Comparison on Some Machine Learning Techniques in Breast Cancer Classification

Nurul Amirah Mashudi
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia, Kuala Lumpur Campus
Kuala Lumpur, Malaysia
nurulamirah@graduate.utm.my

Syaidathul Amaleena Rossli
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia, Kuala Lumpur Campus
Kuala Lumpur, Malaysia
amaleena@graduate.utm.my

Norulhusna Ahmad
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia, Kuala Lumpur Campus
Kuala Lumpur, Malaysia
norulhusna.kl@utm.my

Norliza Mohd Noor*
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia, Kuala Lumpur Campus
Kuala Lumpur, Malaysia
norliza@utm.my

Abstract— **Breast cancer is the second most common cancer after lung cancer and one of the main causes of death worldwide. Women have a higher risk of breast cancer as compared to men. Thus, one of the early diagnosis with an accurate and reliable system is critical in breast cancer treatment. Machine learning techniques are well known and popular among researchers, especially for classification and prediction. An investigation was conducted to evaluate the performance of breast cancer classification for malignant tumors and benign tumors using various machine learning techniques, namely k-Nearest Neighbors (k-NN), Random Forest, and Support Vector Machine (SVM) and ensemble techniques to compute the prediction of the breast cancer survival by implementing 10-fold cross validation. Additionally, the proposed methods are classified using 2-fold, 3-fold, and 5-fold cross validation to meet the best accuracy rate. This study used a dataset obtained from Wisconsin Diagnostic Breast Cancer (WDBC) with 23 selected attributes measured from 569 patients, from which 212 patients have malignant tumors and 357 patients have benign tumors. The performance evaluation of the proposed methods was computed to obtain accuracy, sensitivity, and specificity. Comparison results between all methods show that AdaBoost ensemble methods gave the highest accuracy at 98.77% for 10-fold cross validation, while 2-fold and 3-fold cross validation at 98.41% and 98.24%, respectively. Nevertheless, the result with 5-fold cross validation show SVM produced the best accuracy rate at 98.60% with the lowest error rate.**

*Keywords— **Breast Cancer, Classification, Machine Learning***

I. INTRODUCTION

In 2020, an estimated 42,170 breast cancer deaths were reported in [1]. Breast cancer is the most common disease among women aged 20 to 59 years and the second most common cancer in the United States [2]. Age has a strong influence which may lead to death due to breast cancer. A woman has a 12.5% chance of being diagnosed with breast cancer in the United States [3]. According to the World Health Organization (WHO), an early diagnosis and detection of breast cancer can enhance breast cancer survival.

Generally, breast cancer is initiated by a mutation in a single cell. The expandability of the cell in the breast tissue causes a rapid cell division and the masses are formed [4]. The masses are also known as tumors, which are categorized into a group of malignant and benign. Malignant tumors spread the abnormal cell to the body tissues and damage them. Conversely, benign tumors do not spread metastasize to the other parts of the body [5], yet they can be formed anywhere in the body. Breast cancer is activated by the malignant tumor in the breast. Thus, treatment should be considered for an accurate diagnosis of the tumors [6] to increase the survival rates and build up a chance of recovery significantly.

The survivability prediction of breast cancer is a challenging and intricate research task. An accurate and reliable system is required to early diagnose malignant and benign tumors. Several methods have been performed to diagnose cancer and help reduce the cancer mortality rate. Fine-needle aspiration cytology (FNAC) and mammography are the leading clinical methods used to diagnose breast cancers. Despite the significance of these methods, it has a lack of diagnostic performance satisfaction [7]. The interpretation of mammography screening from the doctor may vary as mammography screening has such limitations of false-positive results [8] and false-negative results [9]. A variety of data mining methods have been adopted to reduce the diagnosis errors and improve the diagnostic performance such as Random Forest, Rotation Forest, Decision Trees, Support Vector Machine (SVM), Bayesian Network, Logistic Regression, Artificial Neural Network (ANN), and the others ensemble of them. Many studies have been carried out to diagnose breast cancer adopting Wisconsin Diagnostic Breast Cancer (WDBC) dataset [10]. Genetic Algorithm (GA) feature selection with data mining methods has been performed to develop a system that differentiated between malignant tumors and benign tumors system [7]. The classification result has shown the Rotation Forest with GA feature selection has achieved the highest accuracy was 99.48%. The C5.0 Decision Tree, SVM, ANN, and ensemble method have been examined to evaluate the classification accuracy [11]. The classification result has shown the ensemble method has the highest accuracy at 98.77% compared to other mentioned methods. The author of [12] proposed a nested ensemble method that executed stacking and voting as a classifier and compared with a single classifier such as BayesNet and Naïve Bayes (NB). The proposed method has achieved the accuracy results of 98.07%.

