

Bagging

Bagging is a "bootstrap" ensemble method that creates individuals for its ensemble by training each classifier on a random redistribution of the training set. Each classifier's training set is generated by randomly drawing, with replacement, *N* examples - where *N* is the size of the original training set; many of the original examples may be repeated in the resulting training set while others may be left out. Each individual classifier in the ensemble is generated with a different random sampling of the training set.

Data Preparation and Preprocessing

As part of data preparation, the data is transformed into a format that is readable by the data mining tool. The dataset was first transformed to comma separated version (CSV) using Excel file.

The original data was taken to Excel and transformed to CSV format before taking it to data mining tool (WEKA). The original data contains categorical and numeric data types; but the UCI data contains the whole attributes in numeric form. Further study was conducted to see how the data can be

formatted to its original form.
The original dataset in CSV format was preprocessed in WEKA. Data preprocessing is applied on the original data to remove noisy and inconsistent data.

- **EXPERIMENTAL RESULTS**

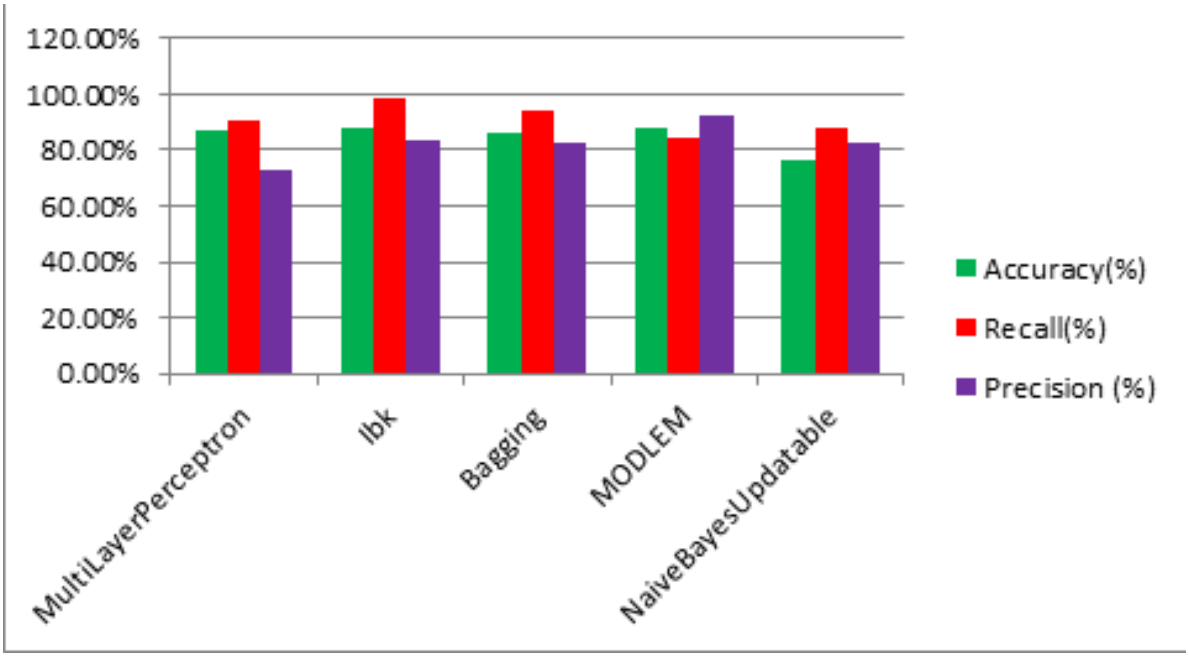
The dataset was taken to WEKA and the experiment was conducted using set of classification algorithms. The result of each classifier is presented in Table 2. A 10-fold cross validation was used to evaluate each classifier and confusion matrix was used to measure the efficiency of the algorithms. The aim of the experiment is to obtain result that will outperform that of RandomForest algorithm which recorded high performance in the literature. The RandomForest has the accuracy of 86.52%, recall of 91.80% and precision of 89.0%.

