EFFECT OF ENSEMBLES ON SELECTED DATA MINING TECHNIQUES IN CLASSIFICATION OF LIVER DISEASES

Ibraheem Oluwashola RAHEEM[1], Abdulrauph Olanrewaju BABATUNDE[2],
Taye Oladele ARO[3], Abdullahi Yola MUSA[4]

[1]University of Ilorin, Department of Computer Science, Ilorin, Nigeria
[2]University of Ilorin, Department of Computer Science, Ilorin, Nigeria
[3]University of Ilorin, Department of Computer Science, Ilorin, Nigeria
[4]ICT Directorate, Federal University, Kashere, Gombe, Nigeria
[1]oluwashola0404@gmail.com, [2]babatundeao@unilorin.edu.ng, [3]taiwo_774@gmail.com,
[4]abdulmygmail.com

Keywords: *Ensembles Bagging, Boosting, Naïve bayes, K-nearest neighbour, Decision trees, Voting and Stacking.*

Abstract: *The liver is a very important organ in the human body as it carries out vital functions such as clearing of toxins from the blood, metabolizing drugs, makes proteins for blood clotting amongst others. The complexity of the liver makes it easily affected by diseases. Data mining is a method used in the classification of diseases including liver diseases. This study applies three classification algorithms; Naïve Bayes, K-nearest neighbour and decision trees, their bagged and boosted versions, then the algorithms were combined together by ensemble methods of stacking and voting on liver diseases dataset using 10-fold cross validation. Results show that bagging, boosting, voting and stacking the algorithms in the classification of liver diseases do not necessarily increase the classification accuracy, but increase their complexity except the boosted version of Naïve Bayes which shows an increase in classification accuracy when compared with Naïve Bayes. The stacking and voting has a reduced root mean squared error as compared to the other algorithms, while it was observed that C4.5 decision tree algorithm gave the best classification accuracy of all the algorithms used.*

1. INTRODUCTION

The liver is one of the largest and important organ in human body [1]. This organ is situated between the gastrointestinal tract and heart in the body with a variety of crucial functions in the body such as production of proteins, enzymes, detoxification, activities related to metabolism, regulation of cholesterol and blood clotting [2].

The major function of liver of human is to take up nutrients, store them and make the nutrients accessible to other organs of the body.

Also the liver can take up potentially damaging substances like bacterial products or drugs delivered by the portal blood or microorganisms [3]. The liver can be damaged by several many factors like alcohol consumption, nutrition, diabetes, obesity amongst others [4] as these things make potentially damaging substances available to it.

The World Health Organisation (WHO) in their research reported that approximately 3% of the world‘s population are infected with hepatitis C, with about 170 million people chronically infected and 3-4 million are newly infected each year [5].

Computer science and medical fields are nested to provide diagnosis of various human diseases.

Information given by different patients to medical personnel in biomedical diagnosis may include redundant, irrelevant, interrelated symptoms and signs especially when a patient suffers from more than one type of disease of the same category.

It becomes a serious issue for physicians to diagnose perfectly. Accurate results and early detection are achievable by doctors using data mining techniques in health care system [6].

Data mining with computational intelligent algorithms can be used to handle prediction in clinical datasets with multiple inputs [7].

The techniques in data mining have contributed immensely in transforming large data into specific and more relevant information for knowledge discovery and prediction purpose [8].

The introduction of different computational models in the medical field is well known and gaining ascendency as researches are being geared towards it [9].

Data mining approaches in medical domains is increasing rapidly due to the improvement effectiveness of these approaches for classification and prediction [10].

A number of studies is on the application of data mining to the diagnosis, prognosis and classification of liver diseases as a way of making the process faster, cheaper and easier [11][12].

Ensemble methods have been used to improve the accuracy of data mining algorithms [13] and thus this paper investigated the major effects of two ensemble methods, bagging and boosting on three data mining classification algorithms; Naïve Bayes (NB), k-nearest neighbour and C4.5 decision trees.

This paper contains several sections, which includes related work, methodology, results/discussion and conclusion.

2. RELATED WORK

[12] conducted a survey on classifying liver diseases using image processing and data mining techniques.

