INTERNATIONAL JOURNAL OF APPLIED METHODS IN ELECTRONICS AND COMPUTERS

www.ijamec.org

International Open Access

Volume 14
Issue 02
June, 2026

Research Article

Performance Analysis of Machine Learning Algorithms on Wisconsin Diagnostic Breast Cancer Data Set Enriched with Data Augmentation Technique

Ayça ACET [a], ***Abdullah Erhan AKKAYA*** [b,*]

[a] *Turkish Aerospace Industry, 06980, Kahramankazan, Ankara, Türkiye*
[b] *İnönü University, Faculty of Engineering, Department of Computer Engineering, Malatya, Türkiye*

ARTICLE INFO

Article history:
Received 14 February 2026
Accepted 26 June 2026

Keywords:
Breast cancer diagnosis,
Classification,
Data augmentation,
Machine learning

ABSTRACT

Breast cancer is a life-threatening cancer worldwide and is commonly seen in women. Early and accurate diagnosis is a key to successful treatment and better survival rates. In this study, using 569 samples from the Wisconsin Diagnostic Breast Cancer dataset, the classification performances of Support Vector Machines (SVM), Adaptive Boosting (AdaBoost), Random Forest (RF), Naïve Bayes (NB) and k-Nearest Neighbor (kNN) algorithms are compared. To improve accuracy, the train dataset was expanded from 381 to 1524 samples using a three-stage method involving random scaling, Gaussian noise, and mathematical transformations. Test dataset remained unchanged. Statistical analyses confirmed that the augmented data presented sufficient variability to enhance generalization while preserving the original distribution. Model performance analysis was conducted using accuracy, precision, sensitivity, F1 score, FDR, MCC and AUC values. Compared to the original dataset, the accuracy of the SVM algorithm improved from 98.94% to 99.47% with the data augmentation techniques. This result shows that data augmentation techniques are effective in improving the classification performance of machine learning models for breast cancer diagnosis.

1. Introduction

The regulation of cell proliferation in healthy tissues is guided by responsive checkpoint networks rather than a static, predetermined schedule. Instead of following a rigid, automated script, cellular growth decisions require the continuous integration of a wide array of inputs, effectively bridging intrinsic transcriptional states with extrinsic microenvironmental cues. Mechanistically, this integration relies on the joint actions of layered transcription factor cascades [1-3], epigenetic machinery [4, 5], localized cytokine signaling [6, 7], and homeostatic metabolic sensors [8, 9]. Because these pathways converge to act as a collective processor, the resulting system is inherently flexible-allowing cellular behavior to shift dynamically based on the cumulative environmental and internal context. Understanding how these diverse signals are synthesized to maintain this delicate balance between proliferation and quiescence remains a foundational challenge in tissue biology.

However, mutations and abnormal gene changes can cause cells to divide and multiply uncontrollably. Over-proliferation causes cells to come together to form solid tissue called a tumor. If a tumor grows only at the place it develops, then it is a benign tumor. On the other hand, if a tumor spreads to different tissues, then it is a malignant tumor [10, 11]. Malignant tumors are also called cancerous tumors; the cells they contain are called cancer cells, Figure 1. The benign stage of a tumor is usually not life-threatening, while the malignant stage can be fatal.

A cancerous tumor located in the breast-an essential

* Corresponding author. E-mail address: *erhan.akkaya@inonu.edu.tr*
DOI: 10.58190/ijamec.2026.171

organ of the female reproductive system-is called breast cancer. This devastating condition is highly influenced by numerous risk factors that can be categorized into two major groups: reproductive and non-reproductive risk factors. The risk factors in the reproductive aspects that predispose a woman to breast cancer are many. First, menarche at an early age has been associated with an increased risk of the development of malignant tumors [12]. Also, late menopause, the natural cessation of menstruation in older age, has been associated with a higher susceptibility to breast cancer [12]. Additionally, factors such as having children at an older age [13] or never having had children [14], as well as breastfeeding less frequently [15], have been identified as potential reproductive risk factors.

Non-reproductive risk factors, on the other hand, are relevant to breast cancer etiology [16]. Of all these risk factors, the main ones are genetic BRCA1 and BRCA2 mutations [17]. People who carry these gene mutations have a much higher risk of developing the disease than the general population. [18]. Obesity [19], considered to be a state characterized by an excess of body weight and/or adipose tissue, has also been implicated as a risk factor contributing etiopathogenetically to the genesis of breast cancer. Postmenopausal obesity is a third non-reproductive risk factor that increases susceptibility to this malignancy. In addition, having a family history of breast cancer is also known as a non-reproductive risk factor since women with blood relatives (mother, sister(s), and children) are more likely to get the disease [19]. Among other significant risk factors are hormone disorders [20], radiation therapy, and alcohol. Only in 2020, the global number of newly diagnosed cancer cases directly due to alcohol consumption was 741,300 [21]. This frightful data makes up 4% of all cancers in the world. People's lifestyle decisions directly affect their individual health. The most striking example of this is the direct link between alcohol and breast cancer.

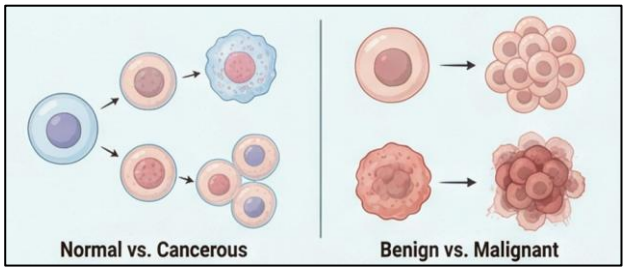

Figure 1. Cellular proliferation, healthy and tumor cells.

Figure 1 shows healthy and cancerous cells. The upper left shows a normal cell's healthy division and differentiation cycle. The lower left shows cancerous cells that have multiplied uncontrollably due to a disruption in the differentiation cycle. The upper right shows a benign tumor structure where the cells are clustered within a defined boundary. The lower right shows a malignant (cancerous) tumor structure where the cells have undergone morphological deformation and spread to surrounding tissues.

The World Health Organization (WHO) has estimated that in the year 2022, approximately 2.3 million new patients were diagnosed with breast cancer worldwide. Unfortunately, this number accounts for 11.6% of all cancer cases diagnosed in 2022. Breast cancer was the most common cancer in women in 157 out of 185 countries in 2022 [22]. In less developed countries, more than half of breast cancer cases occur in women under the age of 50, while in developed countries like the UK, more than a third of cases occur in women over the age of 70 [23]. According to Globocan, 25,249 new cases of breast cancer were diagnosed in Turkey in 2022. This accounts for 23.5% of all female cancers [24]. Early diagnosis and treatment of breast cancer save lives. The Turkish Breast Cancer Foundation, the American College of Radiology, and the National Comprehensive Cancer Network-NCCN recommend that women should start mammography screening at age 40 and continue their screening every two years up to the age of 69. This recommendation is consistent with several international guidelines advocating early detection of breast cancer with regular mammography [25, 26]. Studies show that regular mammography screening reduces deaths from breast cancer [27]. However, it has been reported that participation in mammography screening among women living in Türkiye is low [26, 28].

