© Universiti Tun Hussein Onn Malaysia Publisher's Office

IJIE

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

The International Journal of Integrated Engineering

Breast Cancer Classification: Features Investigation Using Machine Learning Approaches

Nurul Amirah Mashudi[1], Syaidathul Amaleena Rossli[2], Norulhusna Ahmad[3], Norliza Mohd Noor[4*]

[1,2,3,4]Razak Faculty of Technology and Informatics,
Universiti Teknologi Malaysia, Kuala Lumpur, 54100, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/ijie.2021.13.05.012
Received 20 April 2021; Accepted 5 May 2021; Available online 31 July 2021

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. This study used a dataset obtained from Wisconsin Diagnostic Breast Cancer (WDBC) with 23 selected features measured from 569 patients, from which 212 patients have malignant tumors and 357 patients have benign tumors. The analysis was performed to investigate the feature of the tumors based on its mean, standard error, and worst. Each feature has ten properties which are radius, texture, perimeter, area, smoothness, compactness, concavity, concave, symmetry and fractal dimensions. The selection of features was considered a significant influence to the breast cancer. The analysis is compared and evaluated with thirty features to determine the features used for breast cancer classification. The result shown AdaBoost has obtained the highest accuracy for thirty features at 98.95%, ten features of mean at 98.07%, and ten features of worst at 98.77% with a lowest error rate. Additionally, the proposed methods are classified using 2-fold, 3-fold, and 5-fold cross validation to meet the best accuracy rate. 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 shows SVM produced the best accuracy rate at 98.60% with the lowest error rate.

Keywords: Breast cancer, classification, machine learning

1. 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 abnormal cells to the body tissues and

*Corresponding author:norliza@utm.my

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. The malignant tumor activates breast cancer in the breast. Thus, treatment should be considered for an accurate diagnosis of the tumors [6] to significantly increase survival rates and build up a chance of recovery.

The survivability prediction of breast cancer is a challenging and intricate research task. An accurate and reliable system is required to diagnose malignant and benign tumors early. 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 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 that 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 that 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 breast cancer dataset's classification accuracy, sensitivity, and specificity [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 and SVM and C4.5 Decision Tree to evaluate the effectiveness and efficiency based on accuracy, precision, sensitivity, and specificity. SVM has the highest accuracy of 97.13% with the lowest error rate based on the experimental results. 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 of 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 that the performance classification accuracy was 96.43%.

Mammogram images were evaluated using a learning classifier to cope with many features in image classification [18]. As there are large numbers of discrete wavelet transform and binary pattern, the learning method is impractical in this case. The authors have reduced the number of input features to their working solution. Several features and distance conditions were formed to demonstrate the progress of classifier rules. A comprehensive approach to identify benign or malignant breast tumor cell classification was proposed in [19] using breast cell histopathology (BCH) images. The study describes a hybridization of the two classifiers, which are SVM and chain-like techniques such as shape-based and texture-based features, with a classification accuracy of 96.19%. The method can be improved by testing using a large dataset. An adaptive computer-aided-design (CAD) system was used for the size of breast tumor classification [20]. The quantitative morphological and texture features were performed for breast cancer classification. The limitation of the CAD system is that the image may degrade due to speckle noise and other noises. A recursive feature elimination (RFE) and randomized logistic regression (LRL) were used to eliminate unnecessary features [21]. The process of elimination was stopped when the number of features has reached 50 features. The authors have implemented eight machine learning techniques such as SVM, k-NN, decision tree, random forest, logistic regression, Adaboost, and gradient boosting machines for breast cancer classification. The experiment began with the first dataset consisted of 1919 features and 133 samples, and then the second dataset consisted of 24,481 features and 97 samples.

