Turkish Journal of Science & Technology
16(2), 251-260, 2021

Research Paper

Analysis of Mutated RNA-Type Breast Cancer Data with Machine Learning Methods

Rumeysa Hanife KARS[1*]
[1] Biomedical Engineering, School of Engineering and Natural Sciences, Istanbul Medipol University, Istanbul, Turkey
[*]rhkars@st.medipol.edu.tr

(Geliş/Received: 13/07/2021; Kabul/Accepted: 23/08/2021)

Abstract: According to the data for the year 2020, the three most common types of cancer in women are; breast, lung, and colorectal. These types of cancer make up 50% of other types of cancer seen in women. In addition, only breast cancer accounts for 30% of cancer types in women. Early diagnosis and treatment processes of breast cancer patients are important and the correct application of this process increases the survival rate of the patients. Artificial intelligence can contribute to the observational performance of radiologists in breast cancer screening. On the other hand, artificial intelligence-based approaches can also be used to increase the accuracy of digital mammography. The dataset used in this study consists of mutated RNA-type breast cancer data. The dataset includes the clinical and genetic characteristics of the patients. In the approach of the study, it is suggested to use various machine learning methods together. Support Vector Machines method has been decided the best performance with 97.55% in the analyzes performed. It has been observed that the recommended approach in the diagnosis of breast cancer gave successful results.
Key words: Artificial intelligence, Breast cancer, Machine learning, Mutation RNA

Mutasyona Uğramış RNA tipli Göğüs Kanseri Verilerinin Makine Öğrenme Yöntemleri ile Analizi

Öz: Kadınlarda görülen en yaygın üç kanser türü 2020 yılı verilerine göre; göğüs, akciğer ve kolorektaldır. Bu kanser türleri kadınlarda görülen diğer kanser türleri arasında %50'sini oluşturmaktadır. Ayrıca, kadınlarda görülen kanser türleri arasında yalnızca göğüs kanseri %30'unu oluşturmaktadır. Göğüs kanseri hastalarının, erken tanı ve tedavi süreçleri önemlidir ve bu sürecin doğru uygulanması hastaların hayatta kalma oranlarını artırır. Yapay zekâ, radyologların göğüs kanseri taramasındaki gözlemleme performanslarına katkı sağlayabilir. Öte yandan yapay zekâ tabanlı yaklaşımlar, dijital mamografinin doğruluğunu artırmak için de kullanılabilir. Bu çalışmada kullanılan veri kümesi mutasyona uğramış RNA tipi göğüs kanseri verilerinden oluşur. Veri kümesinde hastaların klinik özellikleri ile genetik özellikleri yer alır. Çalışmanın yaklaşımında çeşitli makine öğrenimi yöntemlerinin bir arada kullanılması önerildi. Gerçekleştirilen analizlerde en iyi performansı %97,55 oranında Destek Vektör Makineleri yöntemi verdi. Göğüs kanseri tanısında önerilen yaklaşımın başarılı sonuçlar verdiği gözlemlendi.

Anahtar kelimeler: Göğüs kanseri, Makine öğrenme, Mutasyon RNA, Yapay zekâ

1. Introduction

It is estimated that approximately 1.9 million people will have cancer in 2021 and it is predicted that approximately 608 thousand people will die from cancer patients. Today, breast cancer is the most common type of cancer detected in women, and it is the first cause of death due to cancer in such women. In addition, women living in the USA have an average of 13% risk of developing breast cancer [1]. Once breast cancer is detected early, the chances of survival will be very high. Specialists use certain tests to understand or cure breast cancer. Tests are used to find out if the cancer has spread to a body. Radiologists, on the other hand, are more active in mammography. Conditions such as these tests and treatment can take a lot of time and increase the workload of the doctor and radiologist. [2].

The use of Machine Learning (ML) and Deep Learning (DL) as tools in medical diagnosis has become a very useful application in breast cancer diagnosis [2]. ML methods can be used as an aid to radiologists struggling to improve cell detection performance in a cancer. Several new methods, including ML and DL, have been developed and applied to digital mammography. Preliminary research has shown that the aid of artificial intelligence systems to predict mammograms can increase the productivity of the radiologist in terms of time, specificity, and precision. These developed algorithms and tools have made the work of radiologists and doctors easier and reduced their workload [2].