Diagnosing benign and malignant tumor cells is crucial in determining the type of treatment. To achieve this diagnosis, this study focuses on five commonly used classification algorithms: Adaptive Boosting (AdaBoost), k-Nearest Neighbor (kNN), Näive Bayes (NB), Random Forest (RF), and Support Vector Machines (SVM). These algorithms were used to classify cancer cells based on the provided cellular data, thereby predicting the type of cancer. To evaluate the performance and reliability of the classification algorithms, the Wisconsin Breast Cancer Dataset from the University of California, Irvine (UCI) data mining repository was used. The WDBC dataset contains thirty characteristics of breast cancer cells from 569 patients. The WDBC dataset was chosen for this study because it is lossless.

This study is built upon and significantly extends the preliminary machine learning models developed in the author's Master's thesis [29]. While that previous work focused solely on evaluating the baseline performance of these models on the original dataset, this paper introduces a novel three-stage data augmentation strategy to enhance classification robustness and generalize across small medical datasets. To present these contributions clearly, the paper is structured into six parts: The first part provides a general introduction to the problem. The second part is devoted to solution methods previously applied to the

WDBC dataset in the literature. The third part begins by describing the WDBC dataset in detail, and continues with data augmentation techniques and hyperparameter optimization of machine learning methods. The fourth part presents a comparative analysis of the performance of AdaBoost, kNN, NB, RF, and SVM algorithms in the context of breast cancer diagnosis. The fifth part provides information about the limitations of the study. Lastly, the sixth part concludes the paper and outlines future work.

2. RELATED WORKS

WDBC dataset used in this article consists of 569 samples and 30 features. In this section, the results of the studies that classify the WDBC dataset using AdaBoost, kNN, NB, RF, and SVM methods are given. In the study conducted by Cabanillas-Carbonell and Zapata-Paulini [30], it was observed that AdaBoost 0.97, NB 0.91 and RF had 0.97 accuracies. In another study [31], SVM and RF algorithms gave correct classification rates of 0.986 and 0.965, respectively. Alzboon et al. [32] reported accuracy rates of 0.914 for AdaBoost, 0.807 for kNN, 0.935 for NB, 0.949 for RF and 0.975 for SVM. In a study conducted by Jakhar et al. [33], it was shown that AdaBoost and RF had accuracies of 0.9724 and 0.97. In addition, Chen et al. [34] obtained accuracies of 0.912 and 0.965 for kNN and RF models, respectively. Özcan et al. applied eleven machine learning models [35]. Among the AdaBoost, kNN, NB, RF, and SVM methods, the accuracy rate has been reported to be 0.9806, 0.9385, 0.9737, 0.9736, and 0.9766, respectively. During a similar study, Mashudi et al. [36] assessed five models, with the recorded values being 0.9877.

Imran et al. [37] observed an accuracy of 0.98 when using Random Forest. Random Forest had been previously shown to outperform Naive Bayes and AdaBoost classifiers. They further reported AdaBoost to be at 0.9824, leading the comparative classifiers. They also reported that Naive Bayes could achieve an accuracy of approximately 0.9619. Prastyo et al. [38] compared eight different machine learning models with the application of AdaBoost, kNN, RF, and SVM methods. These resulted in accuracy rates of 0.9561, 0.9649, 0.9561, and 0.9561, respectively. In the work by Aamir et al. [39], five different machine-learning algorithms were used for the classification process. For the 60:40 ratio, the accuracy values in RF and SVM were 0.954 and 0.9755, respectively. When the training and test datasets ratio is 70:30, the accuracy values for RF and SVM are 0.9667 and 0.9721, respectively. When the training and test datasets ratio is 80:20, the accuracy values for RF and SVM are 0.9807 and 0.9776, respectively.

Nuzhat et al. [40] tested six basic algorithms for breast cancer classification on the WDBC dataset. In this context, they applied the KNN, SVM, Logistic Regression, RF, GBM, and XGBoost methods. The researchers developed a new ensemble architecture by integrating these classifiers using the Hard Voting technique. In the performance analysis of individual models, GBM demonstrated the highest success, achieving an accuracy of 98.33% and a recall of 100%. However, the ensemble model-composed of a combination of SVM, RF, and GBM components-achieved an accuracy of 98.89%, precision of 98.08%, recall of 100%, and an F1 score of 99.03%. This developed hybrid structure achieved a higher success rate than individual models across all evaluation metrics [40].

Pristyanto and Wirantanu [41] conducted a comparative analysis of Recursive Feature Elimination (RFE) and Genetic Algorithm (GA)-based feature selection methods on the WDBC dataset, which consists of 30 features and 569 data points. Logistic Regression, SVM, Naive Bayes, KNN, AdaBoost, and Random Forest (RF) algorithms were tested in the classification process. Accuracy analysis revealed that GA-based feature selection outperformed RFE in terms of weighted macro precision, recall, and F1 score metrics.

Tawil et al. [42] analyzed SVM, LightGBM, Random Forest (RF), Logistic Regression (LR), k-NN, and Naive Bayes algorithms on the WDBC dataset. These classifiers were subjected to a comprehensive comparative performance evaluation by integrating them with three different feature selection methods: Pearson correlation coefficient, Lasso, and mRMR. In the baseline scenario without a feature selection filter, LightGBM demonstrated the highest performance with an accuracy rate of 95%. In the mRMR-based feature selection architecture, where the number of features was reduced to 15, LightGBM's classification accuracy rose to 98%. On the other hand, in the Lasso feature selection model, where dimensionality was reduced to 5 features, the LR algorithm recorded the highest performance with an accuracy of 96% [42].

Mezaghrani et al. [43] proposed a hybrid feature selection method (CMGWO) that integrates correlation analysis with the Modified Gray Wolf Optimization (MGWO) algorithm to perform accurate breast cancer classification on the WDBC dataset. The performance of the developed approach was empirically tested using Random Forest (RF), SVM, and Naive Bayes classifiers. The experimental results show that the RF classifier achieved an accuracy rate of 99.12% with the CMGWO-based feature selection architecture, outperforming other alternative approaches in literature.

Salhi et al. [44] evaluated the performance of KNN, RF, MLP, LR, and SVM classifiers on three different datasets, including the WDBC dataset. In this process, they proposed innovative hybrid optimization methods-such as Two-Stage Mutated Gray Wolf Optimization (TMGWO), the Improved Salp Swarm Algorithm (ISSA), and Binary Black Particle Swarm Optimization (BBPSO)-to reduce dimensionality and improve model performance. In the

baseline scenario, where no feature selection procedure was applied, the KNN algorithm achieved a 90% accuracy rate using all 30 features in the WDBC dataset. On the other hand, the proposed TMGWO-based hybrid approach demonstrated the best performance in both feature optimization (feature selection) and classification accuracy compared to the other methods experimentally compared.

