

# The Prediction of Breast Biopsy Outcomes Using Two Data Mining Algorithms Based on Parameter Variations

## Meme Biyopsisi Sonuçlarının İki Veri Madenciliği Algoritması Kullanılarak Parametre Varyasyonlarına Dayalı Tahmin Edilmesi

- Batuhan BAKIRARAR<sup>a</sup>,
- İrem KAR<sup>a</sup>,
- Derya GÖKMEN<sup>a</sup>,
- Atilla Halil ELHAN<sup>a</sup>,
- Volkan GENÇ<sup>b</sup>

Departments of  
<sup>a</sup>Biostatistics,  
<sup>b</sup>Surgery,  
 Ankara University Faculty of Medicine,  
 Ankara, TURKEY

Received: 07.01.2019  
 Received in revised form: 15.03.2019  
 Accepted: 16.03.2019  
 Available Online: 22.03.2019

Correspondence:  
 Batuhan BAKIRARAR  
 Ankara University Faculty of Medicine,  
 Department of Biostatistics, Ankara,  
 TURKEY/TÜRKİYE  
 batuhan\_bakirarar@hotmail.com

**ABSTRACT Objective:** Breast cancer is the type of cancer that causes the most death in women in the United States after lung cancer. The objective of this study was to predict breast biopsy results using age, BI-RADS, Mass Shape, Mass Margin, Mass Density by Multilayer Perceptron and Random Forest algorithms. **Material and Methods:** The dataset contains a BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth for 516 benign and 445 malignant masses. WEKA software was used for data mining. The data mining methods of the Multilayer Perceptron and Random Forest were used to predict the severity of cancer. **Results:** Accuracy, F-measure and Root Mean Squared Error values of the Multilayer Perceptron model were found 0.837, 0.833 and 0.352, respectively while accuracy, F-measure and Root Mean Squared Error values of the Random Forest model were found 0.816, 0.814 and 0.396, respectively. The Multilayer Perceptron method provided a better prediction of breast cancer diagnosis than the Random Forest method and a software was developed based on the best model created by the Multilayer Perceptron method. **Conclusion:** Consequently, a model that was built with the MLP method can be used as an alternative in the diagnosis of patients and be an assistant tool for physicians.

**Keywords:** Breast cancer; data mining; multilayer perceptron; prediction; random forest

**ÖZET Amaç:** Meme kanseri, ABD'de akciğer kanserinden sonra kadınlarda en fazla ölüme neden olan kanser türüdür. Bu çalışmanın amacı yaş, BI-RADS, Kitle Şekli, Kitle Marjı ve Kitle Yoğunluğu değişkenleri yardımıyla Çok Katmanlı Algılayıcı ve Random Forest algoritmaları kullanılarak meme biyopsi sonuçlarını tahmin etmektir. **Gereç ve Yöntemler:** Veri seti 516 iyi huylu ve 445 kötü huylu kitleye yönelik kesin referansla beraber, BI-RADS değerlendirmesi, hastanın yaşı ve üç BI-RADS özelliği değişkenlerinden oluşmaktadır. Veri madenciliği için WEKA yazılımı kullanılmıştır. MLP ve Random Forest veri madenciliği yöntemleri kanser sonuçlarını tahmin etmek için kullanılmıştır. **Bulgular:** Çok Katmanlı Algılayıcı modelinin doğruluk, F-ölçütü ve ortalama hata kareleri kare kökü değerleri sırasıyla 0.837, 0.833 ve 0.352 bulunurken, Random Forest modelinin doğruluk, F-ölçütü ve ortalama hata kareleri kare kökü değerleri sırasıyla 0.816, 0.814 ve 0.396 olarak bulunmuştur. Çok Katmanlı Algılayıcı yöntemi, meme kanseri teşhisi için Random Forest yönteminden daha iyi bir tahmin sağlamıştır, bu yüzden Çok Katmanlı Algılayıcı yöntemi kullanılarak oluşturulan en iyi modele dayalı bir yazılım geliştirilmiştir. **Sonuç:** Sonuç olarak, Çok Katmanlı Algılayıcı yöntemiyle yapılan bir model hastaların teşhisinde alternatif olarak kullanılabilir ve hekimler için yardımcı bir araç olabilir.

