STM JOURNALS

Nano Trends:
A Journal of Nanotechnology and Its Applications

ISSN: 0973-418X
Volume 27, Issue 2, 2025
May–August
DOI (Journal): 10.37591/NTs

https://journals.stmjournals.com/nts/

Research NTs

Computer Aided Diagnosis of Breast Cancer Using Machine Learning Techniques

S. Kavitha[1,*], N. Vaijayanthi[2]

Abstract

Breast cancer is one of the significant health problems that lead to early mortality in women, especially those between 40 and 55 years of age all over the world. In recent years, the number of breast cancer cases among women has risen significantly, making early and accurate diagnosis more important than ever. Computer-Aided Diagnostic (CAD) tools have become valuable in supporting radiologists by enhancing the precision of breast cancer detection. This study introduces a CAD algorithm that uses machine learning techniques to classify breast tumors as either benign or malignant. It also explores how feature selection affects classification accuracy, using the Minimum Redundancy Maximum Relevance (MRMR) method to rank and select the most important features. The algorithm is evaluated using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The performance of machine learning models, such as Linear Discriminant Analysis (LDA), K Nearest Neighbor (KNN), Naive Bayes Classifier (NB), Support Vector Machine (SVM), and Decision Tree (DT) classifier are compared. The experimental results show that when implemented with an LDA classifier with feature selection, the proposed algorithm achieves the maximum accuracy of 97.6% compared to other machine learning (ML) models.

Keywords: Breast cancer, diagnosis, machine learning, feature selection, classification, accuracy

INTRODUCTION

Breast cancer is the uncontrolled growth of abnormal cells in the breast. These cells form a tumor that is malignant if the cells spread to other parts of the body. Most breast cancers can start in the duct that carries milk, called ductal cancer, and some breast cancers begin in the glands that make milk, called lobular cancer. Many types of breasts can cause a lump in the breast, but most breast lumps are not cancer and do not spread outside the breast. Breast cancer is the most common type of cancer in women in the world. It leads to early mortality in women across the globe. In recent years, there has been a significant increase in the occurrence of breast cancer in women [1].

***Author for Correspondence**
S. Kavitha
E-mail: kavithas@jjcet.ac.in

[1]Assistant Professor, Department of Computer Science and Engineering, J.J. College of Engineering and Technology, Tiruchirappalli, Tamil Nadu, India.
[2]Dean, Department of Electronics and Communication Engineering, Indra Ganesan College of Engineering, Tiruchirappalli, Tamil Nadu, India.

Received Date: March 25, 2025
Accepted Date: April 03, 2025
Published Date: May 26, 2025

Citation: S. Kavitha, N. Vaijayanthi. Computer Aided Diagnosis of Breast Cancer Using Machine Learning Techniques. Nano Trends: A Journal of Nanotechnology and Its Applications. 2025; 15(2): 1–11p.

According to the UAE national cancer registry report 2017, nearly 825 women and 9 men are affected by breast cancer. This accounts for 20.23% of all malignant cancer types in the country [2]. In 2023, in the United States of America, there will be an estimated 297,790 new breast cancer cases, which is 31% of all the other cancer types and there will be an estimated deaths of 43,170 cases, which is 15% of all the cancer types [3].

There is a worldwide variation in the survival rates of breast cancer. The survival rates in the most developed countries, like North America range from 80% or more, whereas in Sweden and Japan, the survival rates are around 60%. The

middle-income and low-income countries have survival rates of less than 40%. Survival rates in less developed countries remain low, primarily due to the absence of early breast cancer detection and limited access to adequate diagnostic facilities. This leads to a high ratio of women affected by late-stage breast cancer disease [4]. The early diagnosis can reduce the survival rates of breast cancer through automated methods since manual classification may not always be precise [5].

In this work, a computer aided diagnostic algorithm is proposed to classify breast cancer data into benign and malignant based on machine learning models. As the first step, the Wisconsin diagnostic breast cancer data is obtained from Kaggle, 2016 [6]. Then, the data preparation is done, followed by feature selection using the MRMR algorithm. The selected features are used to train the ML models, such as LDA, KNN, SVM, NB and DT classifiers and the performance of ML models is evaluated by five folds of cross-validation of data. To evaluate the performance of the proposed algorithm, key quantitative metrics, such as accuracy, sensitivity, and specificity are used. The main contributions of the work are: (i) Classification of breast cancer data into benign and malignant, (ii) Implementation of Feature selection algorithm to enhance the classification results, and (iii) Comparative analysis of ML models taken for study.

