Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 11 NO: 2 PAGE: 594-603 YEAR: 2022
DOI: 10.17798/bitlisfen.1065685

Breast Cancer Diagnosis with Machine Learning Techniques

Halime DOĞAN[1], Ahmet Burak TATAR[2], Alper Kadir TANYILDIZI[1],
Beyda TAŞAR[1*]

[1]*Mechatronics Engineering Department, Fırat University, Elazığ, Türkiye*
[2]*Mechanical Engineering Department, Adıyaman University, Adıyaman, Türkiye*
(ORCID: 0000-0002-2825-7479) (ORCID: 0000-0001-5848-443X) (ORCID: 0000-0003-3324-5445)
(ORCID: 0000-0002-4689-8579)

Keywords: Machine Learning, Breast Cancer, Wisconsin Data Set, Classification.

Abstract
Cancer deaths are one of the highest rates of death. Although breast cancer is commonly associated with women, it is sometimes seen in men, and the mortality rate for men with breast cancer may be higher. The importance of early detection and treatment of breast cancer cannot be overstated. Cancer is diagnosed at an early stage thanks to expert systems, artificial intelligence, and machine learning approaches, and data analysis makes life easier for healthcare professionals. The nearest neighbor method, principal component analysis (PCA), and neighborhood component method (NCA) approaches were employed to detect breast cancer in this study. "Breast Cancer Wisconsin Diagnostic" database was used to create and test the approach. According to the results obtained, the highest success rate with 99.42% was obtained by using neighborhood component analysis and the nearest neighbor classification algorithm method.

1. Introduction

Machine learning applications have been used in education, agriculture, health, etc. in recent years. It is widely used in studies in different fields such as Predictive analytical methods using machine learning algorithms to increase the efficiency of studies. Gummadi [1] conducted an analysis to improve students' success rates by using machine learning algorithms. Jagwani [2] stated that machine learning algorithms should be used in education to increase the communication performance between teacher and student and to provide real-time feedback advantage. However, different studies have been carried out using machine learning algorithms in many areas, from the estimation of the bacterial population in agricultural ponds [3] to the analysis and formatting of airline ticket prices in the aviation sector [4]. In addition, machine learning algorithms have been used in the early diagnosis of many diseases in the health sector. Studies have been conducted for the early detection of Alzheimer's disease using machine learning algorithms [5,6].

The International Agency for Research on Cancer (IARC) published its most recent report on worldwide cancer data on December 15, 2020 [7]. The IARC report was prepared in light of the information on 36 cancer types and the frequency and mortality rates of cancer types from 185 countries in the GLOBOCAN 2020 dataset. According to this report, 19.3 million new cases were seen in 2020 and 10.0 million people died due to cancer. Cancer affects one-fifth of the world's population at some point in their lives. Men had a mortality rate of 1/8 and women had a mortality rate of 1/11. The number of cancer patients who survive 5 years following diagnosis has been estimated to reach 50.6 million worldwide. Breast cancer diagnoses will account for around 1/8 of all cancer cases in 2020, or about 2.3 million. Breast cancer claimed the lives of 685,000 persons in 2020. Breast cancer ranks fifth among all cancer types in terms of fatality rate. [8].

[*]Corresponding author: btasar04@hotmail.com

Received: 31.01.2022, Accepted: 20.04.2022

The first way to reduce deaths due to breast cancer is through early diagnosis of the disease. Breast cancer diagnosis by specialists is a time-consuming procedure that demands numerous tests. (mammography and magnetic resonance imaging). The use of machine learning techniques in medicine has become common thanks to advancements in software and imaging technology. Thus, computerized analysis and diagnostic artificial intelligence applications are developed to support doctors. Early detection of cancer allows for more treatment options and a higher survival probability, particularly for breast cancer patients [9,10]. In addition, artificial intelligence, machine learning, and deep learning techniques have achieved successful results in the last 20 years, not only on cancer but also on many diseases (Stroke, Alzheimer's, etc.). This has made the use of artificial intelligence, machine learning, and deep learning algorithms in medical image processing even more important [11,12].