From the different modalities of imaging considered, it was observed that classification of liver diseases more accurate in computed tomography imaging than the ultrasound imaging. As computed tomography imaging provided a good basis for analyzing the texture of the liver where ultrasound images impose some difficulties in analyzing the liver structure thus making the texture analysis a challenge, though ultrasound imaging is cheaper. The need for a hybrid computer aided diagnostic system that will result in a higher classification accuracy was highlighted.

[14] performed a survey of classification techniques in data mining for analyzing liver diseases, the techniques considered are C4.5, Naïve Bayes, Support vector machine, Back propagation neural network and Classification and Regression tree (CART). The algorithms gave various results based on speed, accuracy, performance and cost, and C4.5 was said to give better results in comparison with the other algorithms.

[9] studied the relevance of data mining for identifying negatively influenced factors in sick groups, various symptoms of liver disorders in alcoholic patients were analyzed and negative influence factors were identified especially excessive alcoholic consumption.

[11] deployed random tree algorithm to classify liver based diseases. The liver disease type being classified into fatty liver disease, Wilson disease, Inherited disease, autoimmune disease and Cholestatic disease, while the dataset used contained neurological, psychiatric, pathological, physical and cognitive features all of which were categorized into

either high, medium, low or medium/low categories.

The paper showed that decision trees are used to model actual diagnosis of liver cancer for surgical and non-surgical treatment.

[1] applied common techniques in data mining for diagnosis and treatment of different diseases of liver disease. The study employed Rapid Miner and IBM SPSS Modeler data mining tools together.

The accuracy of different data mining algorithms such as C5.0 and C4.5 Decision Tree and Neural Network were examined by the two data mining tools for predicting the prevalence of these diseases or early diagnosis of them using these algorithms.

According to the experimental results, the C4.5 and C5.0 algorithms using IBM SPSS Modeler and Rapid Miner tools had 72.37% and 87.91% of accuracy respectively. Neural Network algorithm by using Rapid Miner had the ability of showing more details.

[15] investigated various algorithms; Naïve Bayes, Decision Tree, Multi-Layer perceptron, KNN, random forest and Logistic on Indian lover patient datasets containing 414 liver patients and 165 non-liver patients, thus the classification was to either identify one with liver disease or one not having liver disease.

From experimental results, Naïve Bayes performed better in terms of precision, while in terms of recall and sensitivity the logistic and random forest algorithms were preferable.

3.1 METHODOLOGY

The study employed WEKA data mining tool to carry out analysis.

Three data mining algorithms were used in this work; Naïve Bayes, K-nearest neighbour and C4.5 decision trees.

The algorithms were applied on the liver diseases data, Ensembles of the algorithms are then implemented, the boosted version of the algorithms using Adaboost is also be applied, bagging is also done on the algorithms, the three algorithms were also combined using stacking and voting.

In stacking, logistic regression was used as the meta-learner while in voting, the probabilities of the prediction were averaged.

The individual application of the algorithms were taken as the base results and the boosted version, bagged version, the stacked and the voted algorithms is being compared to the base results and to each other to study their influence on the accuracy of classification and the effect on time taken to build their classification models.

All classifications were evaluated using 10-fold cross validation.

The complete approach of the proposed model is shown in Figure 1.