Zhu et al. [45]presented an innovative feature selection approach that integrates the Recursive Feature Elimination (RFE) protocol based on SHAP (Shapley Additive exPlanations) values for the detection of early-stage breast cancer on the WDBC dataset. During the experimental process, KNN, RF, LR, SVM, and LightGBM models were evaluated; the hyperparameter optimization of the classifiers was performed using the Particle Swarm Optimization (PSO) algorithm. Through the proposed hybrid algorithm, 26 critical features were filtered from the dataset. The optimized LightGBM-PSO model achieved 99.0% accuracy, 100% specificity, 100% precision, 97.40% recall, an F-measure of 98.68, and an AUC of 0.9870, demonstrating superior performance compared to the literature.

Ayepeku [46] applied Logistic Regression (LR), KNN, Support Vector Classifier (SVC), Decision Trees (DT), Random Forest (RF), Gradient Boosting (GBM), AdaBoost, and XGBoost algorithms on a comprehensive dataset containing clinical features obtained from the Kaggle platform. The diagnostic performance of the classification models was evaluated using the precision, recall, F1 score, and accuracy metrics. In the performance analysis, when the LR, KNN, SVC, and RF models were compared, it was observed that the accuracy values ranged between 0.91 and 0.93.

3. DATASET, PERFORMANCE METRICS and ANALYSIS TOOL

This section discusses the dataset used in this study and the evaluation metrics used for the analysis. The data description section includes the source and structure of the dataset and its statistical properties. Under the heading of performance metrics, this study discusses accuracy, sensitivity, precision, F1-score, false discovery rate, Matthew's correlation coefficient and area under the curve metrics. These metrics help to determine how well the models perform in classifying or predicting outcomes based on the given data. The hyperparameter optimization strategies and the baseline training configurations for the five machine learning algorithms (AdaBoost, kNN, NB, RF, and SVM) were adopted from [29].

3.1. Dataset Description

This study uses the Wisconsin Diagnostic Breast Cancer dataset from the University of California - Irvine Machine Learning Repository. The breast cancer dataset was created by Dr. William H. Wolberg of the Department of General Surgery and W. Nick Street and Olvi L. Mangasarian of the Department of Computer Science at the University of Wisconsin in 1995. The dataset contains ten essential attributes: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The dataset consists of 10 mean values, 10 standard error values, and ten maximum values of the basic attributes of breast cancer cells examined under the microscope. Together with the patient ID information and class label, the dataset consists of 569 samples with 30 attributes, plus an ID and a class label (total 32 columns). Among the 569 samples in the dataset are 357 benign and 212 malignant breast cancer cells. In WDBC dataset there are no attributes with missing information. Table 1 summarizes the minimum and maximum values of the type, mean, standard deviation, and maximum attributes of the WDBC dataset.

Table 1. Characteristics of the Wisconsin Diagnostic Breast Cancer (WDBC) dataset.

#	Feature	Feature Type	Mean (min-max)	Standard Error (min-max)	Worst (min-max)
1	Patient ID No.	Numeric	–	–	–
2	Diagnosis	Binary	–	–	–
3	Radius	Numeric	6.981 – 28.11	0.112 – 2.873	7.93 – 36.04
4	Texture	Numeric	9.71 – 39.28	0.36 – 4.89	12.02 – 49.54
5	Perimeter	Numeric	43.79 – 188.5	0.76 – 21.98	50.41 – 251.20
6	Area	Numeric	143.5 – 2501.0	6.80 – 542.20	185.2 – 254.0
7	Smoothness	Numeric	0.053 – 0.163	0.002 – 0.031	0.071 – 0.223
8	Compactness	Numeric	0.019 – 0.345	0.002 – 0.135	0.027 – 1.058
9	Concavity	Numeric	0.000 – 0.427	0.000 – 0.396	0.000 – 1.252
10	Concave points	Numeric	0.000 – 0.201	0.000 – 0.053	0.000 – 0.291
11	Symmetry	Numeric	0.106 – 0.304	0.008 – 0.079	0.157 – 0.664
12	Fractal dimension	Numeric	0.050 – 0.097	0.001 – 0.030	0.055 – 0.208

The radius value in the dataset is the radius information of the cells. The radius feature ranges from 6.981 to 28.110 on average. This shows that there are significant variations in cell sizes. This variation in cell size is an essential factor in diagnosing breast cancer. The texture is the rate of change in the grayscale of the inner surfaces of the cells. Texture characterization is vital in distinguishing breast cancer cells from normal cells. Perimeter is the value based on the radius of each cell. The perimeter has the widest range (43.79 - 188.5). The wide range indicates a great variation in the shape and size of the tumors. This variation is a factor to be considered in the diagnosis of breast cancer. Area is the surface area of the cell. The area

features also has a wide range (143.5 - 2501.0). This shows that there is diversity from very small cells to very large cells. Smoothness is the value of the radius of neighboring cells around the cell, i.e. the local variation in radius lengths. Compactness is the average cell density obtained by subtracting the square of the perimeter length of the cell divided by the cell area. Concavity is the magnitude of the indentations and protrusions around the cell. Concave points are the number of indentation and protrusion points around the cell. Smoothness, compactness, concavity, and concave points properties are generally close to zero and measure the surface irregularity of the cells. Symmetry is the value of the elliptical shape change of the cells. The fractal dimension is the ratio of nested irregular cells to all normal cells. The symmetry and fractal dimension properties describe the geometric properties of the cells and vary over a relatively narrow range. Finally, the standard deviation values are quite high, especially for area and perimeter length, indicating large variations in these properties.

3.2. Data Augmentation

In this study, data augmentation techniques were applied to the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Each observation in the original WDBC dataset contains 30 numerical features. To prevent data leakage, the original dataset was first partitioned into training and test subsets, and data augmentation was applied only to the training data. The partitioning ratio was 67% and 33% for the training and test datasets, respectively. Data augmentation was not applied to 188 test datasets out of 569 data. The 381 training datasets were expanded in three ways. Gaussian noise was added to the training dataset via ±5% random variations. Random scaling within a ±10% threshold was applied to the training dataset to capture natural variance. New samples were created in the training dataset through inter-feature mathematical operations. As a result, the original training set was expanded to 1,524 samples, including the original 381 samples. The test dataset remained unchanged. The augmented training samples were designed to preserve the original statistical properties of the data. To maintain physiological realism, the resulting negative feature values were reduced to zero. Statistical validation of the data augmentation process is presented in Figure 2. Histogram analysis based on probability density functions (PDFs) shows that the augmented training data preserves the basic distributional features of the original training set, including mode, skewness, and kurtosis. The observed overlap between the original and augmented distributions indicates that the introduced variations do not distort the basic feature space.

3.3. Analysis Tool

Gentleman and Ihaka originally programmed R. Since about mid-1997 there has been a development core team with write access to the R source [47]. The R language provides an extensive range of statistical and graphical techniques. R language offers methods such as clustering, classification, linear and nonlinear modeling and time series analysis. The R language is also developed as a complete programming language with features to enable users to add new functionality by writing R functions [48]. R users in Turkey can download the R software from http://cran.pau.edu.tr/, the local CRAN server hosted at Pamukkale University. R is used by academic research organizations and many large companies such as Uber, Google, Airbnb, and Facebook [49]. Despite the widespread use of R in industry, academics are the largest user group of the R language.