Table 2 Results

Classifier	Algorithm Used Accuracy(%)	Recall(%)		Precision (%)
Funtion	MultiLayerPerceptron	87.10%	91%	73%
Lazy	Ibk	88.30%	98.80%	83.20%
Meta	Bagging	85.80%	94.10%	82.50%
Rules	MODLEM	87.70%	84.70%	92.30%
Bayes	NaiveBayesUpdatable	76.40%	87.60%	82.30%

From Table 2, the accuracy, recall and precision of all algorithms are presented. It showed that IBk algorithm from Lazy classifier outperformed the rest of the algorithms with overall accuracy of 88.30%, recall of 98.80% and precision 83.20%. In the literature however, it showed that RandomForest algorithm outperformed the rest of the classification algorithms in terms of accuracy. But in our experiment, it was discovered that other algorithms performed better than RandomForest in terms of recall and precision.

The result was presented in the Figure 6.



5. CONCLUSION

In this research, performance of some selected classification algorithms was investigated to predict the survivability of hepatitis patients. The dataset used was obtained from UCI machine learning repository and WEKA data mining tool was used for the experiment. The result showed that IBk algorithm has the highest accuracy of 88.30% over the rests. The precision and recall of each algorithm was also measured to evaluate their efficiency. In terms of recall, IBk still has the highest percentage of 98.80% while MODLEM has outperformed the rest in terms of precision, it has 92.30%. It can be concluded that, RandomForest algorithm can be used in hepatitis disease prediction where accuracy is preferred.

IBk algorithm is recommended in applications that prefer recall while MODLEM is recommended in applications that give preference to precision.

REFERENCES

(D) Park, S. H., Lee, S. M., Kim, Y. J., & Kim, S. (2016). ChARM: Discovery of combinatorial chromatin modification patterns in hepatitis B virus X-transformed mouse liver cancer using association rule mining. BMC bioinformatics, 17(16), 452.

(E) G.Sathyadevi, *Application of CART Algorithm in Hepatitis Disease Diagnosis,*

2011 IEEE, 1283-1287

(F) Fadl Mutaher Ba-Alwi, H. M. (Volume 4, Issue 8, August-2013). *Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach*. International Journal of Scientific & Engineering Research, 680-685.

(G) S. Pushpalatha, and J. Pandya, "Data model comparison for Hepatitis diagnosis," India, July 2014.

(H) Ramasamy, M., Selvaraj, S., & Mayilvaganan, M. (2015, March). An empirical analysis of decision tree algorithms: Modeling hepatitis data. In *Engineering and Technology (ICETECH), 2015 IEEE International Conference on* (pp. 1-4). IEEE.

(I) Kumar, N. K., & Vigneswari, D. (2019). Hepatitis-Infectious Disease Prediction using Classification Algorithms. *Research Journal of Pharmacy and Technology*, *12*(8), 3720-3725.

(J) Javad Salimi Sartakht, J. S. (2011). *Hepatitis disease diagnosis using a novel hybrid method*. Elsevier, 570-579.

(K) Baylis, Philip. "Better health care with data mining", SPSS White Paper, UK, 1999.

(L) Hosseinkhah, Fatemeh, Hassan Ashktorab, Ranjit Veen, and Mohammad Owrang Ojaboni, "Challenges in Data Mining on Medical Databases", pp. 1393-1404, 2009.

(M) Mazaheri, P., Norouzi, A., & Karimi, A. (2015). Using algorithms to predict liver disease Classification. *Electronics Information & Planning*, *3*.

(N) M. H. M. Adnan, W. Husain, N. A. Rashid "Data Mining for Medical Systems: A

Review", *Proc. International Conference on Advances in Computer and Information Technology*, 2012.

(O) S. Bahramirad, A. Mustaphaa, M. Eshraghi "Classification of liver disease diagnosis: A comparative study", *Proc. Second International Conference on Informatics and Applications (ICIA)*, pp. 42-46, 2013.

(P) Thair Nu Phyu, "Survey of Classification Techniques in DataMining", Proceedings of the International Multi Conference of Engineers and Computer Scientists, 2009 (I), IMECS 2009, Hong Kong.

(Q) T. Karthikeyan, "Analysis of Classification Algorithms Applied to Hepatitis Patients," *International Journal of Computer Applications,* vol. 62, 2013.