RESEARCH ARTICLE

Classification of Prostate Cancer and Determination of Related Factors with Different Artificial Neural Network

İpek Balıkcı Çicek[1], Zeynep Küçükakçalı[2]

[1] Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya, Turkey

Abstract

Objective: In this study, it is aimed to classify prostate cancer, compare the predictions of these two models and determine the factors associated with the disease by applying Multilayer Perceptron Neural Network (MLPNN) and Radial-Based Function Neural Network (RBFNN) methods on the open access Prostate cancer dataset.

Methods: In this study, the dataset named "Prostate Cancer Data Set" was used by obtaining from https://www.kaggle.com/sajidsaifi/prostate-cancer address. To classify prostate cancer, MLPNN and RBFNN methods, which are artificial neural network models, is used. The classification performance of the models was evaluated with the sensitivity, specificity, accuracy, negative predictive value and positive predictive value, which are among the classification performance metrics. Prostate cancer related factors were estimated by using MLPNN and RBFNN models.

Results: With the applied MLPNN model, performance metric values were obtained as AUC 0.937, Sensitivity 100%, accuracy 92.5%, Selectivity 84.6%, Positive predictive value 87.5% and Negative predictive value 100%. With the RBFNN model, the performance metric values were obtained as AUC 0.921, Sensitivity 83.3%, accuracy 86.6%, Selectivity 91.6%, Positive predictive value 93.7% and Negative predictive value 78.5%. When the effects of variables in the dataset in this study on prostate cancer are examined; The three most important variables for the MLPNN model were obtained as perimeter, area and compactness, respectively. For the RBFNN model, the three most important variables were obtained as perimeter, area and compactness, respectively.

Conclusion: It was seen that MLPNN and RBFNN models used in this study gave successful predictions in the classification of prostate cancer. In addition, estimating the significance values of factors associated with the disease with these classification models made it different from similar studies with the same dataset.

Key words: Prostate cancer, Multilayer perceptron neural network, Radial-based function neural network, Classification.

Address for correspondence/reprints:

İpek Balıkcı Cicek

Telephone number: +90 (553) 373 24 04

ORCID-ID 0000-0002-3805-9214

E-mail: ipek.balikci@inonu.edu.tr

DOI: 10.19127/mbsjohs.798559

Introduction

The prostate is a small walnut-shaped gland that produces seminal fluid that nourishes and carries sperm in men. Prostate cancer is a type of cancer that occurs in the prostate (Ah 2012). Prostate cancer occurs when cells in the prostate gland begin to grow uncontrollably, and it is a heterogeneous and often multifocal disease that often occurs as adenocarcinoma (Foster et al. 2000). Prostate cancer is the most common urological cancer. Since the prostate is a gland found only in men, prostate cancer is a type of cancer that occurs only in men. Prostate cancer is a disease that significantly affects the health of men and the quality of life due to the problems it causes, and its incidence increases with age. Unlike other organ cancers, early-stage prostate cancer usually develops slowly (Badger et al. 2011). Prostate cancer ranks second among the most common cancers in men worldwide. Prostate Cancer ranks sixth in cancer-related deaths, with approximately 1,600,000 cases and 366,000 deaths annually. Despite this high incidence, information on the etiology and risk factors of prostate cancer is limited. For this reason, mortality and morbidity caused by prostate cancer constitute important health expenses worldwide (Siegel, Miller and Jemal 2019, King et al. 2015).

Artificial neural networks (ANN) are computer systems designed to automatically apply skills such as the ability to obtain new information, create and discover new information through education, which are one of the features of the human brain without any help (Elmas 2016). Similar to the characteristics of the human brain, ANN is applied in areas such as learning, association, classification, generalization, feature determination and optimization with the information obtained from samples (Yildirim 2020). Technically, the most basic task of ANN is to determine a result that can correspond to the examples shown to it. In order to do this, the artificial neural network is trained with examples and gained the ability to generalize. With this generalization, output sets corresponding to similar events are determined (Oztemel 2003). While designing a model in ANN, input and output sets are used. Thus, ANN is able to generate solutions at certain limit ranges for previously unseen, unrecognized, unlearned and unapplied examples by learning all linear and nonlinear relationships between inputs and outputs, and previously acquired past situations. Due to the non-linear feature of ANN, it has many advantages such as its ability to work quickly, learning ability, ability to generalize, adaptable to different problems easily and requiring less information, so it is used in solving different problems in many areas (Oztemel 2003).

