INFLUENCE OF DISCRETIZATION IN CLASSIFICATION OF BREAST CANCER DISEASE

Yakub Kayode SAHEED[1], Abdulsabur Oluseye AKANNI[2], Maruf O ALIMI[3] and F.E. HAMZA-USMAN[4]
[1,2,3]Department of Computer Science, Al-Hikmah University, Ilorin, Nigeria
[4]Department of Computer Science, University of Ilorin, Nigeria.
yksaheed@alhikmah.edu.ng[1], seyeakanni@alhikmah.edu.ng[2], moalimi@alhikmah.edu.ng[3] and usman-hamza.fe@unilorin.edu.ng[4]

Keywords: Breast cancer, Support vector machine-Radial basis function, Adaboost, Discretization

Abstract: *Breast cancer (BC) is one of the leading cancers for women when compared to all other cancers. It is a killer disease prominent and most frequent type of cancer affecting women worldwide and is increasing particularly in Africa. The aim of this paper is to investigate the influence of data preprocessing based on dicretization in the classification of BC. Two different classification algorithms Support vector machine-Radial basis function (SVM-RBF) and Adaboost algorithm were employed. We analyzed the BC data available from the Wisconsin dataset from UCI machine learning repository. The experiment was performed in Waikato Environment For knowledge analysis (Weka) software. The experimental results showed that discretized SVM-RBF and discretized Adaboost algorithms outperforms the non-discretized SVM-RBF and non-discretized Adaboost algorithms in terms of accuracy, precision, recall, f-measure and time taken to build the model.*

1. INTRODUCTION

Breast cancer (BC) is the second and most frequent type of cancer affecting women worldwide [1], and is increasing particularly in developing countries where the majority of cases are diagnosed in late stages. According to [2], an estimated 252,710 new cases of invasive BC was diagnosed among women and 2,470 cases were diagnosed in men. In addition, 63,410 cases of in situ breast carcinoma would be diagnosed among women. Approximately, 40,610 women and 460 men are expected to die from BC in 2017 [2].

About 10% of women from western countries are suffering from BC, millions of women are suffering from this life threatening disease [3]. BC has become a popular and common disease around the world in young women and a leading cause of cancer death [4] [5].

In the past decade, Computer science and medical fields have been involved in providing diagnosis of various human diseases. Information generated from patients to medical personnel in biomedical prognosis and diagnosis may include redundant, irrelevant, and interrelated symptoms most often in the case whereby a patient suffers from more than one type of disease of the same category. Hence, it becomes a serious challenge for physician to diagnose perfectly.

Early detection and accurate prediction is achievable by medical personnel using data mining technique in health care industry [6].

Data mining (DM) also known as knowledge discovery in databases (KDD), is a process that aims to discover relationships between items and hidden information from large datasets [7] [8]. DM has been used recently and successfully in bioinformatics [9] [10] [11], electric load forecasting [12] and educational data mining [13].

The techniques in DM have contributed immensely in transforming large data into specific and more relevant information for knowledge discovery and prediction purpose [14]. Data pre-processing is very crucial in DM process as they directly impact success rate of

the model. There are number of data pre-processing techniques which include aggregation, dimensionality reduction, feature subset selection, discretization and feature creation. This paper focused on the influence of data preprocessing using discretization in classification of BC.

The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 is the methodology. We presented the results and discussion in section 4 and in section 5, the conclusion is presented.

2. RELATED WORKS

[15] concentrated on the detection of breast cancer through a diagnosis system based on RepTree, RBF Network and Simple Logistic. They have used the data provided by the University Medical Centre, Institute of Oncology, and WEKA tool for experiments. The result of their work based on Simple Logistic algorithm achieved 74.47% accuracy for diagnosis of breast cancer.

