CLASSIFICATION OF BREAST CANCER AND DETERMINATION OF RELATED FACTORS WITH DEEP LEARNING APPROACH

E. Guldogan, Z. Tunc and C. Colak

Abstract*— *Aim: In this study, it is aimed to classify breast cancer and identify related factors by applying deep learning method on open access to breast cancer dataset.

Materials and Methods: In this study, 11 variables related to open access to breast cancer dataset of 569 patients shared by the University of Wisconsin were used. The deep learning model for classifying breast cancer was established by a 10-fold cross-validation method. The performance of the model was evaluated with accuracy, sensitivity, specificity, positive/negative predictive values, F-score, and area under the curve (AUC). Factors associated with breast cancer were estimated from the deep learning model.

Results: Accuracy, specificity, AUC, sensitivity, positive predictive value, negative predictive value, and F-score values obtained from the model were 94.91%, 91.47%, 0.988, 96.90%, 95.42%, 95.14%, and 96.03%, respectively. In this study, when the effects of the variables in the dataset on breast cancer were evaluated, the three most important variables were obtained as area mean, concave points mean and symmetry mean, respectively.

Conclusion: The findings of this study showed that the deep learning model provided successful predictions for the classification of breast cancer. Also, unlike similar studies examining the same dataset, the importance values of cancer-related factors were estimated with the help of the model. In the following studies, breast cancer classification performances can give more successful predictions thanks to different deep learning architectures and ensemble learning approaches.

Keywords—Breast cancer, artificial intelligence, deep learning, classification.

1. INTRODUCTION

BREAST cancer is one of the leading causes of death among women in developed and developing countries. Detection and classification of breast cancer development in the early stages allow patients to receive appropriate treatment. Breast cancer is considered a genetically heterogeneous and biologically diverse disease. Long-known clinical and phenotypic differences are associated with differences in gene expression. Previous studies of breast tumors have identified five different types of breast carcinoma subtypes [luminal A (estrogen receptor (ER) +); luminal B (ER +); HER2 overexpression; normal breast-like and basal-like] associated with different clinical outcomes [1, 2].

Artificial intelligence (AI) involves the use of computer systems to achieve set goals by mimicking cognitive abilities. Machine learning (ML) classification is an AI field that allows algorithms or classifiers to learn patterns in large and complex datasets and produce useful predictive outputs. Applying ML algorithms to large datasets can reveal new trends and relationships that may have beneficial effects for clinical practice in medicine. Scientific studies have investigated the application of ML methods in health care and have shown that ML has an important effect on improving health quality and safety [3]. In an actual study, it is reported that artificial intelligence systems that can perform at the level of expert radiologists in digital mammography evaluation increase breast cancer screening accuracy and efficiency [4].

The complex structure of processes such as pretreatment, clustering, feature selection, and extraction, etc. in classical machine learning approaches reduces the performance and accuracy of the system. To solve problems related to traditional machine learning techniques, deep learning strategies are proposed to extract relevant information from raw images and to be used effectively in the classification process In deep learning, features are determined by the training operations performed from data sets with the help of a general-purpose learning approach. [1].

In this study, it is aimed to classify breast cancer and determine related factors by applying deep learning method on open access to breast cancer data set.

2. MATERIAL AND METHODS

2.1. Dataset

To analyze the working principle of the deep learning method and to evaluate the model, the open-access dataset called “Breast Cancer Wisconsin (Diagnostic) Data Set” was obtained from UCI Machine Learning Repository [5]. In the data set used, there are 569 people examined for breast cancer. Of the individuals, 357 (62.7%) were diagnosed as benign and 212 (37.3%) were diagnosed as malignant. The explanations about the variables in the data set and their properties are given in Table 1.

Emek GULDOGAN, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (emek.guldogan@inonu.edu.tr)

Zeynep TUNC, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (zeynep.tunc@inonu.edu.tr)

Cemil COLAK, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (cemil.colak@inonu.edu.tr)

Manuscript received May 08, 2020; accepted May 29, 2020.
Digital Object Identifier:

TABLE I
EXPLANATIONS ABOUT THE VARIABLES IN THE DATASET AND THEIR PROPERTIES

Variable	Variable Description	Variable Type	Variable Role
diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)	Qualitative	Output
radius_mean	Mean of distances from the center to points on the perimeter	Quantitative	Input
texture_mean	The standard deviation of gray-scale values	Quantitative	Input
perimeter_mean	Mean size of the core tumor	Quantitative	Input
area_mean	-	Quantitative	Input
smoothness_mean	Mean of local variation in radius lengths	Quantitative	Input
compactness_mean	mean of perimeter^2 / area - 1.0	Quantitative	Input
concavity_mean	Mean of the severity of concave portions of the contour	Quantitative	Input
concave points_mean	mean for the number of concave portions of the contour	Quantitative	Input
symmetry_mean	-	Quantitative	Input
fractal_dimension_mean	mean for "coastline approximation" - 1	Quantitative	Input