The feature elimination of both datasets has produced results more accurately. Some machine learning and deep learning algorithms were used in [22], such as SVM, random forest, recurrent neural network (RNN), and convolutional neural network (CNN) for breast cancer classification using the Coimbra dataset. The study was performed a feature selection using information gain and ReliefF to select the attributes. The total attributes in Coimbra dataset were ten attributes. The selected attributes with information gain were five attributes, and the selected attributes with ReliefF were six attributes. The selection of attributes seems to have good accuracy results on classification techniques. Feature selection on the classifier appears to be more inefficient since classifiers were built from raw data that did not go through the feature selection process. An attribute selection method based on random forest and principle component analysis (PCA) was proposed to diagnose breast cancer [23]. The breast cancer categorical data

that consists of 30 attributes has been reduced based on random forest, and 21 attributes have been selected based on PCA. With the seven significant features, the PCA was used to construct an ELM classification system with a diverse variety of activation functions. The performance of feature selection on random forest or feature extraction on PCA can minimize the sophistication of the training model and increase the accuracy. The limitations in the study are that the optimum precision of all samples is not achieved. The performance of the model can be improved with high prediction accuracy and speed.

This study aims 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. In addition, this study investigates the feature used for breast cancer on malignant and benign tumors. The following sections structure the paper: Section 2 provides the materials and methods used for each classifier We present the classification methods used in this paper in Section 3, and Section 4 is the result and discussion for each classifier according to the classification performance. Lastly, we conclude the paper in Section 5.

2. Materials and Methods

2.1 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 features. The response class is categorized into two classes: i) malignant tumors with 212 instances and ii) benign tumors with 357 instances. The features 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 feature has mean, standard error, and "worst", which results in 30 features. The description of the WDBC dataset is shown in Table 1.

Table 1 - Description of WDBC dataset

Attribute Number	Attributes	Attribute Range		
		Mean	Standard Error	Worst
1	Radius	6.981 - 28.11	0.1115 - 2.873	7.93 - 36.04
2	Texture	9.71 - 39.28	0.3602 - 4.885	12.02 - 49.54
3	Perimeter	43.79 - 188.5	0.757 - 21.98	50.41 - 251.2
4	Area	143.5 - 2501	6.802 - 542.2	185.2 - 4254
5	Smoothness	0.05263 - 0.1634	0.001713 - 0.03113	0.07117 - 0.2226
6	Compactness	0.01938 - 0.3454	0.002252 - 0.1354	0.02729 - 1.058
7	Concavity	0 - 0.4268	0 - 0.396	0 - 1.252
8	Concave points	0 - 0.2012	0 - 0.05279	0 - 0.291
9	Symmetry	0.106 - 0.304	0.007882 - 0.07895	0.1565 - 0.6638
10	Fractal dimension	0.04996 - 0.09744	0.0008948 - 0.02984	0.05504 - 0.2075

2.2 Feature Selections

Several features in the WDBC dataset are more selective and decisive. Thus, feature selection was implemented to verify the features 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 feature to select for each node in the tree. Entropy is implemented in the decision tree algorithm to search from the features in the dataset that provide information to produce a decision tree. The entropy value of the t set is calculated as follows,

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

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 feature test condition. The equation of Δ info is as follows,

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

where N is the sum of records, k is the value of the feature, 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 feature with a gain ratio value of less than 0.1 is omitted.

2.3 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 2. 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 2 - 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,

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

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

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

3. Classification Methods

3.1 Features Classification

As mentioned in Section 2.1, the breast cancer dataset consists of three types of features: i) mean, ii) standard error and iii) worst. This study evaluates the features of malignant and benign tumors to investigate the features used in breast cancer. The properties of the features such as radius, texture, and area were selected to analyze the features in-depth. These properties provide a significant value on breast cancer classification. Furthermore, an analysis was performed for each of the features that comprise ten properties. The analysis is compared and evaluated with the combination of thirty features to determine the features used for breast cancer classification. Therefore, this study can identify the characteristics of malignant and benign tumors based on the result achieved.

3.2 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 majority of its neighbors classifies the input data. The k-NN algorithm needs to run several times with different K values and choose the K to reduce the number of errors and maintain 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 using the Euclidean distance function for the nearest neighbor search as in Eq. (8). Euclidean distance is used to compute the distance between instances that is good for numeric data on the same scale.