[16] developed a model based on Kernel-Support Vector Machine (K-SVM) for cancer diagnosis using K-means clustering algorithm as feature selection. A Model based on Rough Set (RS) and SVM classifier (RS-SVM) was developed by [17] for breast cancer diagnosis. They used RS as a feature selection which serves as the data pre-processing technique to select the best features of dataset. Further improvement for the accuracy of diagnostic system was obtained by SVM. The effectiveness of RS-SVM was examined on Wisconsin breast cancer (WBCD) dataset. [17] used clustering method along with feature selection technique to develop a hybrid intelligence method for BC analysis.

[18] proposed a method based on Neural Network (NN) technique for solving BC classification problem. [19] proposed a PSO-KDE model using PSO and KDE classifier for breast cancer diagnosis.

[20] proposed a model based on three main phases which are: instance selection phase, feature selection phase and classification phase. [21] developed a NB (weighted NB) classifier for the application breast cancer detection. Using 5 -fold cross validation, their method obtained 99.11%, 98.25%, and 98.54% for sensitivity, specificity and accuracy, respectively.

3. METHODOLOGY

The proposed study employed WEKA to perform the study analysis. Two data mining algorithms were used in this model; Adaboost and SVM. The algorithms were applied on the BC data. Two approaches were followed in implementing the algorithm. The first approach does not involve the data preprocessing stage. The second approach employed the data preprocessing stage. In the first approach, the data set is presented to the model without discretization. In the second approach, data preprocessing known as discretization was employed to eliminate noise and outliers in the data.

In SVM, the RBF was used for the implementation of the model. The individual application of the algorithms was taken as a base results and the discretized part of the algorithms was also noted. The non-discretized version is compared with the discretized part to study their effect and influence on accuracy, sensitivity, specificity, precision, recall, f-measure and time taken to build the model. All classifications were performed based on 10-fold cross validation. The overall system design of the proposed model is shown in the figure 1.

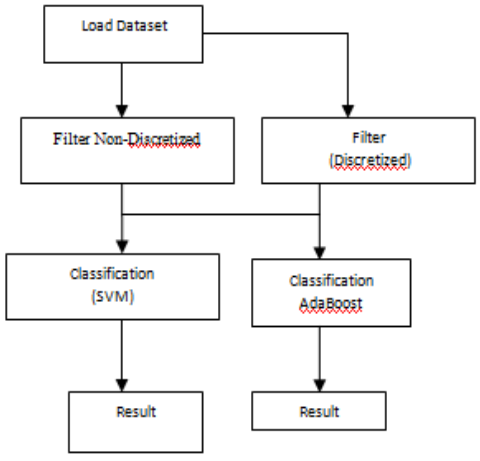

Fig. 1: Proposed framework design

3.1. Tools used for the proposed experiment

The tool used for the experimental analysis was Weka environment. Weka has become one of the most widely used data mining systems.

Weka also became one of the favorite vehicles for data mining research and helped to advance it by making many powerful features available to all [22]. The algorithms can either be called from the users own Java code or be applied directly to the ready dataset.

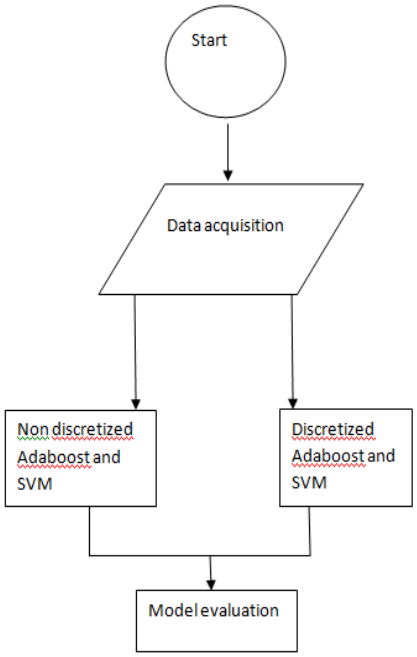

Fig.2: Flowchart of the proposed model

As can be seen from Figure 2, the BC dataset acquisition is the first stage of the proposed model. Then, data preprocessing based on filter discretization method was performed and non-discretization method performed. These two method were then evaluated based on the following performance metrics;

