Journal of Engineering Technology and Applied Sciences, 2022 *e-ISSN:* 2548-0391
Received 25 August 2022, Accepted 13 October 2022 *Vol. 7, No. 3, 195-205*
Published online 22 December 2022, Research Article *doi: 10.30931/jetas.1166768*

Citation: Yilmaz, H., Kuncan, F., "Analysis of Different Machine Learning Techniques with PCA in the Diagnosis of Breast Cancer". Journal of Engineering Technology and Applied Sciences 7 (3) 2022 : 195-205.

ANALYSIS OF DIFFERENT MACHINE LEARNING TECHNIQUES WITH PCA IN THE DIAGNOSIS OF BREAST CANCER

Huseyin Yilmaz[a], Fatma Kuncan[b*]

[a]*Department of Electrical and Electronics Engineering, Faculty of Engineering, University of Siirt, Turkey*
hsynylmz0280@gmail.com

[b]*Department of Computer Engineering, Faculty of Engineering, University of Siirt, Turkey*
*(*corresponding author) fatmakuncan@siirt.edu.tr*

Abstract

In recent years, different types of cancer cases are common. Increasing cancer cases, A rapidly increasing health for countries and humanity becomes a problem. In addition to being the most common cancer among women today, breast cancer has surpassed lung cancer as the most common cancer type in the world since 2021. Early diagnosis greatly reduces the risk of death in breast cancer, and benign tumors are correctly diagnosed, allows the classification of this field to be a new research topic. New developments in the field of Medicine and Technology Machine learning, classification algorithms and computerized diagnosis are used in the correct classification of tumors. increased its use. These systems are extremely important in terms of being an assistant to the expert opinion. In this study, in the Wisconsin Breast Cancer dataset, it is aimed to accelerate the diagnosis of the disease and to reduce the tumors, different machine learning to minimize treatment processes by providing accurate classification techniques were used. In this study, we reduced our dataset to 171 data using Principal Component Analysis (PCA) to accelerate disease diagnosis on the Wisconsin Breast Cancer dataset and 2 different classification processes were performed using 5 different machine learning. The success rate of each algorithm was compared, and it was revealed that Logistic Regression was the most successful method with an accuracy rate of 98.8% after PCA.

Keywords: Breast cancer, PCA, classification algorithms, machine learning, logistic regression

1. Introduction

Cancer is a disease that occurs as a result of uncontrolled proliferation and growth of cells in any organ or tissue of the body. Cancer is named according to the tissue in which it occurs.

More than 200 types have been identified [1]. The uncontrolled proliferation of these cells, which do not have capsules and grow rapidly by entering the surrounding tissues and vessels, causes the tissue and organ to not function or to deteriorate [2]. Breast cancer can start in different areas. There are ducts, adipose tissues, and glands in each breast, right and left. Benign and Malignant breast are shown in Figure-1 [3].

