COMPARATIVE ANALYSIS OF SUPPORT VECTOR MACHINE, C4.5 AND NAÏVE BAYES ALGORITHMS FOR BREAST CANCER DIAGNOSIS

Supported by

Moshood Abiola Hambali[1]*, Morufat Damola Gbolagade[2], Yinusa Ademola Olasupo[1] and Ayodele Odutola Amusan[3]
[1]Department of Computer Science, Federal University PMB 1020, Wukari, Taraba State, Nigeria
[2]Department of Computer Science, Al-Hikmah University, PMB 1601, Ilorin, Kwara State, Nigeria
[3]Rock Security and Compliance, Rochester, Minnesota, USA
*****Corresponding author: hambali@fuwukari.edu.ng

Received: August 26, 2021 **Accepted:** October 14, 2021

Abstract: Breast cancer is one of the leading cancers for women when compared to all other cancers. It is the second highest cause of death in women. Breast cancer risk in Africa revealed that 1 out of 28 women develop breast cancer during their lifetime. This is more prominent in urban areas being 1 out of 22 in a lifetime compared to rural areas where the risk is relatively much lower being 1 out of 60 women developing breast cancer in their lifetime. The aim of this study is to investigate the performance of different classification techniques on the Wisconsin breast cancer dataset from UCI machine learning database. In this experiment, we compare three classification techniques - C4.5 decision tree, Naïve Bayes (NB) and Support Vector Machine (SVM). Two cross validation approaches were used for all the learners, that is, 10 and 20-folds cross validation. The best results were achieved in 20-folds cross validation with NB has accuracy of 97.5%, SVM with accuracy of 97.2%, while C4.5 algorithms with accuracy of 94.8%.

Keywords: Breast cancer, SVM, C4.5, Naïve Bayes, WDBC

Introduction

Breast cancer (BC) is second highest cause of death in women after cancer of lung. It signifies about 12% of all new cases of cancer diagnosed and 25% of all cancers diagnosis in women and there is high death rate every year as a result of BC (Asri *et al.*, 2016). United States statistics for 2014 revealed an estimate population of 2,360 males and 232,670 females' new cases of breast cancer. Out of this population, 40,000 females and 430 males were reported death with this disease (Malla and Bokari, 2017).

BC is a multifaceted heterogeneous disease that arises due to abnormal growth of cells in the breast tissue, with a set of various clinical symptoms. With timely detection and diagnosis of BC will increase survival of patient from 56% to above 86% up to five years after the occurrence of tumor (Montazeri *et al.*, 2016). The unusual growth of cells can either benign or malignant. An expert physician encounters challenges in BC diagnosis due to complex issues surrounding it. The medical approach of diagnosis BC is to use mammography, MRI, blood tests, CT scan and biopsy. However, the results obtain from radiologists are sometimes inconsistent in the way they interpret mammogram and analysis of the result (Hambali *et al*., 2019). Also, Elmore *et al*., (1994) stated that about 90% of radiologists can be able to identify less than 3% of cancer cases.Fine needle aspiration cytology is another alternative approach employed for BC diagnosis with a reasonable prediction accuracy. Though, its correct diagnosis rate is around 90% (Fetiman, 1998).But recent medical diagnoses are built on data acquired from clinical observation or other tests. With this data, it can aid physicians to diagnose BC. Hence, an accurate and dependable system is essential for the timely diagnosis of benign or malignant tumors (Gayathri *et al.*, 2013; Montazeri *et al.*, 2016).

In the last few decades, statistical methods were commonly applied with data mining approaches in building classification models for BC. But, the BC classification task is greatly nonlinear in nature. It is highly tasking to build a reliable model that will consider all independent variables using traditional statistical modeling approaches. Furthermore, integration of typical statistical approaches and data management tools are not adequate for analysing the huge amount of data (Abdulsalam *et al.*, 2015). Data mining is apt method and gaining momentum in many research domains including the medical field to detect and diagnosis various diseases (Hambali *et al.*, 2019).

Machine learning has been playing potential roles in cancer diagnosis and treatment. In fact, advancement in big data is not limited to the size of data but at the same time creating value for it. Big data has becomes a synonymous to business analytics, data mining and business intelligence, and made a big impact in reporting and decision to prediction results. Application of data mining approaches in medical domain has gained a tremendous acceptance due to their highly efficiency in detecting and predicting the outcomes, cost effective of medicine, improving healthcare value, promoting patients' health and quality, and making real time decision to save people's lives. There are various machine learning algorithms for classification and prediction of BC reported in the literature. The objective of this study is to compare the performance of three classifiers: SVM, NB, and C4.5 which are among the most prominent and top 10 ranked data mining methods in the research community (Asri *et al.*, 2016).