Accuracy: this specified how often is the classifier correct? The accuracy is given as;
(TP+TN)/total (1)
Where TP is the true positive and TN is the true negative
Precision: When the classifier predicts yes, how often is it correct? It is given as;
TP/predicted yes (2)
Where TP is the true positive
Recall: When the classifier predicts it's actually yes, how often does it predict yes? It is given as;
TP/actual yes (3)

F-measure: A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional f-measure. It is given as;
2*TP/2*TP) +FP+FN (4)

Time taken to build the model: Is the time taken by the classifier to build the model. It is measured in seconds.

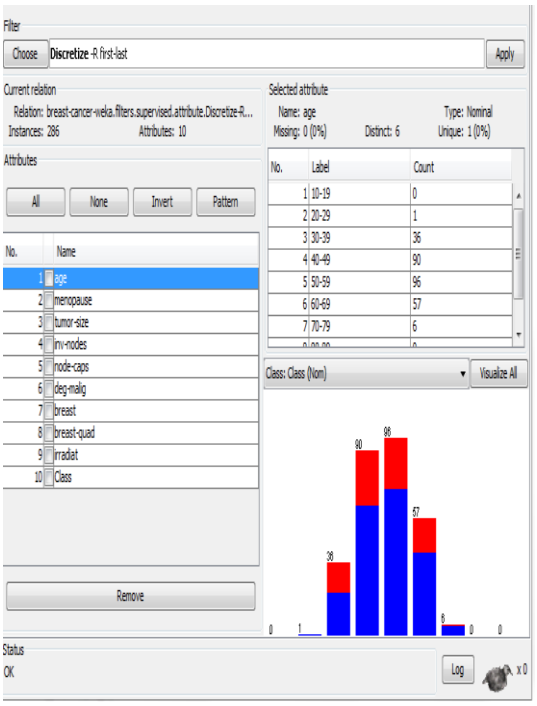

Fig.3: Discretize model of the BC dataset

3.2 SVM-RBF

The Support Vector Machine is one of the most successful classification algorithms in the data mining area [25]. Support Vector Machine (SVM) uses a high dimension space to find a hyper plane to perform binary classification. SVM approach is a classification technique based on Statistical Learning Theory (SLT). It is based on the idea of hyper plane classifier. The goal of SVM is to find a linear optimal hyper plane so that the margin of separation between the two classes is maximized.

3.3 Adaboost

AdaBoost, short for "Adaptive Boosting", is the first practical boosting algorithm proposed by Freund and Schapire in 1996. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one.

3.4 Proposed model Pseudocode

Step 1: Load dataset
Step 2: Train classification algorithms SVM-RBF and Adaboost
Step 3: Apply classification algorithms on test data
Step 4: Train discretize part of the algorithms
Step 5: Apply the discretize part of the algorithms on test data.
Step 6: Evaluate and compare the results obtained.

3.5 System Specifications

The experiment was carried out on a 64-bit operating system with Windows 8.0, Intel(R) Core(Tm) i7-3632QM CPU @ 2.20GHz and 8 GB of RAM. Due to the iterative nature of the experiments and resultant processing power required, the Java heap size for Weka version 3.6.12 was set to 1024MB to assess the effectiveness of the algorithms.

4. RESULTS AND DISCUSSION

The simulation of the proposed model was done by applying two data mining algorithms, SVM-RBF and Adaboost on breast cancer dataset, the discretized part of the algorithms was also applied on the dataset and the 10-fold cross validation was used in the two cases. 10-fold cross validation is a technique used in evaluating the predictive models by partitioning the data into a training sets, it trains the model and test the evaluation. The experimental results are presented in terms of accuracy, specificity, precision, recall, f-measure and time taken to build the model.