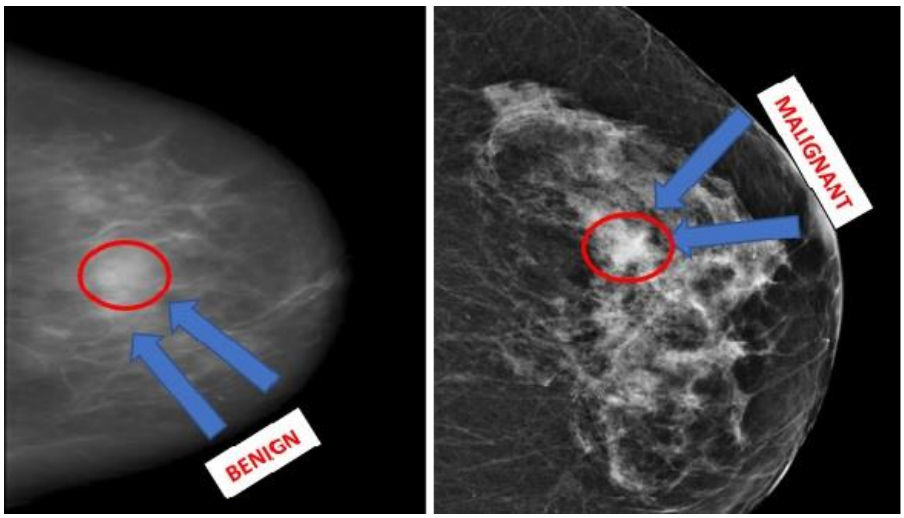

Figure 1. Malignant and Benign breast tissue

According to the data of the Ministry of Health, breast cancer is the most common type of cancer that causes death in women in our country and in the world. Breast cancer is seen in 1 out of every 12 women in our country. The incidence of breast cancer in our country is 45.6 per thousand, which is very close to the world average of 46.3. The incidence of breast cancer is 92.6 per thousand in Northern European countries. In our country, 18.000 women are diagnosed with breast cancer annually. The treatments of breast cancers detected in the early stages before clinical findings appear are more successful and the quality of life of the individual increases significantly [4]. Machine learning, Artificial intelligence and Computerized diagnostic systems are actively used in the detection of cancer in the periods when the early diagnosis of the disease contributes to the healing process very high, thanks to the integration of technology in the field of health.

2. Literature studies

Wisconsin Breast Cancer dataset was created by Dr. William Wolberg from digitized images to be used to determine whether the image obtained with the help of fine needle aspirate from the mass in the breast is benign or malignant [5]. The fact that the data set is suitable for the use of classification algorithms has allowed the data set to be used quite a lot and to make different approaches in this field. Some of the studies in the literature are mentioned below.

Toğacar (2018) performed a deep learning process with a convolutional neural network (CNN) in diagnosing breast cancer using the Wisconsin Breast Cancer data set. He obtained the classification of images as benign and malignant with an accuracy rate of 93.4% using the Support Vector Machines method. In order to obtain a high accuracy in classification of the data set, AlexNet method was used for feature extraction [6].

Yavuz E. – Eyüpoğlu C.(2019) suggested the use of General Regression neural network and Feedforward neural network for classification as a new score fusion approach in the diagnosis of breast cancer. The score fusion approach used works on the principle of obtaining a better score in total by collecting the results obtained by each classification network, and it is aimed to maximize the success achieved. Considering the results obtained, the proposed score fusion method provided a higher rate of classification (95.93) compared to the use of General regression and feedforward neural networks [7].

E. Bayrak et al. (2022) prepared a research article on the comparison of accuracy values and complexity matrix results of different classification algorithms on 2 different Breast cancer datasets in Kaggle. Complexity matrices were prepared by applying k-Nearest Neighbor, Support Vector Machines, Decision Tree, Naive Bayes and Artificial Neural Network algorithms to the data set. Artificial neural networks were applied to 2 different data sets and obtained the highest success rate compared to the others, with 2 different types of 98,2456 and 93,8596 in terms of the obtained classification performances. It has been mentioned that the prepared datasets can be used with a high success rate by using artificial neural networks in the diagnosis of breast cancer disease [8].

S. Motarwar et al. (2022), using the SMOTE technique, compared the pre-SMOTE and post-SMOTE performances of the values obtained from 6 different classification algorithms. The use of SMOTE helps to minimize the clutter and imbalance between classes in the data set. Since personal health is at the forefront in such classifications used in the medical field, the data obtained should be extremely sensitive and directly related. In the classification algorithms used (LR, DTC, RFC, KNN, SVM, ABC), the KNN and SVM algorithms gave the highest results with 94.74 in the data obtained before the SMOTE technique, and a much more sensitive and higher rate in the KNN algorithm with a rate of 95.32 after SMOTE provided [9].

S. Singh et al. (2022) stated that Machine Learning algorithms are used in the diagnosis of diseases in the health field, but there are errors in these classification techniques, and in order to minimize these errors, they are minimized by using various supporting applications such as Features selection or Features extraction to the data sets before classification algorithms. FLANN, one of the simple single-layer neural network models, was used to minimize these errors. Classification was carried out by applying it to two different breast cancer diagnostic datasets existing in the literature, and the experimental results obtained achieved a high accuracy of 99.41% in the diagnosis of early-stage breast cancer [10].