Multilayer Perceptron (MLP) has been performed to measure the classification accuracy, sensitivity, and specificity of the breast cancer dataset [13]. The result of MLP has been compared with other machine learning methods and shown the outperformance of accuracy at 99.04%. In [5], the author has implemented NB and k Nearest Neighbors (k-NN)

for breast cancer classification, and k-NN has obtained the highest accuracy result at 97.51% and NB was 96.19%. Another author in [14] also has implemented NB as well as SVM and C4.5 Decision Tree to evaluate the effectiveness and efficiency based on accuracy, precision, sensitivity, and specificity. Based on the experimental results, SVM has the highest accuracy of 97.13% with the lowest error rate. A proposed method of feature ensemble learning for breast cancer classification based on Sparse Autoencoders and Softmax Regression is executed in [15]. The proposed method has obtained a result of 98.60% according to its accuracy. The fuzzy logic method has been performed to classify breast cancer with Classification and Regression Trees (CART), Expectation-Maximization (EM) as a clustering method, and Principal Component as the proposed knowledge-based system [16]. Based on the result, the proposed method has achieved an accuracy result at 93.2%. Radial Basis Function Network (RBFN), Generalized Regression Neural Network (GRNN), and Feed Forward Neural Network (FFNN) are ensemble neural networks that have been employed to distinguish the malignant tumors and benign tumors [17]. The result of the proposed hybrid method has shown the performance classification accuracy was 96.43%.

The objective of this study is to assess the performance of breast cancer classification in terms of accuracy, sensitivity, and specificity for malignant tumors and benign tumors using machine learning techniques based on Bagging, Random Forest, AdaBoost, SVM, and k-NN as a classifier. The paper is structured by the following sections; Section 2 provides the materials and methods used for each classifier and Section 3 is the result and discussion for each classifier according to the classification performance.

II. MATERIALS AND METHODS

A. Breast Cancer Dataset

This study used Wisconsin Diagnostic Breast Cancer (WDBC) dataset obtained from the UCI Machine Learning Repository [10]. The dataset consists of 569 patients and 30 integer-valued attributes. The response class is categorized into two classes which are malignant tumors with 212 instances, and benign tumors with 357 instances. The attributes included in the dataset have the properties as follows; 1) radius, 2) texture, 3) perimeter, 4) area, 5) smoothness, 6) compactness, 7) concavity, 8) concave, 9) symmetry, and 10) fractal dimensions. Each attribute has mean, standard error, and "worst", which results to 30 attributes.

B. Feature Selections

Several attributes in the WDBC dataset are more selective and decisive. Thus, feature selection was implemented to verify the attributes using the gain ratio as the goodness of split employed by decision tree algorithms. Mainly, the purpose of the decision tree algorithm is to choose which attribute to select for each node in the tree. Entropy is implemented in the decision tree algorithm to search from the attributes in the dataset that provide information to produce a decision tree. The entropy value of the t set is calculated as follows,

$$\text{Entropy}(t) = \sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t) \quad (1)$$

where p($i|t$) demote the element of records refer to class i at a given node t. c is the number of classes. The gain, Δ is also known with the term information gain used to determine the performance of the attribute test condition. The equation of Δ info is as follows,

$$\Delta\text{info} = \text{Entropy}(t) - \sum_{j=1}^{k} \frac{N(v_j)}{N} Entropy(v_j) \quad (2)$$

where N is the sum of records, k is the value of the attribute, and $N(v\_j)$ is the number of records associated with the child node, (v_j). The splitting criterion knows as a gain ratio is used to identify the goodness of a split. The splitting ratio is calculated to normalize the information gain as equation follows,

$$\text{SplitInfo}(S) = \sum_{i=1}^{k} P(v_i) \log_2 P(v_i) \quad (3)$$

where k is the total number of splits. The gain ratio is then calculated as,

$$\text{Gain ratio} = \frac{\Delta\text{info}}{SplitInfo(S)} \quad (4)$$

The attribute with a gain ratio value of less than 0.1 is omitted.