The multilayer perceptron neural network (MLPNN) consists of three layers the input layer where neural network information is entered, hidden layers and output layer. Information's is introduced to the network from the input layer, reaches the output layer from the hidden layers and is transferred from the output layer to the outside world (Haykin 2007). MLPNN is a non-parametric artificial neural network technique that performs many detection and prediction processes (Orhan, Hekim and Ozer 2010). There are transitions called forward and back propagation between layers in MLPNN. Purpose of MLPNN learning method; It is to make the error between the desired output and the output produced by the network minimum. In the forward propagation phase, the output of the network and the error value are calculated. In the back-propagation phase, the connection weight values between the layers are updated to minimize the calculated error value (Arı and Berberler 2017).

The Radial Basis Function neural network (RBFNN) consists of a three-layer structure: an input layer, a single hidden layer using radial functions that give the network its name, and an output layer. The working principle of RBFNN is the process of determining RBFNNs with appropriate width and center values in the hidden layer depending on the input data, creating linear combinations of the outputs produced by these functions in the output layer and determining the relationship between input-output (Kaynar et al. 2016). Radial-based networks are similar in structure to back propagation networks, the difference in RBFNN is that the radial-based activation function (gauss, exponential) in the intermediate layer transforms the inputs. There is also a linear activation function in the output layer (Gemici, Ardiclioglu and Kocabas 2013).

In this study, it was aimed to compare the classification success of prostate cancer and determine the risk factors related to prostate cancer by applying MLPNN and RBFNN methods on the open access prostate cancer dataset.

Methods

Dataset

In the study, an open access dataset named "Prostat Cancer Data Set" was obtained from https://www.kaggle.com/sajidsaifi/prostate-cancer to examine the working principles of MLPNN and RBFNN methods and to determine risk factors. There are 100 patients diagnosed with prostate cancer in this

open access dataset. Of the patients diagnosed with cancer, 38 (38%) were diagnosed as benign, and 62 (62%) were diagnosed as malignant. The variables and the descriptive properties of the variables in the relevant dataset are given in Table 1.

Table 1. Variables in the dataset and their descriptive properties

Variable	Variable Explanation	Variable type	Variable role
Diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)	Qualitative	Dependent/ Target
Radius	Mean distances from the center to perimeter points	Quantitative	Independent/ Predictor
Texture	The standard deviation of gray-scale values	Quantitative	Independent/ Predictor
Perimeter	Mean size of the core tumor	Quantitative	Independent/ Predictor
Area	-	Quantitative	Independent/ Predictor
Smoothness	Mean of local variation in radius lengths	Quantitative	Independent/ Predictor
Compactness	(mean of perimeter)^2 / (area - 1)	Quantitative	Independent/ Predictor
Symmetry	-	Quantitative	Independent/ Predictor
Fractal dimension	mean for "coastline approximation" - 1	Quantitative	Independent/ Predictor

Multilayer Perceptron Neural Network (MLPNN)

Multilayer Perceptron Neural Network (MLPNN) is known as the back propagation model of a feed forward neural network developed by Rumelhart, Hinton and Williams (Rumelhart, Hinton and Williams 1986). MLPNN is an artificial neural network model with a feed forward structure consisting of the input layer, the output layer and the hidden layer(s) between these two layers. The inputs to the neurons in the hidden layer are collected and transmitted to the output layer by multiplying the connection weights between the hidden layer and the output layer in the same way. Neurons in the output layer collect these inputs and produce an output accordingly (Efe and Kaynak 2000).