**Anahtar Kelimeler:** Meme kanseri; veri madenciliği; çok katmanlı algılayıcı; tahmin; random forest

**B**reast cancer is the most common cancer among women. According to the American Cancer Society, it is one of the three most diagnosed cancer in women in the United States. It is the most deadly cancer type in women after lung cancer in the United States.<sup>1</sup> In Turkey, it is also the most common, with 41.6/100,000 incidence.<sup>2</sup> Therefore, early diagnosis of breast cancer has a great importance. Mammography and biopsy (if necessary) are gold standard methods in detecting breast cancer. As indicated by population-based assessments in both randomized trials and screening mammography, early detection of breast cancer by mammography significantly increases survival chances.<sup>3,4</sup>

Although early diagnosis and breast biopsy significantly improve the chances of survival in breast cancer, the high number of unnecessary breast biopsies leads to severe mental and physical discomfort among patients and additional costs for examinations. In the last several years, it has been proposed to utilize computer aided diagnosis systems using lesion descriptions based on the Breast Imaging Reporting and Dated System™ (BI-RADS™) standard lexicon as input attributes so that physicians can use it to support their decision to carry out a breast biopsy or a short follow-up diagnosis on a suspicious region seen in a mammogram.

Today, BI-RADS is the instrument of communication in mammography reports in most countries that establish a breast cancer screening program. Initially, BI-RADS has been used as a quality assurance system that aims to homogenize the data collection and quality of mammographic reports in the United States since 1995.<sup>5</sup> The system, which is popularized by the American College of Radiology, is based on a national database that is regularly enriched with new cases of physicians enrolled in the program.<sup>6</sup> BI-RADS consists of six categories that also include an extra explanation for categories 0 (Table 1).<sup>7</sup>

The term “Big Data” encompasses all the data collections endowed with a sufficient “size” and lack of definition (having been assembled with no a priori hypothesis or specific research task) to be considered as still largely unspoiled territories from where to derive new insights in the form of unforeseen regularities. In analogy with the exploration of the frontier lands during the gold rush, the exploration of such huge data repositories is usually referred to as “Data Mining”.<sup>8</sup>

Data mining is the extraction of the informations or decision-making knowledge from large data sets that can be used to make predictions, forecasting and estimation. The data is often voluminous, but this data has low value when used directly. Therefore, it is tried to reach important confidential information in the data by using data mining.<sup>9</sup>

Nowadays, data mining is widely used in marketing, banking, insurance, retailing, stock market, telecommunications, medicine and industry. It is used in the medical and medical field for the prediction of medical test results, product development, medical diagnosis, and treatment process.<sup>10</sup>

**TABLE 1:** BI-RADS categories.

Categories	Explanation
0	Incomplete - need for an additional imaging evaluation
1	Normal - normal interval follow-up
2	Typically benign - normal interval follow-up
3	Probably benign - a short interval follow-up is recommended
4	Suspicious abnormality - a biopsy should be considered
5	Highly suggestive of malignancy - biopsy or surgery should be performed
6	Histologically proven malignancy - imaging is performed for cancer staging or evaluation after chemotherapy

The objective of this study was to predict breast biopsy results using age, BI-RADS, Mass Shape, Mass Margin, Mass Density by Multilayer Perceptron (MLP) and Random Forest which are common methods in the field of data mining in recent years. At the same time, by examining the underlying logic of data mining methods, it is aimed to show the importance of the parameters and how to achieve the best-performing model by changing only the parameter values of the MLP and Random Forest.