Many studies have been carried out in the literature using the ML and DL approaches. In one study, they observed that the appropriate use of an AI tool increases the diagnostic efficiency of radiologists [3]. In another

[*] Corresponding author: rhkars@st.medipol.edu.tr. ORCID Number of authors: [1] 0000-0002-2865-0414

study, they developed an artificial intelligence system that outperformed radiologists in the clinical processes of breast cancer detection [4]. They observed that the AI system produced more accurate and earlier results against scanning protocols. In another study, they observed a significant difference between the analyzes of radiologists and the analysis of the data obtained by the artificial intelligence-based system they developed. With their proposed approach, they reduced the heavy workload of radiologists and achieved successful results in the detection of cancer [5]. Dembrower et al. performed breast cancer classification on their mammograms with a cancer detector using artificial intelligence. They used machine learning methods in their study. Their proposed approach is able to interpret without radiologist evaluation [6]. The aim of this study is to obtain statistical results by analyzing features containing mutated RNA type breast cancer data and to achieve general success with artificial intelligence-based machine learning methods using this dataset.

2. Materials and Machine Learning Methods

2.1. Dataset

The dataset includes targeted sequencing data of 1,980 primary breast cancer samples. In addition, clinical features, m-RNA levels, z-score and gene mutations were examined for 1904 patients. The dataset contains 693 trait parameters for each patient, and their label groups have one of the values {text}, {number}, {positive, negative}, {0,1}, {0,1,2,3,4} In addition, the "Status" parameter of breast cancer was used in the experimental analyses. There are two label values in the status parameter; "positive" and "negative". If the patient's status parameter is "positive", the patient is breast cancer, otherwise the patient is not breast cancer [7].

Table 1. Parameters and label value of the patients that make up the dataset

The number of the feature	Original parameters available in the dataset	Label Value Type
1	Patient id	{numerical}
2	Age at diagnosis	{numerical}
3	Type of breast surgery	{text}
4	Cancer type	{text}
5	Cancer type detailed	{text}
6	Cellularity	{text}
7	Chemotherapy	{0,1}
8	Pam50 + claudin-low subtype	{text}
9	Cohort	{numerical}
10	Er status measured by ihc	{positive, negative}
11	Er status	{ positive, negative }
12	Neoplasm histologic grade	{numerical}
13	Her2 status measured by snp6	{text}
16	Hormone therapy	{0,1}
24	Mutation count	{numerical}
25	Overall survival months	{numerical}
30	Tumor size	{numerical}
31	Tumor stage	{0,1,2,3,4}
32	Death from cancer	{0,1}
33-693	Genetic features	{numerical, text}

The breast cancer status and count information used in the experimental analyzes are shown in Figure 1. Breast cancer status of 1459 patients is labeled as "positive" and status information of 445 patients is labeled as

"negative". In addition, in all experimental analyzes of the dataset, 70% is reserved as training data and 30% is reserved as test data.