4.1 Accuracy of the proposed model

In SVM-RBF, the discretized version of SVM-RBF has the highest accuracy as shown in Figure 3**,** thus it was able to classify rightly more patient with the disease than non-discretized version while the discretized version results in a better accuracy. In Adaboost showed in figure 2, the discretized part gave higher accuracy than the non-discretized version. This shows that discretized version of Adaboost is better to predicts breast cancer disease.

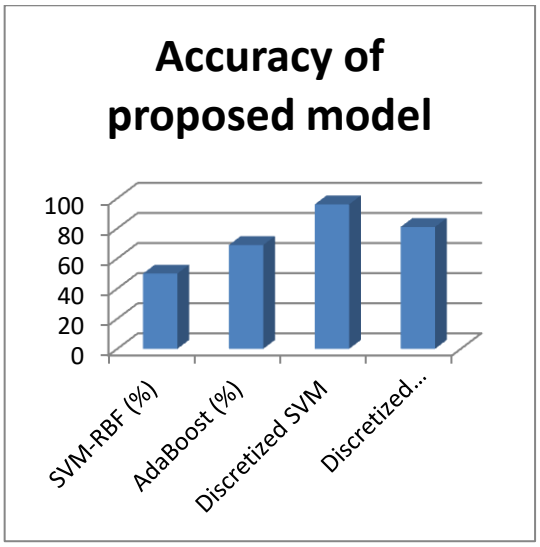

Fig.3. Accuracy of the classification

4.2 False positive of the proposed model

In SVM-RBF, the non-discretized version outperforms the discretized SVM-RBF as illustrated in Figure 4. The Adaboost non-discretized version showed in figure 4 also performed better than the discretized part. The false positive is higher in the non-discretized version of the Adaboost as compared to discretized Adaboost. The results obtained showed that discretization does not influence the false positive of classification of breast cancer disease.

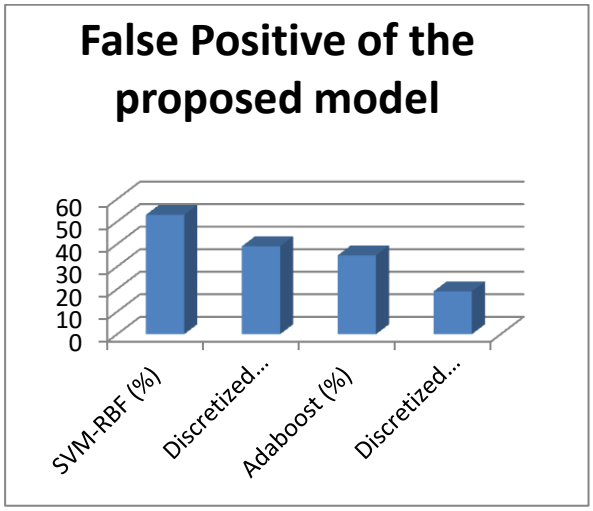

Fig.4.False positive of the classification

4.3. Precision

In SVM-RBF, the discretized part has higher precision than the non discretized SVM-RBF as depicted in Figure 5, this shows that when the model predicts yes, how often is it correct? The discretized SVM-RBF has better ability to classify the patients with breast cancer than the SVM-RBF. The discretized Adaboost has improved precision than the non-discretized Adaboost. This results indicates that discretization influence the classification precision of breast cancer disease.