Prithwish Ghosh (2022) stated that the XGboost algorithm in diagnosing breast cancer provides faster results than other existing algorithms in diagnosing the disease. The author, who claims to include strengthened tree algorithms in the library in order to obtain faster results with the XGBoost algorithm, selected 13 features with optimal features as a result of preprocessing, and achieved a 97.66% success rate as a result of cross validation [11].

M. Mangukiya et al. (2022) compared the performance of different machine learning algorithms on the Wisconsin breast cancer dataset in their Breast cancer detection study with machine learning method. SVM, NB, k-NN, Adaboost, XGboost and Random Forest algorithms were applied to the data set. The main purpose of applying the algorithms was to detect the disease early and accurately by finding the most successful method between the accuracy and precision values of the obtained data. After the study, it was seen that the

XGboost algorithm gave the most successful result of 98.24% with the lowest error and highest accuracy rate after learning machine learning algorithms [12].

B. Nalbant and I. Argun digitized the existing data in the data set and compared two different machine learning algorithms through WEKA and ORANGE software. Although Adaboost and SVm algorithms gave more successful results in previous studies, it was found that the success rate of the SVM algorithm on Oranfge and WEKA was lower than the Random Forest algorithm. In the study, the Random Forest algorithm was successful with an accuracy rate of 89% in Orange software and 96.709% in WEKA, and it was stated that it would facilitate the decision-making processes of experts in the diagnosis of cancer [13].

E. Aydindag et al. (2022), using 5 of the machine learning algorithms that facilitate early-stage detection of cancer disease, compared 2 different datasets on Kaggle, which algorithm gave the higher success and accuracy rate in which dataset. In the study, Artificial Neural Networks, Support Vector Machines, K-nearest neighbor, Naive Bayes and Decision tree algorithms were used. As a result of the classification process, it was found that the Artificial Neural Networks algorithm gave the most successful results with an accuracy rate of 98.24% on the 1st dataset and 93.85% on the 2nd dataset [14].

3. Data set

In this study, the Breast Cancer Wisconsin dataset prepared for the Diagnosis of Breast Cancer on Kaggle will be used. The data set, which was created by Doctor William H. Wolberg and his friends at the University of Wisconsin in 1995, was obtained as a result of digitizing the images taken from the breast with needle aspirate [5]. There are 569 different samples and 32 different attributes in the data set. Among the attributes, "ID" and "DIAGNOSE" are related to the number of entries and results, while the other 30 attributes are the actual values used in the diagnosis of the disease [5]. There are 212 malignant and 357 benign samples in the data set and are shown in Figure-2. Figure-3 shows the correlation matrix of the data set. There is no unspecified or missing data in the data set. We will divide the dataset into 2 as 70% training and 30% testing.