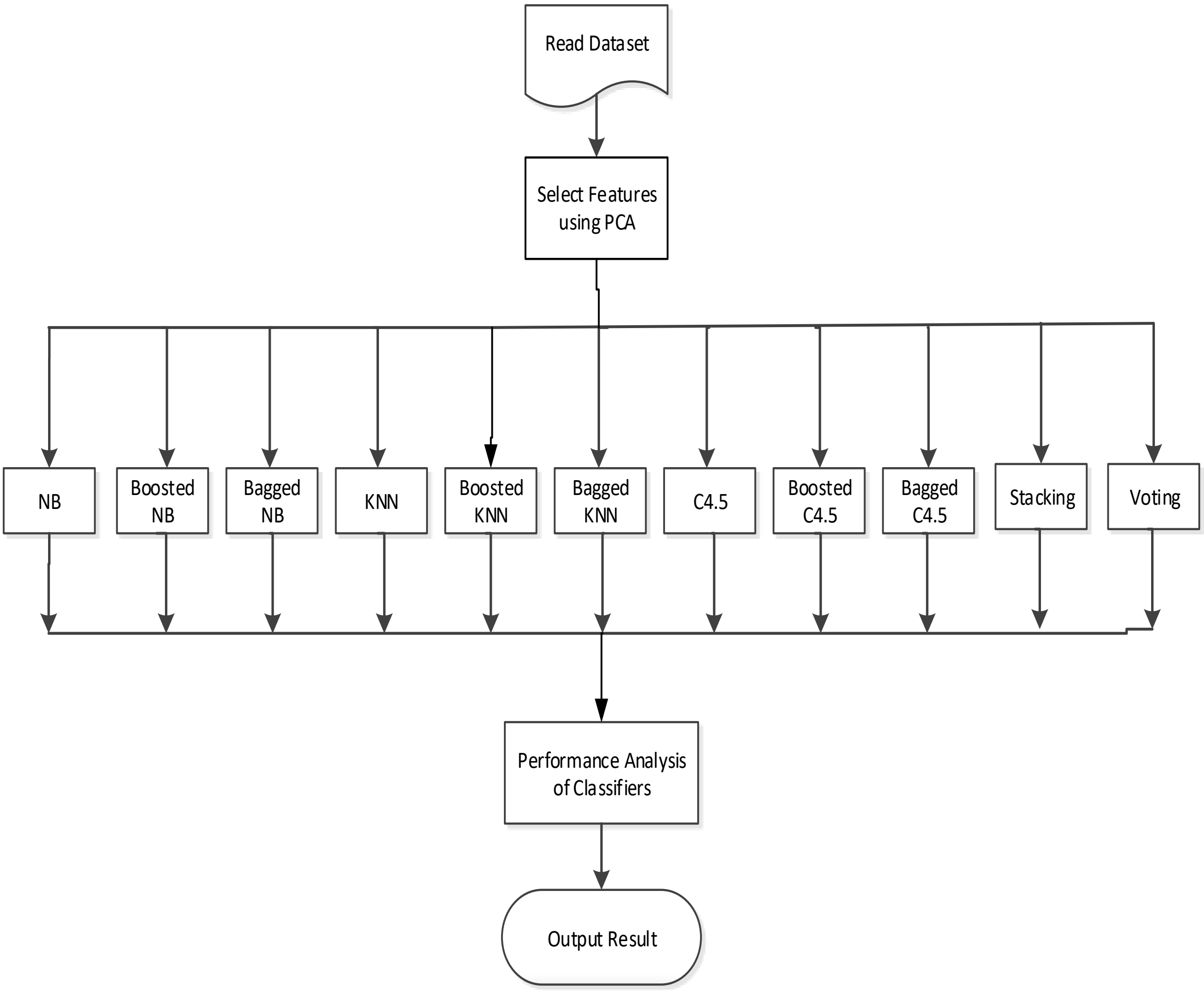

Figure 1: Proposed Model

3.2 Proposed Model Pseudocode

```
Step 1: Collect dataset
Step 2: Transform features using PCA
Step 3: Train classification algorithms NB, KNN and C4.5
Step 4: Apply classification algorithms on test data
Step 5: Train Boosted version of the algorithms
Step 6: Apply Boosted version of the algorithms on the test data
Step 7: Train Bagged version of the algorithms
Step 8: Apply Bagged version of the algorithms on the test data
Step 9: Evaluate and compare results
```

4. RESULTS AND DISCUSSION

Simulations were done by applying the three classification algorithms, NB, KNN and C4.5 on liver diseases dataset, the boosted and bagged version of the algorithms were also applied on the dataset and the algorithms were combined using voting and stacking.

The voting on the three algorithms was combined using their average of probabilities while the stacking was combined using logistic regression and the experimental results are presented in terms of sensitivity, specificity and accuracy as shown in subsection 4.1, 4.2 and 4.3 respectively.

4.1 Sensitivity of Proposed Model

In Naïve Bayes (NB), the boosted version of NB has the highest sensitivity, thus it is able to classify rightly more people with the disease than normal NB or its bagged version, while the boosted version results in a better sensitivity.

The bagged version did not improve the sensitivity significantly.