C. Performance Evaluation

The performance of the classification model is evaluated by the counts of test records which are correctly and incorrectly predicted models. The counts of test records are formulated using a confusion matrix as shown in Table I. The measurement of accuracy, sensitivity, and specificity is calculated from the confusion matrix.

- True Positive (TP) is the data that the patient who has identified with breast cancer.

- False Positive (FP) is the data that the normal patient who has identified with breast cancer.

- True Negative (TN) is the data that the normal patient who has not identified breast cancer.

- False Negative (FN) is the data that the patient who has not identified with breast cancer.

TABLE I.   CONFUSION MATRIX

		Predicted Class	
		Class = 1	**Class = 0**
Actual Class	**Class = 1**	TP	FN
	Class = 0	FP	TN

The equation of accuracy, sensitivity, and specificity are as below,

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (5)$$

$$\text{Sensitivity} = \frac{TN}{FP+TN} \quad (6)$$

$$\text{Specificity} = \frac{TP}{TP+FN} \quad (7)$$

500

III. CLASSIFICATION METHOD

A. K-Nearest Neighbors

k-Nearest Neighbors (k-NN) is a supervised machine learning technique used for classification and regression problems [24]. The initialization of the input parameter K is several classes in the dataset that used a small value and positive integer. The input data is classified by the majority of its neighbors. The k-NN algorithm needs to run several times with different K values and choose the K that reduces the number of errors and maintains the prediction accuracy. Thus, in this case, the input parameter K of the breast cancer dataset is 3. A brute force search algorithm is implemented by using the Euclidean distance function for the nearest neighbor search as in Eq. (8). The function of Euclidean distance is used to compute the distance between instances that is good for numeric data on the same scale.

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (8)$$

B. Support Vector Machine

Support Vector Machine or SVM is a supervised learning technique used for classification [22] and regression. SVM is a well-known technique in machine learning and extensively implemented in cancer diagnosis. Principally, the function of SVM to classify the outcomes by mapping data between input vectors to a huge perspective space. Thus, the main objective of SVM is to determine the optimal hyperplane by dividing the dataset into classes. Linear classifier aims to fully utilize the distance between the decision hyperplane and the marginal distance, which is the nearest data point [23]. In this study, SVM is implemented to obtain the performance accuracy for malignant tumors and benign tumors. The complexity parameter C is used to control the flexibility of the process to draw lines to isolate the classes, and in this case, C-Support Vector Classification (C-SVC) is selected as the SVM type. The type of SVM is used Linear regression is selected as the key parameter in SVM classifier with normalized data.

C. Ensemble Methods – Bagging

Bagging is one of the most popular techniques in ensemble methods and is also known as bootstrap aggregation. Bagging is one of the earliest and simplest algorithms developed by [19]. This method can be used to reduce the variance for the algorithms that have high variance such as decision trees. In this study, bagging is used to predict breast cancer for malignant tumors and benign tumors. The fast decision tree learner algorithm is used as the default classifier to enhance the classification accuracy. The algorithm builds a decision or regression tree using information gain and prunes it using reduced-error pruning with back fitting. The lack of values is coped with by dividing the corresponding instances into bits. The number of iterations to be performed is set to 100 iterations. The final decision trees are obtained as a composition of all base classifiers with the maximum votes.

D. Ensemble Methods – AdaBoost

Adaptive Boosting or also known as AdaBoost was developed by Freud and Robert Shapyr [21]. AdaBoost is an ensemble machine learning algorithm used for classification problems. The main principle of AdaBoost is to fit a sequence of weak learner models that are slightly better than random guessing. Each instance in the training dataset is weighted to determine the accuracy either it is classified correctly or incorrectly. The J48 decision tree is used as a classifier for the AdaBoost model. The main purpose of the J48 decision tree is to predict the target variable of the new dataset record. The number of iterations in AdaBoost is performed for 100 iterations. The final prediction is then obtained from a combination of the predicted models based on a weighted majority vote (classification) or weighted sum (regression).

E. Ensemble Methods – Random Forest

Random forest is a continuation of bagging for decision trees. The method can be either used for regression or classification. The random forest was constructed by [20], which to increase stability to be compared with single decision trees. This method has various decision trees to ensemble the forest of trees. In this study, a random forest can be used to classify the malignant tumors and benign tumors as it has the ability to manage minimum votes. The number of attributes is the key parameter of this method, which the default parameter is set to 0. It selects the value automatically according to the rule of thumb. The number of iterations in the random forest required 100 iterations to obtain a good balance in processing time.