In recent years, many studies have focused on breast cancer diagnosis utilizing machine learning-based methodologies. Gupta P. (2020) [13] used the (WBCD) (357 benign, 212 malignant) dataset. They diagnosed it with 6 different methods: KNN, LR, DT, RF, SVM, Deep learning using Adam Gradient Descent Learning. They used these hyperparameters in their models: Number of trees in RF: 100, Max. depth in DT: 4, k neighbor value in KNN is, 6. The regularization parameter (C value) for SVM is 100, and the Coefficient Value(C)for LR is 1. Epoch value is 150 in DL. They reported the most successful method is deep learning using the gradient descent method with 98.24% accuracy rates. Chaurasia V. (2018) [14] used NB, RBF Neural Network, and J48 algorithms to classify breast cancer diagnosis. They tested their method on the WBCD dataset and they obtained this method’s accuracy rates respectively, 97.36%, 96.77%, and 93.41%. They used 10-fold cross-validation for ML methods. Tafish M. [15] used the KNN, ANN, and SVM methods to solve this problem. And firstly they applied to data scaling and normalization then feature selection methods. He reported that the SVM method is the most successful than other methods with a 77.63% accuracy rate on the Breast Cancer Data in Gaza Strip (BCDG) dataset. Gopal V.N., [16] used PCA for the feature selection method and three different classification methods LR, RF, and MLP. In the study of Gopal et al. [16], the properties of the MLP classifier; It was determined as “learning rate: 0.001, max iteration: 200, tol: 0.0001, 10 cross-fold validation”. These method accuracy rates were 79%, 95%, and 98% respectively. Sawssen et.al. [17] used the PCA dimension reduction method and Differential Evolution (DE) algorithm of Radial-Based Function Kernel Extreme Learning Machines (RBF-KELM) in the classification of breast cancer in their study. The results of the algorithm were tested on the WBCD and MIAS databases and compared. Algorithms accuracy rates were 98.18%, 72.7%, 94.54%, 90.9% for KELM, Poly-KELM, Wav-KELM, SVM respectively. Ismaili et.al. [18] using the MIAS data set, achieved 86.96% and 95.15% accuracy rates as a result of the SVM and ANN methods. Ateş et.al. [19] using the WBCD data set, achieved 96.5%, 92.4%, and 96.5% classification accuracy rates as a result of the NB, DT, and MLP methods. Selvi et.al. [20] using SVM, NB, RF, and KNN methods and tested that performance on the WBCD data set, achieved 93.865%, 94.74%, 96.49%, 91.23%, and 98.24% accuracy rates. Rahman et al [21] created an ANN classification model for the classification of breast cancer. They determined the number of neurons in a single hidden layer of the ANN as 15 using the Taguchi method. The success of the developed model was tested on the Wisconsin Diagnostic Breast Cancer Dataset. It was stated that the proposed model reached 98.8% accuracy in breast cancer classification. Akay et al. [22] proposed an SVM-based method with grid search and combined with feature selection for the diagnosis of breast cancer. They tested their method on different training-test sections of the Wisconsin breast cancer dataset (WBCD). The highest classification accuracy (98.53% for 70-30% training test partition) was obtained with the SVM model with five features. İbrahim et.al. [23] in their study, before passing the features to the classification step, made feature selection using the variance of the input features and correlation analysis. An ensemble method was selected from seven classification algorithms to improve breast cancer classification. They tested performance on the WBCD data set, achieving 99% accuracy rates. Memon, M. H. [24] used the linear SVM method to solve the breast cancer classification problem. They used preprocessing methods such as missing value detection, standard scalar, and min-max scalar. They selected the most relevant features with the REF method. They split the dataset into 70% for training and 30% for testing. Memon and his team [24], which has the closest accuracy to their study, used c=1, γ= 0.0001, time: 0.03 s parameter values in the SVM classifier and achieved 99% accuracy on the WBCD dataset. Bayrak, E.Y. [25] used the multiclass SVM and MLP methods and reported that the both SVM and MLP method accuracy was 95.37% for 66-33% train test splits. Kumar, P.P. [26] proposed a random decision tree algorithm. The weighting of the features according to the feature importance was carried out by the Core Neutrophic C-Means

Clustering method. Larger weights are assigned to applicable features and smaller weights to less applicable features. Performance analysis was tested on the Wisconsin Breast Cancer dataset with a 70%-30% training-test rate. SVM method test accuracy was 80% for the WBCD dataset.

Although there are many applications in the literature for the detection of breast cancer, research on this subject is still continuing since the desired level of success has not been achieved. Considering the methodological approach and results of this study, its contributions to the literature are summarized in articles.

- Unlike the studies summarized above, in this study, first of all, the feature set in the data set was analyzed. The local outlier method was applied for radius mean and tissue mean values, and outliers were removed. Thus, the success of the classification algorithm is increased.
- In the analysis of the WBCD data set used in the study, it was determined that the data in the feature set did not show a natural distribution. For this reason, a standardization process was applied to the data set in order to increase the classification success.
- The train test segmentation rate of 70%-30%, which was reported to have the highest success when the literature was analyzed, was also adopted in this study.
- In this study, two different feature reduction/selection methods (PCA and NCA) were applied to the data set separately, and the performances of the KNN classifier trained with the selected data feature vector were compared.
- Five different performance metrics (Accuracy Sensitivity Specificity Precision F1 MCC Kappa) were used to analyze the experimental results.
- The obtained results proved that the NCA feature selection and KNN classification algorithm reached a 99.42% accuracy rate and a 0.42% higher performance was obtained than the closest high-performance study in the literature.