In KNN, both KNN and its boosted version possessed the same sensitivity while bagged KNN resulted in a reduced sensitivity thus both boosting and bagging did not increase KNN sensitivity.

In C4.5 decision tree classification, boosting and bagging actually reduced the sensitivity of the classification.

Combining the algorithms by voting and stacking the algorithms gave a better sensitivity as compared to NB but lower than KNN and C4.5. Stacking resulted into a much better sensitivity when being compared with the combination through voting as shown in Figure 2.

Overall, C4.5 has the highest sensitivity among all the algorithms.

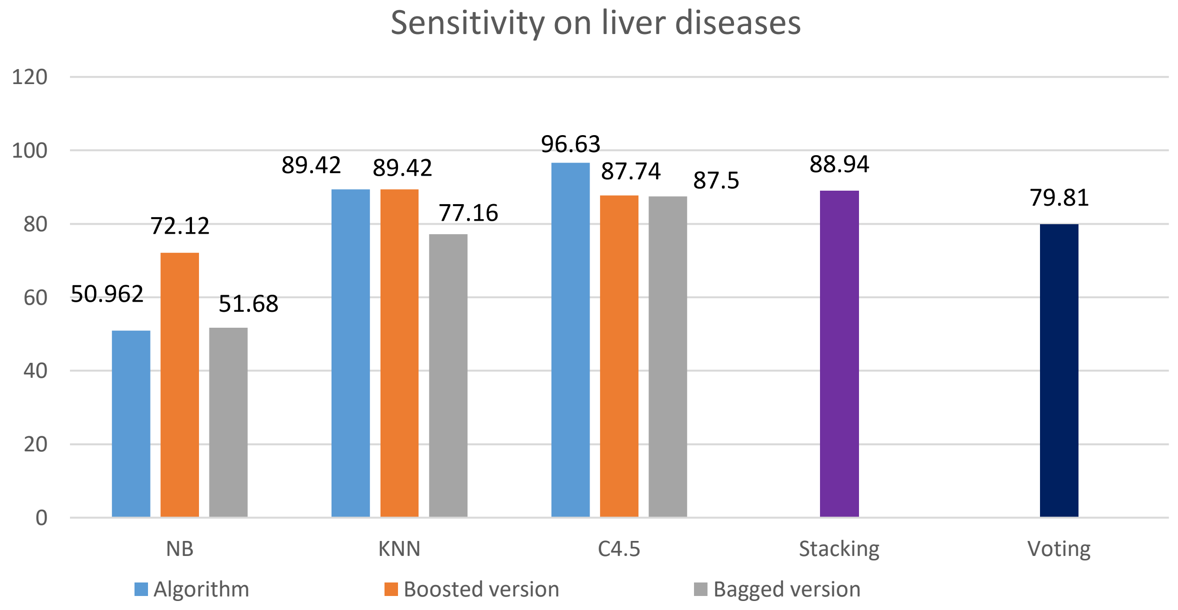

Figure 2: Sensitivity of Proposed Model

4.2. Specificity of Proposed Model

Naïve Bayes (NB) presented the highest specificity among all the classifiers, thus implies that NB predicted better as presented in Figure 3.

In classifying liver disease, comparing the sensitivity and specificity, algorithms with a higher sensitivity tend to have a lower specificity. C4.5 with the highest sensitivity showed a very low specificity.

This shows that it is better classifying a person as having liver diseases at the expense of rightly classifying a person not having the disease.