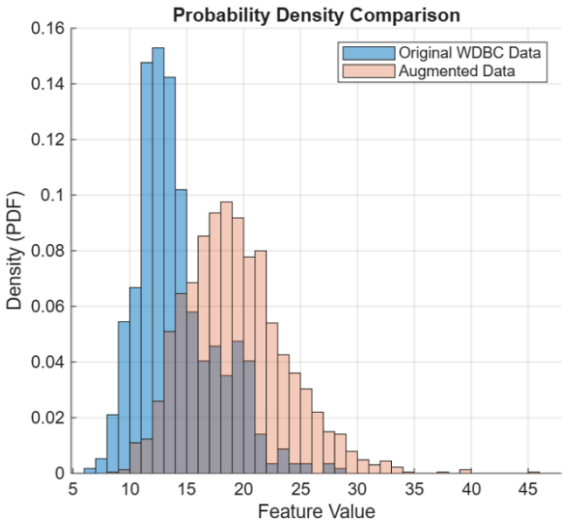

Figure 2. Comparative probability density analysis of original (n=569) and augmented (n=1524) datasets for distribution validation.

3.4. Hyperparameter Optimization Strategy

3.4.1. AdaBoost

The AdaBoost classifier was implemented using the ada package, and the Gentle AdaBoost variant (type = "gentle") was preferred. The boosting iteration count was set to iter = 100. This value represents a parameter range commonly used in the literature that provides a balance between performance and complexity. Higher iteration counts are known to contribute only slightly to accuracy improvement but increase model complexity.

The growth of the decision trees used as weak learners was regulated with the rpart.control() function. The maximum tree depth was set to maxdepth = 14; this value was adjusted to limit excessive complexity while providing sufficient decomposition power appropriate to the data size and feature count. The complexity parameter was set to cp = -1, and internal cross-validation (xval = 0) was disabled. Thanks to this choice, the machine learning model was able to learn the optimal limits independently of early pruning.

3.4.2. k-Nearest Neighbour

In this study, the k-Nearest Neighbor (kNN) algorithm, another machine learning method used, was optimized for the number of neighbors (k) parameter. The k value was systematically tested in the range of 1–100, and the model was trained for each value, with accuracy calculated on an independent test dataset. The obtained accuracy values were compared, and the k value providing the highest performance was determined as the optimal parameter. The final kNN model was reconstructed with this optimal k value, and its performance was reported via the confusion matrix. This ensured a balanced choice between the risks of overfitting that can be seen with small k values and overgeneralization that can occur with large k values.

3.4.3. Naive Bayes (with Laplace optimization)

The Laplace correction parameter for the Naive Bayes model was systematically optimized to mitigate the zero-probability problem and improve classification performance. Model accuracy was measured for each adjustment, with Laplace values evaluated on a range of 1-30. This method is designed to evaluate the effect of the correction on prediction stability, especially in datasets characterized by unbalanced class distributions or low-frequency feature combinations. The Laplace value that yielded the highest accuracy was selected as the optimal parameter for the final model.

3.4.4. Random Forest (RF)

The Random Forest machine learning model consists of 500 decision trees (ntree = 500). Literature indicates that a tree count between 300 and 1000 is generally sufficient for performance stability. A 500-tree structure was chosen to keep computational costs at a reasonable level and reduce variance. The `importance = TRUE` parameter was enabled for variable significance analysis. The proximity matrix between samples was obtained using the `proximity = TRUE` option.

The `mtry` parameter is left at its default value. Using

$$mtry \approx \sqrt{p}$$

as the default in classification problems theoretically increases intertree diversity. Since further testing on `mtry` did not significantly improve performance in preliminary analyses, the default structure has been retained for the simplicity and reproducibility of the model.

3.4.5. Support Vector Machines (SVM)

The Support Vector Machine (SVM) model was run using a radial basis function (RBF) kernel. To optimize model performance, a systematic grid search was applied to two key hyperparameters: the penalty parameter (cost) and the kernel parameter (gamma).

The gamma parameter was scanned in the range of 0-0.1 with increments of 0.005; the cost parameter was scanned in the range of 20-25. These ranges were chosen to prevent the decision boundary from being overly flexible or overly rigid. This chosen approach aims to increase the generalization capacity of the model and ensure that the decision boundary fits the data distribution optimally.

3.5. Performance Metrics

To analyze the classifier's classification efficiency, after the training phase is over, the classifier's prediction accuracy should be measured. The actual performance of the classification model is usually done by making use of the data present in the confusion matrix. A confusion matrix is a matrix of at least 2x2 in size utilized in depicting how much of the predicted value of a classification model created due to machine learning algorithms matches with the actual class value. Once a confusion matrix is developed, the performance of prediction models is compared based on accuracy, sensitivity, precision, F1-score, FDR, MCC and AUC criteria.

3.5.1. Accuracy

It is the number of all correct predictions which includes positive predictions and negative ones over total dataset. Accuracy is a value that ranges from 0 to 1, the lowest possible accuracy and highest, respectively. It can also be computed as (1−error rate).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 1 - error\ rate$$

3.5.2. Sensitivity

Sensitivity is the proportion of positive examples correctly identified (TP) against all positives presented in the data (TP + FN). This metric is also called recall or true positive rate (TPR). The highest sensitivity is 1, while the lowest is set at zero.

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

3.5.3. Precision

Precision indicates how many positive predictions a machine learning model calls positive. In other words, it is the ratio of true positives to all positive predictions. The worst precision is 0, and the best precision is 1. The closer it gets to 0, the more false positive predictions the model makes, and the closer it gets to 1, the more accurate almost all positive predictions are.

$$Precision = \frac{TP}{TP + FP}$$

3.5.4. F1-score

The F1-Score value is a criterion that balances precision and sensitivity values. In the case of unequally distributed datasets, using F1-Score instead of precision will prevent an erroneous model selection. F1-Score is calculated by

taking the harmonic mean of sensitivity and precision values. The reason for using the harmonic mean instead of arithmetic is to consider the extreme cases. In the case of 0 and 1 values, the F1-Score is calculated as 0 with the harmonic mean, while it is incorrectly calculated as 0.5 with the arithmetic mean.

$$F_1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

3.5.5. FDR

False Discovery Rate (FDR) expresses the proportion of samples that a classification model predicts as positive but are actually false positives. In other words, it shows how many of the model's positive predictions are false positives (FP). It is a critical metric for evaluating the false positive rate, especially in fields such as medical diagnosis.

$$FDR = \frac{FP}{TP + FP}$$

3.5.6. MCC

Matthews Correlation Coefficient (MCC) is a metric that measures model performance in binary classification problems. It is known to provide reliable results, especially in datasets with unbalanced class distributions. MCC measures the correlation between actual and predicted classes, evaluating all confusion matrix elements (TP, TN, FP, FN) together.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

3.5.7. Area under the curve (AUC)

AUC stands for Area Under the Curve; that is, the area under the ROC curve. It measures a model's ability to discriminate between classes and typically ranges from 0.5 to 1.0. An AUC value of 0.5 for any model means that the model has not established a recognizable discrimination ability. An AUC value between 0.5 and 0.7 indicates that the model has poor discrimination. AUC values between 0.7 and 0.8 indicate an acceptable model, while values between 0.8 and 0.9 indicate excellent discrimination ability. Values above 0.9 indicate exceptional discrimination. In medical studies, AUC values above 0.80 indicate that the model is clinically valid [50].