There are several works available in literature concern various algorithms that assist healthcare experts in early prediction and accurately diagnosing breast cancer. Zheng *et al*. (2013) developed a model that support the diagnoses of breast cancer using data mining approach by extracting and selected tumor features. The techniques are a combination of K-means and SVM algorithm. The approach was evaluated on Wisconsin Diagnostic Breast Cancer (WDBC) dataset and obtained the accuracy of 97.38%. Asri *et al*. (2016) presented a performance comparison between different machine learning algorithms- SVM, Decision Tree (C4.5), Naive Bayes (NB) and k-Nearest Neighbors (k-NN) on the original WDBC datasets. It was deduced experimentally that SVM gives the highest accuracy of 97.13% with lowest error rate.

Abreu *et al*. (2016) presented a review of machine learning algorithms to forecast the recurrence of breast cancer. It was deduced that despite the hardship to obtain the representative dataset of breast cancer recurrence, by using the combine machine learning technique alongside the definition of standard breast cancer predictor seem to give a good sign to obtain a better result in the future. Montazeri *et al*. (2016) worked on a mixture of rules and various machines learning techniques for breast cancer survival prediction. They use the following machine learning techniques; AdaBoost (AD), Multilayer Perceptron (MLP), NB, 1-Nearest Neighbor

(1NN), RBF Network (RBFN), SVM, Trees Random Forest (TRF) using 10-folds cross technique to give breast cancer survival prediction of which the TRF gave the best result of all. The result of accuracy, sensitivity and the area under receiver operating characteristic (ROC) curve attained are 96, 96, 93%, respectively for TRF. 1NN machine learning technique, being the lowest performance of all, gives accuracy of 91%, sensitivity of 91% and area under ROC curve of 78%. Mohebian *et al*. (2017) proposed an online tool for predicting breast cancer recurrencecalledHybrid Predictor of Breast Cancer Recurrence (HPBCR). The performance of HPBCR yielded a minimum sensitivity, precision, specificity and accuracy of 77, 95, 93 and 85%, respectively. Also, the approach was compared with SVM, DT and MLP.

Malla & Bokari (2017) proposed three models for implementation of early prediction of breast cancer dataset. The models are: Naïve Bayes, Logistic Regression and Random Forest. The results of their study showed that Random Forest performed better with sensitivity of 99% and accuracy of 98%, followed by the Logistic Regression with sensitivity of 98% and accuracy of 96% while NB with sensitivity of 94% and accuracy of 91%.

Huang *et al*. (2017) proposed SVM to predict the breast cancer which is a common disease in women. The SVM classifier was developed using different kernel functions. The result showed that radial basis function (RBF) kernel SVM performance better than other kernel functions for a large dataset with accuracy of 99.41%, ROC of 0.875 and F-measure 0.994. Islam *et al*. (2017) presented a comparison between SVM and K Nearest Neighbors (KNN).The accuracy of the proposed system was realized through the use of 10-folds cross validation. Result revealed that SVM performed better than KNN with accuracy of 98.57% and specificity of 95.65%. Ojha & Goel (2017) worked on three classification algorithms (Decision tree (C5.0), SVM and fuzzy c - means) for the early prediction of breast cancer using dataset contains 194 records. The numbers of non-recurrent were 148 while that of recurrent were 46 cases. The fuzzy c-means showed the lowest result of accuracy of 37% and SVM yield an outstanding performance with accuracy of 81%.

Polat & Senturk (2018) proposed a three steps hybrid system for cancer prediction. The steps involve are: the MAD (median absolute deviation) nominalization method was employed for dataset nominalization. The second steps involve features weighting using k-means clustering and the third step used AdaBoostM1classifier to categorize the weighted dataset. They concluded that their hybrid proposed structure gave an accuracy of 91.37% accomplishment and could safely be used to identify breast cancer. Yue *et al*. (2018) proposed a review on Machine Language (ML) techniques with their application on breast cancer diagnosis and prognosis. The ML algorithms include ANN, SVM, DT and k-NN. Hambali *et al*. (2019) developed an Adaboost ensemble model for the extraction of useful information and make a diagnosisof breast cancer. In their research work, two categories of classifiers were introduced, that is, the homogeneous and heterogenous ensemble classifiers combined with the implementation of Synthetic Minority Over-Sampling Technique (SMOTE) which cater for the class disparity problem and noise in the dataset. The results showed that Adaboost-Random forest performs better with accuracy of 82.52%. Random forest-CART follows with accuracy of 72.73% while the lowest of all in the Naïve Bayes classification with accuracy of 35.70%.