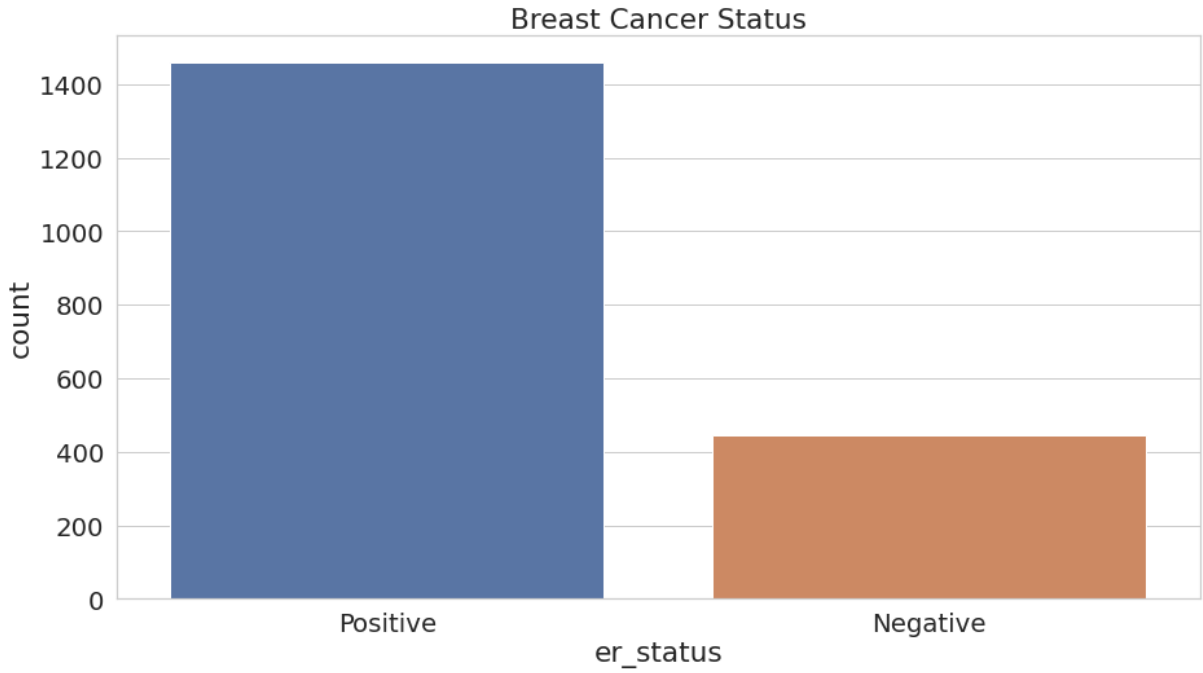

Figure 1. Breast cancer and census statistical information in the dataset.

2.2 Support Vector Machines Method

The Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The states that define the hyperplane are support vectors. The main purpose of SVM is to find the optimal hyperplane that linearly separates the data points in the two components by maximizing the margin [8]. It is still effective when the number of dimensions is greater than the number of samples. Another advantage of SVM is that there is a unique global minimum value if the data can be linearly separable. An ideal SVM analysis should produce a hyperplane that completely separates the vectors into two non-overlapping classes. However, perfect separation may not be possible or may result in a model in so many cases that the model cannot be correctly classified. In this case, the SVM finds the hyperplane that maximizes the margin and minimizes misclassifications. Hyper plane is calculated according to Eq. (1) and distance measurement between plane vectors is calculated according to Eq. (2) [9]. When the equations are examined; w is the vector in the hyperplane, b is the preset variable, x is the input vector, and T is the threshold value. The general design showing the operation of two-class data with the SVM method is shown in Figure 2.

$$H: wT(x) + b = 0 \tag{1}$$

$$\vee w \vee^2 = \sqrt{w_1^2 + w_2^2 + w_3^2 + \ldots w_n^2} \tag{2}$$

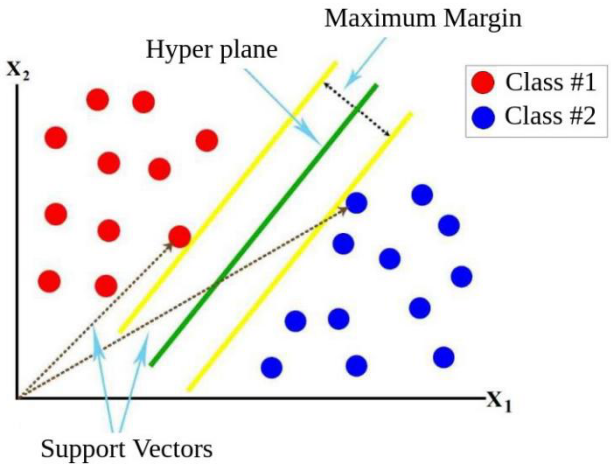

Figure 2. Classification process with SVM method [10].

2.3. Extreme Gradient Boosting Method

Extreme Gradient Boosting (EGB) is an open source approach that enables efficient and effective implementation of gradient boosting algorithm. This approach has led the applied machine learning community to move towards gradient boosting more generally. The EGB method is among the methods used to win solutions for classification and regression problems in machine learning competitions. In the AGA method, the gain parameter is used and it contributes to the transaction performance with this parameter. Gain parameter is calculated according to Eq. (3) and provides convenience in the partitioning process with an approach similar to the tree-root relationship in the classification process. The AGA method scales beyond billions of instances using far fewer resources than existing systems. The process of estimating the output value of the tree structure created in this method is calculated according to Eq. (4) [10].

$$Gain = Gain\ left + Gain\ right - root\ similarity \quad (3)$$

$$New\ prediction = First\ predicted\ value + Learning\ rate\ x\ output\ value \quad (4)$$