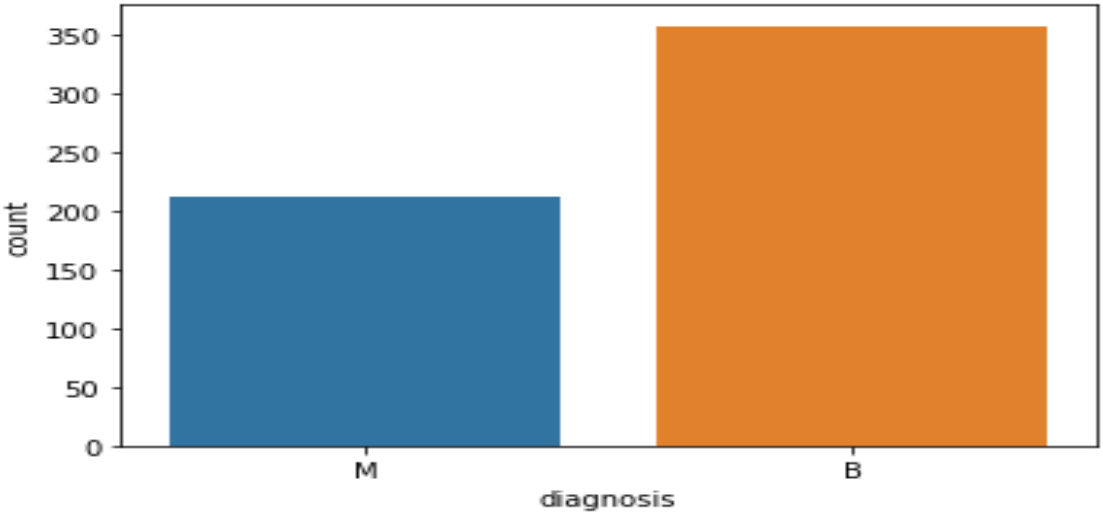

Figure 2. Number of Malicious and Benevolent samples

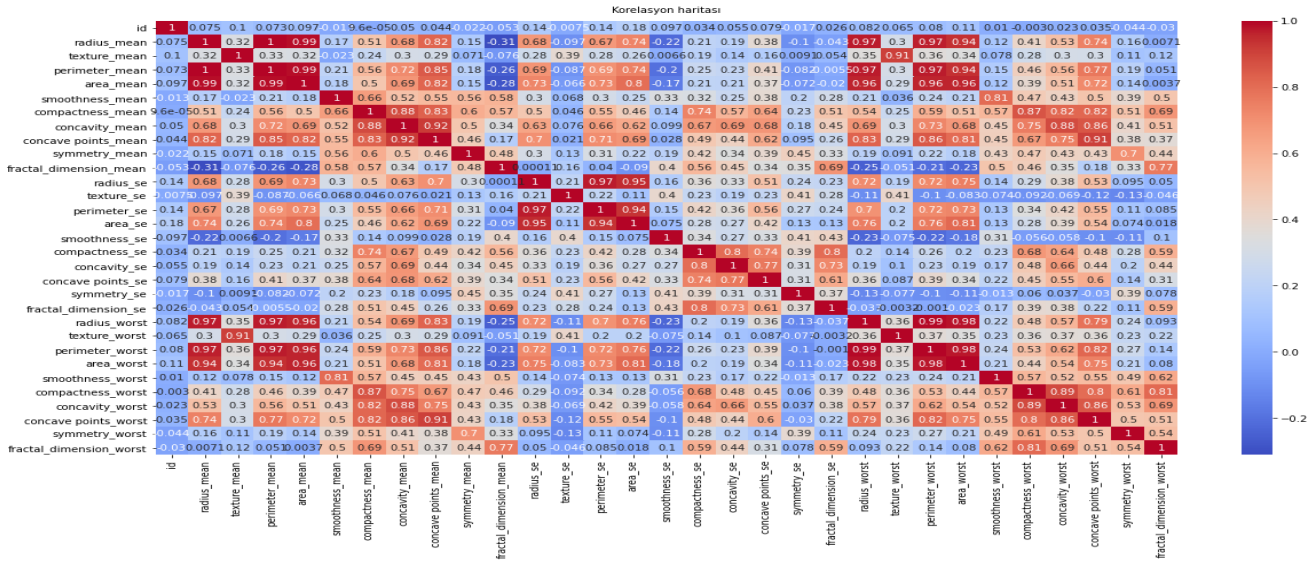

Figure 3. Correlation between features found in the dataset

4. Methods

The study was conducted on the Breast Cancer Wisconsin dataset. The 30 descriptive features in the data set are classified into 2 different categories. In the first category, there are 10 different numerical features belonging to the cell nucleus, and in the second category, there are 20 different numerical values obtained from these 10 measured features. The 9 different features obtained as measurements are [5].

1. Radius
2. Texture
3. Diameter
4. Smoothness
5. Density
6. Concavity
7. Number of concave points
8. Symmetry
9. Fractal dimension

In this study, the success of 5 different machine learning algorithms (Decision Tree, Adaboost, KNN, Random Forest and Logistic Regression), PCA (Principal Component Analysis) using the Python programming language was compared with the success rates obtained without PCA. The effect of PCA on success rates will be examined.

4.1 PCA

PCA (Principal Component Analysis) is a statistical method that helps to reduce the size of datasets by creating a pattern or obtaining a smaller dataset with the data selected from a large dataset [15]. In order to obtain the most suitable line in PCA, the average distances of all points are taken, and the most suitable ones are selected among the perpendicular lines [16]. If we look at it mathematically, it is the problem of finding the Correlation and Covariance values in the data set [17]. It is calculated in 5 steps. These steps can be described as follows.