The main purpose of the MLPNN method is to minimize the error between the expected output of the network and the output it produces. During the training, both the inputs and the (expected) outputs that should be produced against those inputs are shown in these networks (Soylemez 2020). Samples are applied to the input layer, processed in hidden layers, and outputs are obtained from the output layer (Selcuk 2020). According to the training algorithm used, the error between the output of the network and the desired output is spread backwards again and the weight of the network is changed until the error is minimized (Kayna, Tastan and Demirkoparan 2010).

Radial Based Function Neural Network (RBFNN)

The concept of Radial-Based Functions was introduced into the artificial neural networks literature by Broomhead and Lowe in 1988. The artificial neural network model based on Radial Based Functions is inspired by local impulse-response behaviors seen in neurons (nerve cells) in the human nervous system (Poggio and Girosi 1990). RBFNN is a feed forward network consisting of a three-layer structure consisting of an input layer, a single hidden layer using radial functions that give the network its name as a transfer function, and an output layer. While the inputs of this network are not linear, its output is linear (Kim and Kim 2004).

The input layer consists of source nodes and provides the network's connection with the environment. The second layer, the only hidden layer in the network, makes nonlinear conversion from the input area to the hidden area. The transformation from the input layer to the hidden layer is a nonlinear constant transformation with radial based transfer functions (Yildiz, Tasova and Polatci 2020). The output layer is linear and responds to the network, which is the transfer signal applied to the input layer. An adaptive and linear transformation is performed from the hidden layer to the output layer. Thus, the output layer, which provides the response of the network to the transfer signal applied to the input, has the feature of linearity (Haykin 1999).

Performance evaluation metrics

In the performance evaluation of the radial-based artificial neural network and multilayer artificial neural network models, which were created to predict the factors that may be associated with prostate cancer, the performance metrics obtained by using the classification matrix (Table 2) given below were used.

Table 2. Classification 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 performance metrics to be used in the performance evaluation of the models in this study are given below.

Sensitivity = TP/(TP+FP)
Specificity = TN/(TN+FN)
Accuracy = (TP+TN)/(TP+TN+FP+FN)
Negative predictive value =TN/(TN+FP)
Positive predictive value = TP/(TP+FN)

Statistical Analyses

Quantitative data were expressed as mean ± standard deviation, median (minimum-maximum), and qualitative data as number (percentage). Conformity to normal distribution was evaluated using the Shapiro-Wilk test. Whether there is a statistically significant difference between the "Bening" and "Malignant" groups, which are the categories of dependent / target variable (prostate cancer) in terms of independent variables, was examined using the Mann-Whitney U test and the independent samples t test. Values of p <0.05 were considered statistically significant. IBM SPSS Statistics 26.0 package program was used for all analyzes.

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.

Results

Descriptive statistics for the independent variables examined in this study are given in Table 3. There is a statistically significant difference between the dependent / target variable groups in terms of perimeter, area, compactness, symmetry, smoothness variables (p<0.05).

Table 3. Descriptive statistics for quantitative independent variables

Variables	Diagnosis				p-value
	Benign		Malign		
	Median(min-max)	Mean± Standard deviation	Median(min-max)	Mean± Standard deviation	
Radius	18 (9-25)	-	16 (9-25)	-	0.090*
Texture	17 (11-27)	-	18 (11-27)	-	0.450*
Perimeter	78.5 (52-133)	-	104 (72-172)	-	**<0.001***
Area	458.5 (202-1326)	-	790.5 (371-1878)	-	**<0.001***
Compactness	0.0785 (0.038-0.246)	-	0.1405 (0.051-0.345)	-	**<0.001***
Symmetry	0.182 (0.135-0.274)	-	0.193 (0.153-0.304)	-	**0.013***
Fractal_Dimension	0.0635 (0.053-0.09)	-	0.063 (0.053-0.097)	-	0.963*
Smoothness	0.099±0.015		0.105±0.014		**0.049****

*: Mann Whitney U test, **: Independent samples t-test

Classification matrix of MLPNN and RBFNN models are given in Table 4 and Table 5, respectively.