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

3.3 Support Vector Machine

Support Vector Machine or SVM is a supervised learning technique used for classification [25] 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 [26]. 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. In this case, C-Support Vector Classification (C-SVC) is selected as the SVM type. Linear regression is selected as the key parameter in SVM classifier with normalized data.

3.4 Ensemble Method - 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 [27]. 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.

3.5 Ensemble Method - AdaBoost

Adaptive Boosting or also known as AdaBoost was developed by Freud and Robert Shapyr [28]. 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 primary 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).

3.6 Ensemble Method - 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 [29] 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 features 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.

4. 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 feature for the class tag and one feature for the ID variable. The ID feature is a sequence number of the subject and not a feature 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 feature with gain ratio values less than 0.1 has been omitted. Therefore, the number of features used in this study has been reduced to 22 features such as i) mean: radius, texture, perimeter, area, compactness, concavity, concave points, ii) standard error: radius, perimeter, area, compactness, concavity, concave points, and iii) worst: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry. 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. According to its mean based on gain ratio, the graph of all selected features is illustrated in Fig. 1. The analysis has conducted the features for mean, standard error, and worst separately that consist of 10 features. The features were identified based on the radius, texture, perimeter, area, smoothness, compactness, concavity, concave, symmetry, and fractal dimensions.

The classification method is executed afterward using Bagging, Random Forest, AdaBoost, SVM, and k-NN methods. In this study, the libSVM is used for computing the SVM in WEKA [30]. 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 3, it is shown that AdaBoost has succeeded to properly distribute the data and followed by the SVM method with only one difference.

Table 3 - 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
k-NN	556	13

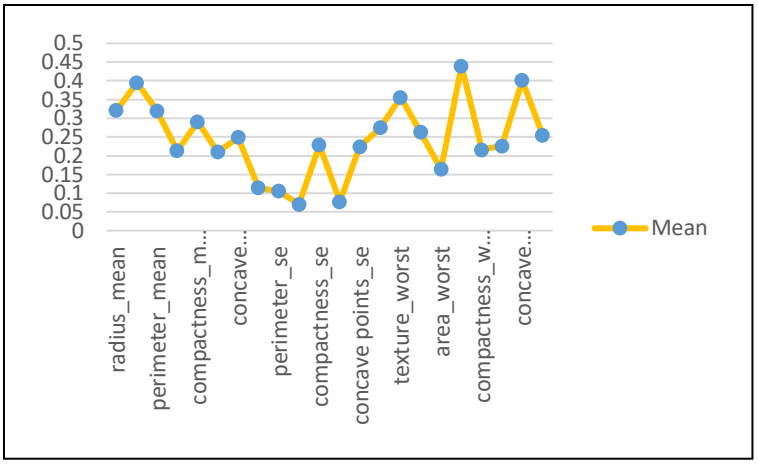

Fig. 1 - Selected features according to its mean based on the gain ratio

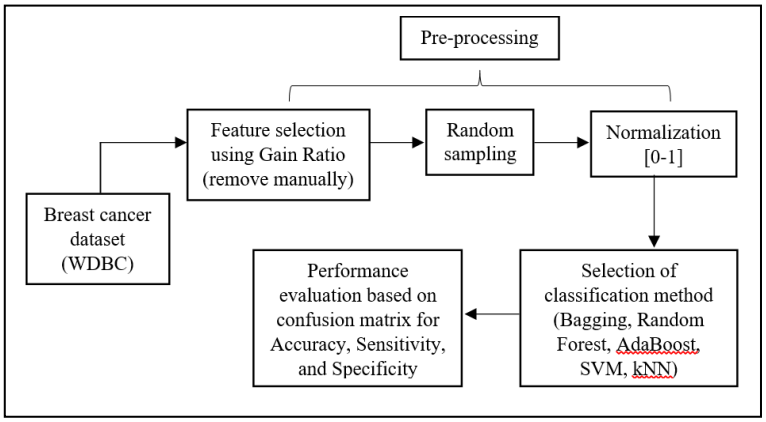

Fig. 2 - Process of the classification system

In Fig. 3, the scatterplot shows 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 features shown that the positive linear relationship exists with a few outliers from malignant data. However, it is presented somewhat scattered in the broader 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 scatter plot in Fig. 4 shows the features of radius mean and worst for malignant and benign tumors. Most of the patients have a radius between 8 mm to 15 mm of the benign tumor, while the number of patients with malignant tumor is scattered and outliers. The highest value of radius mean and worst is 28.11 mm for malignant tumors. Hence, these two features may impact breast cancer classification, particularly for malignant tumors with inconsistent features.