The organization of the paper consists of three main parts. In Section 2, features of Breast Cancer in Wisconsin dataset, outlier detection, standardization and separation of the dataset, designed feature selected methods: PCA, NCA, and classification method: KNN method. And also performance metrics are included in Section 2. Section 3 includes experimental results and discussion. In that part, the confusion matrix and performance table are presented for all methods. In addition, the results of previous studies and the performance of the proposed method are compared in this section. All the important results of the study are summarized in the conclusion section.

2. Material and Method

This section offers details on the dataset and machine learning methods used to construct computerized algorithms for breast cancer detection.

2.1. Dataset for Breast Cancer in Wisconsin

The dataset [27], which is available in the UCI machine learning repository, was used in this study. There are 569 samples in this data collection, each with 32 characteristics (Figure 1). These are ID, diagnostic, and 30 real-valued input properties. The data collection contains 569 samples, 357 of which are benign and 212 of which are malignant. There are no situations where information is missing. To calculate features, a digitized picture of a fine needle aspirate (FNA) of a breast mass was used by Dua et al [27]. Features indicate the characteristics of the cell nuclei in the image. Ten real-valued features are determined for each cell nucleus: Radius, environment, area, evenness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concave points, symmetry, fractal dimension ("coastline approximation" - 1) [27].

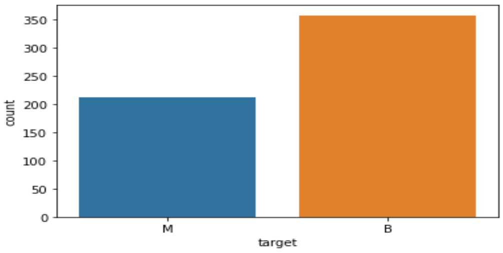


Figure 1. Benign (B) and malignant (M) data distribution

This listed feature set is presented as an input to the classification algorithm proposed in the study and a two-class problem is solved.

2.2. Outlier Detection

Many statistical testing and machine learning algorithms are susceptible to outliers. Outliers reduce classification success. The outlier value may be caused by the error made in the measurements, and for a good performance, these values should be detected in the dataset and taken before the classification step. Within the scope of this study, the local outlier method was applied for radius mean and tissue mean values, and outliers were removed (Figure 2).

The boxplot in Figure 2 provides a visualization of the statistics for the two features. The bottom and top of each rectangular box represent the border of the 25th and 75th percentiles of the data for that sample, respectively, and the data in this range in the rate. The red line in the middle of the box is the median value of that feature. The dashes at the top and bottom of the box and the horizontal line at the end represent the normally distributed maximum and minimum values of the data. The red '+' symbols outside this horizontal line indicate the outliner values.

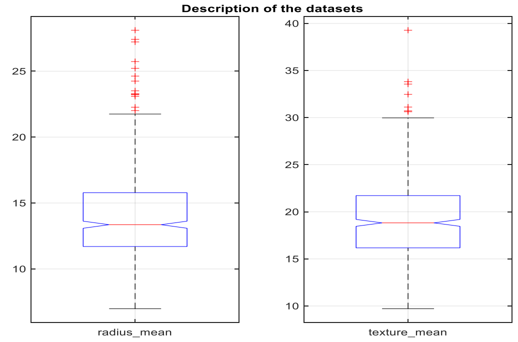


Figure 2. Display of radius mean, tissue mean values, and outliers

2.3. Standardization and Separation of Dataset

If the data set features do not resemble naturally distributed data, the classification may not work properly (such as gaussian with zero mean and unit variance) [28]. As a result, adopting the data standardization method is advantageous. Many machine learning estimators in scikit-learn require dataset standardization. The data is rescaled using a mean of 0 and a standard deviation of 1. The distribution approaches normal. The mean value is subtracted from the obtained value and divided by the standard deviation value and is shown in Eq. (1). Here x is the value in the data. μ is the mean. σ is the standard deviation of the data.

$$z = \frac{x-\mu}{\sigma} \tag{1}$$

Before beginning the model's training phase, the data set was separated into two clusters after standardization. On the training set, the model was trained, and the model's success was tested on the test set. As our training set grows, our model will learn better. As our test set grows, evaluation metrics will produce more reliable test accuracy results. Rahman et.al. [21] Wisconsin dataset was used in their study and they studied the dataset with 60%-40%, 70%-30%, and 80%-20% training-test rates. The highest success was achieved in the application where they divided 70%-30% with an accuracy of 98.5%. In the application where they divided it by 80%-20%, the success decreased to 94.7%. Akay [22], on the other hand, found an accuracy rate of 98.19% when he chose the training-test partition ratio of 50%-50%. The accuracy rate is 99.24% when he chose 70%-30%. The accuracy rate is 98.8% when he chose 80%-20% in his study.