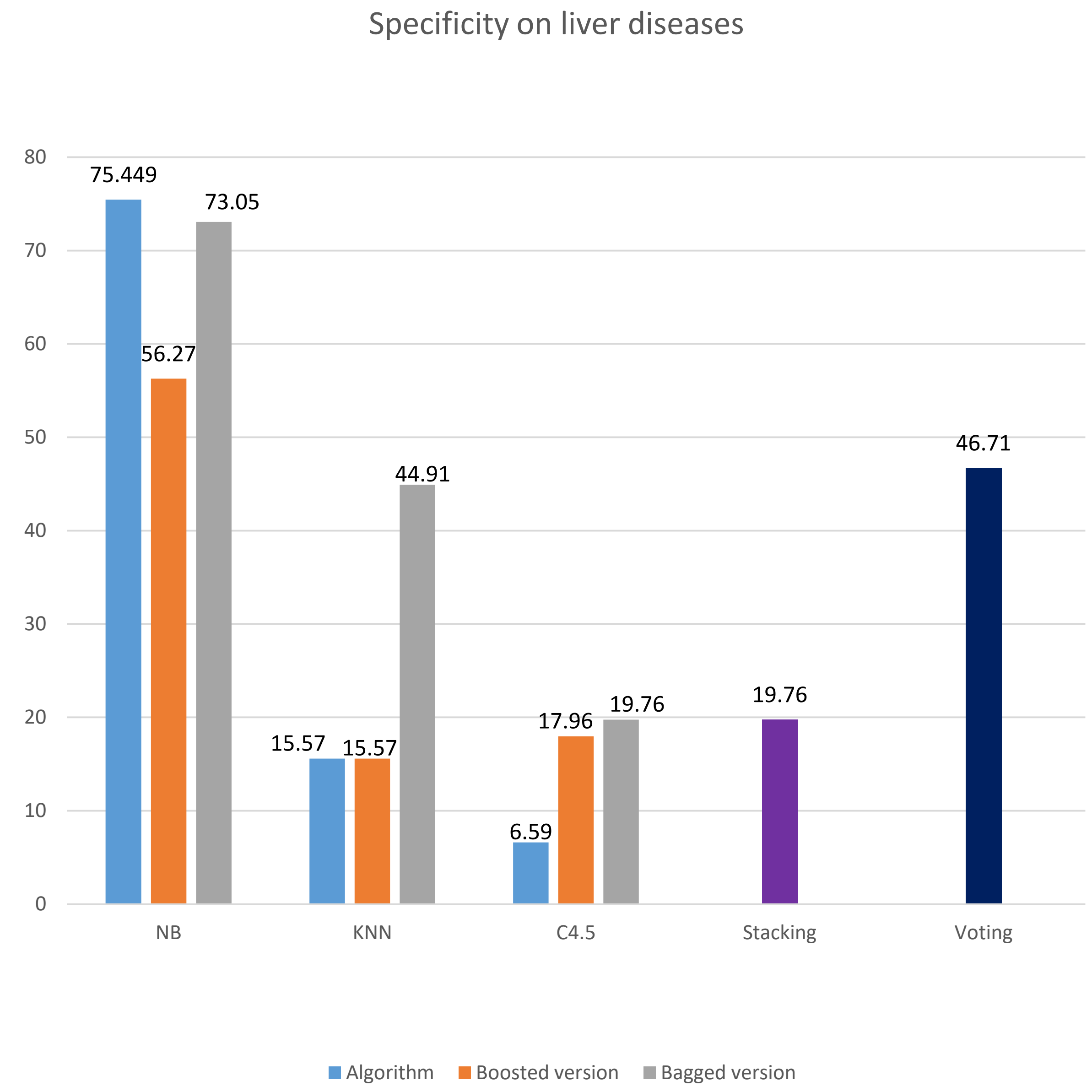

Figure 3: Specificity of Proposed Model

4.3 Classification Accuracy of Proposed Model

An overview of the classification accuracy showed that C4.5 decision tree gave the highest accuracy in classifying whether a patient has liver diseases or not as illustrated in Figure 4.

The bagging and boosting of C4.5 decision tree algorithm actually reduced the classification accuracy of the model, while there was no tangible statistical difference to the classification accuracy of KNN.

In the case of Naïve bayes, Bagging made no difference in the classification accuracy while boosting increased the classification accuracy significantly.

The stacking of the three classifiers was much better in Naïve Bayes, it made no much difference to KNN and with reduction in accuracy as obtained in C4.5 decision tree.

Voting performed slightly better in classification algorithms, their boosted and bagged versions except for C4.5 decision tree.

The incorrectly classified instances can be deduced from the correctly classified instances.