● Standardization: One of the most important steps, data is standardized because PCA is heavily influenced by the variances of the initial variables. Mathematically, for each variable, it is the difference from the mean divided by the standard deviation.

● Calculation of Matrices: It is used to show the deviations of the variables in the data set from the mean and the relationships between them. The co-variances show us that the relationship between the variables is directly proportional if it is positive, and inversely proportional if it is negative.

● Calculating the Eigenvector and Eigenvalues of the Covariance matrix to determine the principal components: The principal component is the mixture or combinations of the initial variables that PCA generates after reduction. While PCA transfers the maximum number of information to the first variable, it transfers the remaining maximum information to the next variable, thus reducing the size without losing too much information. Every eigenvector has an eigenvalue, and they are even. Eigenvectors are the aspects of our principal components that carry the most information, while the eigenvalues are the coefficient added to the eigenvectors giving the variance value.

● Feature Vector: Selecting the principal components with high eigenvalue from the principal components ordered in order of importance and creating a new vector matrix from the selected ones.

● Arranging data along the axis: It is to reorient the components obtained by multiplying the inverse of the feature vector with the inverse of the original data set.

4.2 Classification techniques

4.2.1 Adaboost

It is a boosting algorithm based on ensemble management technique, which is prepared for use in binary classification processes in machine learning [18-20]. It was formulated in 1995 by Y. Freund-R. Schapire. Adaboost is based on the principle of creating a strong classification algorithm by combining classifier algorithms that appear weak in the classification process. Performing the classification process with a single type of algorithm may adversely affect the result. Adaboost iteratively selects and retrains the previous training set. Each trained algorithm is given a weight score, the scores given are calculated as an alpha (α) value for each classifier, based on the classifier error rate. Calculation of alpha (α) value is shown in Equation 1. The lower the error rate, the higher the alpha.

$$\alpha_t = \frac{1}{2} ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \quad (1)$$

The alpha value found; the weight ratios change after each wrong iteration. These models, whose weights change, are more effective in the next classification process [21]. Weight update is given in Equation 2.

$$D_{t+1}(i) = \frac{D_t(i) exp(-\alpha_t y_i h_t(x_i))}{Z_t} \quad (2)$$

After the iterations are completed, the weak samples are weighted as shown in Equation 3 to form a strong classifier and the final classifier is obtained [21].

$$H(x) = \text{sign}(\sum_{t=1}^{T} a_t h_t(x)) \quad (3)$$

4.2.2 Decision trees

It is a non-parametric, machine learning algorithm used in decision trees, Classification and Regression methods [22-23]. The decision trees algorithm is an inverted tree view, starting from the root and extending to the branches. New branches that multiply after each branch test the attribute from the parent root and have it parse it repeatedly. There are several steps to decide how to divide the data obtained in decision trees into different branches. These steps are briefly described below.

● Gini Pollution: It is an impurity criterion that gives us an idea about the possibility of the model dividing a data incorrectly, allowing us to see the purity of the decision tree, that is, whether it is in the most ideal decomposition. Equation 4 has a mathematical representation of the Gini impurity measure.

$$\text{Gini}=1-\sum_{i=1}^{n} pi^2 \quad (4)$$

● Entropy: It is used to calculate whether the data descending from the root to the branch is homogeneously distributed as a sample [24]. Mathematical representation of Entropy is given in Equation 5. Entropy of zero in a sample indicates a homogeneous division.

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i \quad (5)$$

● Information Gain: Finding out which branch has the highest information gain with a high homogeneous distribution is called information gain. It is related to the decrease of entropy. In Equation 6, the information gain is expressed mathematically.

$$\text{Information Gain} = E(S) - \sum_{vevcalues(\alpha)} \frac{|Sv|}{|s|} E(Sv) \quad (6)$$

4.2.3 Random forest

Basically, it predicts the class on the density of the resulting vector with its neighbors, while estimating about the independent variables [25]. It is an algorithm that can be used in Classification and Regression studies [26]. Calculating nearest points using Euclidean connection works. The result values and estimates are obtained over the K value. The K value usually consists of odd numbers. The k value was taken as 3 in the study.