2.4. Gradient Boosting Method

Gradient Boosting (GB) can produce a prediction model, typically in the form of a collection of weak prediction models such as decision trees; preferred machine learning method for solving classification and regression problems. This method is based on performing the processing steps by combining the best possible model approach with the previous model approaches, thus minimizing the overall estimation error. The outcome for each sample input in the dataset depends on how much the change in the prediction affects the overall prediction error. The steps to be applied in my classification process in the GB method are as follows;

- Placing a simple linear regression or decision tree on the data,
- Calculating error residuals according to Eq. (5),
- Placing a new model on the error residues as the target variable with the same input variables,
- Adding predicted residuals to previous forecasts,
- Placing another model on the remaining residues [11].

$$Error = True\ target\ value - predicted\ target\ value \quad (5)$$

2.5. Decision Tree Method

Decision tree (DT) is the popular machine learning method of choice for solving classification and prediction problems. DT is a flowchart-like tree structure where each internal node represents a test on an attribute, each branch represents a result of the test, and each leaf node holds a class label. The overall design of the KA is shown in Figure 3. Decision trees can handle both continuous and categorical variables. This method provides a clear indication of which areas are most important for forecasting or classification. As with all analytical methods, the decision tree method has limitations that users should be aware of. Its main disadvantage is that it may not work well, especially when using a small dataset. This problem may limit the generalizability and robustness of the emerging models [13].

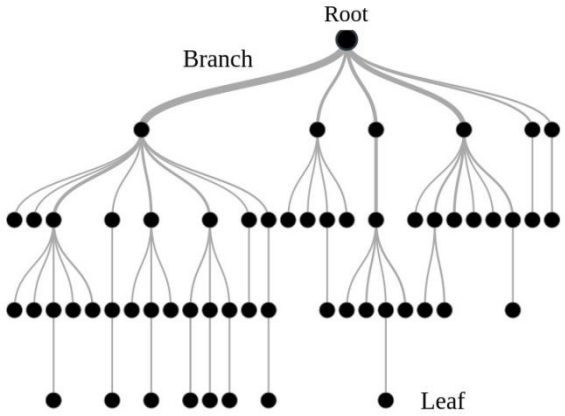

Figure 3. The general view of DT method [14].

2.6. Random Forest Method

Random Forest (RF) is a machine learning method that contains more than one decision tree in its algorithm structure. The overall design of the RF method is shown in Figure 4. To classify a new object from an input vector, insert the input vector into each of the trees in the forest. Each tree in the RF approach performs a classification and the result is calculated by performing voting. In the RF method, each input data chooses the classification with the most votes. This method generally works more efficiently on large datasets. The margin of error is important in calculating the efficiency, and in the RF method, the mean square error (MSE) formula given in Eq. (6) is used. In this equation; ,the variable N represents the number of data points, the variable fi represents the value returned by the method. The actual value for each i in the data point is represented by the variable yi. It also makes predictions about which variables are important in the classification, and the forests created are recorded for use in other data entries [15].

$$\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{f_i} - \mathrm{y_i})^2 \tag{6}$$

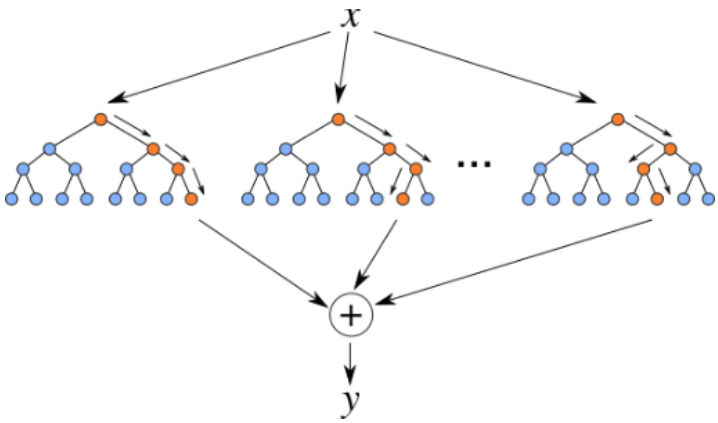

Figure 4. General representation of RT method [16].