## MATERIAL AND METHODS

### DATASETS

The UCI Machine Learning Repository is a databases and contains data that can be used for the empirical analysis of machine learning algorithms. The dataset used in our study is from the UCI database. The data is used for predicting the severity (benign or malignant) of a mammographic mass lesion. It contains a BI-RADS assessment, the patient's age, three BI-RADS attributes and severity (516 benign and 445 malignant masses) variables. The data were collected at the Institute of Radiology of the University of Erlangen-Nuremberg between 2003 and 2006 (<https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass>).<sup>11</sup>

WEKA (version 3.6) software was used for data mining.<sup>12</sup> Using the data mining methods of MLP and Random Forest, our study aimed to predict the severity of cancer. Missing data percentages of diagnosis, BI-RADS, age, shape, margin and density parameters in the dataset are 0.0%, 2.0%, 1.0%, 3.0%, 5.0%, and 8.0%, respectively.

### METHODS

#### Multilayer Perceptron

MLP is developed by Rumelhart et al. and is also called error propagation network or backpropagation network model. Ability to use several teaching algorithms to train this network is why it is commonly used. MLP model is composed of one input layer, one or more hidden layers and one output layer.<sup>13,14</sup>

■ **Input Layer:** This is where Artificial Neural Network (ANN) receives information. Number of neurons in this layer is as much as the number of input parameter. Each neuron belongs to one input. Inputs (parameters) are relayed to the hidden layer without any process.

■ **Hidden Layers:** These are where the actual process takes place. In general, the number of neurons in these layers are more than the number of neurons in input and output layers. Numbers of hidden layers and processors in hidden layers are found through trial and error.

■ **Output Layer:** It processes the information coming from hidden layers and sends the output produced in compliance with the input from the input layer. Number of neurons is as much as the number of the outcome parameter class.

#### *Learning Rate*

Dataset is divided into two groups in data mining: training and test data. As training data learns the data on the level of training rate, training rate is a very important concept.<sup>14</sup>

■ Large learning rate: The system learns the data very fast, number of total error increases.

■ Low learning rate: The system learns very slowly; training duration extends but number of error decreases.

#### *Momentum*

It adds some of the weight changes already made to the current weight change and speeds the training of any MLP. But if your momentum constant is too large, the network will not converge, and it will probab-

ly get stuck in local minima. So, it is important to choose best momentum value for your model. Values between 0.6 and 0.8 are recommended.<sup>13,14</sup>

### ***ANN Learning Properties - Epoch***

- It is necessary for the system to be able to train itself according to the training data.
- The system comes to the point where it can train itself with few errors as a result of implementing the same inputs for numerous times.
- With a large number of epochs, the system is accepted as fully trained.<sup>14</sup>

### ***ANN Learning Properties - Stopping Criteria***

- Training may be over when the time is at the threshold.
- If there is a predefined epoch value, it can be over when algorithm arrives at this epoch value.
- It can be also over when the predefined error value is achieved.
- Training can be over when the error value between two epochs decrease.<sup>13,14</sup>

### **Random Forest**

Random Forest method is composed of the combination of multiple decision trees. In this algorithm developed by Breiman, one tree is formed with the decisions of several multi-parameter decisions rather than forming a single tree. This method utilizes classification and regression trees (CART) algorithm to create multivariate decision trees. Qualities are identified by calculating for every level of the tree, and then these qualities are combined to choose attribute for the decision tree, and this process is repeated for every level of the tree. CART algorithm calculates information gain and decides the parameters from which the tree will be branched. After creating the tree, the data are analyzed according to this tree model, and classification or clustering are performed.<sup>10</sup> In the study, it was first shown how to improve the performance of the methods. Afterwards, performance comparison was made according to the best models obtained from MLP and Random Forest methods.

In the MLP method, different scenarios were created by using the values of 70%-30%, 80%-20% and ten-fold cross validation options to divide data into training and test data; one hidden layer with 6 neurons and three hidden layers with 6, 4, and 2 neurons, respectively; 250 miliseconds(ms) and 500 ms for the training duration; and 0.2, 0.6, 0.7 values for momentum.

In the Random Forest method, different scenarios were created by using the values of 70%-30%, 80%-20% and ten-fold cross validation options, 100 and 150 trees and the values of 1, 2, 3 for seed.