The remainder of this paper is structured as follows: the next section presents a review of existing literature on breast cancer classification. Section III outlines the proposed methodology, while Section IV covers the results and discussion. Finally, Section V concludes the study and suggests directions for future research.

RELATED WORK

This section discusses the existing research carried out for breast cancer classification on breast cancer datasets using machine learning algorithms. Even though there are much research on this breast cancer diagnosis, there are only a few papers focused on the effect of feature selection on diagnosis results.

In the experimental study by Chaurasia et al. [7], the Wisconsin breast cancer dataset was used for cancer diagnosis using ML models, such as NB, SVM, neural networks and DT classifiers, and their performance was compared. The results show that the maximum accuracy of 96.84% was achieved with the SVM classifier. Aruna et al. [8] made an experimental study to classify the breast cancer data using SVM, NB and DT classifiers. The best results were achieved with SVM with an accuracy of 96.99%. Delen et al. [9] investigated the breast cancer data of 202,932 records for the classification based on NB, c4.5 DT classifier and neural networks. The maximum accuracy was achieved with c4.5 DT classifier.

In the study performed by Shen et al. [10], the XGBoost ML model was used for the classification of breast cancer, and an accuracy of 97.86% was achieved with this method.

In the Monirujjaman Khan et al. [11] study, different ML models, such as random forest, DT, KNN and logistic regression techniques were used. The highest score was achieved with a random forest classifier with an F1 score of 96%. In the work done by Bharadwaj et al. [12], to classify breast cancer cells into benign and malignant, various ML models, such as multilayer perceptron, KNN, genetic algorithm and random forest classifier were used. The maximum accuracy of 96.24% was achieved with a random forest classifier. Dong and Ma [13] have examined triple negative breast cancer predictions using ML algorithms. The SVM model has achieved the highest accuracy of 97.8%.

Zheng et al. [14] proposed a hybrid algorithm using SVM and K which meant clustering and tested the algorithm on Wisconsin breast cancer data. The experimental results show that the proposed algorithm has achieved an accuracy of 97.38%. An optimization algorithm called Whale optimization algorithm was proposed by Jia et al. [15], which is used to adjust the optimization parameters of the SVM model. The proposed model has obtained a classification accuracy of 97.5%. Mahest et al. [16]

addressed the data imbalance issues in breast cancer data using synthetic minority oversampling technology. ML models, such as NB, DT and random forest classifiers combined with XGBoost were used to classify breast cancer. The XGBoost with random forest classifier has achieved the maximum accuracy of 98.29%.

In the proposed work, feature selection based on minimum redundancy maximum relevance is implemented to improve the classification accuracy. The most important features are selected prior to applying them to machine learning models. The performance of the models is then compared with and without feature selection to assess its impact.

The performance of the ML models is cross-validated with five folds of cross-validation of data.

PROPOSED WORK

In this study, the breast cancer data is classified into benign and malignant using machine learning techniques and their performance is compared. For this to be accomplished, as the first step, the Wisconsin breast cancer data is collected and data preprocessing is done to remove the unwanted data, such as patient IDs in the database. In the next step, significant features are selected based on the MRMR algorithm. The selected features are then used to train the ML models, and the performance of the models are evaluated by five folds of cross-validation of the data. The performance of the ML models is compared in terms of accuracy, sensitivity, and specificity.

Data Collection and Preprocessing

- The breast cancer tumor data from Dr. William of the University of Wisconsin Hospital is collected from Kaggle (2016) [6]. This is the public data, which consists of the features extracted from the digital image of finite needle aspirate of the solid breast mass. These features are the characteristics of cell nuclei present in the digital image. There are 569 observations. Each observation consists of 32 columns, of which the first column consists of the patient ID number, the second column is the diagnosis type: malignant and benign The remaining 30 features are derived by calculating the mean, standard error, and worst values of 10 real-valued characteristics – namely radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension – for each digital image.
- After data collection, the first column is removed since it consists of the patient ID number, which is not required for the task. Then the second column, which consists of categorical classes benign and malignant, are separated from the feature data and stored separately as response variable. The categorical classes are converted to binary labels with label “0” representing benign class and label “1” representing malignant class. Now, the size of the cancer data is 569 x 30 with 569 observations and 30 predictor variables or features. Of 569 samples, 357 are benign and 212 are malignant. The size of the response vector is 569 x 1, consisting of class labels 0 or 1 for each of observation.