3. Proposed Approach

The proposed method was used to determine the characteristics such as cancer status, age determination, cancer stage of breast cancer patients. The dataset contains mutated RNA type breast cancer data. To conclude the correlation between the proposed approach and the features in the mutated dataset with graphical analyzes and using machine learning methods, it is determined whether the patient has cancer or not. The overall design of the proposed approach is shown in Figure 5. Performance comparisons are made using SVM, EGB, GB, DT and RF methods, and the most appropriate method is determined to determine the patient's cancer status (Positive: Cancer-containing, Negative: Cancer-free).

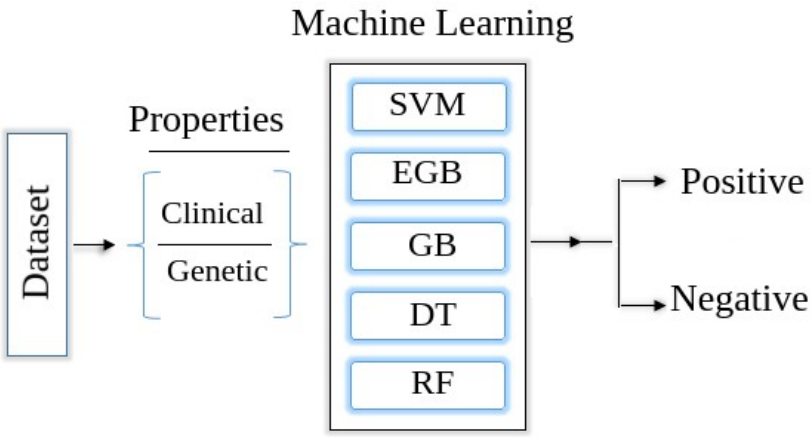

Figure 5. Design of the proposed approach.

4. Results

The analyzes of this study were carried out using the Google Colaboratory server installed on the Ubuntu-Linux operating system [17,18]. Python software language libraries (Sklearn, Numpy, Pandas) [19] were used for statistical analysis and Jupyter Notebook interface was used for compiling the codes. The confusion matrix was preferred as a criterion in the evaluation of breast cancer. The metric parameters of the confusion matrix are; accuracy, specificity, sensitivity and f-score. The formulas between Eq. (7) and Eq. (10) are used to calculate these metric parameters. For the validity and reliability of the analyzes, it is necessary to look at the True(T) and False(F) of positive (P), negative (N) decisions. The equations given below perform these operations [18,20].

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (8)$$

$$\text{f score} = \frac{2xTP}{2xTP+FN+FP} \quad (9)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (10)$$

Experimental analyzes were carried out in two parts. In the first analysis, the statistical information of the dataset was extracted and some clinical-genetic features were used for this. The open source codes were compiled using the "seaborn" library and the results obtained in the first analysis are shown in Figure 6. Figure 6 shows the four trait parameters and their correlation with patient conditions. The incidence of breast cancer increases with age. Mutation amounts were observed to be more effective in patients with positive cancer status. Tumor size was not observed to be affected by positive and negative cases. Finally, tumor stages have been observed to be effective in positive cases.