IV. RESULTS AND DISCUSSION

Data pre-processing is the first stage to be completed before the simulation on all models. According to the breast cancer dataset, there is one attribute for the class tag and one attribute for the ID variable. The ID attribute is a sequence number of the subject and not an attribute to be evaluated. Thus, it has been manually removed. A further step is gain ratio as a feature selection method. The gain ratio is implemented into the dataset, whereas the highest ratio is selected. The attribute with the gain ratio values less than 0.1 have been omitted. Therefore, the number of attributes used in this study has been reduced to 23 attributes. After the feature selection is selected, random sampling is carried out to solve the unbalance class distributions. Then, normalization is performed with the range [0-1] to enhance the success rate of the classification method. The graph of all selected attributes according to its mean based on gain ratio is illustrated in Fig. 1.

The classification method is executed afterward using Bagging, Random Forest, AdaBoost, SVM, and kNN methods. In this study, the libSVM is used for computing the SVM in WEKA [25]. The process of the classification system is demonstrated in Fig. 2. The feature selection and data-preprocessing have shown a successful result in classifying the dataset either correctly or incorrectly. Based on the result in Table II, it is shown that AdaBoost has succeeded to properly distribute the data and followed by the SVM method with only 1 difference.

TABLE II.   CLASSIFICATION RESULTS BASED ON CORRECTLY AND INCORRECTLY CLASSIFIED DATA

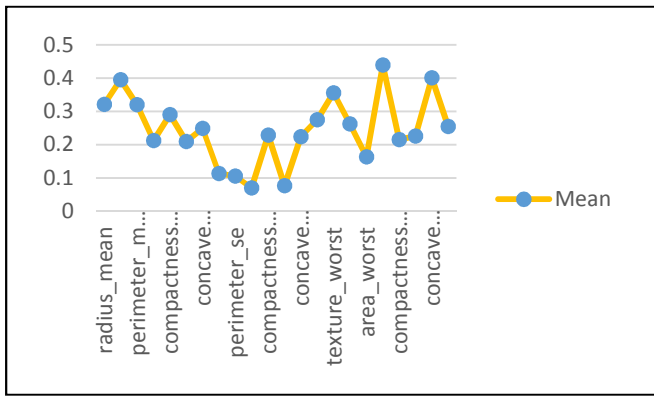| Methods	Correctly Classified	Incorrectly Classified
Bagging	545	24
Random Forest	556	13
AdaBoost	562	7
SVM	561	8
kNN (K=3)	556	13

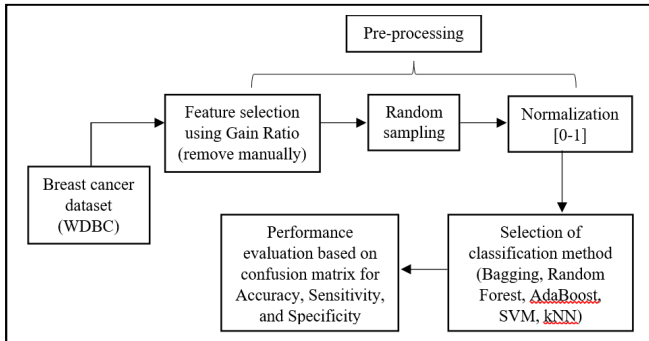Fig. 1. Selected attributes according to its mean based on the gain ratio.

Fig. 2. Process of the classification system.

In Fig. 3, the scatterplot is illustrated the perimeter mean versus concave point worst according to data points of malignant and benign tumors which consist of 569 instances. Based on the observations, the relationship between those attributes shown the positive linear relationship exists with a few outliers from malignant data. However, it is presented somewhat scattered in a wider band. The correlation of two variables for the data indicated strong correlations as evidenced by the much cleaner line formed by both data points, albeit the one outlier from malignant data scattered around benign data. The outliers from the data may lead to a high error rate of breast cancer data.

The confusion matrix was computed for each model to obtain the prediction of class significantly. Confusion matrices associated with the five different machine learning techniques are given in Table III. AdaBoost has the highest predicted class for malignant which the patient has been identified with breast cancer. However, AdaBoost has the same predicted class for benign with the SVM method which the patient has not been identified with breast cancer. The number of predicted classes are computed based on the confusion matrices to obtain accuracy, sensitivity, and specificity.