2.4. Principal Components Analysis

Principal Components Analysis (PCA) is a multivariate size reduction technique that uses orthogonal transformations to condense a large number of linked variables into a smaller set. [29,30]. The equations used to calculate the principal components are presented in Equations (2-5). *For t= (1,2, …, n)*, *{x(t)}* represents *a* random dataset containing zero-mean features. In this case, the covariance matrix is R;

$$R = \frac{1}{n-1}\sum_{t=1}^{n}[x(t)x(t)^T] \tag{2}$$

It has been calculated as (3) was used to calculate linear combinations of variables in the original data, namely the linear transformation from x(t) to y(t).

$$y(t) = M^T x(t) \tag{3}$$

M is an orthogonal matrix of size n × n. Column, *i*. is equal to the eigenvector. In this case, at the point, the eigenvalue problem can be written as in (4). q_i denotes the eigenvector and R an eigenvalue of the covariance matrix. ($\lambda_1>\lambda_2>\cdots>\lambda_n$).

$$\lambda_i q_i = R q_i \tag{4}$$

The principal components were calculated with the help of (5). Here $y_i(t)$, *i*. symbolizes the basic component [29,30].

$$y_i(t) = q_i{}^T x(t), i = 1,2,3,\dots,n \tag{5}$$

2.5. Neighborhood Components Analysis

Neighborhood component analysis (NCA) is a supervised learning method that uses a distance measure to classify multivariate data [31]. It is a metric learning machine learning algorithm. It learns a supervised linear transformation to improve the stochastic nearest neighbors' rule's classification accuracy in the modified space [32]. NCA is a statistical method that is used to improve accuracy. NCA's goal is to create an optimal objective function by utilizing a gradient-based optimizer. NCA selects a neighbor at random and returns the predicted decision for each class in which each point *i* selects a neighboring *j* point with *a* distance d_{ij} between points *i* and *j* and a probability of P_{ij} as Eq. (7) [33].

The sum of the absolute differences between the feature vectors given in Eq. (6) is multiplied by the weight of the vector to find the distance between them. The distance between x_i and x_j is d_{ij}.

$$d_{ij} = \sum_{l=1}^{p} w_l^2 \left| x_{il} - x_{jl} \right| \qquad (6)$$

The weight of the first feature is represented by $w_{l.}$ Unlike the K-Nearest Neighbor algorithm, the NCA selects a random neighbor for each class and makes the predicted judgment. Here, each point x_i chooses a neighboring point x_j with probability P_{ij}.

$$P_{ij} = \frac{k(d_{ij})}{\sum_{k=i} k(d_{ik})};\ P_{ij} = 0 \qquad (7)$$

$k(z) = \exp(-z/\sigma)$ is a kernel function. The possibility of each point being chosen as the reference point is influenced by the kernel width *d*. The predicted fraction of time is represented by Probability Pi. The point *i* will be labeled in the correct class (Eq. 8) denoting the set of points in the same class as *i*. To maximize the objective function, (Eq. 9) is applied.

$$P_i = \sum_{j \in C_i} P_{ij} \qquad (8)$$

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} P_i \qquad (9)$$

2.6. k-Nearest Neighbor Classification Method

Because of its simplicity and efficiency, the k-Nearest Neighbor Classification Method (KNN) is a commonly used pattern classification technique. It is one of the effective classification algorithms among supervised learning methods [34]. Also, KNN, a powerful multivariate statistical method that uses distance and assesses differentiating features [35], employs distance and analyzes distinguishing features. It makes no assumptions about the data's statistical properties. KNN predicts the class attribute based on the k nearest training instances in the feature space. It selects the class by taking into account the most representative samples given a data set, using the k closest samples from the classified training data. The Euclidean distance similarity metric was employed to pick neighbors in this study. The calculation of Euclidean distances is done as in Eq. (10) [36].

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (10)$$

In this case, x_i and y_i are two Euclidean n-space points. The classification accuracy was evaluated by dividing the number of successfully classified samples by the total number of samples after all test samples were categorized by KNN. Eq. (11) is used to calculate the mean absolute error (MAE) [36].

$$Mean\ Absolute\ Error = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i| \qquad (11)$$

In this study, the design model type of KNN is Fine. The distance metric is Euclidean. The number of neighbors (*k*) is 3. Distance weight is equal.

Table 1. Performance Evaluation metrics

Performance Metric	Acronym	Equation	Explanation
Precision *(Positive Prediction Value)*	PPV	$\frac{TP}{TP + FP}$	From all samples projected to be positive, the proportion of accurately predicted positive samples.
Negative Prediction Value	NPV	$\frac{TN}{TN + FN}$	From all samples projected to be negative, the proportion of accurately predicted negative samples.
Sensitivity –Recall *(True Positive)*	TPR	$\frac{TP}{TP + FN}$	The ratio of TP results in the total number of true positive samples.
Specificity *(True Negative)*	TNR	$\frac{TN}{TN + FP}$	The ratio of TP results to the total number of true negative samples.
Accuracy	ACC	$\frac{TP + TN}{TP + TN + FP + FN}$	The ratio of the number of correct guesses made by the method to the total number of guesses made.
F-Score	F-Score	$\frac{2 \times precision \times recall}{precision + recall}$	The F-Score value shows us the harmonic average of the Sensitivity and Recall values.
MCC-Matthews correlation coefficient	MCC	$\max([(TP * TN) - FP * FN)/((TP + FP) * P * N * (TN + FN))\char`^0.5],\ [((TP + FP) * P * N * (TN + FN))\char`^0.5])$	
Cohen's Kappa	Kappa	$\mathrm{po} = \mathrm{ACC}$ $\mathrm{pe} = ((\mathrm{P} * (\mathrm{TP} + \mathrm{FP}) + (\mathrm{FN} + \mathrm{TN}))/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})\char`^2$ $\mathrm{kappa} = \max(\left[\frac{\mathrm{po} - \mathrm{pe}}{1 - \mathrm{pe}}; \frac{\mathrm{pe} - \mathrm{po}}{1 - \mathrm{po}}\right])$	