Proposed Approach

The techniques and methods employ for this study are briefly explained in this section and proposed framework is shown in Fig. 1. In this work, we have investigated performance of three data mining techniques: SVM, Naïve Bayes, and C4.5 decision tree algorithms in classifying breast cancer dataset into benign (normal) and malignant (diseased) class. The implementation was done on Orange (3.13.0 version) data mining environment.

Data source

In this work, experiments were performed on WBCD database taken from UCI machine learning repository (Bache & Lichman, 2013). The WBCD contains 699 samples obtained from Fine Needle Aspirates (FNA) of human breast tissue. The dataset consists of 9 features and its class (malignant or benign) corresponding to individual record. The value of each feature is an integer value range from 1 to 10, 10 designates the most abnormal state. Out of the 699 samples, 16 instances contain missing value attributes which were discarded and make uses of the remaining 683 instances. 444 out of used instances belong to benign class while 239 were malignant class. Dataset distributions were presented in Fig. 2 where blue colour indicates benign class while red colour indicates malignant class.

Data pre-processing and feature selection

Data pre-processing is a data mining approach that involves transforming raw data into a comprehensible format, normalization of data, take care of missing value and soon. In this research work, the data preprocessing was first performed by discarding the instances that contain missing attributes. Then, selection of relevant features was performed using information gain technique with threshold fixed to 10%.

Support vector machine

A support vector machine (SVM) is a supervised learning algorithm method used for classification and regression. SVM is a powerful classification algorithm; due to it efficient performance in pattern recognition domain. SVM constructs a hyperplanes or a set of hyperplanes in a high-dimensional space that can help in classification, regression, or other tasks. SVM has the capability to deal with linear and nonlinear datasets. In linear data, SVM tries to find an optimal separating hyperplane that maximizes the margin between the training examples and the class boundary. In nonlinear data, we need to define a feature mapping function X → φ(x). This mechanism that defines feature mapping process is called kernel function. The most popular among them are three:

- Polynomial kernel

$$k(x_i, x_j) = (x_i \cdot x_j + a)^b \qquad (1)$$

- Radial basis kernel

$$k(x_i, x_j) = e^{-(\|x_i - x_j\|^2 /_{2\sigma^2})} \qquad (2)$$

- Sigmoidal kernel

$$k(x_i, x_j) = \tanh(\propto x_i \cdot x_j - b) \qquad (3)$$

Where: a and b are parameters define the kernel's behavior. SVM has the capability to deal with linear and nonlinear datasets. In linear data, SVM tries to find an optimal separating hyperplane that maximizes the margin between the training examples and the class boundary. In this experiment, we explore sigmoidal kernel function.