Feature Selection Using MRMR Algorithm

Before applying the data to the ML models, feature selection must be performed to remove the redundant features to improve the overall classification accuracy. In this study, minimum redundancy maximum relevance (MRMR) feature selection method is implemented.

MRMR Algorithm

MRMR is a multivariate filtering technique in which the features’ correlation is maximized while reducing redundancy. In MRMR, feature subset is found where the difference between the correlation, D of the features and redundancy, R between the features is maximized. That is D-R, should have the maximum value.

The MRMR feature selection method combines redundancy and relevance between the features and the target classification activity. For selecting the features based on MRMR algorithm, there are two criteria which optimize both redundancy R and relevance C simultaneously and this criterion is defined by the following equations [17].

$$\max \alpha_1(F, c), \alpha_1 = C - R \text{ or } \max \alpha_2(F, c), \alpha_2 = \frac{C}{R} \quad (1)$$

where, F: Feature Subset, c: Target Classification Problem.

In this work, it is examined that the top 18 features ranked based on MRMR algorithm have significant impact on the classification accuracy and hence these features are selected to train the ML models.

CLASSIFICATION BASED ON MACHINE LEARNING

The ML models that are taken for this study are LDA, KNN, SVM, NB and DT classifiers. The breast cancer data with selected 18 features are used to train the ML models. The techniques adopted in these ML models are described as follows:

Linear Discriminant Analysis Classifier

In LDA, the optimal transformation is found by minimizing the distance of within-class sample features and maximizing the distance of between-class sample features simultaneously. It was developed by Fisher in 1936 [18]. In LDA, a linear decision function in the attribute space is defined based on the empirical technique. The main concept in LDA is to search for a linear combination of predictors that maximally discriminates between the two classes. This linear combination of predictors with N classes is defined in Equation (2).

$$Z = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots \beta_N x_N \quad (2)$$

Then based on Fisher discriminant rule a suitable value of β is found. Fisher is since it is easier to separate the distribution with greater variances between class features than within class features. This is achieved by maximizing the score function. For the separation between two classes (benign and malignant), the score function is described in Equation (3).

$$Z(\beta) = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T W \beta} \quad (3)$$

where β are linear coefficients, μ_1, μ_2 are the mean vectors and W is given in the Equation (4), with covariance matrices W_1, W_2.

$$W = \frac{1}{N_1 + N_2}(N_1 W_1 + N_2 W_2) \quad (4)$$

Thus, the linear discriminator rule for class j, k is stated in Equation (5).

$$x \in j \; if \; \left|\beta^T x - \beta^T \mu_j\right| < \left|\beta^T x - \beta^T \mu_k\right|; \; j \neq k \quad (5)$$

By performing five folds of cross validation LDA achieves the average accuracy of 96.31% with all the features included and 97.55% with top feature selection. For the discriminant type of "diaglinear" and "diagquadratic" the average accuracy of 94.32%, 92.42% is achieved with all features and 94.55%, 93.45% with feature selection.

K Nearest Neighbor Classifier

In the classification model using KNN, the samples are projected into large dimensional space based on the values of the predictor variables [19]. The distance between the k-nearest neighbors is measured by Euclidean distance which is calculated as follows:

$$D(X_i, X_j) = \left(\sum \left|X_i^m - X_j^m\right|\right)^{1/2} \quad (6)$$

where X_i,X_j are the two samples, X_i^m is the sample value on the feature vector "m". In the KNN algorithm, the classification is done based on the selection of k value, distance measure and the decision rule. In this work, for k = 1, the accuracy of 92.1% is obtained both for without feature and with feature selection. For k = 5, the average accuracy after five folds of cross validation is 94.8% both for with and without feature selection.