4.2.4 Logistic regression

It is a regression model whose dependent variables are limited between 0 and 1. In this model, the independent variables show continuity. The consistency of the model is related to the correct determination of the dependent variable [27]. The purpose of logistic regression is to model the relationship between independent variables in the most accurate way.

5. Performance evaluation

In total, in the data set consisting of 212 malignant and 357 benign samples, classification success rates were obtained in 5 different algorithms by distinguishing between 70% and 30% as training and testing. The obtained accuracy rates are also seen in the Complexity matrices given in Figure 4. Table 1 shows the locations of the values that should be found in the Complexity matrix value table. Accuracy, Sensitivity, F1 and Precision values of each algorithm run using the obtained values were calculated.

Table 1. Performance Criteria

Predicted Values	**Real Values**	
	Positive	**Negative**
Positive	TP (True Positive)	FP (False Positive)
Negative	FN (False Negative)	TN (True Negative)

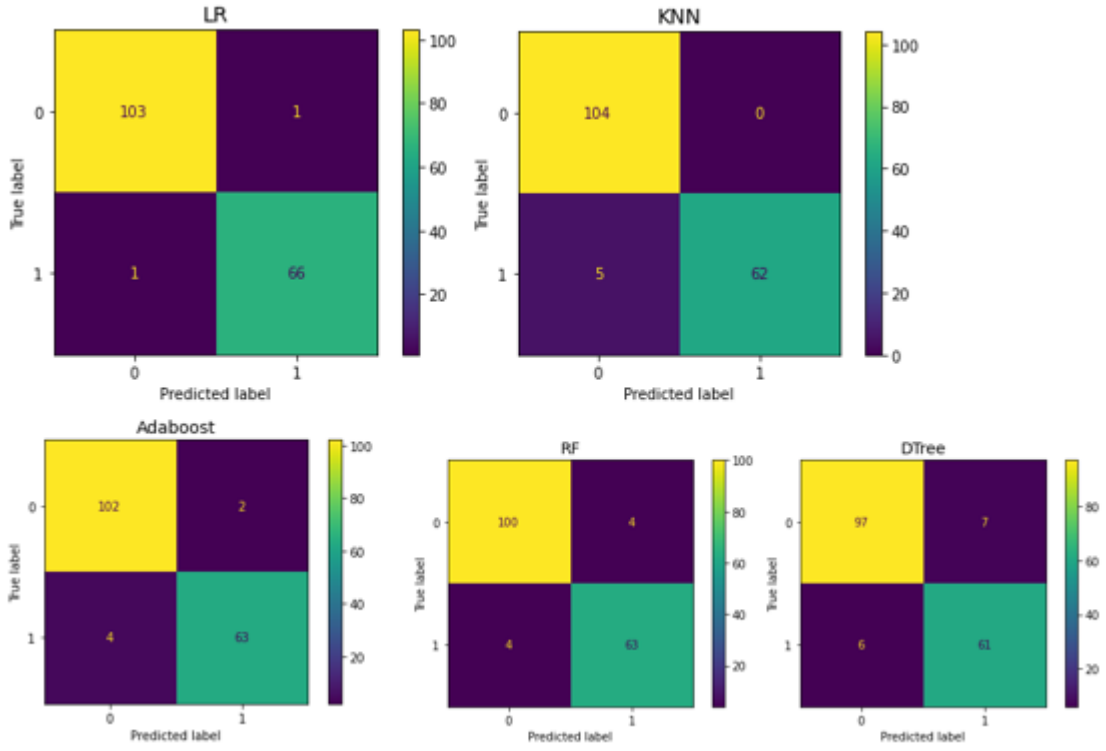

Figure 4. Confusion matrices

Accuracy, Precision, Sensitivity and F1 values are obtained with the performance criteria TP, FP, TN and FN as shown in the equations below.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (9)$$

$$F1 = 2 \text{ x } \frac{\text{Sensibility} * \text{Precision}}{\text{Sensibility} + \text{Precision}} \quad (10)$$

6. Findings and Discussion

569 samples containing breast cancer metrics were first run 5-fold cross validation in Matlab environment, and the Classification algorithm was run without applying PCA on the data set. Table-2 shows the accuracy values, F-1 scores, Sensitivity, and acuity values obtained from the classification algorithms without applying PCA. Table-3 shows the accuracy values, F-1 scores, Sensitivity, and acuity values obtained from the post-PCA classification algorithms. After applying PCA to the data set, the number of samples from 569 was reduced to 171.