The pattern of the scatter plot in Fig. 5 demonstrates an assembled texture feature between the mean and worst for a malignant and benign tumor. The features are varied for all patients. Only two patients with a malignant tumor have a high texture of worst feature at 44.87 and 45.41. Meanwhile, the mean and worst texture for benign tumors has diverged since each patient has different types of textures in the tumor. The corresponding pattern is different for the area feature in which the amount of the area is assembled. The area feature is one of the significant effects for breast cancer which the sizing of the tumor will affect the condition of the patient.

The worst malignant tumor has an area sizing of 4254 cm as illustrated in Fig. 6. Generally, a malignant tumor has a more extensive tumor area than a benign tumor due to this tumor being harmful. Therefore, many patients with malignant tumors will have a large area scale.

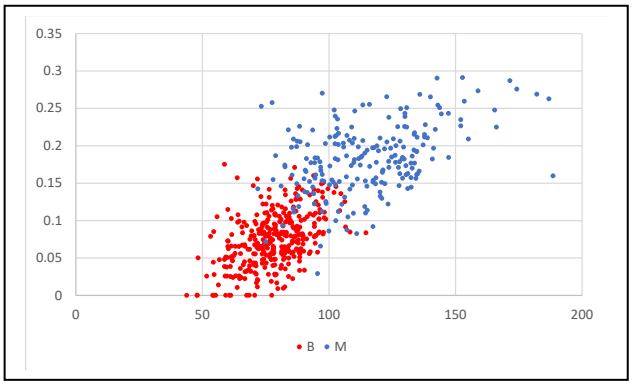

Fig. 3 - Scatter plot of perimeter mean and concave point worst for benign (B) and malignant (M)

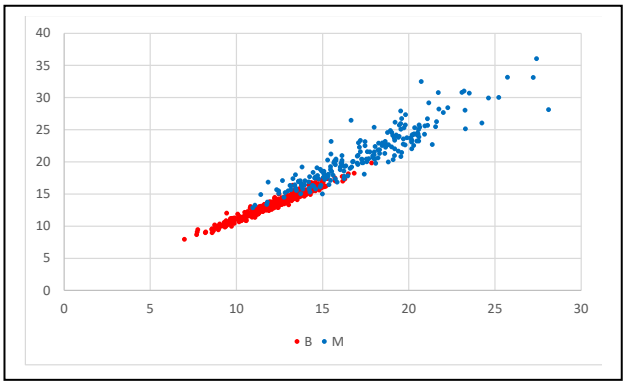

Fig. 4 - Scatter plot of radius mean and radius worst for malignant and benign tumors

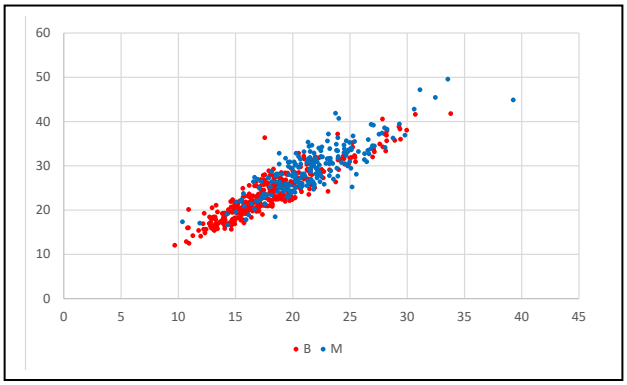

Fig. 5 - Scatter plot of texture mean and texture worst for malignant and benign tumors