### **Performance Evaluation Criterias**

The Accuracy value, which is used as the evaluation criterion in the study, shows how accurately the model is predicting the lesion of a new patient. The F-measure value is a function of Precision and Recall criteria and it is interpreted as the higher the F-measure value, the more accurate the model is. RMSE is a criterion that indicates an error and a lower value means that the model is better. These values were calculated according to MLP's and Random Forest's scenarios and their formulas are also given in Table 2. These values were separately evaluated to compare the performances of the models.

Total Number of Examples (N) = TP + FP + FN + TN

True Classification Rate (Accuracy) =  $\frac{(TP + TN)}{N}$

**TABLE 2: Classification matrix.**

Class Prediction of Model	Actual Class		
	Positive	Negative	Total
	Positive	Number of True Positive (TP)	Number of False Positive (FP)
Negative	Number of False Negative (FN)	Number of True Negative (TN)	FN + TN
Total	TP + FN	FP + TN	N

$$\text{True Positive Rate (Recall)} = \frac{TP}{(TP + FN)}$$

$$\text{Positive Predictive Value (Precision)} = \frac{TP}{(TP + FP)}$$

$$F - \text{measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{(\text{Recall} + \text{Precision})}$$

The RMSE is the square root of the average value of the square of the residual (actual - predicted):

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{N}}$$

where  $Y_i$  is actual and  $\hat{Y}_i$  is the predicted value of class variable.

## RESULTS

The descriptive statistics regarding the independent variables which form the dataset of the study on the level of dependent parameter were given in Table 3.

**TABLE 3: Descriptive statistics of variables in the dataset.**

Variable	Category	Diagnosis	
		Benign	Malignant
Age	≤40	129 (86.0)	21 (14.0)
	41-50	131 (68.9)	59 (31.1)
	51-60	141 (56.4)	109 (43.6)
	61-70	87 (39.2)	135 (60.8)
	≥71	28 (19.4)	116 (80.6)
BI-RADS	2	13 (92.9)	1 (7.1)
	3	30 (83.3)	6 (16.7)
	4	427 (78.1)	120 (21.9)
	5	40 (11.6)	306 (88.4)
Shape	Round	186 (83.0)	38 (17.0)
	Oval	176 (83.4)	35 (16.6)
	Lobular	50 (52.6)	45 (47.4)
	Irregular	85 (21.2)	315 (78.8)
Margin	Circumscribed	316 (88.5)	41 (11.5)
	Microlobulated	9 (37.5)	15 (62.5)
	Obscured	43 (37.1)	73 (62.9)
	Ill-Defined	89 (31.8)	191 (68.2)
	Spiculated	22 (16.2)	114 (83.8)
Density	High	9 (56.2)	7 (43.8)
	Iso	41 (69.5)	18 (30.5)
	Low	405 (50.8)	393 (49.2)
	Fat-Containing	7 (58.3)	5 (41.7)

\* Values in the cells correspond to n (%).

### OUTCOMES OF MULTI LAYER PERCEPTRON

Attributes of 10 scenarios which provided the best outcome according to the MLP method using different scenarios were shown in Table 4.

MLP model in the Scenario 10 provided the best result, and MLP was created with 1 input layer, 3 hidden layers and 1 output layer and trained with different learning algorithms in this model. Five input variables were used in the input layer. These input variables were age, BI-RADS, shape, margin and density parameters. Diverse numbers of artificial neurons were used in the hidden layers. The value in the output layer was that diagnosis for breast cancer was either benign or malign according to the severity values. The established model was an MLP model with 5 inputs, 12 artificial neurons in its hidden layers (6 in the first, 4 in the second, and 2 in the third hidden layers) and 1 output layer. Number of epochs was taken 250 or 500 in the steps. It was seen how better model's performance became when we changed only a few parameters in Table 4. For example, the model with the best performance was achieved in terms of all criteria (Accuracy, F-measure, RMSE) when momentum was 0.7 and epoch was 250 in Scenario 10 (Table 4).

### OUTCOMES OF RANDOM FOREST

Attributes of 10 scenarios which provided the best outcome according to the Random Forest method using different scenarios were shown in Table 5.