Table 2. Classification results without using PCA

Algorithms	TP	FN	TN	FP	Accuracy	F-1	Sensitivity	Precision
Adaboost	348	17	195	9	0,9543	0,962	0,953	0,974
Decision Tree	322	27	185	30	0,891	0,916	0,922	0,914
KNN	355	38	174	2	0,929	0,944	0,903	0,994
Logistic Regression	340	14	198	17	0,945	0,954	0,960	0,952
Random Forest	348	17	195	9	0,954	0,962	0,953	0,974

Table 3. Classification results using PCA

Algorithms	TP	FN	TN	FP	Accuracy	F-1	Sensitivity	Precision
Adaboost	102	4	63	2	0,964	0,970	0,962	0,980
Decision Tree	97	6	61	7	0,923	0,936	0,941	0,932
KNN	104	5	62	0	0,970	0,976	0,954	1,00
Logistic Regression	103	1	66	1	0,988	0,988	0,990	0,990
Random Forest	100	4	63	4	0,953	0,974	0,961	0,990

When the data obtained in Table 2 and Table 3 are compared, it is seen that the accuracy rates in classification algorithms increase after PCA. This increase rate is also clearly seen in the Sensitivity, Sharpness and F-1 score data.

7. Results

Machine learning and artificial intelligence are used actively in many fields, as well as in the field of medicine, they have been increasing their usage rates recently. Breast cancer, which is one of the biggest threats to women in our country and in the world, continues to increase day by day. In this period when we know that early diagnosis is vital and contributes positively to the healing process of the person, machine learning has once again shown us that the diagnosis of breast cancer in people can be diagnosed and treated early with a high degree of accuracy and sensitivity. As a result of the study, it was seen that the addition of PCA to the classification algorithms increased the accuracy performance of the algorithms. As a result of the study, it was revealed that the Logistic Regression algorithm diagnosed the disease with 94.5% accuracy from 171 data. After applying PCA, the success rate of the Logistic Regression algorithm increased to 98.8%.

The ability of machine learning processes to provide experts with the ability to diagnose and make autonomous decisions will have an impact on the acceleration of decision-making processes in many areas, including disease. As can be seen in the study, the success rate of algorithms increases after PCA, and this increasing rate accelerates the decision-making

processes of the experts and allows the patients to be less psychologically worn out in this process. In addition, early diagnosis of cancer with machine learning will reduce health system expenses.

Acknowledgement

This article study was carried out in Siirt University Engineering Faculty Human Computer Interaction Laboratory. I would like to thank the Human Computer Interaction Laboratory staff for their support.

References

[1] Choi, Y.K., Woo, S.M., Cho, S.G., Moon, H.E., Yun, Y.J., Kim, J.W., Ko, S.G., "Brain-metastatic triple-negative breast cancer cells regain growth ability by altering gene expression patterns", Cancer Genomics & Proteomics 10(6) (2013) : 265-275.

[2] Waks, A.G., Winer, E.P., "Breast cancer treatment: a review", Jama 321(3) (2019) : 288-300.

[3] Yancik, R., Ries, L.A., "Aging and cancer in America: demographic and epidemiologic perspectives", Hematology/oncology clinics of North America 14(1) (2000) : 17-23.

[4] Goldstein, A.J., Harmon, L.D., Lesk, A.B., "Identification of human faces", Proceedings of the IEEE 59(5) (1971) : 748-760.

[5] Agarap, A.F.M., "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset", In Proceedings of the 2nd international conference on machine learning and soft computing (2018) : 5-9.

