HOMOGENOUS ENSEMBLES ON DATA MINING TECHNIQUES FOR BREAST CANCER DIAGNOSIS

*Taye Oladele Aro[1], Hakeem Babalola Akande[2], Muhammed Besiru Jibrin[3], Usman Abubakar Jauro[3]
[1]Department of Mathematical and Computing Sciences, KolaDaisi University, Ibadan, Nigeria
[2]Department of Computer Science, University of Ilorin, Ilorin, Nigeria
[3]Department of Computer Science, Federal University, Kashere, Gombe, Nigeria.

Email: taiwo774@gmail.com

Abstract: Breast cancer is a disease usually found in women which poses serious health challenges and can be fatal if not diagnosed quickly and treated immediately. Techniques of data mining have been defined to play a significant role in the diagnosis of numerous diseases in which breast cancer diagnosis is a good example. This paper employed homogenous ensemble on methods of data mining for breast cancer diagnosis. Three data mining classification algorithms: k-nearest neighbour, Decision Tree (C4.5) and Support Vector Machines (SVM) with their homogenous ensembles of Bagging and Boosting were applied. The experimental result showed that support vector machines possess the highest classification accuracy, homogenous ensembles of bagging and boosting does not affect classification accuracy greatly as they either slightly increase or reduce the accuracy of classification while increasing the time it takes for the algorithm to build its model.

Keywords: *Breast cancer, Homogenous, Data mining, Healthcare, Prediction*

1. INTRODUCTION

In recent years, breast cancer in the developing countries is on the high increase, which is a great threat [1]. Breast cancer is a type of cancer which happens due to change in normal cells at the breast region of the body as a result of uncontrolled growth of cells which then give rise to tumour(s). The classification of cancer as a disease in healthcare is based on the specific region where the cell or tissue cancer is formed. Early detection remains the most reliable and effective method to reduce death cause by breast cancer. Prompt diagnosis involves a correct and dependable diagnosis method that permits medical doctors knowing the different between benign breast tumours and malignant ones without going for surgical biopsy [2].

The study of cancer in nature can either be biological or medical, data driven statistical study has become a common complement. The prediction of disease outcome is one of the most interesting and challenging tasks on where to develop data mining systems. As computers are being used to handle large volumes of medical data collected by medical researchers, this makes discovery of knowledge in databases which is the data mining method a popular search tool in medicine to recognize and exploit patterns. This makes relationship among large volume of variables able to predict the outcome of a disease based the old cases stored within datasets.

Computational intelligent algorithms with data mining can be employed to handle prediction in clinical datasets with multiple inputs [3]. Techniques in data mining have contributed immensely in transforming large data into specific and more relevant information for knowledge discovery and prediction purpose [4]. Data mining approach in healthcare is a significant component of knowledge discovery in database that is used for the extraction of data. associated with several diseases from dataset in order to facilitate easier prognosis of diseases [5].

This paper investigated the influence homogenous ensembles of data mining techniques has on the diagnosis of breast cancer. Ensemble methods of bagging and boosting which are homogenous because it uses only one algorithm were considered, while the base algorithms used were C4.5 decision tree, k-Nearest Neighbour (kNN) and Support Vector Machines using the radial basis function kernel

2. RELATED WORK

Several studies have been conducted on the performance analysis of data mining techniques in medical diagnosis of breast cancer. These include:

[2] performed a survey on the prominence and usefulness of techniques of data mining. The study applied diverse techniques of data mining such as classification, clustering, Decision Tree, Naïve Bayes. Comparison was done on different data mining techniques from clinical dataset with different accuracy. From the several literatures reviewed, it was observed that existing performance analysis of data mining techniques do not consider the feature selection.

[6] developed a system which used discernibility nearest neighbour classifier, K-means clustering algorithm, and fuzzy rough feature set. The proposed model was compared with previous studies and shown to perform better than others with an accuracy of 98.9%.