The Random Forest method which provided the best outcome in our study created 150 separate trees, and the weighted mean of the values regarding these trees was calculated to achieve a single value. When using the Random Forest model with these parameters, the accuracy value was found 0.816 and F-measure value was found 0.814. It was seen how better model's performance became when we changed only a few parameters in Table 5. For example, the model with the best performance was achieved in terms of accuracy and F-measure criteria when number of trees was 150 and seed was 2 in Scenario 10 (Table 5). The tree diagram of one of the 150 trees created with Random Forest was in Figure 1.

**TABLE 4:** Model performances of the Multilayer Perceptron under different scenarios.

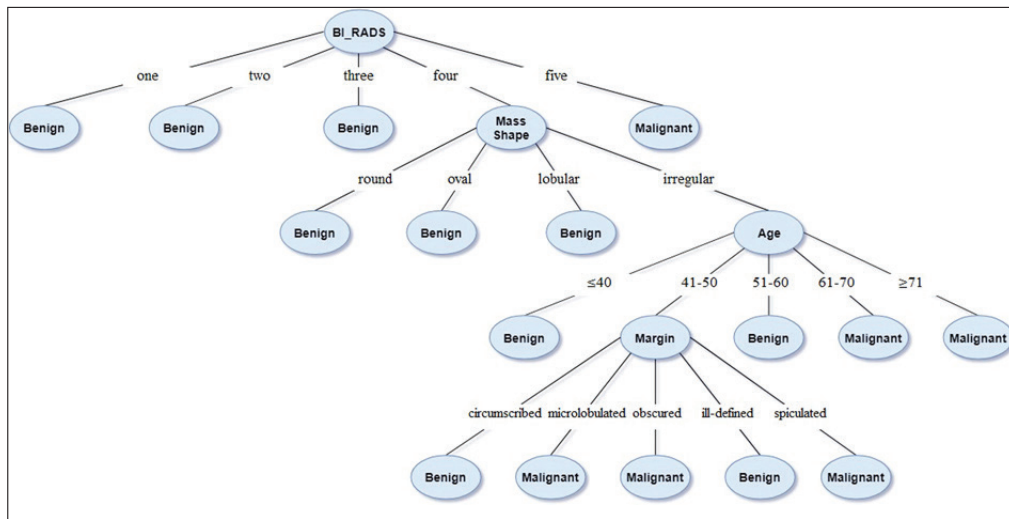
S.	Number of Neurons in Layers			Division Criterion	T.D.	L.R.	M.	Outcomes		
	I.	H <sub>1</sub> /H <sub>2</sub> /H <sub>3</sub>	O.					Acc. (S.E.)	F-m. (S.E.)	RMSE (S.E.)
1	5	6/-/-	1	70%-30%	250	0.3	0.2	0.816 (0.003)	0.815 (0.002)	0.369 (0.001)
2	5	6/-/-	1	80%-20%	250	0.3	0.2	0.815 (0.002)	0.814 (0.002)	0.360 (0.001)
3	5	6/-/-	1	10 Fold C.V.	250	0.3	0.2	0.802 (0.002)	0.801 (0.002)	0.363 (0.001)
4	5	6/-/-	1	70%-30%	500	0.3	0.2	0.823 (0.003)	0.815 (0.002)	0.369 (0.001)
5	5	6/-/-	1	80%-20%	500	0.3	0.2	0.803 (0.003)	0.800 (0.002)	0.371 (0.002)
6	5	6/-/-	1	10 Fold C.V.	500	0.3	0.2	0.803 (0.002)	0.808 (0.003)	0.376 (0.002)
7	5	6/4/2	1	70%-30%	250	0.3	0.6	0.826 (0.003)	0.831 (0.003)	0.353 (0.001)
8	5	6/4/2	1	80%-20%	250	0.3	0.6	0.814 (0.002)	0.810 (0.003)	0.380 (0.001)
9	5	6/4/2	1	10 Fold C.V.	250	0.3	0.6	0.809 (0.002)	0.810 (0.002)	0.381 (0.002)
10	5	6/4/2	1	70%-30%	250	0.3	0.7	0.837 (0.002)	0.833 (0.002)	0.352 (0.001)

S: Scenario, I: Input Layer, H: Hidden Layer, O: Output Layer, T: Training, Duration: D, L.R.: Learning Rate, M: Momentum, Acc: Accuracy, F-m.:F-measure, RMSE: Root Mean Squared Error, C.V.: Cross-Validation, S.E.: Standard error of mean.