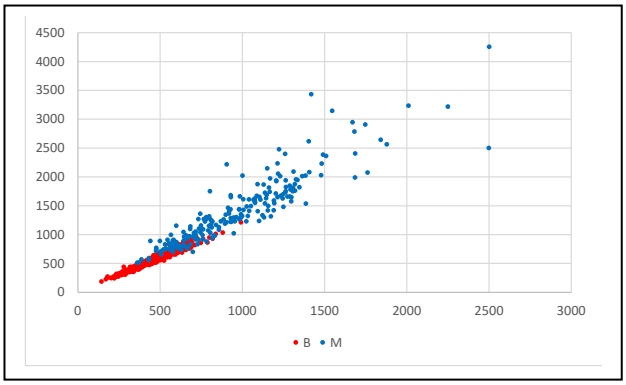

Fig. 6 - Scatter plot of area mean and area worst for malignant and benign tumors

The accuracy result for each feature's properties is compared to the combined set of thirty features, as illustrated in Fig. 10. We conducted this study with the ratio of training and testing of 90:10. The ten features of mean were performed as illustrated in Fig. 7 shows AdaBoost has the highest accuracy of the feature at 98.06% with 0.02 error rate followed by Random Forest at 97.89% with 0.05 error rate, respectively. Nevertheless, the accuracy of AdaBoost has slight decreased for the ten features of standard error at 96.13% with an error rate of 0.04 as shown in Fig. 8. The Random Forest has produced a high accuracy result for the ten features of standard error at 96.66% with an error rate at 0.09, respectively. Therefore, AdaBoost was evaluated as a better classification accuracy with the lowest error rate. The classification accuracy results for AdaBoost and Random Forest are equally estimated at 98.77%, respectively, for the ten features of worst, as demonstrated in Fig. 9. However, AdaBoost has a lower error rate at 0.01 compared to the Random Forest at 0.04, and it is shown that the accuracy rate of AdaBoost is better than Random Forest. Therefore, it can be concluded that AdaBoost has the highest classification accuracy for the feature of mean, standard error, and worst, with the lowest error rate compared to the other methods in this study.

Furthermore, the classification accuracy results for SVM were poor of all features at 92.27% for mean, 78.38% for standard error, and 95.61% for worst. However, the accuracy result of SVM for the 30 features such as mean, standard error, and worst has produced slightly higher than worst feature at 95.96%, as shown in Fig. 10. The reason may vary due to the characteristics of the features as in Fig. 4, Fig. 5, and Fig. 6, especially the ten features of worst. The malignant tumor shows a varied pattern in the area mean and area worst, leading to decreased accuracy.

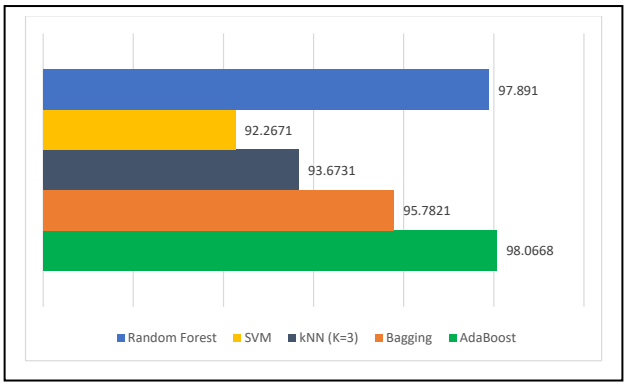

Fig. 7 - Accuracy result of 10 features of 'mean'

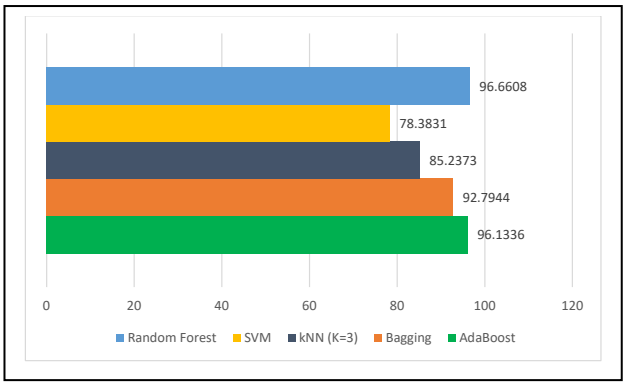

Fig. 8 - Accuracy result of 10 features of 'standard error'