[7] presented a new multi layered system by the combination of clustering and decision tree technique was used to build a cancer risk prediction system to provide a cost effective and early warning to the users. A prediction of lung, breast, oral, cervix, stomach and blood cancers was done by the system. The study made use of techniques in data mining like clustering, classification, and prediction to identify potential cancer patients. The developed predictive model projected the breast cancer risk in the earlier stage and also confirmed by comparison of the predicted results with previous medical information of patients.

[8] designed prediction of cancer by application of data mining techniques, the study classified dataset of colon cancer microarray in bioinformatics using five diverse algorithms in classification: Naïve Bayesian, K-Nearest Neighbours, Support Vector Machine, Random Forest and Neural Network. The evaluation for the performance of classification algorithms were achieved in terms of classification accuracy, precision and recall. The experimental result

showed that the highest accuracy was found in both KNN and Neural Network classifier among all other classification algorithms.

[9] conducted a survey comparatively on techniques of data mining for breast cancer diagnosis using neural network, naïve bayes, and C4.5 Decision tree algorithms which was implemented in Weka toolkit. Experimental result indicated that C4.5 Decision tree outperformed other methods.

[10] used concepts and techniques of data mining predominantly in the health care and imaging showing diverse tasks of data mining that are of benefit to diagnosis, decision making, screening, monitoring, therapy support, patient management amongst others thus improving quality and decreasing cost

[11] summarized literatures on diagnosis and prognosis of some diseases. The current research with data mining techniques is to enhance the disease(s) prediction process. The future trends of current techniques of KDD were discussed in using data mining tools for healthcare, important issues and problems applying data mining and healthcare. The result discovered the growing number of applications of data mining, including analysis of health care centre for effective policy-making in health, disease outbreaks detection and preventable hospital deaths.

[12] proposed an approach for detecting breast cancer using techniques of data mining. The study investigated the effectiveness in classification techniques. The breast cancer data with a total 683 rows and 10 columns was to be evaluated using classification accuracy. The study considered the data of breast cancer in the Wisconsin data from UCL machine learning with the purpose of developing accurate predictive model for breast cancer

[13] used classifiers in data mining on the database of breast cancer, by using classification accuracy with and without techniques of feature selection. The feature selection contributes to better accuracy of classifier due to fact that it removes irrelevant features. The experimental result showed that the selection of feature improves the accuracy of all three classifiers, reduces the Mean Standard Error (MSE) and increase Receiver Operating Characteristics (ROC).

[14] improved t the decision-making system for breast cancer management in the Kingdom of Saudi Arabia. This was accomplished through several association rule mining algorithms on the cancer information system in Saudi Arabia. It study provided information useful about predicted distribution and cancer segmentation in Saudi Arabia, which may be linked to possible risk factors. From the extracted patterns, the information require to be reflected in the decision- making phase can be identified as well, which yields to knowledge based decisions. Consequently, knowing health risk behaviour among target group of patients and taking preventive measures can be initiated to reduce the cancer of breast incidence and prevalence ultimately.

[15] presented techniques of diagnosis and prognosis of cancer diseases. The study mentioned that the disease prediction remains the most exciting and challenging tasks to build data mining applications. A survey of the current study was carried out on various types of breast cancer datasets. The accuracy of the three data mining techniques was compared. Experimental results capable for the application of the data mining methods into the survivability prediction problem in medical databases. The performance of decision tree outperformed the other two techniques.

[161] developed breast cancer diagnostic system by applying the fuzzy systems and evolutionary algorithms. Fuzzy rules are desirable because of their human experts interpretability.. Ant colony algorithm was used as evolutionary algorithm to optimize the obtained set of fuzzy rules. Results on breast cancer diagnosis dataset from UCI machine learning repository showed that the system identified cancer instances with high accuracy rate in addition to adequate interpretability of extracted rules.