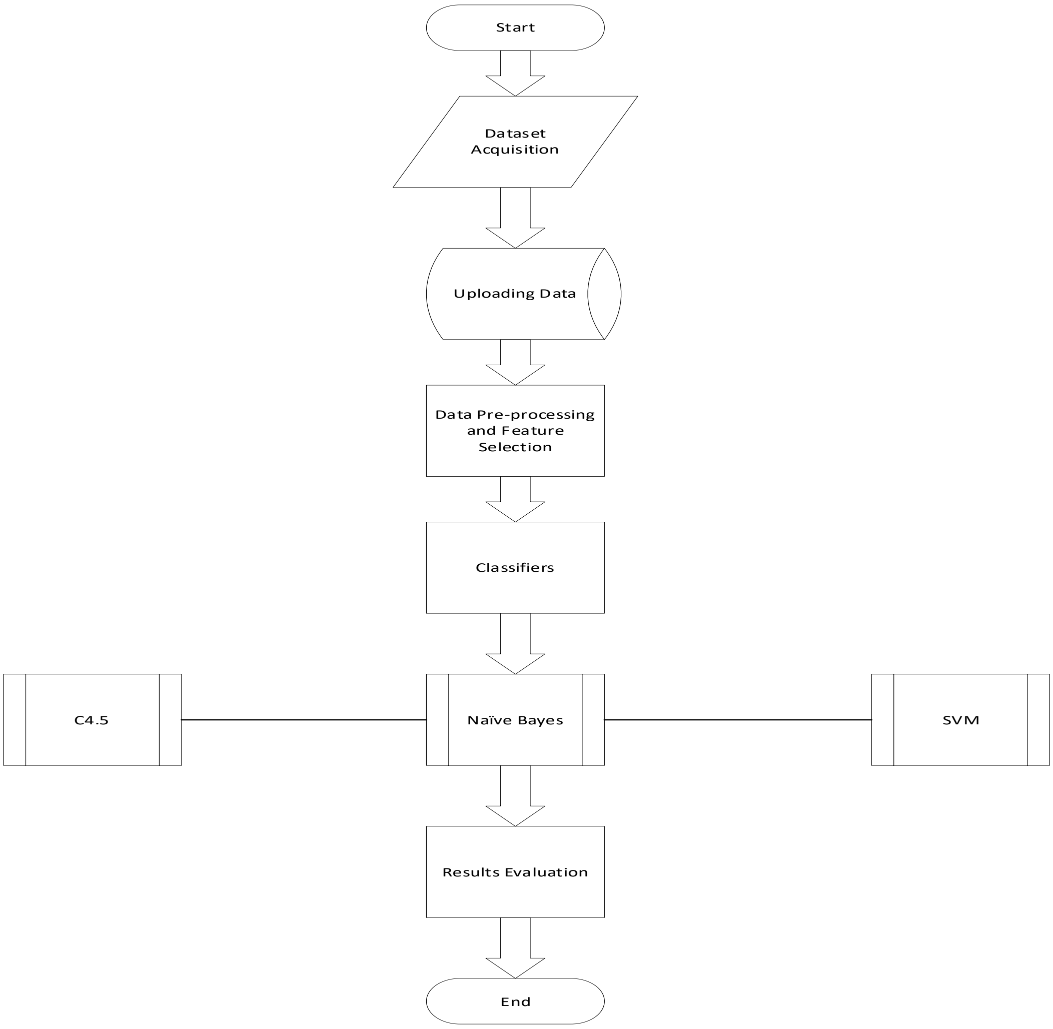
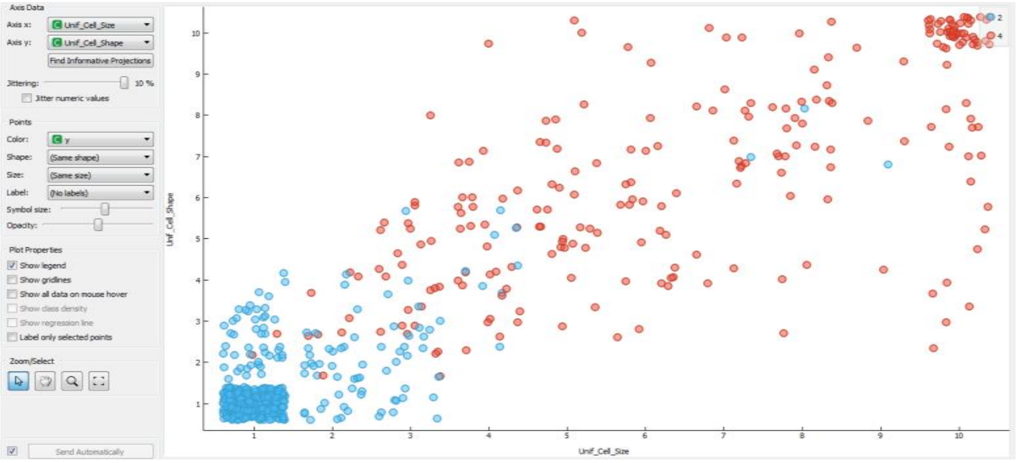


Fig. 1 Proposed System frame work



Axis Data
Axis x: Unif_Cell_Size
Axis y: Unif_Cell_Shape
Find Informative Projections
Jittering: 10 %
□ Jitter numeric values

Points
Color: y
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
Symbol size:
Opacity:

Plot Properties
☑ Show legend
□ Show gridlines
□ Show all data on mouse hover
□ Show class density
□ Show regression line
□ Label only selected points

Zoom/Select

☑ Send Automatically

Fig. 2: Data visualization

C4.5 algorithm

The C 4.5 algorithm applies divide and conquer method in order to construct a decision tree. It is a decision tree induction algorithm works as top-down approach. It utilizes heuristics method for pruning, built on the statistical significance of splits. The algorithm is used to implement classification and reduce the influence of biasing.