Table 4. Classification matrix of MLPNN model

Predicted \ Real	Malign	Benign	Total
Malign	14	2	16
Benign	0	11	11
Total	14	13	27

Table 5. Classification matrix of the RBFNN model

Predicted \ Real	Malign	Benign	Total
Malign	15	1	16
Benign	3	11	14
Total	18	12	30

In Table 6, the values of performance metric calculated from the models created to classify prostate cancer in the test stage are given below.

Table 6. Performance metric values calculated from created models in the testing stage

Model Performance Metrics	MLPNN Value	RBFNN Value
Accuracy (%)	92.5	86.6
Specificity (%)	84.6	91.6
AUC	0.937	0.921
Sensitivity (%)	100	83.3
Positive predictive value (%)	87.5	93.7
Negative predictive value (%)	100	78.5

AUC: Area under the ROC curve; MLPNN: Multilayer perceptron neural network; RBFNN: Radial-based function neural network

In Figure 1, values of performance metric obtained from MLPNN and RBFNN models are plotted.

In this study, while the significance values of the factors associated with prostate cancer are given in Table 7, the values for these significance percentages are shown in Figure 2.
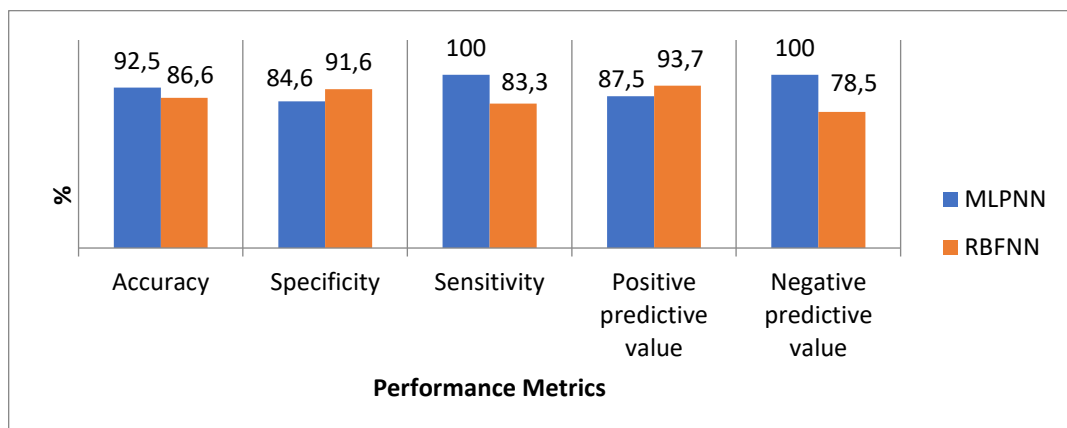

Figure 1. Performance metric values obtained from MLPNN and RBFNN models in the testing stage

Table 7. Importance values of explanatory variables according to MLPNN and RBFNN models