The comparison for accuracy, sensitivity, and specificity of the classification methods in this study are illustrated in Fig. 4. The classification accuracy results for k-NN and random forest are equally estimated at 97.72% respectively. However, k-NN has a lower error rate at 0.0255 compared to the random forest at 0.0448 and it is shown that the accuracy rate of k-NN is better than random forest. Furthermore, bagging has the lowest accuracy rate at 95.78% and lowest specificity at 92.52% for overall performance evaluation. AdaBoost has the highest accuracy at 98.77%, and thus it is the most successful method to be compared with other classification methods.

This is followed by SVM with an accuracy of 98.59% and the difference is 0.18% with AdaBoost. Based on the sensitivity results, AdaBoost and SVM have the same highest sensitivity at 99.44%, which the number of patients who is correctly identified as identified with breast cancer is high. While AdaBoost has a huge percentage of specificity at 97.66% respectively, which has the highest number of patients who are not identified with breast cancer. Based on the performance evaluation, it is concluded that AdaBoost has the highest classification accuracy for the breast cancer data at 98.77% respectively.
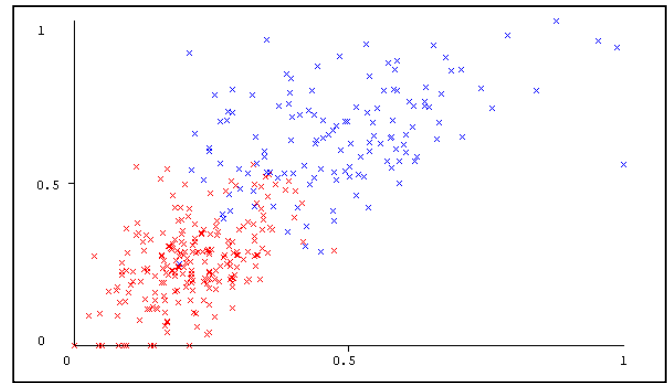

Fig. 3. Scatter plot between perimeter mean (x) and concave point worst (y).

TABLE III. CONFUSION MATRICES OF THE FIVE MACHINE LEARNING TECHNIQUES

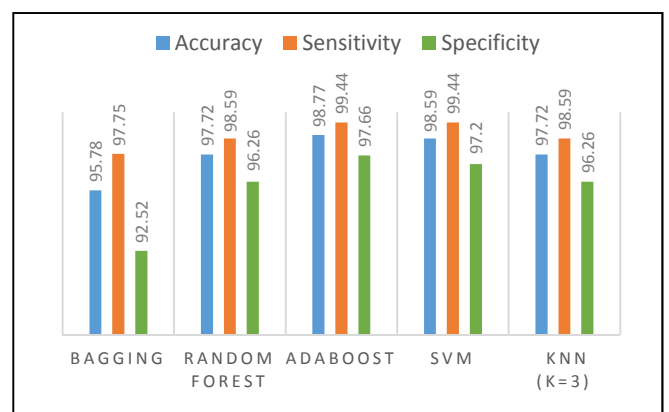| Methods	Predicted Class		Actual Class
	Malignant (M)	Benign (B)	
Bagging	198	16	M
	8	347	B
Random Forest	206	8	M
	5	350	B
AdaBoost	209	5	M
	2	353	B
SVM	208	6	M
	2	353	B
kNN (K=3)	206	8	M
	5	350	B

Fig. 4. Comparison of performance evaluation for classification methods.

The author in [14] has implemented SVM, NB, C4.5, and k-NN methods into the breast cancer dataset. The result found that SVM has the superlative classification accuracy at 97.13% compared to the other methods in the proposed study. The proposed study has applied 10-fold cross validation and data pre-processing. Regardless of applying the data pre-processing into the dataset, the classification accuracy of the method is slightly lower than our proposed method. Furthermore, the dataset used was an original Wisconsin breast cancer dataset that contained 699 patients and 10 attributes. The author has eliminated the ID attribute from the dataset and not omitted attributes that comprise the gain ratio value less than 0.1. Therefore, our proposed method has shown better results in terms of accuracy and low error rate. The comparison of the methods used in this study and the literature is tabulated in Table IV according to the accuracy rate.