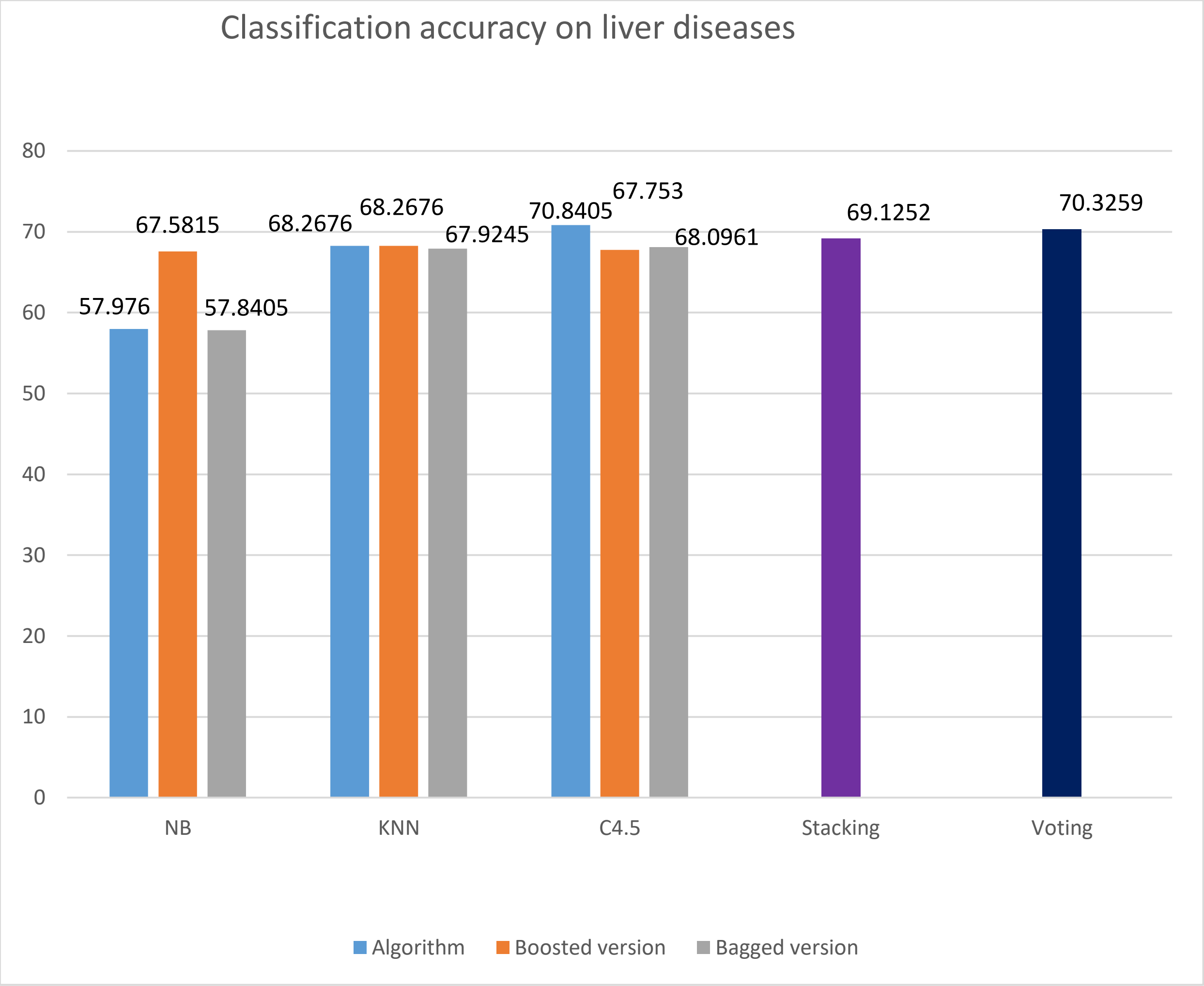

Figure 4: Classification Accuracy of Proposed Model

4.4 Root Mean Squared Error

In Naïve bayes, the root mean squared error was reduced in the boosted version, implying that it has the least error and coupled with its accuracy performance as illustrated in Figure 5.

In KNN, the algorithm and its boosted version produced the same error rate while the bagged version has a higher error rate, while decision tree and its boosted version and bagged versions gave similar error rates though slightly lower.