3. DEEP LEARNING MODEL

Deep Learning is based on the multi-layer feed-forward neural network trained with stochastic slope landing using the back-propagation approach. Related network; the hyperbolic tangent (tanh), rectifier, and maxout (a generalization of ReLU and leaky ReLU functions) can contain many hidden layers of neurons with activation functions. Advanced features such as adaptive learning speed, rate annealing, momentum training, dropout, and L1 or L2 regulations can provide high predictive accuracy. L1 is a regularization technique that restrains the absolute valuation of the weights and has the net influence of dropping some weights (setting them to zero) from a model to decrease complexity and refrain overfitting problems. L2 is another regularization technique that restrains the sum of the squared weights. This technique presents bias into the estimates of the parameter; however, it frequently performs considerable gains in modeling as the variance of the estimate is decreased. Each computes node trains a copy of global model parameters on local data in multiple threads (asynchronously) and periodically contributes to the global model through the model average across the network [6].

For the validity of the model, a 10-fold cross-validation method was used. In the 10-fold cross-validation method, all data is divided into 10 equal parts. One part is used as a test set and the remaining 9 parts are used as a training data set and this process is repeated 10 times. Hyperparameters related to the deep learning model were selected as activation function (Maxout linear unit), hidden layer sizes (50), the number of revolutions (10), epsilon ($1.0\ e^{8)}$ and rho (0.99). Table 2 shows the hyperparameters used in building the deep learning model [7]. RapidMiner Studio software was used in all modeling and analysis [8].

TABLE II
HYPERPARAMETERS USED TO CONSTRUCT A DEEP LEARNING MODEL

Hyperparameter name	Hyperparameter selection
Activation function	Maxout linear unit
Hidden layer sizes	50
Number of revolutions	10
Epsilon	$1.0\ e^{8}$
Rho	0.99

3.1. Performance evaluation criteria

The classification matrix for the calculation of performance metrics is given in Table 3.

TABLE III
CONFUSION MATRIX FOR CALCULATING PERFORMANCE METRICS

		Real		
		Positive	**Negative**	**Total**
Predicted	Positive	True positive (TP)	False negative (FN)	TP+FN
	Negative	False positive (FP)	True negative (TN)	FP+TN
	Total	TP+FP	FN+TN	TP+TN+FP+FN

The metrics considered in the performance evaluation of the models in this study are given below.

Accuracy = (TP+TN)/(TP+TN+FP+FN)
Sensitivity = TP/(TP+FP)
Specificity = TN/(TN+FN)
Positive predictive value = TP/(TP+FN)
Negative predictive value =TN/(TN+FP)
F-score = (2*TP)/(2*TP+FP+FN)

4. DATA ANALYSIS

Quantitative data are summarized by median (minimum-maximum) and qualitative variables are given by number and percentage. Normal distribution was evaluated with the Kolmogorov-Smirnov test. Whether there is a statistically significant difference between categories of the dependent

variable in terms of input variables, the Mann-Whitney U test was used for the analyses. p<0.05 values were considered statistically significant. In all analyzes, IBM SPSS Statistics 26.0 for the Windows package program was used.

5. RESULTS

Descriptive statistics related to the target variable examined in this study are presented in Table 4. There is a statistically significant difference between the dependent variable classes in terms of other variables other than the fractal_dimension_mean variable (p<0.001).

TABLE IV
DESCRIPTIVE STATISTICS ABOUT INPUT AND OUTPUT VARIABLES

Variables	Diagnosis		p* value
	Benign (n=357) Median (min-max)	Malignant (n=212) Median (min-max)	
radius_mean	12.2 (6.98-17.85)	17.33 (10.95-28.11)	**<0.001**
texture_mean	17.39 (9.71-33.81)	21.46 (10.38-39.28)	**<0.001**
perimeter_mean	78.18 (43.79-114.6)	114.2 (71.9-188.5)	**<0.001**
area_mean	458.4 (143.5-992.1)	932 (361.6-2501)	**<0.001**
smoothness_mean	0.09 (0.05-0.16)	0.1 (0.07-0.14)	**<0.001**
compactness_mean	0.08 (0.02-0.22)	0.13 (0.05-0.35)	**<0.001**
concavity_mean	0.04 (0-0.41)	0.15 (0.02-0.43)	**<0.001**
concave points_mean	0.02 (0-0.09)	0.09 (0.02-0.2)	**<0.001**
symmetry_mean	0.17 (0.11-0.27)	0.19 (0.13-0.3)	**<0.001**
fractal_dimension_mean	0.06 (0.05-0.1)	0.06 (0.05-0.1)	0.537

*: Mann Whitney U test

In this study, the classification matrix for the deep learning model used to classify breast cancer is given in Table 5 below.