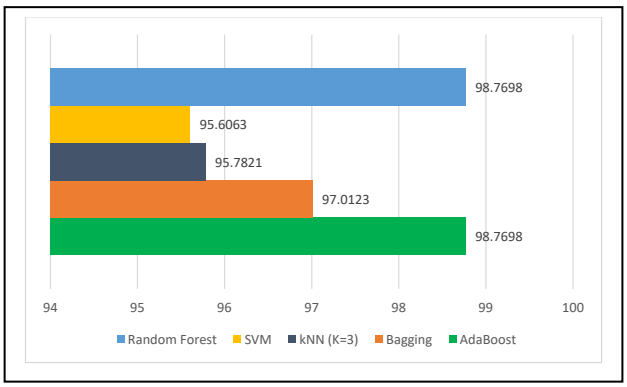

Fig. 9 - Accuracy result of 10 features of 'worst'

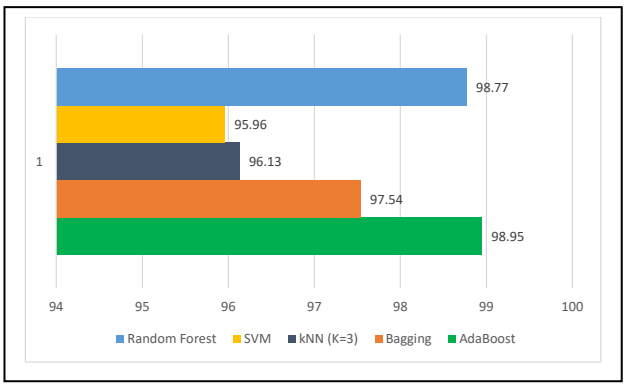

Fig. 10 - Accuracy result of 30 features

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 4. AdaBoost has the highest predicted class for malignant, which the patient has been identified with breast cancer [31]. 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 is 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. 11 with 22 features. 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 accuracy is followed by SVM with 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 are correctly identified as identified with breast cancer is high. While AdaBoost has a more significant 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.

Table 4 - Confusion matrices of the five machine learning techniques

Methods	Predicted Class		Actual Class
	Correctly Classified	Incorrectly Classified	
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
k-NN	206	8	M
	5	350	B

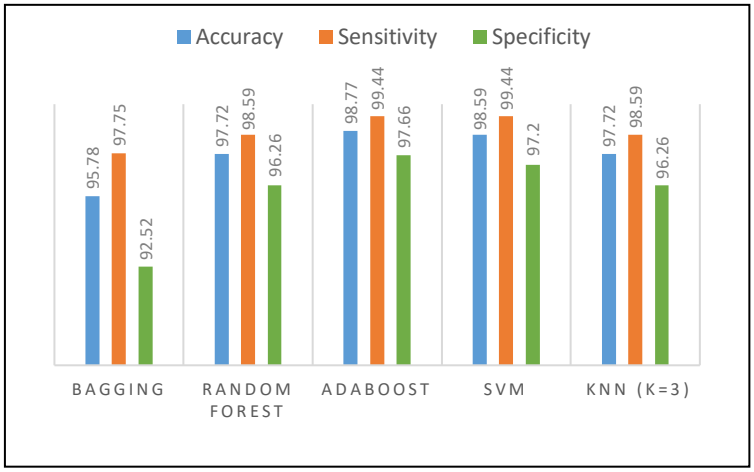

Fig. 11 - Comparison of performance evaluation for classification methods with 22 features

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 ten features. The author has eliminated the ID feature from the dataset and not omitted features 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. Comparing the methods used in this study and the literature is tabulated in Table 5 according to the accuracy rate.