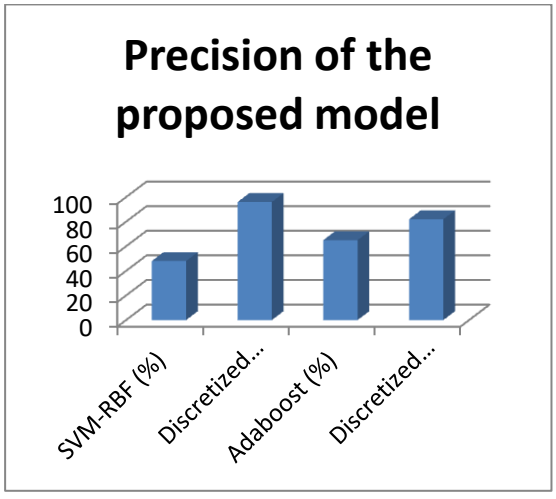

Fig.5. Precision of the classification

4.4 Recall

In SVM-RBF, the discretized part has better recall than the non discretized SVM-RBF as depicted in Figure 6, this show that when the model predicts yes, how often is it correct? The discretized SVM-RBF has better ability to classify the patients with breast cancer than the SVM-RBF. Also, the discretized Adaboost has higher recall than the non-discretized Adaboost. This results indicates that discretization influence the classification recall of breast cancer disease.

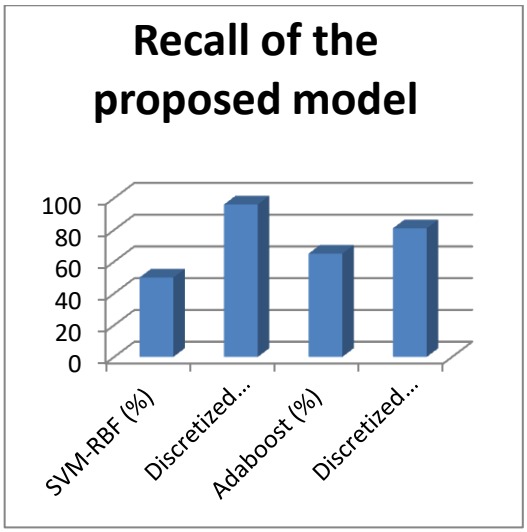

Fig.6. Recall of the classification

4.5 F-measure

In SVM-RBF, the discretized part has greater recall than the non discretized SVM-RBF as depicted in Figure 7, The discretized SVM-RBF has better ability to classify the patients with breast cancer than the SVM-RBF. Also, the discretized Adaboost has higher f-measure than the non-discretized Adaboost. The experimental results revealed that discretization influence the classification recall of breast cancer disease.

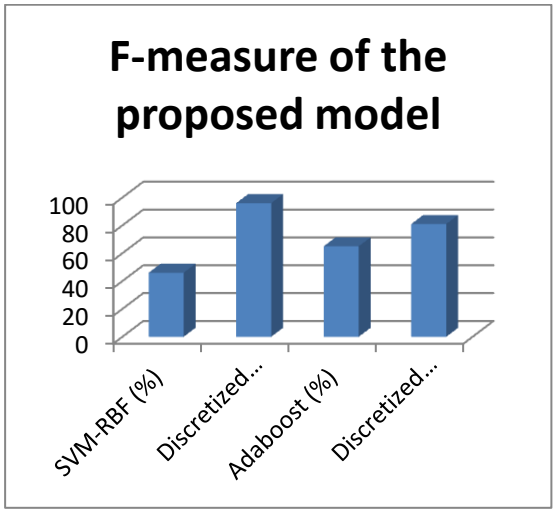

Fig.7. F-measure of the classification

4.6 Time taken to build the model

The SVM-RBF took more time than the discretized SVM-RBF as shown in Figure 8. The reason is that outliers and noise in the dataset can degrade the time taken to build the model in non discretized SVM-RBF which results in high time to build the model as compared to discretized SVM-RBF in which the noise and outliers has been eliminated by the discretization process and results in less time to build the model.

Also, the discretized Adaboost has lesser time taken to build the model as compared to the non discretized SVM-RBF as shown in Figure 8. The discretization process greatly influenced the time taken to build the model which results in less time to build the model.