**TABLE 5:** Model performances of the Random Forest under different scenarios.

Scenario	Division Criterion	Number of Trees	Seed	Outcomes		
	Training-Test			Accuracy (S.E.)	F-measure (S.E.)	RMSE (S.E.)
1	70%-30%	100	1	0.786 (0.002)	0.793 (0.003)	0.401 (0.003)
2	80%-20%	100	1	0.801 (0.002)	0.799 (0.002)	0.398 (0.002)
3	10 Fold C.V.	100	1	0.786 (0.003)	0.788 (0.002)	0.387 (0.002)
4	70%-30%	100	2	0.781 (0.002)	0.783 (0.002)	0.400 (0.002)
5	80%-20%	100	2	0.792 (0.002)	0.789 (0.003)	0.391 (0.002)
6	10 Fold C.V.	100	2	0.804 (0.002)	0.806 (0.003)	0.380 (0.002)
7	70%-30%	150	1	0.787 (0.002)	0.790 (0.002)	0.403 (0.003)
8	80%-20%	150	1	0.796 (0.002)	0.793 (0.003)	0.412 (0.003)
9	10 Fold C.V.	150	1	0.801 (0.002)	0.795 (0.002)	0.393 (0.002)
10	80%-20%	150	2	0.816 (0.003)	0.814 (0.003)	0.396 (0.002)

RMSE: Root Mean Squared Error, C.V.: Cross-Validation, S.E.: Standard error of mean.



**FIGURE 1:** Tree diagram of Random Forest model.

The tree structure in Figure 1 as follows, and accordingly, accuracy value was calculated as 0.809 and F-measure value was calculated as 0.811. These values, calculated from the single tree structure, were close to the values calculated by Random Forest by weighting 150 trees.

By using this tree structure it can be concluded that, when BI-RADS is 1 or 2 or 3 the lesion is probably benign, when BI-RADS is 5 the lesion is probably malignant and when BI-RADS is 4 it is better to reach a decision of the lesion by evaluating the other variables in the data set (Figure 1).

Classification achievements of the methods were evaluated based on the criteria of accuracy, F-measure and RMSE. The results showed that accuracy, F-measure and RMSE values of the best MLP model were

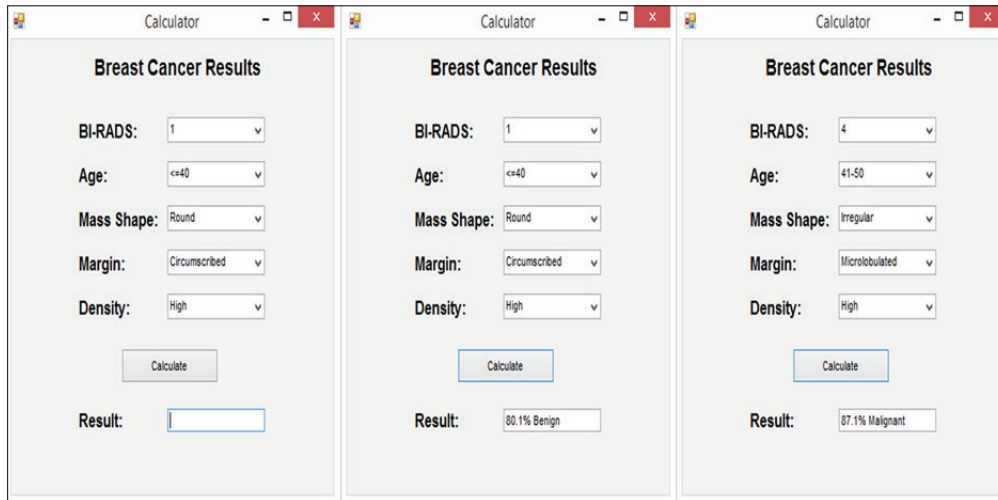


FIGURE 2: Outputs of software that was developed based on the best model in MLP.

found 0.837, 0.833 and 0.352, respectively while accuracy, F-measure and RMSE values of the best Random Forest model were found 0.816, 0.814 and 0.396, respectively. It could be understood that the MLP method provided a better prediction of breast cancer diagnosis than the Random Forest method for this dataset. In this study, a software was developed based on the best model created by MLP method and the sample outputs were given in Figure 2. This software was designed to be a web application using the asp.net software language in the visual studio environment. As a future work, the web application will be activated and used actively for both patient registration as a database and for cancer prediction.