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

$$AUC = \frac{TPR + (1 - FPR)}{2}$$

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

This study presents a detailed analysis and performance comparison of machine learning classifiers used in breast cancer diagnosis. The WDBC dataset used in the study consists of a total of 569 data points. 381 samples (67%) from the original WDBC dataset were used as training data, while 188 samples (33%) formed the test dataset. The test dataset consisted of 188 cell samples, 121 of which were benign and 67 malignant. To ensure an unbiased assessment and prevent data leakage, the original WDBC dataset (n = 569) was first divided into training (67%, n = 381) and test (33%, n = 188) subsets. Data augmentation techniques were applied only to the training set, which was expanded to 1,524 samples, while the test set remained entirely the same. Classifiers were trained using the augmented training data and evaluated on the original test data, which consisted of 188 cell samples containing 121 benign and 67 malignant cases. Figures 3 and 4 show the confusion matrices derived from the original and augmented training scenarios, respectively.

Figures 3 and 4 show the confusion matrices obtained from the original and augmented WDBC datasets, respectively. In the original dataset, SVM achieved the highest classification performance with 66 true positives, 120 true negatives, only 1 false positive, and 1 false negative. After data augmentation, SVM further improved its performance by completely eliminating false positives (FP = 0) and retaining only one false negative.

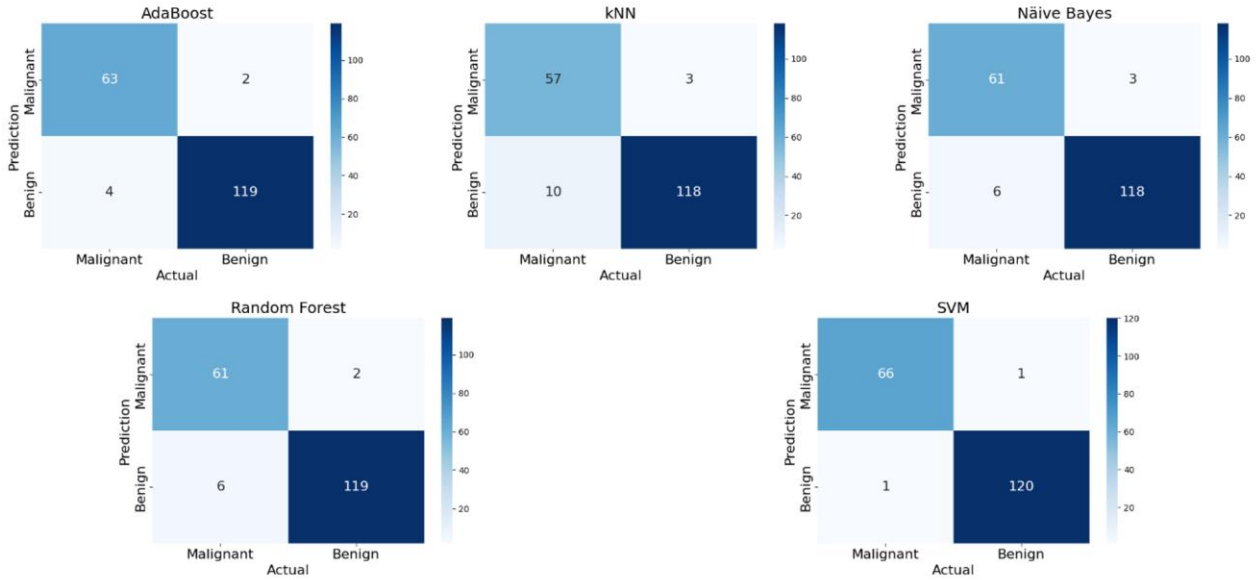

Figure 3. Confusion matrices of five machine learning algorithms based on the WDBC dataset.

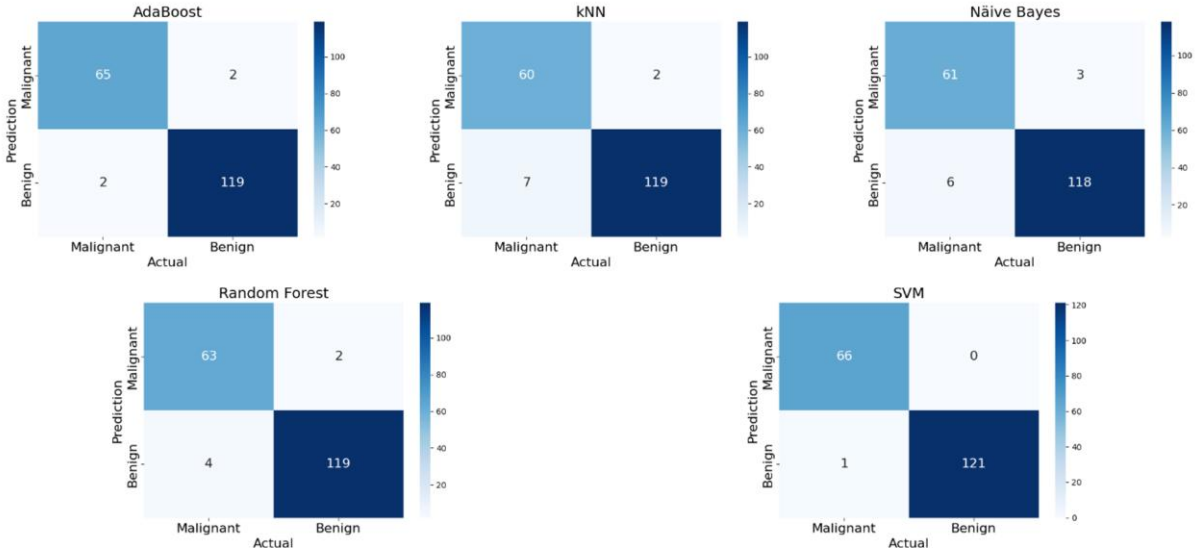

Figure 4. Confusion matrices of five machine learning algorithms based on the augmented dataset.

The AdaBoost and Random Forest algorithms also showed improvement in true positive classifications after data augmentation, while kNN showed a moderate improvement in sensitivity. Naive Bayes performance remained stable in both datasets. Overall, the data augmentation strategy contributed to a reduction in misclassification rates, especially for clinically more critical malignant cases.

Table 2. Performance of models trained with the original dataset on a test dataset.