Support Vector Machine Classifier

The SVM Classifier is used for classification in applications that do not need large training data sets. Linear discrimination is achieved in SVM, in which a hyperplane is calculated to separate the feature samples into different classes. The hyperplane is calculated in such a way that the distance between the marginal feature samples of data of each class is maximized. Thus, the optimization problem consists of (i) Finding hyperplane for separation of samples into classes. (ii) Maximizing the distance between the marginal feature samples of each class [20]. The solution to this optimization problem is provided by the small percentage of marginal samples called support vectors. The decision function gives the distance of a feature sample from the hyperplane:

$$\hat{f}(x) = b + \sum_{i=1}^{n} \alpha_i \, y_i K(x_i, x) \quad (7)$$

where $i = 1, 2, n$ are the multipliers which are non-zero for the support vectors, x_i is the feature vector, b is a biasing term, $K(x_i, x)$ is a kernel function, α_i is the Lagrangian multiplier that are determined by optimization. The kernel function is defined as:

$$K(x_i, x) = exp\ (-\frac{\|x_i - x\|}{2\gamma^2}). \quad (8)$$

In this work, Gaussian radial basis function is used as kernel with γ = 0.5 as width of the radial basis function kernel. On performing five folds of cross validation the average accuracy of 63.14%, 95.3% is achieved without and with feature selection, respectively. For the Gaussian kernel function the average accuracy of 62.34%, 92.3% is obtained without and with feature selection, respectively.

Naive Bayes Classifier

It is the simplest of all the classification. NB is a model based on conditional probability. If a problem is to be classified in which there are n feature vectors given by x = (x1, x2, xn). These vectors represent features extracted from the image. Hence, a probabilistic assignment is based on p(Bk|x1,x2, xn) for each k possible outcome or class. The main problem in this model is that if the number of feature vectors is large, it is infeasible to base such a model on the probability table [21]. Thus, the model is reformulated by using Baye's theorem as follows:

$$P(B_k|x) = \frac{P(Bk)P(x|Bk)}{P(x)} \quad (9)$$

In the above equation, the numerator is called joint probability, and the denominator is constant and does not depend on either B or the feature vector values.

The classifier can be constructed from the probabilistic model which is Naive Bayes classifier in which the decision rule is combined with the probabilistic model. The decision rule is based on Maximum A posteriori Probability (MAP) criteria. The Bayes classifier thus assigns a label of class, $\tilde{y}$ = Bk, which is given as:

$$\tilde{y} = \underset{k}{argmax}\, P(Bk) \prod_{i=1}^{n} P(xi|Bk) \quad (10).$$

There are five different types of Naive Bayes models and the most used models are: Gaussian NB, Bernoulli NB and Multinomial NB.

In this work, Gaussian NB classifier is implemented and upon five folds of cross validation the average accuracy of 95.68%, 96.57% is achieved without and with feature selection.

Decision Tree Classifier

In the decision tree algorithm, a model is created to predict the value of the response variable. The prediction is made by the decision rules that are learned from the feature vector [22]. The classification is based on the tree structure in which there are two nodes decision nodes and leaf nodes. The decision nodes have many branches where each branch represents the decision rules, and the leaf node represents the outcome. A decision tree classifier is a machine learning algorithm that is commonly used for classification tasks. It is particularly effective for datasets with categorical or numerical features, making it suitable for classifying the cancer dataset into benign or malignant.

The decision tree classifier builds a tree-like structure where internal nodes represent features, and leaf nodes indicate class labels – either benign or malignant. The tree is constructed recursively by choosing the most suitable feature at each step to split the dataset, using criteria, such as information gain, Gini index, or entropy. Information gain, Gini index, or entropy are used to measure the quality of a split at each node. Information gain measures the reduction in uncertainty in the target variable or class label after splitting on a particular attribute.

Once the decision tree is built, it can be used to classify new instances (e.g., patients with unknown labels) by traversing the tree from the root node to a leaf node. The instance is assigned to one of the child nodes at each internal node based on the attribute value, following the corresponding decision rule. At the leaf node, the predicted class label (benign or malignant) is assigned to the instance.

In this work, the split criterion of cross entropy is implemented and upon five folds of cross validation the average accuracy of 94.55%, 95.43% is achieved without and with feature selection, respectively. On implementing the split criterion of Gini index, the average accuracy of 92.35%, 93.45% is achieved without and with feature selection, respectively.