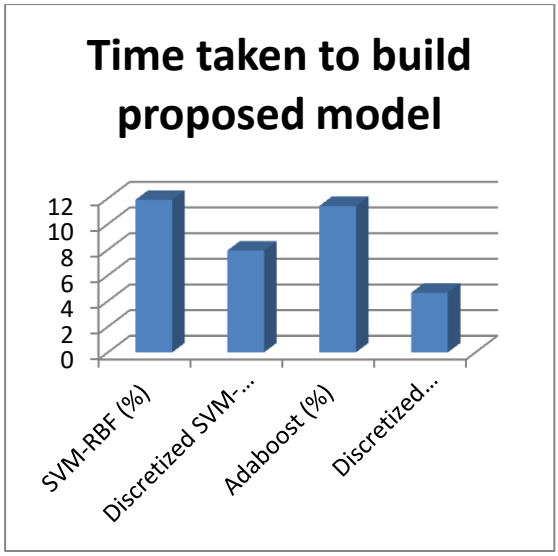

Fig.8. Time taken to build the model

Table 1.Comparison with existing methods

Authors/Year	**Algorithms**	**Accuracy obtained**
Subrata [23]	Naïve Bayes	94.40
Siddhant and Mangesh [24]	Bayes Net	47.67
Siddhant and Mangesh [24]	Naïve Bayes	46.1
Siddhant and Mangesh [24]	IBK	45.3
Proposed method	Adaboost	81
Proposed method	SVM-RBF	96

As can be seen from Table 1, the proposed Adaboost and SVM-RBF outperformed the existing methods in terms of accuracy.

5. CONCLUSION

In this paper, it was revealed that data preprocessing by discretization method increase the accuracy, precision, recall, f-measure and time taken to build the model of SVM-RBF and Adaboost algorithms in determining whether a patient has breast cancer disease or not. Except in the case of false positive in which the model predicted yes that patient have the breast cancer disease but they don't actually have the disease. The results obtained in this instance depicts that discretization method does not influence the false positive in classification of breast cancer disease. From the experimental analysis, it is better to apply data preprocessing based on discretization in determining BC disease than using algorithms directly without data preprocessing. This shows that data preprocessing is a crucial step in BC classification as it directly influence the success rate of breast cancer classification.

Finally, the overall experimental results showed that discretized SVM-RBF and discretized Adaboost algorithms perform better than the non-discretized SVM-RBF and non-discretized Adaboost algorithms in terms of accuracy, precision, recall, f-measure and time taken to build the model. In future work, other data-preprocessing techniques such as aggregation, sampling and dimensionality reduction will be an interesting data reduction approaches to explore and ascertain if they can influence BC classification.

4. REFERENCES

[1] Chaurasia, V. and Pal, S. (2017). A Novel Approach for Breast Cancer Detection Using Data Mining Techniques. June 29, 2017.International Journal of Innovative Research in Computer and Communication Engineering.

[2] American cancer society, Facts and figures 2017-2018. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2017-2018.Accessed 22/7/2018.

[3] Mehrbakhsh Nilashi, Othman Ibrahim, Hossein Ahmadi and Leila Shahmoradi (2017). A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics and Informatics 34 (2017) pp.133–144.

[4] McCarthy, A.M., Yang, J., Armstrong, K., (2015). Increasing disparities in breast cancer mortality from 1979 to 2010 for US black women aged 20 to 49 years. Am. J. Public Health, e1–e3.

[5] Kharazmi, E., Försti, A., Sundquist, K., Hemminki, K., (2016). Survival in familial and non-familial breast cancer by age and stage at diagnosis. Eur. J. Cancer 52, pp.10–18.

[6] Himanshu, S. and M. A. Rizvi., (2017). Prediction of Heart Disease using Machine Learning Algorithms: A Survey. International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 5 Issue: 8.

[7] Abdulsalam, S. O., Hambali, M. A. Salau-Ibrahim, T.T. Saheed, Y.K. Akinbowale, N.B. (2017). Knowledge Discovery From Educational Database Using Apriori Algorithm. GESJ: Computer Science and Telecommunications 2017.No.1(51).

[8] Tseng, W.T., Chiang, W.F., Liu, S.Y., Roan, J., and Lin, C.N., (2015). The application of data mining techniques to oral cancer prognosis. J Med Syst. 39(5):59, 2015. doi:10.1007/s10916-015-0241-3.