Metrics	AdaBoost	kNN	Näive Bayes	RF	SVM
Accuracy	0.9681	0.9309	0.9521	0.9574	0.9894
Precision	0.9692	0.9500	0.9531	0.9683	0.9851
Sensitivity	0.9403	0.8507	0.9104	0.9104	0.9851
F1-Score	0.9545	0.8976	0.9313	0.9385	0.9851
FDR	0.0308	0.0500	0.0469	0.0317	0.0149
MCC	0.9302	0.8486	0.8951	0.9070	0.9768
AUC	0.9619	0.9130	0.9428	0.9470	0.9884

The performance metrics listed in Table 2 represent the baseline classification results on the original WDBC dataset, which were originally presented in [29] and recalculated here focusing on malignant cases as the target class to align with clinical standards. Table 2 shows the classification performance of five machine learning algorithms trained with 381 training data points from the

original WDBC dataset, on 188 test data points. Among all classifiers, SVM yielded the highest overall performance with 98.94% accuracy, 98.51% sensitivity, and the highest MCC value (0.9768). These results for SVM demonstrate a strong correlation between predicted and actual classes. The low FDR (0.0149) confirms that the false positive prediction rate is minimal. AdaBoost ranked second with 96.81% accuracy. kNN showed a relatively lower sensitivity of 85.07%, performing poorer in detecting malignant cases. Naive Bayes produced moderate results across all metrics. Overall, the results show that SVM has the most reliable classification performance on the original dataset, particularly in terms of robustness (MCC) and discrimination ability (AUC = 0.9884).

Table 3. Performance of models trained with an augmented dataset on a test dataset.

Metrics	AdaBoost	kNN	Näive Bayes	RF	SVM
Accuracy	0.9787	0.9521	0.9521	0.9681	0.9947
Precision	0.9701	0.9677	0.9531	0.9692	1.0000
Sensitivity	0.9701	0.8955	0.9104	0.9403	0.9851
F1-Score	0.9701	0.9302	0.9313	0.9545	0.9925
FDR	0.0299	0.0323	0.0469	0.0308	0.0000
MCC	0.9536	0.8954	0.8951	0.9302	0.9884
AUC	0.9768	0.9395	0.9428	0.9619	0.9925

Table 3 shows the classification performance of five machine learning algorithms trained with 1524 augmented training data obtained by applying data augmentation techniques to 381 training data points of the original WDBC dataset, on 188 test data points. SVM achieved the highest accuracy (99.47%), sensitivity (0.9851), and zero false discovery rate (FDR = 0.0000), demonstrating the complete elimination of false positive predictions. The MCC (0.9884) and AUC (0.9925) values of SVM show improved discrimination after augmentation. AdaBoost also showed a significant improvement, increasing its accuracy to 97.87%. Random Forest showed an improvement in sensitivity (0.9403) and provided better malignant tumor detection after augmentation. Naive Bayes remained largely unchanged and showed limited sensitivity to augmentation techniques. Overall, the results demonstrate that the proposed data augmentation strategy, especially for SVM, improves classifier generalization and reduces false classification rates.

Table 4. Comparison of original and augmented WDBC dataset results with studies in the literature.

Author (s)	Method	Accuracy
Cabanillas-Carbonell [30]	AdaBoost	0.9700
Zhao et al. [31]	SVM	0.9860
Alzboon et al. [32]	SVM	0.9750
Jakhar et al. [33]	AdaBoost	0.9724
Chen et al. [34]	RF	0.9650
Ozcan et al. [35]	AdaBoost	0.9806
Mashudi et al. [36]	AdaBoost	0.9877
Imran et al. [37]	AdaBoost	0.9824
Prastyo et al. [38]	kNN	0.9649
Aamir et al. [39]	RF	0.9807
Nuzhat et al. [40]	SVM + RF + GBM	0.9808
Tawil et al. [42]	LightGBM	0.9800
Mezaghrani et al. [43]	CMGWO+RF	0.9912
Salhi et al. [44]	MLP+BBPSO	0.9600
Zhu et al. [45]	LightGBM-PSO	0.9900
This study WDBC dataset	SVM	0.9894
This study augmented WDBC dataset	SVM	0.9947

Cabanillas-Carbonell and Zapata-Paulini compared AdaBoost, NB, and RF algorithms. The accuracy rates of the algorithms were 0.97, 0.91, and 0.97, respectively. In another study by Zhao et al., the accuracy rates reported for SVM and RF methods were 0.986 and 0.965. In the analysis of Alzboon et al., the accuracy rates of AdaBoost, kNN, NB, RF, and SVM methods were found to be 0.914, 0.807, 0.935, 0.949, and 0.975, respectively. Similarly, SVM and RF algorithms have high accuracy rates in other studies. The SVM accuracy rate obtained in this study is 0.9894, one of the highest values in literature. In conclusion, the SVM algorithm demonstrates the best performance in breast cancer diagnosis, making it a promising candidate for decision support systems.

Several machine learning models were trained on the augmented dataset. The SVM algorithm achieved the highest accuracy rate with 99.47%. This was followed by AdaBoost (97.87%) and RandomForest (96.81%). Compared to the original dataset, there is a general improvement in the results obtained using data augmentation techniques. In particular, the performance of the SVM algorithm improved from 98.94% accuracy on the original dataset to 99.47% on the augmented dataset. Similarly, the AdaBoost algorithm showed better results with data augmentation techniques.

5. LIMITATIONS OF THE STUDY

This study has some limitations. The WDBC dataset originally contained 569 instances, and although this number was increased using data augmentation techniques, it is still limited compared to real-world data. Data augmentation techniques have helped to reduce the imbalance between the 357 benign and 212 malignant samples in the original dataset, but the distribution between classes is still not exactly equal and there is still a

risk that models are biased towards the majority class. Furthermore, artificial samples generated by data augmentation techniques may not be fully representative of real patient data. Although data augmentation techniques provide significant improvements, larger and more diverse data sets are needed to evaluate the performance of models in real clinical settings.

6. CONCLUSION

In this study, we focus on the comparative results of five different machine learning algorithms (AdaBoost, k-NN, NB, RF and SVM) for breast cancer diagnosis. Experiments on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset and the results obtained using data augmentation techniques revealed that the SVM algorithm outperformed the others. The SVM algorithm achieved the highest success with data augmentation techniques, reaching 99.47% accuracy. The AdaBoost algorithm showed the second-best performance with 97.87% accuracy. Random Forest algorithm ranked third with 96.81% accuracy. The data augmentation techniques significantly improved the results on the original dataset. The accuracy of the SVM algorithm increased from 98.94% to 99.47%. Similarly, the AdaBoost algorithm also showed better results with data augmentation techniques. These results show that data augmentation techniques are an effective method to improve the performance of machine learning models used in breast cancer diagnosis and the SVM algorithm is a promising candidate for preliminary diagnostic aid.

Declaration of Ethical Standards

The article does not contain any studies with human or animal subjects.