Explanatory Variables	MLPNN	RBFNN
radius	0.042	0.027
texture	0.054	0.036
perimeter	0.279	0.233
area	0.198	0.210
smoothness	0.087	0.106
compactness	0.189	0.187
symmetry	0.083	0.113
fractal_dimension	0.068	0.088
Total	1	1
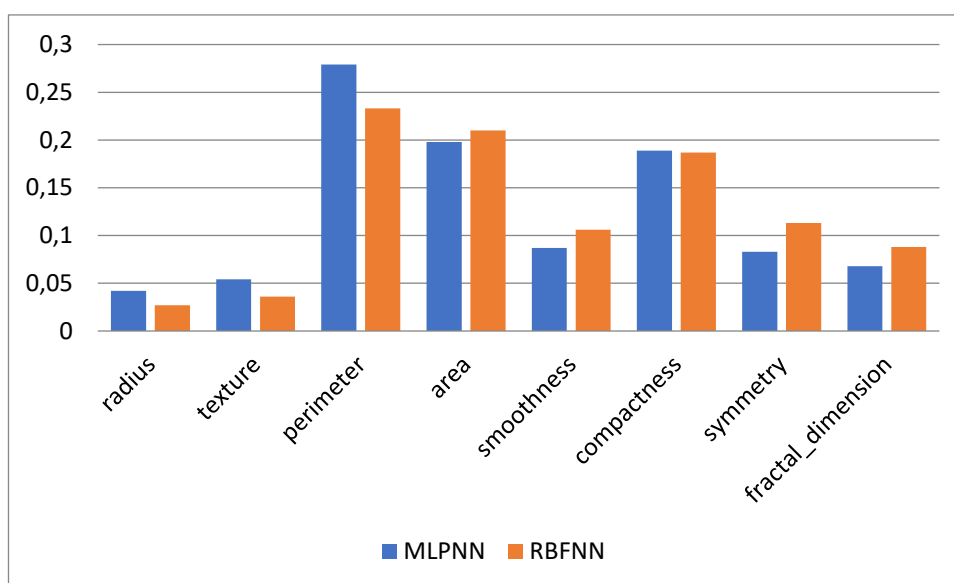
Figure 2. The importance values for possible risk factors

Discussion

Prostate cancer, in the world as of 2018 and is the second most common cancer in men in Turkey. The increase in the prevalence due to the aging of the population will bring along an increase in health expenditures. Considering its potential impact on health expenditures, it has become necessary to evaluate factors related to prostate cancer (Arslan and Esatoglu, 2018).

Artificial neural networks is a model that have the ability to obtain solutions for large and complex datasets, have no distribution requirements, can be applied to multivariate nonlinear problems, can detect complex nonlinear relationships between dependent and independent variables, can detect all possible interactions between predictive variables (Etikan et al. 2009). First, artificial neural networks perform training on the determined dataset. Second, the model examined is validated to determine the classification of a new dataset. The performance of the created

models is evaluated using different metrics (Nasser and Abu-Naser, 2019).

In this study, multilayer artificial neural network and radial-based artificial neural network models, which are among the artificial neural network models, were applied on an open access prostate cancer dataset and it was aimed to compare the classification estimates of these two models. In this context, different factors (explanatory variables) that may be associated with prostate cancer (dependent variable) have been estimated with multilayer artificial neural network and radial based artificial neural network models. Thus, it has been shown that artificial intelligence models can be used in the classification problem. In addition, significance levels of factors that may be associated with prostate cancer were obtained from these models for use in preventive medicine practices.

In this study, the MLPNN model gave better predictive results than the RBFNN model in the

classification of prostate cancer according to the performance criteria AUC, negative predictive value, and accuracy results used to compare classification performances. However, considering the positive predictive value and specificity criteria, the RBFNN model gave better predictive results than the MLPNN model. The three most important risk factors that may be associated with prostate cancer were obtained as perimeter, area and compactness according to MLPNN model. In the RBFNN model, the three most important risk factors that may be associated with prostate cancer were estimated as perimeter, area and compactness.

In a study using the same dataset, the accuracy results obtained with different machine learning methods were compared. According to the results of the current study, the highest accuracy was obtained as 0.80 with the k-Nearest Neighbor and Naive Bayes Classification models. In this study, an accuracy of 0.9 was obtained, and the rules related to the disease were also obtained (https://www.kaggle.com/alihantabak/prostate-cancer-predictions-with-ml-and-dl-methods).

Conclusion

In conclusion, considering the findings of this study, it was seen that classification of prostate cancer diagnosis gave successful predictions. At the same time, estimating the significance values of factors associated with the disease with the classification models used in this study made it different from similar studies with the same dataset.

References

Ah T. New markers and Phi score in prostate cancer. Turk Urol Sem, 2012; 3: 61-69.

Ari A & Berberler ME. Interface design for the solution of prediction and classification problems with artificial neural networks. Acta Infologica, 2017; 1: 55-73.