[6] Toğaçar, M., Ergen, B., "Deep learning approach for classification of breast cancer", In 2018 International Conference on Artificial Intelligence and Data Processing (IDAP) (2018) : 1-5. IEEE.

[7] Yavuz, E., Eyüpoğlu, C., "Meme Kanseri Teşhisi İçin Yeni Bir Skor Füzyon Yaklaşımı", Düzce Üniversitesi Bilim ve Teknoloji Dergisi 7(3) (2019) : 1045-1060.

[8] Bayrak, E.A., Kırcı, P., Ensari, T., Seven, E., "Dağtekin, M., Göğüs Kanseri Verileri Üzerinde Makine Öğrenmesi Yöntemlerinin Uygulanması", Journal of Intelligent Systems: Theory and Applications 5(1) (2022) : 35-41.

[9] Ganggayah, M.D., Taib, N.A., Har, Y.C., Lio, P., Dhillon, S.K., "Predicting factors for survival of breast cancer patients using machine learning techniques", BMC medical informatics and decision making 19(1) (2019) : 1-17.

[10] Singh, S., Jangir, S.K., Kumar, M., Verma, M., Kumar, S., Walia, T.S., Kamal, S.M., "Feature Importance Score-Based Functional Link Artificial Neural Networks for Breast Cancer Classification", BioMed Research International (2022) : 1-8.

[11] Ghosh, P., "Breast Cancer Wisconsin (Diagnostic) Prediction", Available online: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic) (accessed on 1 October 2022).

[12] Mangukiya, M., Vaghani, A., Savani, M., "Breast Cancer Detection with Machine Learning", International Journal for Research in Applied Science and Engineering Technology 10(2) (2022) : 141-145.

[13] Argun, İ.D., Nalbant, B., "Using Classification Algorithms in Data Mining in Diagnosing Breast Cancer", Advances in Artificial Intelligence Research 2(2) (2022) : 65-70.
[14] Bayrak, E.A., Kırcı, P., Ensari, T., Seven, E., Dağtekin, M., "Göğüs Kanseri Verileri Üzerinde Makine Öğrenmesi Yöntemlerinin Uygulanması", Journal of Intelligent Systems: Theory and Applications 5(1) (2022) : 35-41.
[15] Jolliffe, I.T., Cadima, J., "Principal component analysis: a review and recent developments", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374(2065) (2016) : 20150202.
[16] Ringnér, M., "What is principal component analysis?", Nature biotechnology 26(3) (2008) : 303-304.
[17] Ding, C., Zhou, D., He, X., Zha, H., "R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization", In Proceedings of the 23rd international conference on Machine learning (2006) : 281-288.
[18] Schapire, R.E., "Explaining adaboost", In Empirical inference (2013) : 37-52. Springer, Berlin, Heidelberg.
[19] Wang, R., "AdaBoost for feature selection, classification, and its relation with SVM, a review", Physics Procedia 25 (2012) : 800-807.
[20] Gao, L., Cheng, W., Zhang, J., Wang, J., "EEG classification for motor imagery and resting state in BCI applications using multi-class Adaboost extreme learning machine", Review of scientific instruments 87(8) (2016) : 085110.
[21] Quinlan, J.R., "Learning decision tree classifiers", ACM Computing Surveys (CSUR) 28(1) (1996) : 71-72.
[22] Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., "An introduction to decision tree modeling", Journal of Chemometrics: A Journal of the Chemometrics Society 18(6) (2004) : 275-285.
[23] Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K., "KNN model-based approach in classification", In OTM Confederated International Conferences On the Move to Meaningful Internet Systems (2003) : 986-996. Springer, Berlin, Heidelberg.
[24] Sha'Abani, M.N.A.H., Fuad, N., Jamal, N., Ismail, M.F., "kNN and SVM classification for EEG: a review", InECCE2019 (2020) : 555-565.
[25] Biau, G., Scornet, E., "A random forest guided tour", Test 25(2) (2016) : 197-227.
[26] More, A.S., Rana, D.P., "Review of random forest classification techniques to resolve data imbalance", In 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM) (2017) : 72-78. IEEE.
[27] Wright, R.E., "Logistic regression", (1995) : 217-244.