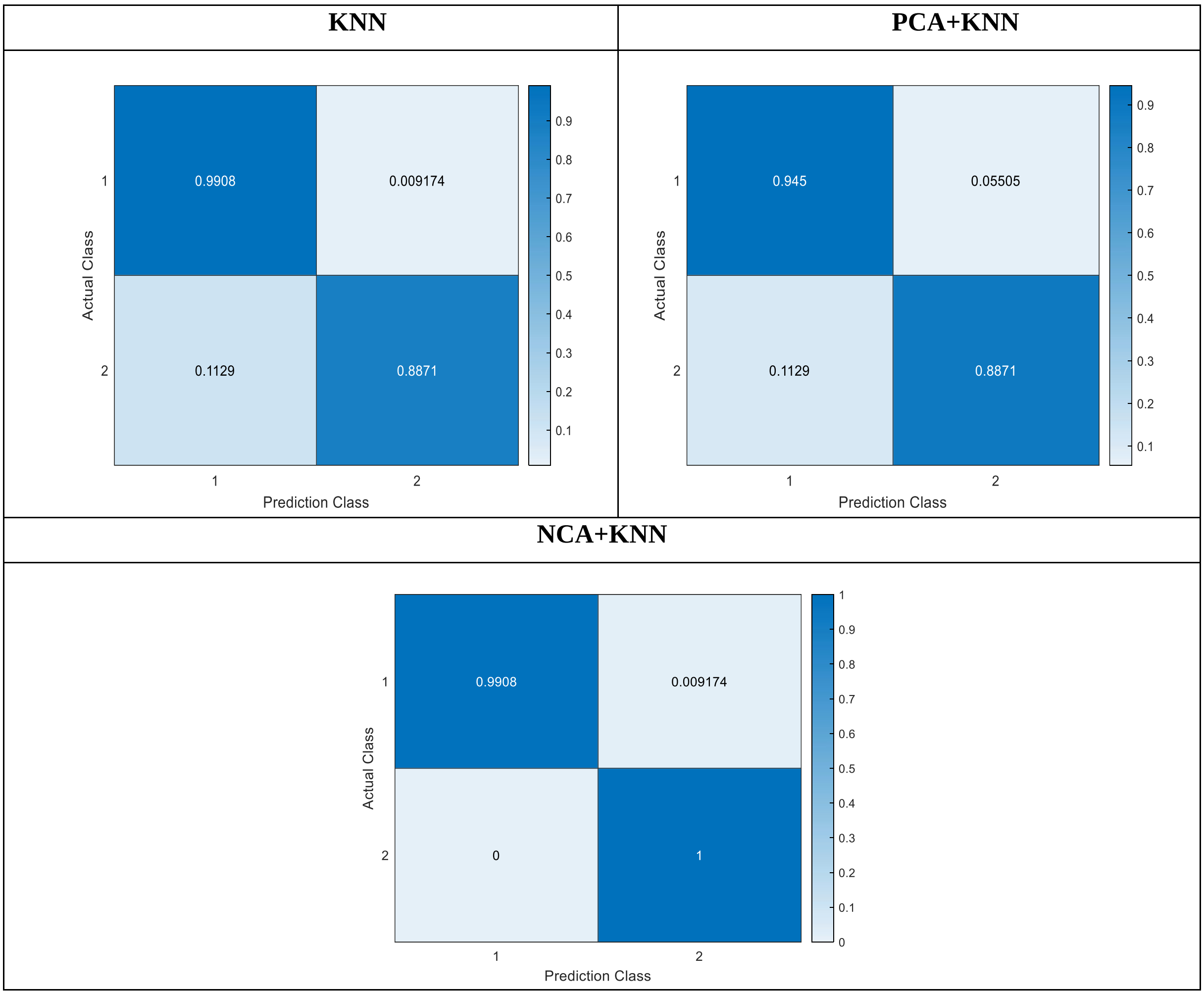
2.7. Performance Evaluation Methods

A range of performance criteria was used to compare the performance of early detection of breast cancer using machine learning techniques [37-39]. In Table 1, TP is actual value positive and prediction positive. FP is actual value negative and prediction positive. TN is actual value negative and prediction negative. FN is actual value positive and prediction negative. The terms recall, sensitivity, precision, and specificity values are described and all these metrics' formulas are in Table 1 [37-40]. The most frequently used performance criteria in machine learning applications, F1 score, MCC, Kappa, and Percentage accuracy values are calculated with the formulas in Table 1 and presented in the results section. Table 1 also includes the theoretical explanation and abbreviation of each criterion.