[9] Ayomikun, K.O., T.O. Oladele, & Y. K. Saheed, (2018). Comparative Evaluation of Linear Support Vector Machine and K-Nearest Neighbour Algorithm using Microarray Data On Leukemia Cancer Dataset, Afr. J. Comp. & ICT, Vol.11, No.2, pp. 1 - 10.

[10] R.G. Jimoh, R.M. Yusuf, Yusuf, O.O. Saheed, Y.K. (2018). Application of Dimensionality Reduction on Classification of Colon cancer Using ICA and K NN Algorithms. Annals. Computer Science Series. 16th Tome 1st Fasc. – 2018.

[11] Arowolo, M.O., R.M. Isiaka, S.O. Abdulsalam, Y.K. Saheed and K.A. Gbolagade (2017).A Comparative Analysis of Feature Extraction Methods for Classifying Colon Cancer Microarray Data. EAI Endorsed Transactions on Scalable Information Systems 07 2017 - 09 2017. Volume 4 Issue 14. doi: 10.4108/eai.25-9-2017.153147.

[12] Hambali, M. A., Saheed, Y. K, Gbolagade, M. D, Gaddafi M. (2017). Artificial Neural Network Approach For Electric Load Forecasting in Power Distribution Company. Volume 6 Issue 2 2017, 80-90.e-Academia Journal. http://journale-academiauitmt.uitm.edu.my Universiti Teknologi MARA Terengganu.

[13] Abdulsalam, S.O. Saheed, Y.K. Hambali, M.A. Salau-Ibrahim, T.T. Akinbowale, N.B. (2017). Student's Performance Analysis Using Decision Tree Algorithms. Annals. Computer Science Series. 15th Tome 1st Fasc. – 2017

[14] Agrawal, N.R., and P.A. Chopde, A Survey on Heart Disease Prediction Using Soft Computing. Int. J.Eng. Sci. Res., Vol.5, no.3, pp.582-587, 2016.

[15] Chaurasia, V. and Pal, S. (2014). A Novel Approach for Breast Cancer Detection Using Data Mining Techniques. International Journal of Innovative Research in Computer and Communication Engineering Vol. 2, Issue 1, January 2014.

[16] Zheng, B., Yoon, S.W., Lam, S.S., (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst. Appl. 41 (4), pp.1476–1482.

[17] Chen, C.H., (2014). A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. Appl. Soft Comput.2014, pp.4–14.

[18] Bhardwaj, A., Tiwari, A., (2015). Breast cancer diagnosis using genetically optimized neural network model. Expert Syst. Appl. 42 (10), pp.4611–4620.

[19] Sheikhpour, R., Sarram, M.A., Sheikhpour, R., (2015). Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. Appl. Soft Comput. Vol. 40. March 2016 pp.113-131.

[20] Onan, A., (2015). A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Syst. Appl. Vol.42.Issue 20, pp.6844-6852.

[21] Karabatak, M., (2015). A new classifier for breast cancer detection based on Naïve Bayesian. Measurement 72, pp.32–36.

[22] Zdravko Markov and Ingrid Russell, An Introduction to the WEKA Data Mining System,http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf , accessed: 2018-11-20

[23]Subrata Kumar Mandal (2015).Performance Analysis Of Data Mining Algorithms For Breast

Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree. International Journal Of Engineering And Computer Science ISSN: 2319-7242.Volume 6 Issue 2 Feb. 2017, pp.20388-20391

[24] Siddhant Kulkarni and Mangesh Bhagwat (2015). Predicting Breast Cancer Recurrence using Data Mining Techniques.International Journal of Computer Applications (0975 – 8887) Volume 122 – No.23, July 2015

[25] Bhavsar, Yogita B., and Kalyani C. Waghmare (2013). Intrusion Detection System Using Data Mining Technique: Support Vector Machine. International Journal of Emerging Technology and Advanced Engineering 3.3 (2013).