Naive Bayes classification

Naïve classifier is principally built on Bayes theorem with independence assumptions between predictors. Bayesian classification model is simple to construct, without composite iterative parameter estimation. This makes it specifically good for extremely large datasets. Naive Bayesian (NB) is based on two approach phases: the training phase and prediction phase. The training phase is also known as the learning phase where the input data is used to evaluate the parameter of a probability distribution. In this case, the predictors are assuming to be conditionally independent. The prediction phase predicts any unknown dataset and evaluates the posterior probability of each class from the sample.

Distribution of training and test datasets

After data preprocessing, the dataset was partitioned into train and test sets with 70 – 30%. That is, training set is 70% while 30% for testing.

Results and Discussion

In this work, the results of proposed comparative of C4.5, NB and SVM to diagnose WBCD breast cancer were presented in this section.

Experiment results

The experiment was implemented on Orange 3.13.0 version data mining environment. The dataset was first uploaded into the explorer, and information gain algorithm for feature selection was used to reduce the effect of irrelevant data. In our experiment, we used random average of 10-folds and 20-folds cross validation approach for each learner approach.

Performance metrics used for the classification algorithms

The following are the metrics used to evaluate the performance of classification algorithms used in this study.

Confusion matrix: Confusion matrix is not a performance metric but the easiest means to find and compute accuracy of the model. It is usually employed in Classification task where the outcome can be more than two classes. Most of the metrics used to evaluate the classification performance are based on the Confusion Matrix. Table 1 depicts the confusion matrix.

Table 1: Confusion Matrix

		Predicted	
		Positive	**Negative**
Actual	**Positive**	**TP**	**FP**
	Negative	**FN**	**TN**

i. Classification accuracy (CA): Classification **Accuracy** is the percentage of correctly classifies instances out of all instances, that is, the number of correct predictions achieved by the model over all types of predictions made. Accuracy is a good measure when the target variable classes in the data are nearly balanced.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \qquad (4)$$

ii. Recall or sensitivity: Sensitivity is the true positive rate (also known as recall). It is the number of instances from the positive class that actually predicted correctly. In cancer classification, it is a measure of the proportion of the patients that actually had cancer and was predicted by the model as having cancer.

$$Sensitivity = \frac{TP}{(TP+FN)} \qquad (5)$$

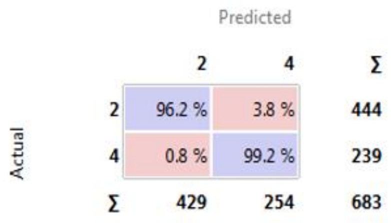
iii. F1 score rate: F1 score is the compute weighted mean of both precision and recall. Thus, this score considers both false positives and false negatives.

iv. Precision: Precision is the ratio of correctly predicted positive samples to the total predicted positive samples.

v. Area under curve (AUC): AUC is a measure of how well a parameter can differentiate between two diagnostic classes (normal/diseased). AUC range lies between 0 and 1. AUC with value close to 1 shows a very reliable diagnostic result.

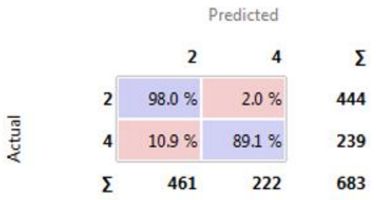
Note: TP –True Positive, TN - True Negative, FP – False Positive, FN – False Negative

The result of the experiments is presented in Tables 2 and 3, and graphical representation in Figs. 3 – 7.



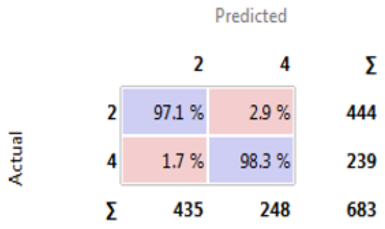
		Predicted		
		2	4	Σ
Actual	2	96.2 %	3.8 %	444
	4	0.8 %	99.2 %	239
	Σ	429	254	683

Fig 3: Confusion matrix of SVM



		Predicted		
		2	4	Σ
Actual	2	98.0 %	2.0 %	444
	4	10.9 %	89.1 %	239
	Σ	461	222	683