Credit Authorship Contribution Statement

All authors contributed to the design and execution of the study. Ayça Acet was responsible for the conceptualization of the study, the application of machine learning models, the data augmentation process, data preparation, statistical analysis, and the writing of the article. Abdullah Erhan Akkaya supervised the research process, contributed to the development and validation of the methodology, ensured the critical review of the intellectual content, and supported the overall structure and scientific rigor of the study. All authors contributed to the interpretation of the findings, critically reviewed the article, and approved the final version of the article.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Funding / Acknowledgements

No funding or research grants were received during the preparation of this study.

Data Availability

The dataset can be downloaded from https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic .

References

[1] C. Fang *et al.*, "Regulatory Mechanisms of O6-Methylguanine Methyltransferase Expression in Glioma Cells," *Sci. Prog.,* vol. 108, no. 2, 2025, doi: 10.1177/00368504251345014.

[2] S. Zhu, N. Zhou, Q. Li, and X. Liu, "Rewiring Immune Suppression in NSCLC: Roles and Plasticity of Tregs and Th17 Cells," *Front. Immunol.,* vol. 16, 2025, doi: 10.3389/fimmu.2025.1658848.

[3] T. Zhu, X. Teng, Q. Jiao, and Y. Ren, "T-cell Exhaustion From a Multiomics Perspective: Differentiation Mechanisms and Regulatory Networks in the Journey From Progenitor-Exhausted T cells to Terminally Exhausted T Cells," *Clin. Transl. Med.,* vol. 16, no. 2, 2026, doi: 10.1002/ctm2.70609.

[4] Z. Lv *et al.*, "The Dual Immunomodulatory Role of B Cells in Tumorigenesis: Mechanisms, Microenvironment Crosstalk, and Therapeutic Implications," *Front. Immunol.,* vol. 16, 2025, doi: 10.3389/fimmu.2025.1649812.

[5] F. P. Fabrizio, "Unlocking Lung Cancer Cell Dormancy: An Epigenetic Perspective," *Int. J. Mol. Sci.,* vol. 26, no. 22, p. 10997, 2025, doi: 10.3390/ijms262210997.

[6] M. Kos *et al.*, "From Defense to Disease: How the Immune System Fuels Epithelial–Mesenchymal Transition in Ovarian Cancer," *Int. J. Mol. Sci.,* vol. 26, no. 9, p. 4041, 2025, doi: 10.3390/ijms26094041.

[7] Y. Wang, A. Lin, Z. Liu, Q. Cheng, J. Zhang, and P. Luo, "Tumor Microenvironment Onmyoji: Cytokines With Dual Protumor and Antitumor Roles," *Cancer Commun.,* vol. 46, 2026, doi: 10.34133/cancomm.0008.

[8] M. Ma and D. Q. Li, "Autophagy-Related Proteins in Triple-Negative Breast Cancer: From Molecular Insights to Therapeutic Applications," *Int. J. Mol. Sci.,* vol. 26, no. 18, p. 9231, 2025, doi: 10.3390/ijms26189231.

[9] X. Huang, J. He, H. Sun, Y. Wu, R. Gu, and Z. Li, "PKM2-driven Metabolic Reprogramming in Digestive System Tumors: Mechanisms, Therapeutic Advances, and Clinical Challenges," *Front. Immunol.,* vol. 16, 2025, doi: 10.3389/fimmu.2025.1634786.

[10] L. A. Torland *et al.*, "Benign breast tumors may arise on different immunological backgrounds," *Mol. Oncol.,* vol. 18, no. 10, pp. 2495-2509, 2024.

[11] A. Sucharitha and D. Bhuvana, "An overview of cancer," *Adv. Canc. Chemther. Pharmacol.,* vol. 1, no. 1, pp. 1-8, 2023.

[12] X. Mao *et al.*, "Association of reproductive risk factors and breast cancer molecular subtypes: a systematic review and meta-analysis," *BMC Cancer 2023 23:1,* vol. 23, no. 1, 2023-07-10, doi: 10.1186/s12885-023-11049-0.

[13] E. I. Obeagu and G. U. Obeagu, "Breast cancer: A review of risk factors and diagnosis," *Medicine,* vol. 103, no. 3, January 19, 2024, doi: 10.1097/MD.0000000000036905.

[14] J. D. Figueroa *et al.*, "Reproductive factors and risk of breast cancer by tumour subtypes among Ghanaian women: A population-based case-control study," *Int. J. Cancer,* vol. 147, no. 6, 2020/09/09, doi: 10.1002/ijc.32929.

[15] A. Bothou *et al.*, "Breastfeeding and Breast Cancer Risk: Our Experience and Mini-review of the Literature," *Mater. Sociomed.,* vol. 34, no. 1, 2022/03, doi: 10.5455/msm.2022.33.28-32.

[16] Z. Wang *et al.*, "CTHRC1 is a Potential Prognostic Biomarker and Correlated with Macrophage Infiltration in Breast Cancer," *Int. J. Gen. Med.,* vol. 15, 2022, doi: 10.2147/IJGM.S366272.

[17] M. Bieuville *et al.*, "Frontiers | Number of lifetime menses increases breast cancer occurrence in postmenopausal women at high familial risk," *Front. Ecol. Evol.,* vol. 11, 2023/02/02, doi: 10.3389/fevo.2023.912083.

[18] L. Odén *et al.*, "Plasma osteoprotegerin and breast cancer risk

in BRCA1 and BRCA2 mutation carriers," *Oncotarget,* vol. 7, no. 52, 2016/12/12, doi: 10.18632/oncotarget.13417.

[19] S. Nag *et al.*, "Risk factors for the development of triple-negative breast cancer versus non-triple-negative breast cancer: a case–control study," *Sci. Rep.,* vol. 13, no. 1, p. 13551, 2023.

[20] F. A. C. d. Luz *et al.*, "Frontiers | Survival differences between women and men in the non-reproductive cancers: Results from a matched analysis of the surveillance, epidemiology, and end-results program," *Front. Public Health,* vol. 10, 2023/01/06, doi: 10.3389/fpubh.2022.1076682.

[21] H. Rumgay *et al.*, "Global burden of cancer in 2020 attributable to alcohol consumption: a population-based study," *Lancet Oncol.,* vol. 22, no. 8, 2021/08/01, doi: 10.1016/S1470-2045(21)00279-5.

[22] F. Bray *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J Clin,* vol. 74, no. 3, pp. 229-263, 2024, doi: https://doi.org/10.3322/caac.21834.

[23] L. Wilkinson and T. Gathani, "Understanding breast cancer as a global health concern," *Br J Radiol,* vol. 95, no. 1130, p. 20211033, 2022.

[24] J. Ferlay *et al.* "Global Cancer Observatory: Cancer Today." Lyon, France: International Agency for Research on Cancer. https://gco.iarc.who.int/today (accessed 19/08/2024, 2024).

[25] I. Guvenc, G. Guvenc, S. Tastan, and A. Akyuz, "Identifying Women's Knowledge about Risk Factors of Breast Cancer and Reasons for Having Mammography," *Asian Pac J Cancer Prev,* vol. 13, no. 8, 2012, doi: 10.7314/APJCP.2012.13.8.4191.