3. Results and Discussion

Figure 3 shows the confusion matrices obtained for the application of breast cancer detection using the KNN, PCA+KNN, and NCA+KNN methods. As can be seen, the approach with the highest TP and FP rates is the combined use of NCA+ KNN.

At the end of the study, 108 of 109 benign cells were predicted correctly and 1 was incorrectly predicted by the KNN method. Out of 62 malignant cells, 55 were predicted correctly and 7 were incorrectly predicted. With the PCA+ KNN method, 103 of 109 benign cells were correct and 6 were incorrect; Out of 62 malignant cells, 55 were predicted correctly and 7 were incorrectly predicted. By the NCA + KNN method, 108 of 109 benign cells were predicted correctly and 1 was predicted incorrectly. Table 2 shows the performance metric table. After feature selection with PCA, the use of the KNN classifier has a 92.39% accuracy rate, although it reduces the success. In the application where only the KNN algorithm was used, an accuracy rate of 95.32% was achieved. The highest success was obtained in feature selection with NCA and classification with KNN, and the accuracy rate was 99.42%. The F1 score performance values of these three methods are 0.9643, 0.9406, and 0.9954, respectively. Sensitivity, Specificity, Precision, Kappa, and MCC values are also included in the table.


KNN

PCA+KNN

NCA+KNN

Figure 3. Confusion matrices of classification methods

Table 2. Performances of classification methods

Classifier	Accuracy	Sensitivity	Specificity	Precision	F1	MCC	Kappa
KNN	0.9532	0.9908	0.8871	0.93.91	0.9643	0.8966	0.8966
PCA + KNN	0.9240	0.9450	0.8871	0.9364	0.9406	0.8350	0.8350
NCA + KNN	0.9942	0.9908	1	1	0.9954	0.9875	0.9874

Table 3 summarizes the results of chest cancer detection studies using the WDBC dataset. It is seen that the classification accuracies of the existing studies in the literature vary between 72% and 99%, the precision value varies between 78% and 100%, the recall value varies between 79% and 99.24%, and F1 scores vary between 75% and 99%. In this study, the 99.42% accuracy, 100% precision, 99.08% recall, and 99.54% F1 score values obtained with the NCA + KNN method are relatively higher than the studies in the literature.

Table 3. Comparison of the results of some breast cancer diagnosis studies using the WBCD dataset

Study	Outliner Detection	Data selection	Classification Algorithms	Accuracy (%)	Precision	Recall / Sensitivity	F1-Score
Gupta (2020)[13]	X	√	*KNN*	95.8	93.5	93.5	93
			LR	95.8	96.5	95	95.5
			DT	95.8	96.5	95	95.5
			RF	97.2	97	97.5	97
			SVM	97.2	97.5	97	97
			DLL	98.24	98	98	98
Chaurasia (2018) [14]	X	X	*NB*	97.36	-	97.4	-
			RBF Neural Network	96.77	-	97.07	-
			J48 Algorithm	93.41	-	93.4	-
Gopal (2021) [16]	X	√	*LR*	79	78	79	78
			RF	95	90	94	92
			MLP	98	98	97	96
Sawssen (2022) [17]	X	√	PCA+KELM	98.18	-	94.83	92.80
			PCA+Poly-KELM	72.7	-	93.82	90.51
			PCA+Wav-KELM	94.54	-	100	91.26
			PCA+SVM	90.9	-	85.91	91.04
Ateş (2021) [19]	X	X	*NB*	96.5	96.70	93.70	95.20
			DT	92.4	87.90	92.10	89.90
			MLP	96.5	95.20	95.20	95.20
Sevli (2019) [20]	X	X	*SVM*	93.86	97.37	86.05	-
			NB	94.74	100	86.06	-
			RF	96.49	95.35	95.35	-
			KNN	91.23	94.59	81.39	-
			LR	98.24	100	95.35	-
Rahman (2020) [21]	X	X	*ANN*	98.8	98.40	99.10	98.80
Akay (2009) [22]	X	√	*SVM*	98.53	99.55	99.24	-
Ibrahim (2021) [23]	X	√	*Ensemble based*	99.00	99.29	96.00	97.51
Memon (2019) [24]	√	√	*SVM-linear*	99.00	99	98	99
Bayrak (2019) [25]	X	X	*SVM*	95.37	95.40	95.40	-
			MLP	95.37	95.50	95.40	-
Kumar (2021) [26]	X	X	*GA-SVM*	80.00	92.50	80.00	75.00
This Study	√	√	KNN	95.32	93.91	99.08	96.43
			PCA + KNN	92.40	93.64	94.50	94.06
			NCA + KNN	**99.42**	**100**	**99.08**	**99.54**