RESULT ANALYSIS

For each machine learning model taken for study, the experiments are conducted in two ways: One without applying feature selection algorithm and the other with MRMR feature selection algorithm. The ML models are trained and tested with five folds of cross validation of data. The data is splitted into training and tests randomly in every fold. The classification results of the ML models on test data is observed and the performance metrics, such as accuracy, sensitivity or recall, and specificity are used to assess the model.

The experiments for this work are conducted in MATLAB R2021 64-bit environment on a computer with Intel Core i5 processor with 8 GB RAM.

PERFORMANCE METRICS

- *Accuracy:* Accuracy is the measure to find how far the ML could classify the cancer data into benign or malignant correctly, which is defined as follows:

$$Accuracy = \frac{\text{TP+ TN}}{\text{Total number of messages in the SMS (TP+TN+FP+FN)}} \quad (11)$$

- *Sensitivity or Recall:* Recall is defined as the ratio of the number of TP results to the total number of TP and False Negative (FN) results, that is the number of malignant cases missed by the ML model, given as follows:

$$Sensitivity = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}} \quad (12)$$

If the sensitivity is high, it shows that the ML model has the potential to correctly detect the malignant cases.

- *Specificity:* It is the measure to show the number of benign cases that are correctly detected by the ML model which is defined by the ratio of the number of true negative to the total number of true negative and false positive cases. If the specificity is high, then the ML model has the potential to correctly detect benign cases with very less false positives.

$$Specificity = \frac{TN}{TN+FP} \quad (13)$$

K-Folds of Cross-Validation

K-Fold cross-validation is a common technique used in machine learning to assess the performance of a model. The original dataset is randomly partitioned into K equally sized subsets which are called folds and this approach, the model is trained on K-1 folds of the data and tested on the remaining fold. This process is repeated K times, with each fold taking a turn as the validation set while the rest are used for training. The overall performance of the model is then calculated by averaging metrics, like accuracy, recall (or sensitivity), and specificity, across all K validation runs. In this study, five folds of cross validation is done by dividing the data into five equally sized subsets. On running the experiment, every time the train and test data size changes and one random split up on five folds is shown in Table 1.

Table 1. Five-fold cross-validation on breast cancer data.

Fold No	Train Set	Test Set
1	456	113
2	454	115
3	456	113
4	455	114
5	455	114

Performance Comparison

The classification results obtained after every fold of cross validation for each ML model are compared using performance metrics. The accuracy of the ML models is determined for each fold and the average is calculated. Table 2 shows the accuracy achieved by the ML models on every fold of cross validation and the average accuracy is calculated which is given in the last row of Table 2.

Table 2. % Accuracy in five folds of cross validation.

Fold No	LDA		KNN		SVM		NB		DT	
	All	FS	All	FS	All	FS	All	FS	All	FS
1	95.6	99.1	97.4	97.4	63.2	97.4	98.2	99.1	94.7	95.6
2	97.4	98.7	94.7	94.7	63.4	96.0	96.5	96.9	93.4	94.7
3	96.2	97.1	94.4	94.4	63.0	94.7	94.7	95.6	93.8	94.7
4	96.5	97.1	94.1	94.1	63.0	94.9	94.9	95.8	95.2	95.6
5	95.9	95.8	93.7	93.7	63.1	93.5	94.0	95.4	94.7	95.1
Avg	96.3	97.6	94.8	94.8	63.1	95.3	95.7	96.6	94.4	95.1

ALL: ALL FEATURES INCLUDED FS: FEATURE SELECTION

From Table 2, it is observed that the average accuracy of LDA is maximum with 97.6% when feature selection algorithm is implemented. The next best forming classifier is Naive Bayes classifier with the average accuracy of 96.6%. Similarly, the sensitivity, specificity obtained during five folds of cross validation is shown in Tables 3 and 4, respectively.

Table 3. % Sensitivity in five folds of cross validation.

Fold No	LDA		KNN		SVM		NB		DT	
	All	FS	All	FS	All	FS	All	FS	All	FS
1	99.0	100	97.2	97.2	100	95.8	100	100.	93.0	94.4
2	99.0	100	97.8	97.8	100	93.7	96.5	96.5	93.0	94.4
3	99.0	100	97.3	97.3	100	93.9	96.7	97.2	94.9	95.8
4	99.0	100	96.1	96.1	100	94.0	96.1	96.5	96.1	96.5
5	99.1	99.4	96.6	96.6	100	92.4	95.5	96.4	95.8	96.1
Avg	99.0	99.9	97.0	97.0	100	94.0	97.0	97.3	94.6	95.4

Table 4. % Specificity in five folds of cross validation.