[26] T. Anuk and H. Çantay, "Do women living in northeast Anatolia get mammography screening? A hospital-focused cross-sectional study," *J Surg Med,* vol. 6, no. 3, 2022/03/01, doi: 10.28982/josam.982615.

[27] R. Ali, A. England, M. F. McEntee, C. Mercer, A. Tootell, and P. Hogg, "Effective lifetime radiation risk for a number of national mammography screening programmes," *Radiography,* vol. 24, no. 3, pp. 240-246, 2018/08/01, doi: 10.1016/j.radi.2018.02.001.

[28] M. E. Akın, "Rate and reasons of missed screening mammography in the COVID-19 pandemic from Turkey," *J Health Sci Med,* vol. 5, no. 4, 2022-07-20, doi: 10.32322/jhsm.1110424.

[29] A. Acet, "Prediction of breast cancer using SVM, NB, KNN, AdaBoost and Random Forest classification algorithms," Master Thesis, Computer Engineering, İnönü University, Malatya, Türkiye, Institute of Science and Technology, 2022. [Online]. Available: https://tez.yok.gov.tr/UlusalTezMerkezi

[30] M. Cabanillas-Carbonell and J. Zapata-Paulini, "Improving the Accuracy of Oncology Diagnosis: A Machine Learning-Based Approach to Cancer Prediction," *Int. J. Online Biomed. Eng.,* vol. 20, no. 11, 2024/08/08, doi: 10.3991/ijoe.v20i11.49139.

[31] J. Zhao, T. Y. Pan, W. Yao, H. Lu, and Z. Liu, "Analysis of classification algorithms: Insights from MNIST and WDBC datasets," *Applied and Computational Engineering,* vol. 79, no. 1, 2024/07/25, doi: 10.54254/2755-2721/79/20241622.

[32] M. S. Alzboon, S. Qawasmeh, M. Alqaraleh, A. Abuashour, A. F. Bader, and M. Al-Batah, "Machine Learning Classification Algorithms for Accurate Breast Cancer Diagnosis | IEEE Conference Publication | IEEE Xplore," *2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA),* 2023.

[33] A. K. Jakhar, A. Gupta, M. Singh, A. K. Jakhar, A. Gupta, and M. Singh, "SELF: a stacked-based ensemble learning framework for breast cancer classification," *Evol. Intell.,* vol. 17, no. 3, 2023-01-31, doi: 10.1007/s12065-023-00824-4.

[34] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, "Classification Prediction of Breast Cancer Based on Machine Learning," *Comput. Intell. Neurosci.,* vol. 2023, pp. 1-9, 2023, doi: 10.1155/2023/6530719.

[35] I. Ozcan, H. Aydin, and A. Cetinkaya, "Comparison of Classification Success Rates of Different Machine Learning Algorithms in the Diagnosis of Breast Cancer," *Asian Pac J Cancer Prev,* vol. 23, no. 10, 2022/10, doi: 10.31557/APJCP.2022.23.10.3287.

[36] N. A. Mashudi, S. A. Rossli, N. Ahmad, and N. M. Noor, "Comparison on Some Machine Learning Techniques in Breast Cancer Classification | IEEE Conference Publication | IEEE Xplore," *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES),* 2021, doi: 10.1109/IECBES48179.2021.9398837.

[37] B. Imran, H. Hambali, A. Subki, Z. Zaeniah, A. Yani, and M. R. Alfian, "Data Mining Using Random Forest, Naïve Bayes, and Adaboost Models for Prediction and Classification of Benign And Malignant Breast Cancer," *J. Pilar Nusa Mandiri,* vol. 18, no. 1, 2022/03/09, doi: 10.33480/pilar.v18i1.2912.

[38] P. H. Prastyo, I. G. Y. Paramartha, M. S. M. Pakpahan, and I. Ardiyanto, "Predicting Breast Cancer: A Comparative Analysis of Machine Learning Algorithms," *Proceeding International Conference on Science and Engineering,* vol. 3, 2020/04/30, doi: 10.14421/icse.v3.545.

[39] S. Aamir *et al.*, "Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques," (in English), *Comput Math Method M,* vol. 2022, Aug 16 2022, doi: Artn 5869529, 10.1155/2022/5869529.

[40] N. Nuzhat, F. Islam, A. S. Sikder, and N. R. Chakraborty, "Comparative Performance Analysis of Ensemble Models for Breast Cancer Classification," *International Journal of Imminent Science & Technology,* vol. 2, no. 2, 2024, doi: 10.70774/ijist.v2i2.27.

[41] Y. Pristyanto and D. Wirantanu, "Comparison of Genetic Algorithm and Recursive Feature Elimination on High Dimensional Data," *Jurnal Resti (Rekayasa Sistem Dan Teknologi Informasi),* vol. 8, no. 2, pp. 189-198, 2024, doi: 10.29207/resti.v8i2.5375.

[42] A. A. Tawil, L. Almazaydeh, B. Alqudah, A. Z. Abualkishik, and A. A. Alwan, "Predictive Modeling for Breast Cancer Based on Machine Learning Algorithms and Features Selection Methods," *International Journal of Electrical and Computer Engineering (Ijece),* vol. 14, no. 2, p. 1937, 2024, doi: 10.11591/ijece.v14i2.pp1937-1947.

[43] A. Mezaghrani, M. Debakla, and K. Djemal, "Novel Feature Selection Method for Accurate Breast Cancer Classification Using Correlation Coefficient and Modified GWO Algorithm," *Computer Science Journal of Moldova,* vol. 32, no. 2(95), pp. 175-198, 2024, doi: 10.56415/csjm.v32.10.

[44] A. Salhi, R. Alshamrani, A. Althbiti, A. Ismail, M. Abd-ElRahman, and B. M. Hassan, "Optimizing High Dimensional Data Classification With a Hybrid AI Driven Feature Selection Framework and Machine Learning Schema," *Scientific Reports,* vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-08699-4.

[45] J. Zhu, B. Yin, C. Wu, C. Yin, R. Chen, and Y. Ding, "An Integrated Approach of Feature Selection and Machine Learning for Early Detection of Breast Cancer," *Scientific Reports,* vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-97685-x.

[46] O. F. Ayepeku, "Analysis and Visualization of Breast Cancer Prediction Through Machine Learning Models," *Sistemasi,* vol. 13, no. 3, p. 1178, 2024, doi: 10.32520/stmsi.v13i3.4100.

[47] R. Ihaka and R. Gentleman, "R: A language for data analysis and graphics," *J. Comput. Graphical Stat.,* vol. 5, no. 3, p. 299, 1996.

[48] L. Tierney, "The R Statistical Computing Environment," *Lect. Notes Stat.,* 2012, doi: 10.1007/978-1-4614-3520-4_41.

[49] N. Vismam and K. S. Rao, "R Progamming in Different Fields," *Int. J. Comput. Sci. Eng. Technol.,* vol. 9, pp. 1-8, 2019.

[50] K. Smith and S. Climer, "Heterogeneity impacts biomarker discovery for precision medicine," *medRxiv,* 2022-08-01, doi: 10.1101/2022.02.14.22270972.