[17] studied feature selection techniques: rank search, genetic search and greedy step wise search methods to identify the potential features from the Breast Cancer dataset using the diagnosis of heart attack using data mining techniques. Wisconsin Breast Cancer Dataset (WBCD) with. features of 2 through 10 used to represent instances. Each instance has one of 2 possible classes: benign or malignant. Number of instances: 699. The study investigated classification data mining techniques such as BayesNet, Attribute Selected Classifier, J48, Classification via Regression, Logistic, and OneR. The result showed that the three different set of potential attributes were selected using genetic search, rank search and greedy step. It was clearly shown that the performance and time taken by each classification algorithms were significantly enhanced after feature selection and the bayes net classifier outperformed the remaining algorithms used.

[18] used ensemble learning algorithm based on SVM to reduce the variance of diagnosis and increase accuracy of diagnosis. Twelve SVMs, based on the proposed Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE) method, were hybridized. The performance of the proposed model was evaluated using Wisconsin Breast Cancer database. The experimental results showed that model achieved a higher accuracy with a significantly lower variance for breast cancer diagnosis compared to five other ensemble mechanisms and two common ensemble models (adaptive boosting and bagging classification tree). The model reduced the variance by 97.89% and increased accuracy by 33.34%, compared to the best single SVM model on the SEER dataset.

3. METHODOLOGY

Three data mining algorithms; decision tree, KNN and Support Vector Machine were used as the base algorithms while ensemble method of and boosting (Adaboost) were combined with the base algorithms. The dimensionality reduction was done using Principal Component Analysis by selecting and transforming the most relevant features. The Wisconsin original Breast Cancer Dataset was used to evaluate the developed diagnostics model, the dataset contains 669 instances and 10 attributes, it was acquired from the University of California, Irvine repository. All implementations were done using WEKA toolkit, a tool for data mining. The system architecture is shown in Fig. 1.

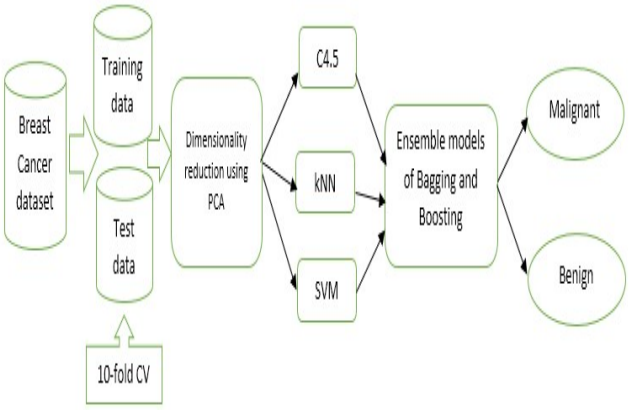

Fig. 1. System Architecture

4. RESULTS AND DISCUSSION

The results of the developed system for a homogenous ensembles using data mining techniques for breast cancer diagnosis are shown in Fig. 2, 3, 4, and 5.