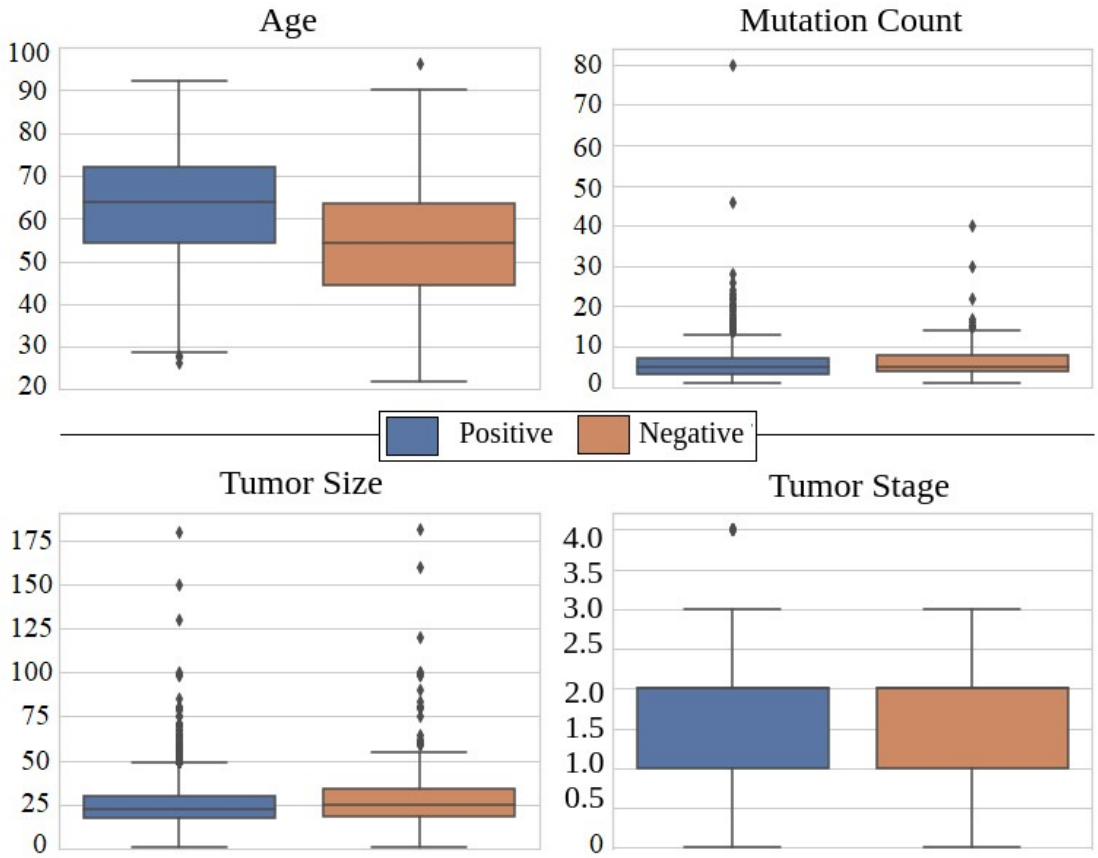

Figure 6. Statistical results of the dataset using some clinical-genetic features.

In the second analysis of the experiment, machine learning methods (SVM, EGB, GB, DT and RF) were used to determine whether the patients had breast cancer or not. The open source codes of the "Sklearn" library were used in machine learning methods and the preferred parameters in this study are the default values. The confusion matrix graphs obtained as a result of the second analysis are shown in Figure 7. Accordingly, 97.55% with the SVM method; 96.32% with EGB and GB methods; An accuracy rate of 92.65% was obtained with the DT method and 94.75% with the RF method. The SVM method provided the best performance in the recommended approach. With the SVM method, the sensitivity of the data with negative disease status was 93.83%; specificity success was 98.82% and f-score success was 95.13%. With the SVM method, the sensitivity of the data with positive disease status was 98.82%; specificity success of 93.83%; f-score success was 98.36%. The success information of the metric parameters obtained from the confusion matrices are given in Table 2.