Fold No	LDA		KNN		SVM		NB		DT	
	All	FS	All	FS	All	FS	All	FS	All	FS
1	96.3	97.6	97.6	97.6	58.8	100	95.2	97.6	97.6	97.6
2	91.3	96.4	92.9	92.9	11.9	100	96.4	97.6	94.1	95.3
3	91.3	92.1	91.3	91.3	78.7	96.1	91.3	92.9	92.1	92.9
4	90.5	92.3	90.5	90.5	59.2	96.4	92.9	94.7	93.5	94.1
5	88.7	89.6	88.7	88.7	94.3	95.3	91.5	93.9	92.9	93.4
Avg	91.6	93.6	92.2	92.2	60.6	97.6	93.5	95.4	94.1	94.7

From Table 3, it is seen that the SVM classifier has achieved the maximum sensitivity of 100% which means that this model has detected the malignant class in all five folds correctly. However, after implementing the feature selection algorithm the sensitivity is reduced to 94%. The optimum sensitivity of 99% is achieved with LDA classifier. On observing the specificity of the ML models given in Table 4, the SVM classifier has achieved maximum specificity of 97.6% with feature selection which shows that this model can detect the benign class better than the other ML models.

The comparison of average accuracy, sensitivity, and specificity achieved by the ML models are shown in Figures 1, 2, and 3, respectively.

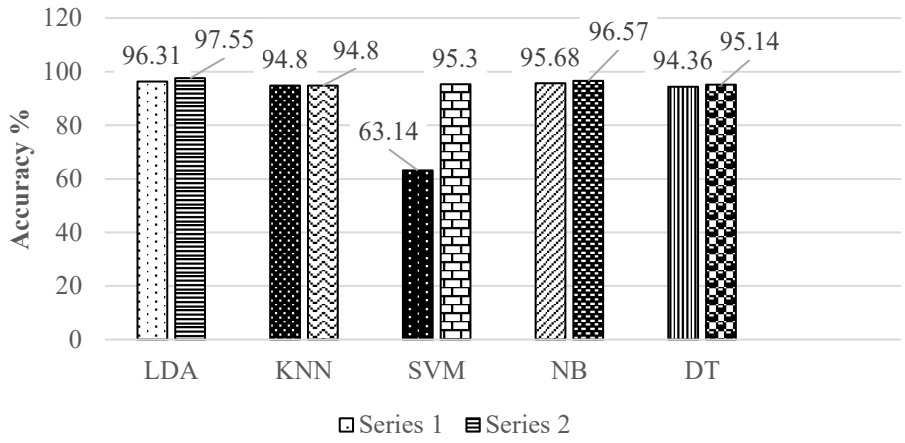

Figure 1. Performance comparison: Accuracy.

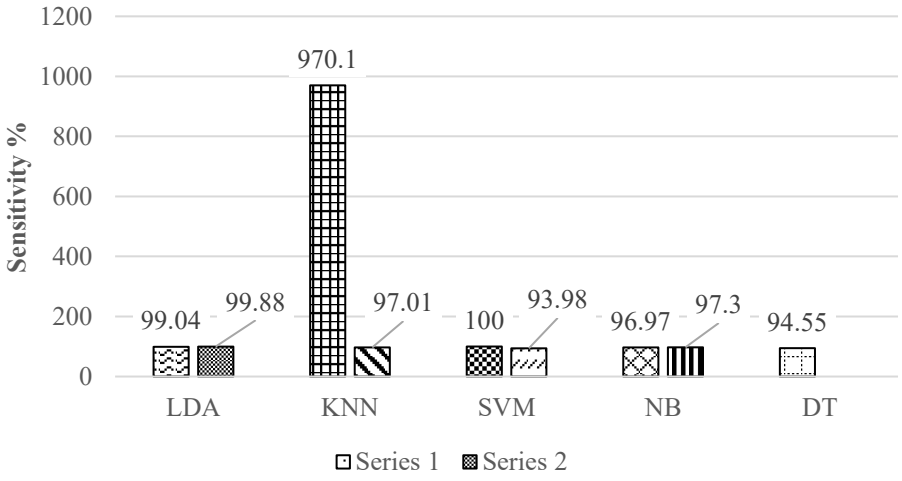

Figure 2. Performance comparison: Sensitivity.