## DISCUSSION

There are several data mining methods that can be used for classification and prediction. In this study, MLP and Random Forest methods were utilized for the prediction of breast cancer biopsy outcomes from BI-RADS input attributes and these methods were examined with all their details in an effort. Parameters were altered for the MLP model, several scenarios were tried, and the model providing the best result was chosen for the study. In the Random Forest, one of the 150 trees was shown as an example, and how calculations were done in the background was demonstrated in an effort. However, these results were obtained from approximately one thousand cases. Therefore, it is recommended to confirm the results by a clinical prospective study involving larger cases. The results achieved in the method comparison in this study are valid only for datasets that contain the variables that are used in this study; it would not be right to make a sweeping statement like MLP method has a better classification performance than Random Forest.

In our study, classification achievements of the methods were evaluated based on the criteria of accuracy, F-measure and RMSE. The results show that accuracy, F-measure and RMSE values of the MLP model were found 0.837, 0.833 and 0.352, respectively while accuracy, F-measure and RMSE values of the Random Forest model were found 0.816, 0.814 and 0.396, respectively. It can be understood in the light of these results that the MLP method provided a better prediction of breast cancer diagnosis than the Random Forest method for this dataset according to all three criteria.

In literature, there is only one data mining study conducted by using this dataset and in this study Elter et al. presented two computer-aided diagnosis systems, the first approach is based on decision-tree learning and the second on case-based reasoning using an entropic distance measure. They compared their results with a state of the art ANN approach and reported that their second system had better performance than



ANN.<sup>11</sup> Asri et al. used Wisconsin Breast Cancer (original) datasets with 699 patients, used the methods of Decision Tree (C4.5), Support Vector Machine, Naive Bayes and Nearest Neighbors in their study and made a performance comparison according to the Accuracy and RMSE metrics.<sup>15</sup> They found the accuracy values 95.13%, 97.13%, 95.99%, and 95.27% respectively and RMSE values 0.21, 0.16, 0.19, and 0.21, respectively for these methods and stated that the model providing the best result according to both values was Support Vector Machine. Mandal used the Wisconsin Diagnosis Breast Cancer dataset with 569 patients, utilized the methods of Logistic Regression, Naive Bayes and Decision Tree in the study and found the accuracy values 94.4%, 97.9% and 96.5% respectively for these methods.<sup>16</sup> Similarly, Shrivastava et al. used the J48 Decision Tree methods in their study using the Wisconsin Breast Cancer (original) datasets with 699 patients and found the accuracy value 98.14 for this method.<sup>17</sup> In their study with 3000 patients, Akinsola Adeniyi et al. used the methods of Decision Tree (C4.5), Naive Bayes and MLP and found the accuracy values 93.9%, 76.5% and 83.7% respectively and RMSE values 0.223, 0.417 and 0.335, respectively for these methods.<sup>14</sup> They decided the model providing the best result to be the Decision Tree (C4.5) method according to the two performance criteria.

## CONCLUSION

In this study, MLP method provided a better prediction of breast cancer diagnosis than the Random Forest method for this dataset according to accuracy, F-measure and RMSE criterias. Consequently, when the clinical parameters used in breast cancer prediction are known, a model can be created with the MLP method, this model can be used as an alternative in the diagnosis of patients and be an assistant tool for physicians.