The input parameter K of the 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, in which the feature characteristics are an integer with 699 instances compared to our proposed method that used feature 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 pre-processing data stage. If the pre-processing data 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 the 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 features, and thus the method has outperformed among others. While the classification accuracy of MLP is 99.04%, which the second-highest accuracy in the breast cancer classification for 22 features. 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, the AdaBoost model has improved the accuracy rate equitably and sensitivity rate significantly.

Table 5 - 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	95.43
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, has 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 Fig. 12. According to 2-fold cross-validation, AdaBoost has the highest accuracy of 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 than 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 illustrated in Fig. 12 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.

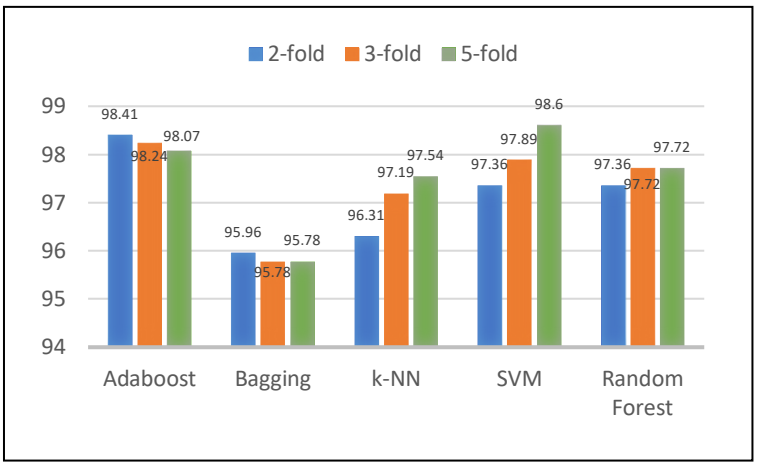

Fig. 12 - Comparison according to k-fold cross validation

5. Conclusion

Wisconsin Diagnostic Breast Cancer (WDBC) dataset is used in the proposed study to predict the classification model of breast cancer that consists of malignant and benign tumors. The features used was investigates based on the mean, standard error, and worst. Each feature has ten properties such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave, symmetry and fractal dimensions. These features were performed to obtain breast cancer classification accuracy. This study investigated ten features for mean, standard error, and worst, and the results were compared with 30 features and 22 features to determine the features used for breast cancer classification. 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 features 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. This research counts precision, specificity, and sensitivity in addition to accuracy to determine the estimated number of patients with and without breast cancer. The result also shown AdaBoost has obtained the highest accuracy for thirty features at 98.95%, ten features of mean at 98.07%, and ten features of worst at 98.77% with a lowest error rate. 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 the 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 breast cancer diagnostic accuracy. Furthermore, numerous feature selection techniques to manage complexity and a considerable number of breast cancer data can be extended in the future.

Acknowledgment

This study was funded by Universiti Teknologi Malaysia (Vote: Q.K.130000.2856.00L66) and the Ministry of Higher Education Malaysia.