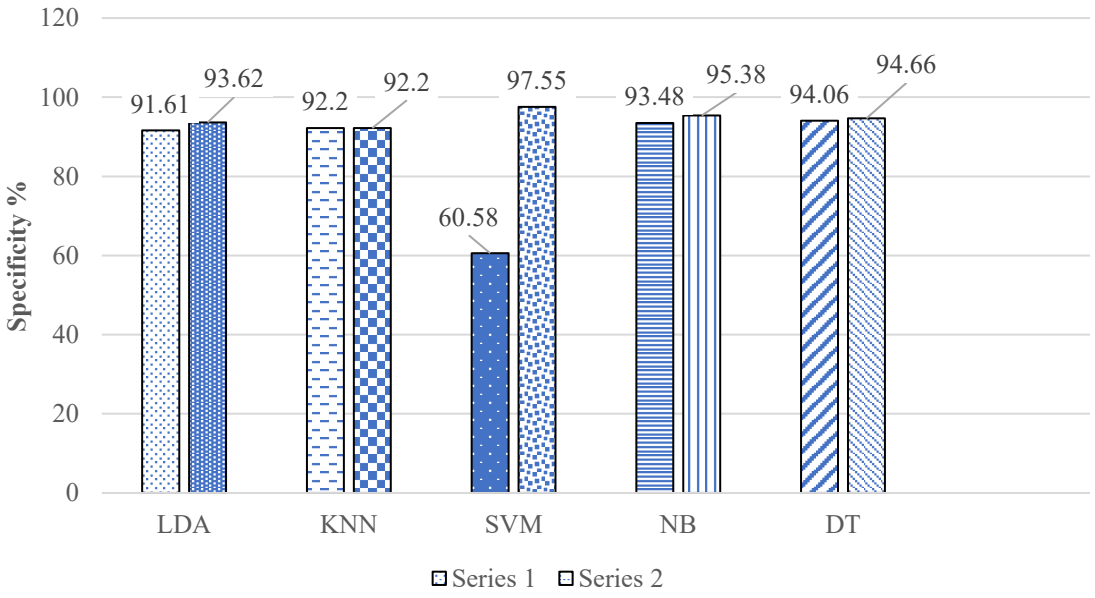

Figure 3. Performance comparison: Specificity.

Figures 1, 2, and 3 show that the classification results are better achieved if the most significant features are selected while ignoring the redundant features based on the MRMR algorithm. Thus, the feature selection significantly impacts the classification accuracy except for the SVM model. The SVM model's performance is achieved better without feature selection. However, the sensitivity and specificity of SVM classifier is high when the feature selection algorithm is implemented. By comparing the results, the linear discriminant analysis classifier has achieved the highest performance with the accuracy of 97.6% and sensitivity of 99%. However, the specificity of LDA is only 93.6% whereas the SVM classifier has achieved the maximum specificity of 97.6%.

The performance of the Naive Bayes classifier is also equally high when compared to LDA classifier with the accuracy of 96.6%, 97.3% and 95.4%. Both Naive Bayes classifier and LDA have achieved optimum results in classifying breast cancer data.

CONCLUSIONS

In this work, the Wisconsin breast cancer data is classified into benign or malignant using machine learning techniques. The data preprocessing is done to remove the unwanted information irrelevant to the intended task. The most significant features are selected using MRMR feature selection algorithm. The top ranked 18 features are selected and are applied to ML models for training. The ML models are trained and validated with five folds of cross validation of data and the average results are calculated. The performance of the ML models is compared in terms of accuracy, sensitivity, and specificity. The experimental results show that LDA classifiers have outperformed the other models with the accuracy of 97.6% is achieved with feature selection algorithm. The sensitivity and specificity of LDA classifier is 99% and 93%. Even though LDA has achieved less specificity when compared to the other models, the sensitivity is the most important parameter in assessing the ML model to find its potential in detecting the malignant cases. Since LDA has achieved sensitivity of 99% with maximum accuracy, this model is best suitable for the classification of Wisconsin breast cancer data. The next better performing classifier is Naive Bayes classifier which has achieved accuracy, sensitivity, and specificity of 96.6%, 97.3% and 95.4%. The results also show that the feature selection algorithm has improved the classification results.