TABLE V
CLASSIFICATION MATRIX FOR DEEP LEARNING MODEL

Real / Predicted	Malignant	Benign	Total
Present	194	11	205
Absent	18	346	364
Total	212	357	569

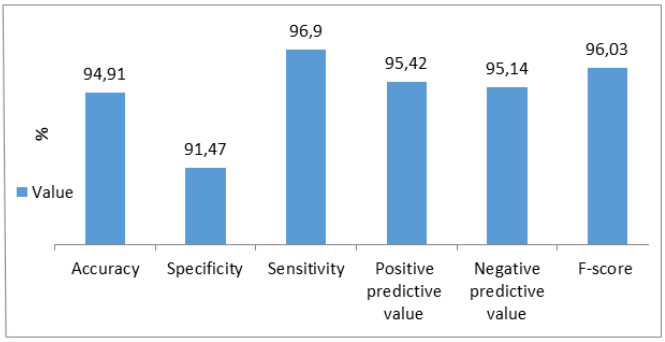

Fig. 1. Values related to performance criteria

The values related to performance criteria are given graphically in Figure 1.

Accuracy, specificity, AUC, sensitivity, positive/negative predictive value, F-score metrics for the deep learning model are summarized in Table 6. Accuracy, specificity, AUC, sensitivity, positive predictive value, negative predictive value, and F-score values obtained from the model were 94.91%, 91.47%, 0.988, 96.90%, %95.42%, %95.14%, and %96.03%, respectively.

TABLE VI
THE VALUES OF PERFORMANCE METRICS

Performance Criterion	Value
Accuracy (%)	94.91
Specificity (%)	91.47
AUC	0.988
Sensitivity (%)	96.90
Positive predictive value (%)	95.42
Negative predictive value (%)	95.14
F-score (%)	96.03

 ISSN: 2548-0650 http://dergipark.gov.tr/jcs

In this study, the importance values of the factors related to breast cancer are given in Table 7. Among the three most important factors, the most important variable is the area_mean , followed by the concave points_mean and ssymmetry_mean, respectively.

TABLE VII
SEQUENCE OF VARIABLES IN ORDER OF IMPORTANCE

Variables	Importance (%)
area_mean	10.99
concave points_mean	10.78
symmetry_mean	10.77
perimeter_mean	10.67
fractal_dimension_mean	10.04
concavity_mean	10.03
radius_mean	9.32
smoothness_mean	9.22
compactness_mean	9.13
texture_mean	9.06

According to the World Health Organization (WHO), early diagnosis of cancer greatly increases the chances of making the right decision on a successful treatment plan [9, 10]. Computer-Aided Diagnosis (CAD) systems are widely applied in the detection and differential diagnosis of many different types of diseases. Therefore, increasing the accuracy of a CAD system has become one of the main research areas. In this study, using the deep learning approach on open access breast cancer data set, computer-aided classification of breast cancer and related factors were determined. Thus, it is possible to prevent the progression of the disease and to implement alternative treatment protocols by diagnosing breast cancer in the early stages [11, 12].

When similar studies were examined, Wisconsin Original Data Set consisting of 569 records and 31 (30 predictors, 1 target) feature/variable was used to increase the accuracy of the diagnosis of breast cancer in different machine learning methods. The accuracy of the proposed support vector machine model was found to be 0.9766 and the study results showed that the proposed model has a high-performance rate and will contribute to improving breast cancer diagnosis accuracy, which is an important problem of today. In this study, the accuracy value was calculated as 0.9491 in the breast cancer classification made using only 11 (10 predictors, 1 dependent) feature/variable on the same data set [13]. In this study, breast cancer classification was made successfully by using fewer variables/features, and similar performance criteria were obtained in the study described. Thus, in this study, the breast cancer classification accuracy rate was obtained very high by using fewer variables, and the importance values related to the investigated features were also revealed with the deep learning technique. Clinicians can evaluate the risk factors that may be effective in the development of breast cancer clinically more effectively with the help of the importance values related to the variables obtained from the deep learning model created. In a similar study, a CAD was developed for the detection of breast cancer using a back-propagation supervised approach following deep belief networks unsupervised learning. In the model used in this process, weights were obtained from the deep belief network and backpropagation neural network was used with the learning function of Liebenberg Marquardt. The model created was tested on the Wisconsin Breast Cancer Data Set and gave a 99.68% accuracy rate showing promising results compared to previously published studies [14]. In the study summarized, only deep learning algorithms were used to detect breast cancer and no risk factor analysis that could be associated with breast cancer was performed. In this respect, this study shows significant differences from similar studies examining the same data set.