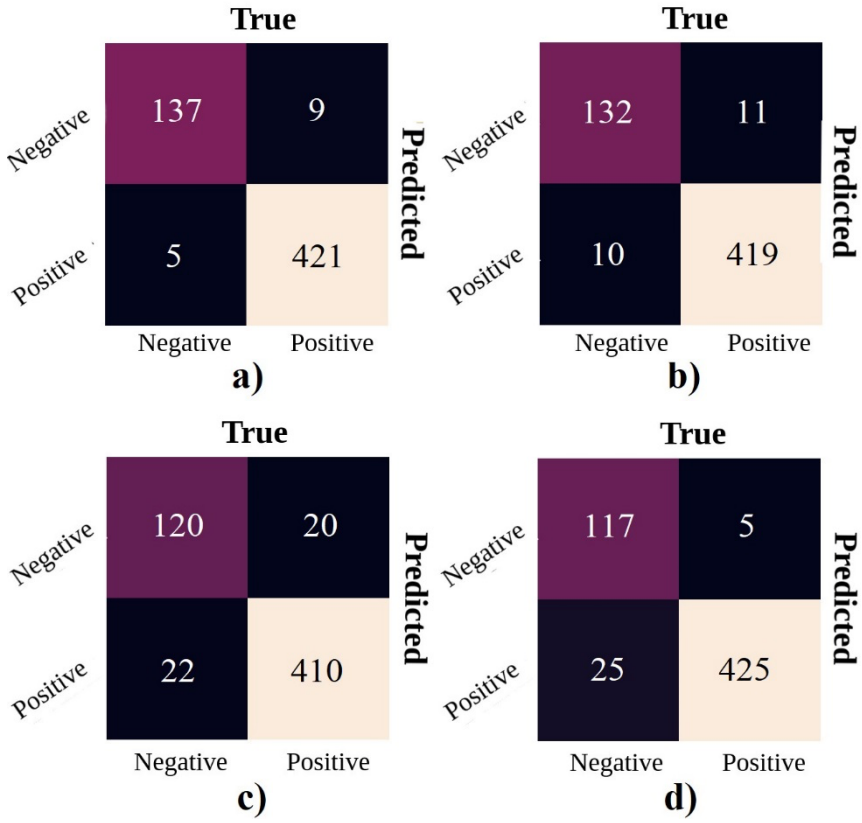

Figure 7. In the proposed approach, confusion matrices obtained by machine learning methods; a) SVM, b) EGB and GB, c) DT, d) RF.

Table 2. Analysis from confusion matrices of breast cancer (%).

Method	Class	Sensitivity	Specificity	f-score	Accuracy
SVM	Negative	93,83	98,82	95,13	97,55
	Positive	98,82	93,83	98,36	
EGB	Negative	92,30	97,66	92,63	96,32
	Positive	97,66	92,30	97,55	
GB	Negative	92,30	97,66	92,63	96,32
	Positive	97,66	92,30	97,55	
DT	Negative	85,71	94,90	85,10	92,65
	Positive	94,90	85,71	95,12	
RF	Negative	95,90	94,44	88,63	94,75
	Positive	94,44	95,90	96,59	

5. Discussion

Breast cancer is a disease that is frequently seen among women and causes an increase in the number of cases worldwide every year. To benefit from technological stakeholders in the diagnosis of this disease; It is important to increase the rate of accurate and fast decision making in early diagnosis processes. In this study, analyzes were performed using mutated breast cancer data. The use of various machine learning methods in the proposed

approach has given us an advantage over the analysis of datasets with different characteristics. Thus, analyzes were carried out without relying on a single machine learning method. I achieved the best performance with the SVM method in mutated breast cancer data. The disadvantage of the proposed approach is that it does not have an end-to-end algorithm. The analyzes of different studies that performed analyzes using the dataset are given in Table 3.

Table 3. Results of studies using the same dataset.

Study	**Year**	**Method**	**Accuracy** (%)
Felonneau et al. [21]	2018	Radial Basis Function (RBF)	66,20
Proposed Method	2021	SVM	97,55

In one study, they performed analyzes using two methods. These methods are; It was RBF and SVM. They also used different datasets in their study and presented a pipeline tool for easier analyzes across diseases on other datasets [21]. The analysis results were more productive than the work of Felonneau et al. In order for them to increase their classification success, they should have used various machine learning methods in experimental analysis. In addition, it cannot be said that the parameter values and optimization methods they use affect the classification results positively. The SVM method used in our study provided convenience for data that are not regularly distributed and whose distribution is unknown. As a result of this situation, an accuracy rate of 97.55% was achieved.

In this article, higher rate of accuracy has been achieved when it is compared to the work of Felonneau et al. bu utilizing SVM method. Machine learning methods have contributed to this kind of success. Importantly, SVM method gave the highest rate of success after investigating the dataset.

6. Conclusion

According to studies, the most common type of cancer in women is breast cancer. Thousands of women die every year because of this situation. In order to prevent this situation and to facilitate the work of doctors and radiologists, various artificial intelligence applications are being developed. Recently, interest in the classification of cancer datasets with deep learning methods has increased. In this study, mutated RNA type breast cancer data were examined. For this purpose, five different machine learning models were used to evaluate whether the patient had breast cancer, and criteria such as age, number of mutations, and tumor stage were examined. According to the results obtained, the SVM machine learning method came to the fore by showing the highest performance. Using this method, a success rate of 97.55% is achieved. In this study, the availability of statistical results by data analysis and the general success of general success with artificial intelligence-based machine learning methods were investigated using this dataset.

References

[1] R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer Statistics, 2021, CA: A Cancer Journal for Clinicians. 71 (2021) 7–33. https://doi.org/10.3322/caac.21654.

[2] A. Aloraini, Different Machine Learning Algorithms for Breast Cancer Diagnosis, International Journal of Artificial Intelligence & Applications. 3 (2012) 21–30. https://doi.org/10.5121/ijaia.2012.3603.