The input parameter K of k-NN method in [5] used the value of 3, which was equal to the input parameter K in our proposed method. According to the experimental studies, the authors had divided the cross validation into 60% training sets and 40% testing sets. The study used WBCD which the attribute characteristics are an integer with 699 instances as compared to our proposed method that used attribute characteristics of real with 569 instances. The accuracy comparison has noticed that our proposed method for k-NN is somewhat higher than the authors in [5] by the difference of 0.21% respectively. However, the accuracy result is also dependent on the data pre-processing stage, and if the data-preprocessing has been completed successfully into the dataset, the result may increase or decrease.

The proposed method used in this study almost surpassed the other methods proposed by the author based on literature. Nonetheless, this is not including Rotation Forest with the highest accuracy rate at 99.48%, respectively. The author has implemented the feature selection namely genetic algorithm-based to identify the best attributes, and thus the method has outperformed among others. While the classification accuracy of MLP is 99.04%, the second-highest accuracy respectively. The author has conducted the ratio of training and testing at 60:40 and 70:30, respectively. Based on the accuracy result, the training and testing of 70:30 have produced the best accuracy rate. AdaBoost has the same accuracy rate as the author in [11] that implemented C5.0, SVM, ANN methods. Moreover, it is the third-highest classification accuracy for the breast cancer dataset. Thus, AdaBoost model has improved the accuracy rate equitably and sensitivity rate significantly.

TABLE IV.   COMPARISON OF MACHINE LEARNING METHODS USED IN THIS STUDY AND THE LITERATURE

Author(s)	Methods	Accuracy (%)
[14]	SVM	97.13
[5]	k-NN	97.51
[7]	Rotation Forest with GA	99.48
[11]	C5.0, SVM, ANN	98.77
[12]	Nested Ensemble (Stacking, Voting)	98.07
[13]	MLP	99.04
[15]	Sparse Encoders, Softmax Regression	98.60
[16]	Fuzzy Logic	93.2

[17]	RBFN, GRNN, FFNN	96.43
This study	AdaBoost	98.77
	SVM	98.59
	k-NN	97.72
	Random Forest	97.72
	Bagging	95.78

Furthermore, k-fold cross validation namely 2-fold, 3-fold, and 5-fold have been implemented to indicate the accuracy rate and error rate of machine learning methods used in this study. The comparison of the methods according to the 2, 3, 5-fold cross validation is shown in Figure 6. According to 2-fold cross validation, AdaBoost has the highest accuracy as compared to the other machine learning methods in this study with a 98.41% accuracy rate and a 0.04 error rate. SVM and Random Forest have produced the same accuracy rate at 97.36%, however, SVM has shown a better accuracy result due to the lowest error rate compared to Random Forest. Thus, SVM has placed as a second highest accuracy rate with 0.03 error rate, respectively.

Not standing with that, AdaBoost has still shown a better accuracy rate for 3-fold cross validation, which at 98.24% accuracy rate and 0.02 error rate. Nevertheless, the method has come to the second-highest accuracy for 5-fold cross validation at the accuracy rate of 98.07%, respectively. The implementation of 5-fold cross validation to the machine learning methods in this study has presented SVM as a higher accuracy method at 98.60% accuracy rate with the lowest error rate at 0.01, respectively. The graph as illustrated in Fig. 5 has indicated the differentiation of increasing accuracy rate of Adaboost and SVM, and the comparison according to k-fold cross validation for all methods in this study.
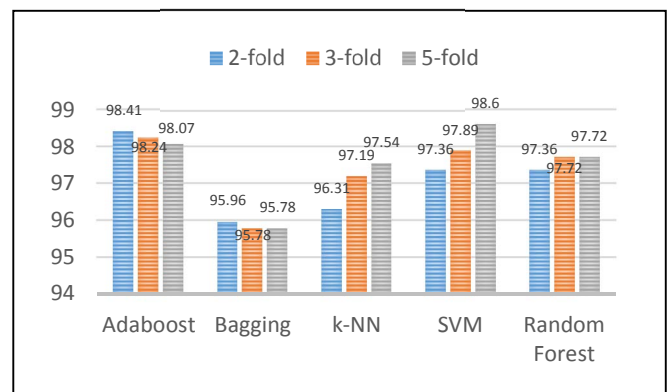

Fig. 5.   Comparison according to k-fold cross validation.