Arslan AGUDT & Esatoglu AE. Systematic Compilation of studies investigating radical prostatectomy costs and cost effectiveness. Journal of Academic Value Studies,2018; 4: 143-162

Badger TA, Segrin C, Figueredo AJ, Harrington J, Sheppard K, Passalacqua S, Pasvogel A & Bishop M. Psychosocial interventions to improve quality of life in prostate cancer survivors and their intimate or family partners. Quality of Life Research, 2011; 20: 833-844.

Efe O & Kaynak O. Artificial neural networks and applications. Istanbul: Bogazici University Publishing House, 2000

Elmas C. Artificial intelligence applications. Seckin Publishing,2016

Etikan I, Cumurcu BE, Celikel FC & Erkorkmaz U. Artificial neural networks method and classification of psychiatric diagnoses using this method. Turkey Journal of Medical Sciences, 2009; 29: 314-320.

Foster C, Bostwick D, Bonkhoff H, Damber JE, Van der Kwast T, Montironi R & Sakr W. Cellular and molecular pathology of prostate cancer precursors. Scandinavian Journal of Urology and Nephrology, 2000; 34: 19-43.

Gemici E, Ardiclioglu M & Kocabas F. Modeling of flow in rivers with artificial intelligence methods. Erciyes University Institute of Science Journal of Science, 2013; 29: 135-143.

Haykin S. Neural networks: a comprehensive foundation, Prentice-Hall Inc. Upper Saddle River, New Jersey, 1999; 7458: 161-175.

Haykin S. Neural networks: a comprehensive foundation. Prentice-Hall, Inc. 2007

Kayna O, Tastan S & Demirkoparan F. Estimation of crude oil prices by artificial neural networks.Ege Academic Review, 2010; 10.

Kaynar O, Gormez Y, Isık YE & Demirkoparan F. Intrusion detection with radial based artificial neural networks trained with different clustering algorithms.In International Artificial Intelligence and Data Processing Symposium (IDAP'16). 2016

Kim KB & Kim CK. Performance improvement of RBF network using ART2 algorithm and fuzzy logic system. In Australasian Joint Conference on Artificial Intelligence, 2004; 853-860. Springer.

King A, Evans M, Moore T, Paterson C, Sharp D, Persad R & Huntley A. Prostate cancer and supportive care: a systematic review and qualitative synthesis of men's experiences and unmet needs. European journal of cancer care, 2015; 24: 618-634.

Nasser IM & Abu-Naser SS. Lung cancer detection using artificial neural network. International Journal of Engineering and Information Systems (IJEAIS), 2019; 3: 17-23.

Orhan U, Hekim M & Ozer M. Discretization approach to EEG signal classification using multilayer perceptron neural network model. In Biomedical Engineering Meeting (BIYOMUT), 2010; 1-4.

Oztemel E. Artificial neural networks. Daisy Publishing, Istanbul. 2003

Poggio T & Girosi F. Regularization algorithms for learning that are equivalent to multilayer networks. Science, 1990; 247: 978-982.

Rumelhart DE, Hinton GE & Williams RJ. Learning representations by back-propagating errors. Nature, 1986; 323: 533-536.

Selcuk M. Modeling of Voice Quality in VoIP Network with Multi-Layered Artificial Neural Networks. European Journal of Science and Technology, 2020; 679-690.

Siegel RL, Miller KD & Jemal A. Cancer statistics. CA: a cancer journal for clinicians, 2019; 69: 7-34.

Soylemez Y. Prediction of Gold Prices Using Multi-Layer Neural Networks Method. Socioeconomics, 2020; 28: 271-291.

Yildirim AN. Artificial neural networks with threshold single multiplicative neuron models for the prediction problem. Master Thesis. Giresun University, Institute of Science,2020

Yildiz AK, Tasova M and Polatci H. The Use of Radial-Based Artificial Neural Networks in Modelling Drying Kinetics. Turkish Journal of Agriculture-Food Science and Technology, 2020; 8: 511-514.