References

[1] Health, N. I. o. (2018). National Cancer Institute. Surveillance, Epidemiology, and End Results Program. Cancer stat facts: female breast cancer.
[2] Siegel, R. L., Miller, K. D., and Jemal, A., (2019). Cancer statistics, 2019. CA: a cancer journal for clinicians, 69(1), 7-34.
[3] Bellaachia, A. and Guven, E., (2006). Predicting breast cancer survivability using data mining techniques. Age, 58(13), 10-110.
[4] Saygılı, A., (2018). Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers. International Scientific and Vocational Studies Journal, 2(2), 48-56.
[5] Amrane, M., Oukid, S., Gagaoua, I., and Ensarİ, T. (2018). Breast cancer classification using machine learning, in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), pp. 1-4, IEEE.
[6] Montazeri, M., Montazeri, M., Montazeri, M., and Beigzadeh, A., (2016). Machine learning models in breast cancer survival prediction. Technology and Health Care, 24(1), 31-42.
[7] Aličković, E. and Subasi, A., (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications, vol. 28(4), 753-763.
[8] Hubbard, R. A., Kerlikowske, K., Flowers, C. I., Yankaskas, B. C., Zhu, W., and Miglioretti, D. L., (2011). Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: a cohort study. Annals of internal medicine, 155(8), 481-492.
[9] Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H., and Feinstein, A. R., (1994). Variability in radiologists' interpretations of mammograms. New England Journal of Medicine, 331(22), 1493-1499.
[10] Wolberg, W. H., Street, W. N., and Mangasarian, O. L., (1992). Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/].
[11] Zorluoglu, G. and Agaoglu, M., (2017). Diagnosis of breast cancer using ensemble of data mining classification methods. International Journal of Oncology and Cancer Therapy, vol. 2.
[12] Abdar, M., Zomorodi-Moghadam, M., Zhou, X., Gururajan, R., Tao, X., Barua, P. D., and Gururajan, R., (2018). A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognition Letters.
[13] Agarap, A. F. M. (2018). 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, pp. 5-9
[14] Asri, H., Mousannif, H., Al Moatassime, H., and Noel, T., (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.

[15] Kadam, V. J., Jadhav, S. M., and Vijayakumar, K., (2019). Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression. Journal of medical systems, 43(8), 263.
[16] Nilashi, M., Ibrahim, O., Ahmadi, H., and Shahmoradi, L., (2017). A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics and Informatics, 34(4), 133-144.
[17] Yavuz, E., Eyupoglu, C., Sanver, U., and Yazici, R. (2017). An ensemble of neural networks for breast cancer diagnosis, in 2017 International Conference on Computer Science and Engineering (UBMK), pp. 538-543, IEEE.
[18] Siddique, A., Iqbal, M., and Browne, W. N. (2016). A comprehensive strategy for mammogram image classification using learning classifier systems, in 2016 IEEE congress on evolutionary computation (CEC), pp. 2201-2208, IEEE.
[19] Wang, P., Hu, X., Li, Y., Liu, Q., and Zhu, X., (2016). Automatic cell nuclei segmentation and classification of breast cancer histopathology images. Signal Processing, 122, 1-13.
[20] Moon, W. K., Chen, I.-L., Chang, J. M., Shin, S. U., Lo, C.-M., and Chang, R.-F., (2017). The adaptive computer-aided diagnosis system based on tumor sizes for the classification of breast tumors detected at screening ultrasound. Ultrasonics, 76, 70-77.
[21] Turgut, S., Dağtekin, M., and Ensari, T. (2018). Microarray breast cancer data classification using machine learning methods, in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), pp. 1-3, IEEE.
[22] Basarslan, M. S. and Kayaalp, F. (2021). Performance evaluation of classification algorithms on diagnosis of breast cancer and skin disease, in Deep Learning for Cancer Diagnosis, pp. 27-35, Springer.
[23] Bian, K., Zhou, M., Hu, F., and Lai, W., (2020). RF-PCA: A New Solution for Rapid Identification of Breast Cancer Categorical Data Based on Attribute Selection and Feature Extraction. Frontiers in genetics, 11, 1082.
[24] Hall, P., Park, B. U., and Samworth, R. J., (2008). Choice of neighbor order in nearest-neighbor classification. The Annals of Statistics, 36(5), 2135-2152.
[25] Cortes, C. and Vapnik, V., (1995). Support-vector networks. Machine learning, 20(3), 273-297.
[26] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I., (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.
[27] Breiman, L., (1996). Heuristics of instability and stabilization in model selection. The annals of statistics, 24(6), 2350-2383.
[28] Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting, in European conference on computational learning theory, pp. 23-37, Springer.
[29] Breiman, L., (2001). Random forests. Machine learning, 45(1), 5-32.
[30] Chang, C.-C. and Lin, C.-J., (2011). LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.
[31] Mashudi, N. A., Rossli, S. A., Ahmad, N., and Noor, N. M. (2021). Comparison on Some Machine Learning Techniques in Breast Cancer Classification, in 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), pp. 499-504, IEEE.