******* k-Nearest Neighbor (*KNN*), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes Algorithm (NB), RBF Neural Network, Kernel Extreme Learning Machine (KELM)

As can be seen in Table 3, studies have been conducted in the literature on the prediction of breast cancer diagnosis by many different machine learning techniques. Although similar machine learning classification methods are used in the literature, the reason for achieving success in different accuracy ranges is the pre-processing and feature selection/reduction methods applied to the data before the classification step. In order to design an effective machine learning model, it is necessary to analyze the data well, detect and remove outliers, and select the most effective features. In this study, unlike previous studies, the step of detection and exclusion of outliers was added to the data set. When the table is examined, it is seen that the results of the studies with feature selection [13,16-17,22-24] are considerably higher than the results of the studies without feature selection [14,16-21,25-26]. PCA, KPCA, etc. in the literature. Although these methods were used, the feature selection method with NCA was applied to WBCD data for the first time in this study. And the obtained results showed that feature selection with NCA is much more efficient than feature selection with PCA.

4. Conclusion

In this study, three models for the diagnosis of breast cancer with machine learning methods have been developed. The dataset was used for training and testing the developed method. According to the results obtained, it was seen that the highest classification success was obtained by using the NCA+KNN technique with 99.42%.

It is thought that this method can be used as an assistive system for physicians in the early-stage detection of breast cancer, which is an important problem of our time, with high accuracy. Early diagnosis and treatment of breast cancer prolong patient survival. It also shortens the treatment process with early diagnosis, which significantly reduces healthcare costs.

Acknowledgment

This study was supported by Firat University within the scope of the MF 21.14 Graduate BAP Project.

Contribution of Authors

Every author contributed equally to this work.

Conflict of Interest

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

Research and publication ethics complied with the study.

References

[1] P. Gummadi, “Analysis of machine learning in education sector”, 1-5, 2020.

[2] A. Jagwani, “A review of machine learning in education”, *Journal of Emerging Technologies and Innovative Research (JETIR)*, 6(5), 384-386, 2019.

[3] G. Buyrukoğlu, S. Buyrukoğlu, Z. Topalcengiz, “Comparing regression models with count data to artificial neural network and ensemble models for prediction of generic escherichia coli population in agricultural ponds based on weather station measurements”, *Microbal Risk Analysis*, 19, 100171, 2021.

[4] S. Buyrukoğlu, Y. Yılmaz, “An approach for airfare prices analysis with penalized regression methods”, *Data Science*, 4(2), 57-61, 2021.

[5] S. Buyrukoğlu, “Early detection of Alzheimer’s disease using data mining: comparison of ensemble feature selection approaches”, *Konya Journal of Engineering Sciences*, 9(1), 50-61, 2021.

[6] S. Buyrukoğlu, “Improvement of machine learning models’ performances based on ensemble learning for the detection of Alzheimer Disease”, *6th International Conference on Computer Science and Engineering*, 102-106, 2021.

[7] World Health Organzation, “International agency for research on cancer-iarc”, 2020, https://gco.iarc.fr/today/home.

[8] L. Çelik, “Meme kanseri taramasinda yapay zeka”, 2020, https://www.drozdogan.com/turkiye-kanser-istatistikleri-2020/

[9] C. Eyupoglu, “Breast cancer classification using k-nearest neighbors’ algorithm”, *The Online Journal of Science and Technology*, 8(3), 29-34, 2018.

[10] Ł. Jeleń, A. Krzyżak, T. Fevens, M. Jeleń, “Influence of feature set reduction on breast cancer malignancy classification of fine needle aspiration biopsies”, *Computers in Biology and Medicine*, 79, 80- 91, 2016.

[11] S. Savaş, “Detecting the stages of Alzheimer’s disease with pre-trained deep learning architectures”, *Arabian Journal for Science and Engineering*, 47, 2201-2218, 2022.

[12] S. Savaş, N. Topaloğlu, Ö. Kazcı, P. Koşar, “Comparison of deep learning models in carotid artery intima-media thickness ultrasound images: CAIMTUSNet”, *Journal of Information Technologies*, 15(1), 1-12, 2022.

[13] P. Gupta, S. Garg, “Breast cancer prediction using varying parameters of machine learning models”, *Procedia Computer Science*, vol. 171, pp. 593–601, 2020.

[14] V. Chaurasia, S. Pal, B. Tiwari, “Prediction of benign and malignant breast cancer using data mining techniques”, *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 119–126, 2018.

[15] M.H. Tafish, A.M. El-Halees, “Breast cancer severity degree predication using data mining techniques in the gaza strip”, *in 2018 International conference on promising electronic technologies (ICPET), Deir El-Balah*, pp. 124–128, 2018.

[16] V.N. Gopal, F.A. Turjman, L. Anand, M. Rajesh, “Feature selection and classification in breast cancer prediction using IoT and machine learning”, *Measurement*, 178, 109442, 2021.