Stacking and voting produced similar and low root mean squared error meaning their error in prediction is lower as compared to others implying that the confidence in their prediction is more compared to the others, C4.5 decision tree; it's boosted and bagged version error rate is comparable to that of stacking and voting.

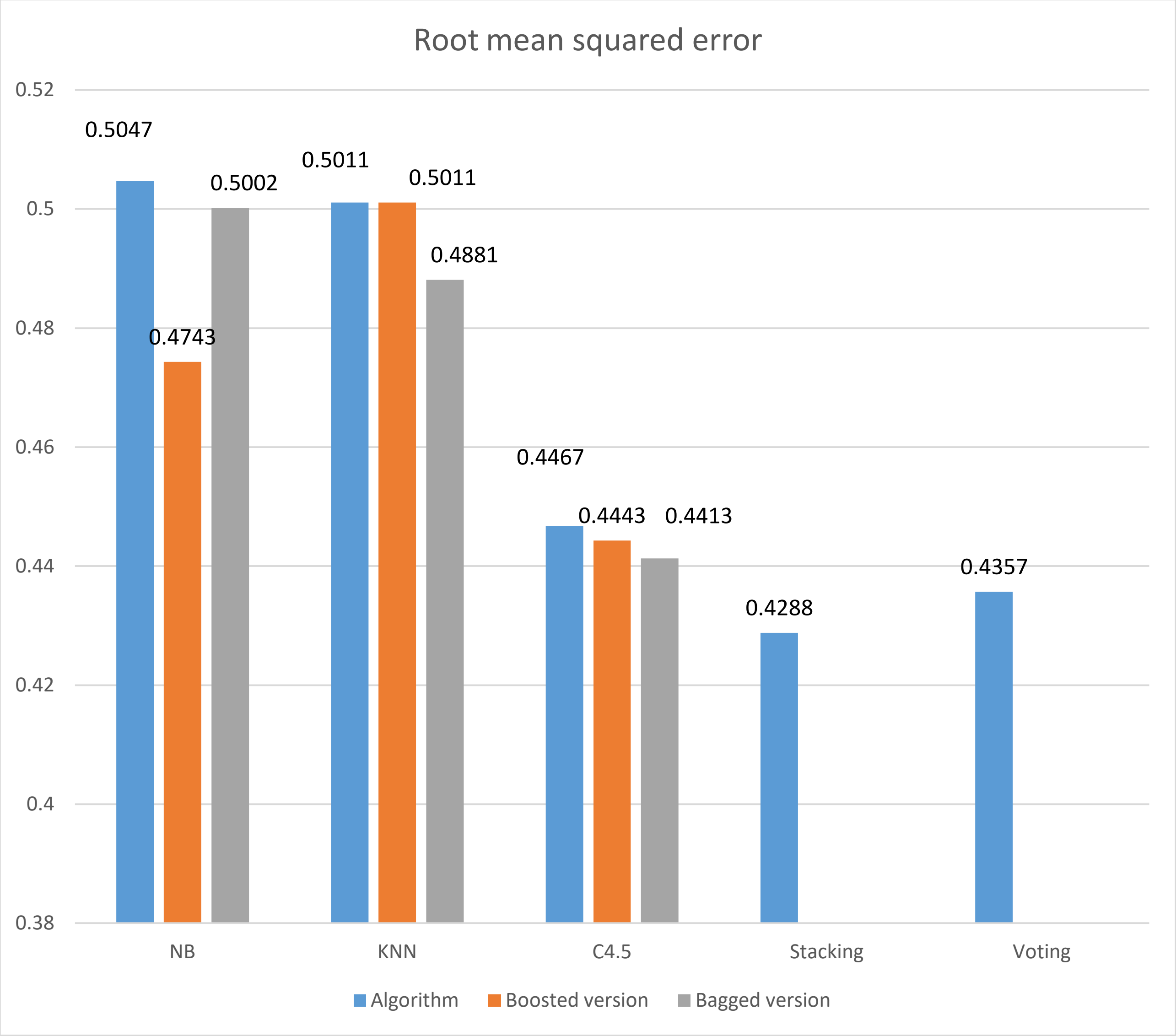

Figure 5: Root Mean Square Error of Proposed Model

5. CONCLUSION

In this paper, it is concluded that ensemble methods by bagging and boosting do not increase the accuracy of algorithms in determining whether a patient has liver diseases or not.