Breast cancer risk prediction provides systematic identification of individuals at the highest and lowest risk. Thus, the detection of high levels of breast cancer risk factors in the general society and women with a family history provides a more accurate decision on disease prevention therapies and screening [15-17]. When the effects of the variables in the data set examined in this study on breast cancer are examined; the three most important variables are as area_mean (10.99%), concave points_mean (10.78%), and symmetry_mean (10.77%) were obtained as a result of calculations.

To sum up, the findings obtained from this study showed that the deep learning model created gave successful predictions in classifying breast cancer. Besides, unlike similar studies examining the same data set, the significance values of cancer-related factors were estimated from the model created. In further studies, the classification performances of different types of deep learning architectures and ensemble learning approaches can provide more successful predictions.

REFERENCES

[1] S. Khan, N. Islam, Z. Jan, I. U. Din, and J. J. C. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognition Letters*, vol. 125, pp. 1-6, 2019.

[2] A. C. Peterson and H. Uppal, "Method for predicting response to breast cancer therapeutic agents and method of treatment of breast cancer," ed: Google Patents, 2019.

[3] Q. D. Buchlak et al., "Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review," Neurosurgical review, pp. 1-19, 2019.

[4] A. Rodriguez-Ruiz et al., "Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists," JNCI: Journal of the National Cancer Institute, vol. 111, no. 9, pp. 916-922, 2019.

[5] D. Dua and C. J. U. h. a. i. u. e. m. Graff, "UCI machine learning repository, 2017," vol. 37, 2019.

[6] M. Hofmann and R. Klinkenberg, RapidMiner: Data mining use cases and business analytics applications. CRC Press, 2016.

[7] G. O. TEMEL, S. ERDOĞAN, and H. ANKARALI, "Sınıflama Modelinin Performansını Değerlendirmede Yeniden Örnekleme Yöntemlerinin Kullanımı," Bilişim Teknolojileri Dergisi, vol. 5, no. 3, pp. 1-8, 2012.

[8] I. Mierswa and R. Klinkenberg, "RapidMiner Studio Version 9.5," ed, 2019.

[9] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 68, no. 6, pp. 394-424, 2018.

[10] WHO. (2018). Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. Available: https://www.who.int/cancer/PRGlobocanFinal.pdf
[11] N. ALPASLAN, "MEME KANSERİ TANISI İÇİN DERİN ÖZNİTELİK TABANLI KARAR DESTEK SİSTEMİ," Selçuk Üniversitesi Mühendislik, Bilim Ve Teknoloji Dergisi, vol. 7, no. 1, pp. 213-227, 2019.
[12] V. Bajaj, M. Pawar, V. K. Meena, M. Kumar, A. Sengur, and Y. Guo, "Computer-aided diagnosis of breast cancer using bi-dimensional empirical mode decomposition," Neural Computing and Applications, vol. 31, no. 8, pp. 3307-3315, 2019.
[13] H. Kör, "Classification of Breast Cancer by Machine Learning Methods."
[14] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," Expert Systems with Applications, vol. 46, pp. 139-144, 2016.
[15] A. Lee et al., "BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors," 2019.
[16] A. Brédart et al., "Clinicians' use of breast cancer risk assessment tools according to their perceived importance of breast cancer risk factors: an international survey," Journal of community genetics, vol. 10, no. 1, pp. 61-71, 2019.
[17] S. Karadag Arli, A. B. Bakan, and G. Aslan, "Distribution of cervical and breast cancer risk factors in women and their screening behaviours," European journal of cancer care, vol. 28, no. 2, p. e12960, 2019.

BIOGRAPHIES

Emek Güldoğan obtained his BSc. degree in Computer Engineering from Middle East Technical University in 2001. He received MSc. degree in biostatistics and medical informatics from the Inonu University in 2005, and Ph.D. degrees in biostatistics and medical informatics from the Inonu University in 2017. He is currently working as an assistant professor of the Department of Biostatistics and Medical Informatics at Inonu University and as the information processing manager at Turgut Özal Medical Center. His research interests are cognitive systems, data mining, machine learning, deep learning.

Zeynep Tunç obtained her BSc. degree in mathematics from Çukurova University in 2010. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning.

Cemil Çolak obtained his BSc. degree in statisticsfrom Ondokuz Mayıs University in 1999. He received MSc. degree in Biostatistics from the Inonu University in 2001, and Ph.D. degree in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. His research interests are cognitive systems,data mining, reliability, and biomedical system, genetics, and bioengineering. In 2016, he joined the Department of Biostatistics and Medical Informatics at Inonu University as a Professor, where he is presently a professor. He is active in teaching and research in the general image processing , artificial intelligence, data mining, analysis.