[17] B. Sawssen, T. Okba, “A novel machine learning approach for breast cancer diagnosis”, *Measurement*, 187, 110233, 2022.

[18] F. Ismaili, L. Shabani, B. Raufi, J. Adjari, X. Zenuni, “Enhancing breast cancer detection using data mining classification techniques”, *Press Academia Procedia, 2nd World Conference on Technology, Innovation and Enterpreunership, 2017, İstanbul, Turkey*, 2017.

[19] İ. Ateş, T.T. Bilgin, “The investigation of the success of different machine learning methods in breast cancer diagnosis”, *Konuralp Medical Journal*, 13(2), 347-356, 2021.

[20] O. Sevli, “Performance Comparison of Different Machine Learning Techniques in Diagnosis of Breast Cancer”, *European Journal of Science and Technology*, 16, 176-185, 2019.

[21] A. Rahman, R.C. Muniyandi, D. Albashish, M. Rahman, O.L. Usman, ”Artificial neural network with taguchi method for robust classification model to improve classification accuracy of breast cancer”, *PeerJ Computer Science*, 2-27, 2020.

[22] M.F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis”, *Expert Systems with Applications*, 36, 3240-3247, 2009.

[23] S. Ibrahim, S. Nazir, S.A. Velastin, “Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis”, *Journal of Imaging*, 7(225), 3-16, 2021.

[24] M.H. Memon, J.P. Li, A.U. Haq, M.H. Memon, W. Zhou, “Breast cancer detection in the IOT health environment using modified recursive feature selection”, *Hindawi Wireless Communication and Mobile Computing*, 1-19, 2019.

[25] E.Y. Bayrak, P. Kırcı, T. Ensari, “Comparison of machine learning methods for breast cancer diagnosis”, *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 2019.

[26] P.P. Kumar, M.A.V. Bai, G.G. Nair, “An efficient classification framework for breast cancer using hyperparameter tuned random decision forest classifier and Bayesian optimization”, *Biomedical Signal Processing and Control*, 68, 102682, 2021.

[27] D. Dua, C. Graff, “UCI machine learning repository”, *Irvine, CA: University of California, School of Information and Computer Science*, 2019 http://archive.ics.uci.edu/ml

[28] A. Zang, A. Casari, “Feature engineering for machine learning: principles and techniques for data scientist”, *O’Relly Media Publishing, Sebastopol, USA*, 2018.

[29] F. Salo, A.B. Nassif, A. Essex,” Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection”, *Computer Networks*, 148, 164-175, 2019.

[30] H.J. Chiu, T.H.S. Li, P.H. Kuo, “Breast cancer–detection system using PCA, multilayer perceptron, transfer learning, and support vector machine”, *IEEE Access*, 8, 204309-204324, 2020.

[31] S. Laghmati, B. Cherradi, A. Tmiri, O. Daanouni, S. Hamida, “Classification of patients with breast cancer using neighborhood component analysis and supervised machine learning techniques”, *In 2020 3rd International conference on advanced communication technologies and networking (CommNet), IEEE* pp. 1-6, 2020.

[32] S.F. Khorshid, A.M. Abdulazeez, “Breast cancer diagnosis based on k-nearest neighbors: a review”, *PalArch's Journal of the archaeology of Egypt/Egyptology*, 18(4), 1927-1951, 2021.

[33] R. Massafra, A. Latorre, A. Fanizzi, R. Bellotti, V. Didonna, V. Giotta, V. Lorusso, "A clinical decision support system for predicting invasive breast cancer recurrence: preliminary results", *Frontiers in Oncology*, 11, 284, 2021.

[34] T. Savaş, S. Savaş, "Feature-based performance comparison of machine learning algorithms for phishing detection through uniform resource locator", *Journal of Polytechnic, early access,* 2021.

[35] T.A. Assegie, "An optimized k-nearest neighbor-based breast cancer detection", *Journal of robotics and control (JRC)*, 2(3), 115-118, 2021.

[36] C.J. Willmott, K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE)", *in assessing average model performance. Climate Research,* 30, 79-82, 2005.

[37] G. Kou, Y. Peng, G. Wang, "Evaluation of clustering algorithms for financial risk analysis using MCDM methods", *Inf Sci (Ny),* 275, 1–12, 2014

[38] H. Xiao, K. Yi, R. Peng, G. Kou, "Reliability of a Distributed Computing System with Performance Sharing", *IEEE Trans Reliab,* 1–12, 2021.

[39] G. Kou, H. Xiao, M. Cao, L.H. Lee, "Optimal computing budget allocation for the vector evaluated genetic algorithm in multi-objective simulation optimization", *Automatica,* 129, 109599, 2021.

[40] H. Xiao, Y. Yan, G. Kou, S. Wu, "Optimal Inspection Policy for a Single-Unit System Considering Two Failure Modes and Production Wait Time", *IEEE Trans Reliab,* pp:1–13, 2021.