Future Scope: Machine learning models are implemented without hyper parameter tuning in this work. Hence if the hyper parameters are properly tuned with k folds of cross validation the classification results will be improved.

The SVM classifier has achieved the maximum specificity of 97.6%, which shows that it has the potential to detect the benign cases correctly. But whereas the LDA has achieved the maximum sensitivity of 99%. If sensitivity and specificity of the ML model is high, then accuracy will also be increased. Hence, a hybrid algorithm combining the potential of LDA and SVM classifiers can be proposed in the future to achieve enhanced classification results.

REFERENCES

1. Al-Shamsi HO, Coomes EA. Higher and increasing incidence of cancer between the age of 20–49 years in the UAE population: A focus analysis of the UAE National Cancer Registry Data 2015–2017. J Oncol Res Rev Rep. 2021;126(2):2–5. doi:10.47363/JONRR/2021(2).
2. Arya C. Design of structural elements: Concrete, steelwork, masonry and timber design to British standards and Eurocodes [Internet]. London: E & FN Spon; 1994 [cited 2017 Oct 26]. Available from: http://www.dawsonera.com/abstract/9780203926505.
3. Lim DH, Nawy EG. Behavior of plain and steel-fibre reinforced high strength concrete under uniaxial and biaxial compression. Mag Concr Res. 2005;57(10):603–10.
4. World Health Organization. Breast cancer [Internet]. Geneva: WHO; [cited 2023 May 3]. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
5. Dora L, Agrawal S, Panda R, Abraham A. Optimal breast cancer classification using Gauss–Newton representation based algorithm. Expert Syst Appl. 2017;85:134–45.
6. Kaggle. Breast Cancer Wisconsin (Diagnostic) Data Set [Internet]. [cited 2023 Mar 13]. Available from: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.
7. Chaurasia V, Pal S. Data mining techniques: To predict and resolve breast cancer survivability. Int J Comput Sci Mob Comput. 2014;3(1):10–22.
8. Aruna S, Rajagopalan SP, Nandakishore LV. Knowledge based analysis of various statistical tools in detecting breast cancer. Comput Sci Inf Technol. 2011;2:37–45.
9. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif Intell Med. 2005;34(2):113–27.
10. Shen Q, Shao F, Sun R. Prediction model of breast cancer based on XGBoost. J Qingdao Univ Nat Sci Ed. 2019;32:95–100.
11. Khan MM, Islam S, Sarkar S, Ayaz FI, Kabir M, Tazin T, et al. Machine learning based comparative analysis for breast cancer prediction. J Healthc Eng. 2022;2022:1–10.

12. Bhardwaj A, Bhardwaj H, Sakalle A, Uddin Z, Sakalle M, Ibrahim W. Tree-based and machine learning algorithm analysis for breast cancer classification. Comput Intell Neurosci. 2022;2022:1–12.
13. Dong H, Ma L. Prediction model of triple negative breast cancer based on machine learning. J Yunnan Univ. 2017;39(1):111–5.
14. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst Appl. 2014;41(4):1476–82.
15. Jia X, Sun X, Zhang X. Breast cancer identification using machine learning. Math Probl Eng. 2022;2022:1–11.
16. Mahesh TR, Vinoth Kumar V, Muthukumaran V, Shashikala HK, Swapna B, Guluwadi S. Performance analysis of XGBoost ensemble methods for survivability with the classification of breast cancer. J Sensors. 2022;2022:1–8.
17. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38.
18. Izenman AJ. Linear discriminant analysis. In: Modern multivariate statistical techniques: Regression, classification, and manifold learning. New York: Springer. 2008; p. 237–80.
19. Rukmawan SH, Aszhari FR, Rustam Z, Pandelaki J. Cerebral infarction classification using the k-nearest neighbor and naive Bayes classifier. J Phys Conf Ser. 2021;1752(1):012045.
20. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
21. Mitchell TM. Machine learning and data mining. Commun ACM. 1999;42(11):30–6.
22. Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. J Appl Sci Technol Trends. 2021;2(1):20–8.