### Acknowledgement

*I would like to thank to the translator and interpreter Mustafa Yelken for all his support.*

### Source of Finance

*During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.*

### Conflict of Interest

*No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.*

### Authorship Contributions

**Idea/Concept:** Batuhan Bakırarar; **Design:** Batuhan Bakırarar; **Control/Supervision:** Derya Gökmen, Atilla Halil Elhan; **Data Collection and/or Processing:** Batuhan Bakırarar, İrem Kar; **Analysis and/or Interpretation:** Batuhan Bakırarar, İrem Kar, Derya Gökmen, Atilla Halil Elhan, Volkan Genç; **Literature Review:** Batuhan Bakırarar, İrem Kar, Derya Gökmen; **Writing The Article:** Batuhan Bakırarar, İrem Kar; **Critical Review:** Derya Gökmen, Atilla Halil Elhan, Volkan Genç.

## REFERENCES

1. Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, et al. Cancer statistics. *CA Cancer J Clin.* 2005;55(1):10-30. PMID: 15661684 [[Crossref](#)] [[PubMed](#)]
2. Pirhan Y, Sucu BB. [Complication of heparin used during hemodialysis in breast patient with cancer: case report]. *Türk Nefroloji, Diyaliz ve Transplantasyon Hemşireleri Derneği Nefroloji Hemşireliği Dergisi.* 2018;1(13):32-5.
3. Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. *Ann Intern Med.* 2002;137(5 Part 1):347-60. PMID: 12204020 [[Crossref](#)] [[PubMed](#)]
4. Tabar L, Yen MF, Vitak B, Chan HH, Smith RA, Duffy SW. Mammography service screening and mortality in breast cancer patients: 20-years follow-up before and after introduction of screening. *Lancet.* 2003;361(9367):1405-10. PMID: 12727392 [[Crossref](#)]

5. Illustrated Breast Imaging Reporting and Data System (BI-RADS TM). Reston, VA: Am Coll Radiol. 1998.
6. Liberman L, Menell JH. Breast imaging reporting and data system (BI-RADS). *Radiol Clin North Am.* 2002;40(3):409-30. PMID: 12117184 [[Crossref](#)]
7. Balleyguier C, Ayadi S, Van Nguyen K, Vanel D, Dromain C, Sigal R. BIRADS classification in mammography. *Eur J Radiol.* 2007;61(2):192-4. PMID: 17164080 [[Crossref](#)] [[PubMed](#)]
8. Todde V, Giuliani A. Big Data. A briefing. *Annali Dell'Istituto Superiore Di Sanita.* 2018;54(3):174-5.
9. Nagabhushana S. *Data Warehousing OLAP and Data Mining.* 1<sup>st</sup> ed. New Delhi, India: New Age International; 2006. p.251-2.
10. Han J, Kamber M, Pei J. *Data Mining Concepts and Techniques.* 3<sup>rd</sup> ed. Waltham, MA, USA: Morgan Kaufmann/Elsevier; 2011. p.2-5.
11. Elter M, Schulz-Wendtland R, Wittenberg T. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Med Phys.* 2007;34(11):4164-72. PMID: 18072480 [[Crossref](#)] [[PubMed](#)]
12. Eibe F, Mark AH, Ian HW. *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques.* 4<sup>th</sup> ed. Morgan Kaufmann; 2016.
13. Hand D, Mannila H, Smyth P. *Principles of Data Mining (Adaptive Computation and Machine Learning).* 1<sup>st</sup> ed. Cambridge, MA, USA: The MIT Press; 2001. p.151-6.
14. Akinsola AF, Sokunbi MA, Okikiola FM, Onadokun IO. Data mining for breast cancer classification. *International Journal of Engineering and Computer Science (IJCSSE).* 2017;6(8):22250-8.
15. Asri H, Mousannif H, Moatassime HA, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Comput Sci.* 2016;83:1064-9. [[Crossref](#)]
16. Mandal SK. Performance analysis of data mining algorithms for breast cancer cell detection using naïve bayes, logistic regression and decision tree. *International Journal of Engineering and Computer Science (IJCSSE).* 2017;6(2):20388-91.
17. Shrivastava SS, Sant A, Aharwal RP. An overview on data mining approach on breast cancer data. *International Journal of Advanced Computer Research (IJACR).* 2013;3(4):256-62.