V.   CONCLUSION

Wisconsin Diagnostic Breast Cancer (WDBC) dataset used in the proposed study to predict the classification model of breast cancer that consists of malignant and benign tumors. Cross validation is implemented 10-fold into the dataset. The data-preprocessing stage is run through the dataset to random sample and normalize the dataset. Few attributes are omitted which have no significant value for the classification process. Variation of machine learning techniques has been proposed to correctly classify the data namely as bagging, random forest, AdaBoost, SVM, and k-NN. According to the results, AdaBoost has achieved the highest accuracy at 98.77%. The

accuracy results in this study are compared to the previous research works that used the breast cancer dataset. In addition to accuracy, specificity and sensitivity are also counted in this study to find the number of patients with breast cancer and without having breast cancer, approximately. The experiment with 2-fold, 3-fold, and 5-fold cross validation also has been implemented to the proposed methods. The result shows AdaBoost has produced best accuracy with 98.41% and 98.24%, respectively, for 2-fold and 3-fold cross validation. However, the accuracy rate is decreased with 5-fold cross validation, whereas SVM shows the highest accuracy at 98.60% with a 0.01 error rate. In future work, various ensemble techniques may be employed on the newly proposed methods to enhance the diagnostic accuracy of breast cancer. Furthermore, numerous feature selection techniques to manage complexity and a huge number of breast cancer data can be extended in the future.

ACKNOWLEDGMENT

REFERENCES

[1] N. I. o. Health, "National Cancer Institute. Surveillance, Epidemiology, and End Results Program. Cancer stat facts: female breast cancer," ed, 2018.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," CA: a cancer journal for clinicians, vol. 69, no. 1, pp. 7-34, 2019.

[3] A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques," Age, vol. 58, no. 13, pp. 10-110, 2006.

[4] A. Saygılı, "Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers," International Scientific and Vocational Studies Journal, vol. 2, no. 2, pp. 48-56, 2018.

[5] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensarİ, "Breast cancer classification using machine learning," in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2018, pp. 1-4: IEEE.

[6] M. Montazeri, M. Montazeri, M. Montazeri, and A. Beigzadeh, "Machine learning models in breast cancer survival prediction," Technology and Health Care, vol. 24, no. 1, pp. 31-42, 2016.

[7] E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest," Neural Computing and Applications, vol. 28, no. 4, pp. 753-763, 2017.

[8] R. A. Hubbard, K. Kerlikowske, C. I. Flowers, B. C. Yankaskas, W. Zhu, and D. L. Miglioretti, "Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: a cohort study," Annals of internal medicine, vol. 155, no. 8, pp. 481-492, 2011.

[9] J. G. Elmore, C. K. Wells, C. H. Lee, D. H. Howard, and A. R. Feinstein, "Variability in radiologists' interpretations of mammograms," New England Journal of Medicine, vol. 331, no. 22, pp. 1493-1499, 1994.

[10] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast cancer Wisconsin (diagnostic) data set," UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/], 1992.

[11] G. Zorluoglu and M. Agaoglu, "Diagnosis of breast cancer using ensemble of data mining classification methods," International Journal of Oncology and Cancer Therapy, vol. 2, 2017.

[12] M. Abdar et al., "A new nested ensemble technique for automated diagnosis of breast cancer," Pattern Recognition Letters, 2018.

[13] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," in Proceedings of the 2nd International Conference on Machine Learning and Soft Computing, 2018, pp. 5-9.

[14] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," Procedia Computer Science, vol. 83, pp. 1064-1069, 2016.

[15] V. J. Kadam, S. M. Jadhav, and K. Vijayakumar, "Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression," Journal of medical systems, vol. 43, no. 8, p. 263, 2019.

[16] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," Telematics and Informatics, vol. 34, no. 4, pp. 133-144, 2017.

[17] E. Yavuz, C. Eyupoglu, U. Sanver, and R. Yazici, "An ensemble of neural networks for breast cancer diagnosis," in 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 538-543: IEEE.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10-18, 2009.

[19] L. Breiman, "Heuristics of instability and stabilization in model selection," The annals of statistics, vol. 24, no. 6, pp. 2350-2383, 1996.

[20] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[21] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in European conference on computational learning theory, 1995, pp. 23-37: Springer.

[22] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273-297, 1995.

[23] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and structural biotechnology journal, vol. 13, pp. 8-17, 2015.

[24] P. Hall, B. U. Park, and R. J. Samworth, "Choice of neighbor order in nearest-neighbor classification," The Annals of Statistics, vol. 36, no. 5, pp. 2135-2152, 2008.

[25] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM transactions on intelligent systems and technology (TIST), vol. 2, no. 3, pp. 1-27, 2011.