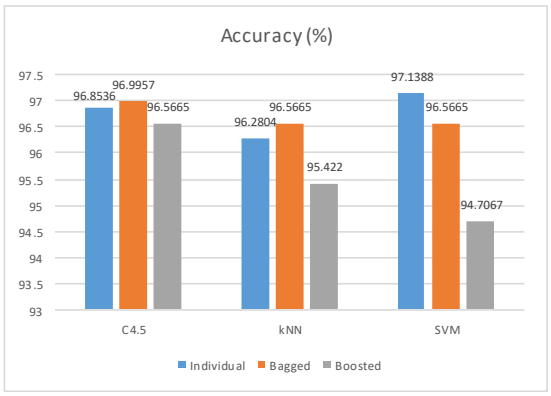

Fig. 2. Accuracy

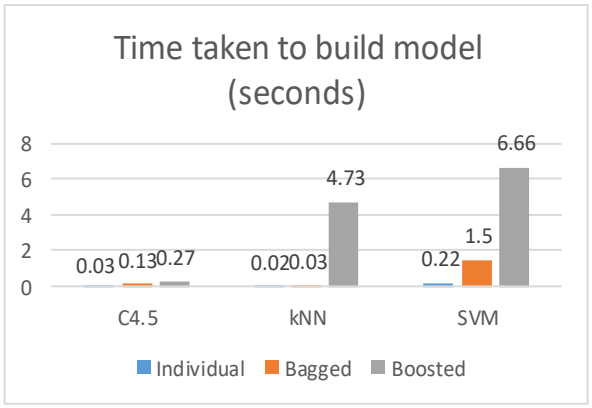

Fig. 3. Time Taken to build model

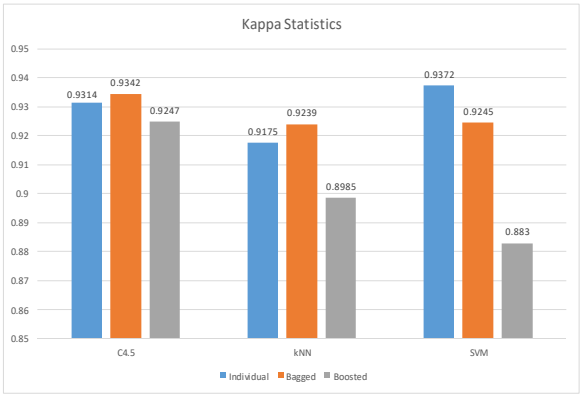

Fig. 4. Kappa Statistics

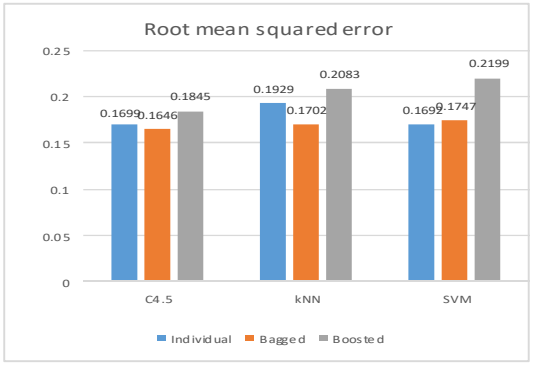

Fig. 5. Root mean squared error

A. Summary and Discussion

Comparing the algorithms individually, SVM has a higher classification accuracy though it slightly perform better than the other two algorithms as the difference between their accuracy is not up to 1%. For C4.5 Decision Tree Algorithm, while the accuracy was increased marginally by bagging, it was reduced a little by boosting but both increases is negligible even at 1% statistical significance, while both bagging and boosting increased the time taken by C4.5 to build its classification model. The effect of time taken to build classification model as that of C4.5 Decision Tree. In the case of SVM, both bagging and boosting reduced the classification accuracy of SVM while increasing the time taken to build its classification model. Holistically, boosting was found to increase the time taken by an algorithm to build its classification model much more than bagging. SVM took the longest time in building its classification model as compared to the other algorithms. The kappa statistic which is a measure of the confidence of how well the classifier is doing better than chance shows a corresponding increase in the classification accuracy. While the root mean squared error shows a decrease with an increasing kappa statistic, implying that the higher the classification accuracy of an algorithm, the higher the kappa statistic and the lower the root mean squared error. Bagging had better kappa statistics for both C4.5 and kNN as the cases brought about a slightly higher classification accuracy with a lower statistic for SVM. Boosting in all cases reduced classification accuracy and thus has a lower kappa statistic. Comparatively, the {18] in their study recorded accuracy of 97.89% as against the developed homogenous ensemble breast cancer diagnosis system, but other evaluation parameters such as Kappa statistic, Time taken to build a model and RMSE were not considered.

5. CONCLUSION

This paper shows that homogenous ensembles of Bagging and Boosting does not bring about an increase in accuracy of algorithms in the diagnosis of breast cancer while it increases the time taken by the algorithms to build its classification model. Bagging was shown to result in a slight increase for C4.5 and kNN while it slightly reduced that of SVM while Boosting reduced classification accuracy in all cases. Thus homogenous ensembles can be said to increase complexity of classification algorithms without bringing about a corresponding increase in classification accuracy of Breast Cancer and thus should not be used by researchers. More work

should focus on heterogeneous ensembles as a way to increase classification accuracy for Breast Cancer.