[3] S. Pacilè, J. Lopez, P. Chone, T. Bertinotti, J.M. Grouin, P. Fillard, Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool, Radiology: Artificial Intelligence. 2 (2020) e190208. https://doi.org/10.1148/ryai.2020190208.

[4] S.M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, and et al. International evaluation of an AI system for breast cancer screening, Nature. 577 (2020) 89–94. https://doi.org/10.1038/s41586-019-1799-6.

[5] I. Sechopoulos, J. Teuwen, R. Mann, Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art, Seminars in Cancer Biology. (2020). https://doi.org/10.1016/j.semcancer.2020.06.002.

[6] K. Dembrower, E. Wåhlin, Y. Liu, M. Salim, K. Smith, P. Lindholm, M. Eklund, F. Strand, Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study, The Lancet Digital Health. 2 (2020) e468–e474. https://doi.org/10.1016/S2589-7500(20)30185-0.

[7] B. Pereira, S.F. Chin, O.M. Rueda, H.K.M. Vollan, and et al., The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes, Nature Communications. 7 (2016) 1–16. https://doi.org/10.1038/ncomms11479.

[8] M. Marjanović, M. Kovačević, B. Bajat, V. Voženílek, Landslide susceptibility assessment using SVM machine learning algorithm, Engineering Geology. 123 (2011) 225–234. https://doi.org/10.1016/j.enggeo.2011.09.006.

[9] T. Shon, Y. Kim, C. Lee, J. Moon, A machine learning framework for network anomaly detection using SVM and GA, in: Proceedings from the 6th Annual IEEE System, Man and Cybernetics Information Assurance Workshop, SMC 2005, 2005: pp. 176–183. https://doi.org/10.1109/IAW.2005.1495950.

[10] Support Vector Machine Machine learning algorithm with example and code - Codershood, (n.d.). https://www.codershood.info/2019/01/10/support-vector-machine-machine-learning-algorithm-with-example-and-code/ (accessed February 17, 2021).

[11] A. Ibrahem Ahmed Osman, A. Najah Ahmed, M.F. Chow, Y. Feng Huang, A. El-Shafie, Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia, Ain Shams Engineering Journal. (2021). https://doi.org/10.1016/j.asej.2020.11.011.

[12] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Frontiers in Neurorobotics. 7 (2013). https://doi.org/10.3389/fnbot.2013.00021.

[13] Y.Y. Song, Y. Lu, Decision tree methods: applications for classification and prediction, Shanghai Archives of Psychiatry. 27 (2015) 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044.

[14] An Introduction to Machine Learning, (n.d.). https://bioinformatics-training.github.io/intro-machine-learning-2017/decision-trees.html (accessed February 18, 2021).

[15] K. Fawagreh, M.M. Gaber, E. Elyan, Random forests: from early developments to recent advancements, Systems Science & Control Engineering. 2 (2014) 602–609. https://doi.org/10.1080/21642583.2014.956265.

[16] Random Forest Regression. Random Forest Regression is a… | by Chaya Bakshi | Level Up Coding, (n.d.). https://levelup.gitconnected.com/random-forest-regression-209c0f354c84 (accessed February 18, 2021).

[17] T. Carneiro, R.V.M. da Nobrega, T. Nepomuceno, G. bin Bian, V.H.C. de Albuquerque, P.P.R. Filho, Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications, IEEE Access. 6 (2018) 61677–61685. https://doi.org/10.1109/ACCESS.2018.2874767.

[18] S. Walker, W. Khan, K. Katic, W. Maassen, W. Zeiler, Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings, Energy and Buildings. 209 (2020) 109705. https://doi.org/10.1016/j.enbuild.2019.109705.

[19] GitHub - scikit-learn-contrib/sklearn-pandas: Pandas integration with sklearn, (n.d.). https://github.com/scikit-learn-contrib/sklearn-pandas (accessed February 21, 2021).

[20] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics. 21 (2020) 1–13. https://doi.org/10.1186/s12864-019-6413-7.

[21] G. Dubourg-Felonneau, T. Cannings, F. Cotter, H. Thompson, N. Patel, J.W. Cassidy, H.W. Clifford, A Framework for Implementing Machine Learning on Omics Data, ArXiv. (2018) 1–5.