Fig 4: Confusion matrix of C4.5 Tree algorithm



		Predicted		
		2	4	Σ
Actual	2	97.1 %	2.9 %	444
	4	1.7 %	98.3 %	239
	Σ	435	248	683

Fig. 5: Confusion matrix of Naïve Bayes algorithm

Table 2 revealed that the C4.5 algorithm yielded an AUC of 92.3%, CA of 94.5%, F1 score of 94.5% and precision of 94.5% and recall of 94.5%. The SVM algorithm had 99.2% AUC, CA of 96.7%, F1 of 96.7% and precision of 96.4% and recall 96.5%, while NB algorithm had 99.2% AUC, CA of 97.4%, F1 of 97.4%, precision of 97.5% and recall 97.4%. Graphical comparative is shown in Fig. 6. This results show that NB out performed both SVM and C4.5.

Table 2: Performance evaluation of the proposed three classifiers using 10-folds cross validation

Methods	AUC	CA	F1	Precision	Recall
C4.5	92.3	94.5	94.5	94.5	94.5
SVM	99.2	96.7	96.7	96.9	96.7
NB	99.2	97.4	97.4	97.5	97.4

Table 3: Performance evaluations of the proposed three classifiers using 20- folds cross validation

Methods	AUC	CA	F1	Precision	Recall
C4.5	92.2	94.9	94.8	94.9	94.9
SVM	99.4	97.2	97.2	97.4	97.2
NB	99.4	97.5	97.5	97.6	97.5

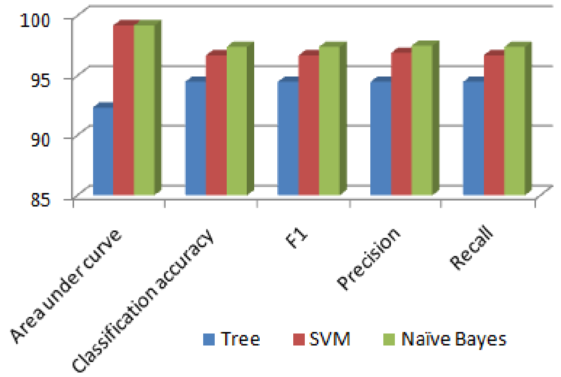


Fig. 6: Performance of the three classifiers on 10-folds cross validation

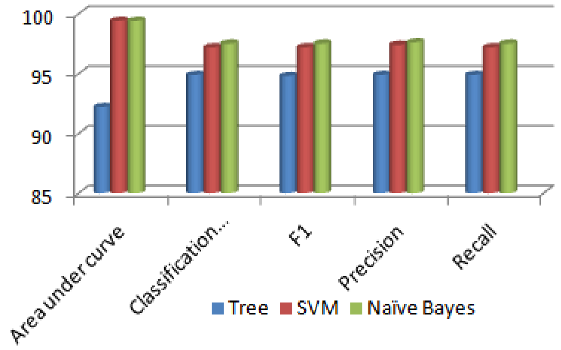


Fig. 7: Performance evaluations of the classifiers on 20-folds cross validation

Table 3 revealed that the C4.5 algorithm had an AUC of 92.3%, CA of 94.9%, F1 score of 94.8% and precision of 94.9% and recall 94.9%. The SVM algorithm had 99.4% AUC, CA of 97.2%, F1 of 97.2% and precision of 97.4% and recall 97.2%, while NB algorithm had 99.4% AUC, CA of 97.5%, F1 of 97.5%, precision of 97.5% and recall 97.6%. Graphical comparative is shown in figure 7. This results show that NB out performed both SVM and C4.5.

From the experimental analysis, it was observed and revealed that the NB outperforms other classifiers on both cross validation approaches used. It achieved a higher classification accuracy than the other classifiers with the same value of AUC with SVM. It was observed that there is performance improvement from 10-folds to 20-folds cross validation, CA has increment of about 0.2 - 0.5%, AUC has improvement of about 0.1 - 0.2%, F1 score has significant improvement of 0.1 - 0.5%, Precision of about 0.3 – 1.3% improvement while Recall has about 0.1 – 0.5% improvement. Therefore, there is performance improvement in all classifiers from 10 – 20-folds cross validation.