REFRENCES

[1] V. Chaurasia and S. Pal, "Data Mining Techniques : To Predict and Resolve Breast Cancer Survivability," vol. 3, no. 1, pp. 10–22, 2014.

[2] S. Shukla, D. L. Gupta, and B. R. Prasad, "Comparative Study of Recent Trends on Cancer Disease Prediction using Data Mining Techniques," *Int. J. Database Theory Appl.*, vol. 9, no. 9, pp. 107–118, 2016.

[3] 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.

[4] D. Chandna, "Diagnosis of Heart Disease Using Data Mining Algorithm," *Inetrnational Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 1678–1680, 2014.

[5] O. O. Adeyemo and T. O. Adeyeye, "Comparative Study of ID3 / C4 . 5 Decision tree and Multilayer Perceptron Algorithms for the Prediction of Typhoid Fever," *African J. Comput. ICT*, vol. 8, no. 1, pp. 103–112, 2015.

[6] I. M. El-hasnony, H. M. El-bakry, and A. A. Saleh, "Classification of Breast Cancer Using Softcomputing Techniques," *Int. J. Electron. Inf. Eng.*, vol. 4, no. 1, pp. 45–54, 2016.

[7] K. Arutchelvan and R. Periyasamy, "Cancer Prediction System Using Datamining Techniques," *Int. Res. J. Eng. Technol.*, vol. 2, no. 8, pp. 1179–1183, 2015.

[8] G. Porkodi, R & Suganya, "A Comparative Study of Different Deployment Models in a Cloud," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 5, pp. 512–515, 2015.

[9] H. Karim and K. Zand, "A Comparitive Survey on Data Mining Techniques for Breast Cancer Diagnosis and Prediction," *Indian J. Fundam. Appl. Life Sci.*, vol. 5, no. 1, pp. 4330–4339, 2015.

[10] S. Demigha, "Data Mining for Breast Cancer Screening," in *The 10th International Conference on Computer Science & Education*, 2015, no. Iccse, pp. 65–69.

[11] S. Kaur and R. K. Bawa, "Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System," *Int. J. energy, Inf. Commun.*, vol. 6, no. 4, pp. 17–34, 2015.

[12] V. Chaurasia and S. Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques," *InternatioInterain Comput. Commun. Eng.*, vol. 2, no. 1, pp. 2456–2465, 2014.

[13] A. Lebbe, S. Saabith, E. Sundararajan, and A. A. Bakar, "Comparative Study on Different Classification Techniques for Breast Cancer Datsset," *Int. J. Comput. Sci. Mob. Comput.*, vol. 3, no. 10, pp. 185–191, 2014.

[14] A. Omari, "A knowledge discovery approach for breast cancer management in the kingdom of saudi arabia," *Heal. Informatics Int. J.*, vol. 2, no. 3, pp. 1–7, 2013.

[15] S. Kharya, "Using Data mining Techniques for Diagnosis and Prognosis of Cancer Disease," *Int. J. Comput. Eng. Inf. Technol.*, vol. 2, no. 2, p. 2012, 2012.

[16] A. Einipour, "A Fuzzy-ACO Method for Detect Breast Cancer," *Glob. J. Health Sci.*, vol. 3, no. 2, pp. 195–199, 2011.

[17] G. Devi, "Breast Cancer Prediction System using Feature Selection and Data Mining Methods," *Int. J. Adv. Res. Comput. Sci.*, vol. 2, no. 976, pp. 81–87, 2011.

[18] H. Wang, Zheng, B. Yoon, S., and Ko, H. S " A Support Vector Machine-Based Ensemble Algorithm for Breast Cancer Diagnosis, Eur. *J. operational Research. Comput. Sci.*, vol. 262, no.2, pp. 687-699, 2018.