Except in the case of boosting in Naïve bayes which produced better classification accuracy than Naïve Bayes and its bagged version, thus bagging and boosting generally can be said to increase the complexity of the classification of liver diseases without increasing the classification accuracy.

It's better to apply algorithms directly in determining liver diseases than using their ensemble method of bagging and boosting.

Stacking and voting did not improve the algorithms significantly except in comparison with Naïve bayes, which had a lower accuracy compared to the others.

Finally, the overall results showed that C4.5 decision tree algorithm outperformed all other algorithms in term of classification accuracy.

REFERENCES

[1] M. Abdar, "A Survey and Compare the Performance of IBM SPSS Modeler and Rapid Miner Software for Predicting Liver Disease by Using Various Data Mining Algorithms," in *The Second National Conference on Applied Research in Science and Technology*, 2015, vol. 36, no. 2, pp. 3231–3241.

[2] S. Vijayarani and S. Dhayanand, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms," *Int. J. Sci. Eng. Technol. Res.*, vol. 4, no. 4, pp. 816–820, 2015.

[3] G. Ramadori, F. Moriconi, I. Malik, and J. Dudas, "Physiology and pathophysiology of liver inflammation, damage and repair," *J. Physiol. Pharmacol.*, vol. 59, pp. 107–117, 2008.

[4] D. Sindhuja and R. Priyadarsini, "A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder," *Int. J. Comput. Sci. Mob. Comput.*, vol. 5, no. 5, pp. 483–488, 2016.

[5] V. Kirubha and S. M. Priya, "Survey on Data Mining Algorithms in Disease Prediction," *Int. J. Comput. Trends Technol.*, vol. 38, no. 3, pp. 124–128, 2016.

[6] M. Sharma, H. & Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms : A Survey," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 5, no. 8, pp. 99–104, 2017.

[7] D. C. Bindushree, "Prediction of Cardiovascular Risk Analysis and Performance Evaluation Using Various Data Mining Technioques: A Review," *Int. J. Enginnering Res.*, vol. 5013, no. 5, pp. 796–800, 2016.

[8] N. R. Agrawal, P. A & Chopde, "International journal of engineering sciences & research technology a survey on heart disease prediction using soft computing," *Int. J. Eng. Sci. Res.*, vol. 5, no. 3, pp. 582–587, 2016.

[9] C. J. Aneeshkumar, A. S & Venkateswaran, "An Approach of Data Mining for Predicting the Chances of Liver Disease in Ectopic Pregnant Groups." In The International Conference on Communication, Computing and Information Technology , , 19 – 22.," 2012, pp. 19–22.

[10] A. Lebba, A., Sayeth, S., Elankovan, S., & Azuraliza, "Comparative Study on Different Classification Techniques for Breast Cancer Dataset," *Int. J. Comput. Sci. Mob. Comput.*, vol. 3, no. 10, pp. 185–191, 2014.

[11] K. Kaur, P. & Aditya, "Classification of Liver Based Diseases Using Random Tree," *Int. J. Adv. Eng. Technol.*, vol. 8, no. 3, pp. 306–313, 2015.

[12] V. S. Patil, R. V., Sannakki, S. S & Rajpurohit, "A Survey on Classification of Liver Diseases Using Image Processing and Data Mining Techniques," *Int. J. Comput. Sci. Eng.*, vol. 5, no. 3, pp. 23–34, 2017.

[13] S. P. Chaudhari, A. A & Akarte and A. A. Chaudhari, "Fuzzy & Datamining based Disease Prediction Using K-NN Algorithm," *Int. J. Innov. Engineeering Technol.*, vol. 3, no. 4, pp. 9–14, 2014.

[14] D. & P. Sindhuja, "A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder," *Inetrnationa J. Comput. Sci. Mob. Comput.*, vol. 5, no. 5, pp. 483–488, 2016.

[15] H. Jin, S. Kim, and J. Kim, “Decision Factors on Effective Liver Patient Data Prediction,” *Int. J. Bio-science Bio-Technology,* vol. 6, no. 4, pp. 167–177, 2014.