Table 4: Comparison of proposed approach with other approach in literature

Performance Metrics (%)	Proposed Approach	Zheng *et al.* (2013)	Asri *et al.* (2016)	Montazeri *et al.* (2016)	Mohebian *et al.* (2017)	Malla and Bokari (2017)	Ojha and Goel (2017)	Polat and Senturk (2018)
AC	**97.5**	97.3	97.13	96	85	91	81	91.37
AUC	**99.4**	-	-	93	-	-	-	-
F1	97.5	-	-	-	-	-	-	-
Recall	**97.5**	-	97	96	77	94	-	-
Precision	97.6	-	**98**	-	95	-	-	-

The proposed approach was compared with other approaches in literature as shown in Table 4. The basis of comparison is on those using the same dataset with other algorithms or using the NB algorithm. It was observed that the proposed approach performed better that other approach compared with, except in the work of (Asri *et al.*, 2016) that has better performance than proposed approach in precision metric.

Conclusion

BC prediction is very noteworthy task in Biomedical and Medicare, because BC is very serious disease that has result in death of lot women all over the world. Thus, early detection and diagnosis of this cancer will be of help to save a lot of valuable life and increase the survival rate. We compared the performance of three data mining techniques (C4.5, NB and SVM) on Wisconsin Breast cancer disease. The experimental results showed that NB outperformed other approached used and its performance was compared with other approaches in the literature.The proposed model will be a supportive tool for the medical staffs to aid in early detection of BC.

References

Abdulsalam SO *et al*. 2015. Comparative analysis of decision tree algorithms for predicting undergraduate students: Performance in computer programming. *J. Advan. in Scient. Res. & Its Applic.*, 2: 79–92.

Abreu PH *et al.* 2016. Predicting breast cancer recurrence using machine learning techniques: A systematic review PEDRO. *ACM Comput. Survey*, 49(3): 52:1–40.

Asri H *et al.* 2016. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia - Procedia Computer Science*. Elsevier Masson SAS, 83(Fams), pp. 1064–1069. doi: 10.1016/j.procs.2016.04.224.

Bache K & Lichman M 2013. *UCI machine learning repository*. Available at: http://archive.ics.uci.edu/ml/machine-learning-databases/ breast-cancerwisconsin%0A/.

Elmore J *et al.* 1994. Variability in radiologists interpretation of memograms. *New England Journal of Medicine*, 331(22): 1493–1499.

Fetiman IS 1998. *Detection and Treatment of Breast Cancer*. 2nd edn. London: Martin Duntiz.

Gayathri B, Sumathi C & Santhanam T 2013. Breast cancer diagnosis using machine learning algorithms–A survey. *Int. J. Distrib. and Parallel Systems*, 4(3): 105.

Hambali MA *et al*. 2019. ADABOOST ensemble algorithms for breast cancer. *J. Advan. in Comp. Res.*, 10(2): 31–52.

Huang M *et al.* 2017. SVM and SVM ensembles in breast cancer prediction. *PloS one*, 12(1): 1–14. doi: 10.1371/journal.pone.0161501.

Islam M *et al.* 2017. Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors', in *7 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. Dhaka, Bangladesh: IEEE, pp. 226–229.

Malla YA & Bokari MU 2017. A machine learning approach for early prediction of breast cancer. *Int. J. Engr. and Comp. Sci.*, 6(5): 21371–21377. doi: 10.18535/ijecs/v6i5.31.

Mohebian MR *et al.* 2017. A Hybrid Computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning. *Computa. and Structural Biotechn. J.*, 15: 75–85. doi: 10.1016/j.csbj.2016.11.004.

Montazeri Mitra *et al.* 2016 'Machine learning models in breast cancer survival prediction', *Technology and Health Care*, 24: 31–42. doi: 10.3233/THC-151071.

Ojha U & Goel S 2017. A study on prediction of breast cancer recurrence using data mining techniques. in *7th Int. Conf. on Cloud Comp., Data Sci. & Engr. – Confluence*. IEEE, pp. 527–530.

Polat K & Senturk U 2018. A novel ML approach to prediction of breast cancer: Combining of mad normalization , KMC based feature weighting and AdaBoostM1 classifier', in *2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE,*. IEEE, pp. 1–4.

Yue W *et al.* 2018. Machine Learning with applications in breast cancer diagnosis and prognosis. *Designs*, 2(13): 1–17. doi: 10.3390/designs2020013.

Zheng B, Yoon SW & Lam SS 2013. Expert systems with applications breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*. Elsevier Ltd, (September). doi: 10.1016/j.eswa.2013.08.044.