THE JOURNAL OF COGNITIVE SYSTEMS
Journal Homepage: https://dergipark.org.tr/en/pub/jcs
Vol.7, No.2, 2022

Machine Learning Approach for Classification of Prostate Cancer Based on Clinical Biomarkers

Onural Ozhan[1], **Fatma Hilal Yagin**[2]

[1] Department of Medicinal Pharmacology, Faculty of Medicine, İnonu University, Malatya, 44210, Turkey (e-mail: onural.ozhan@inonu.edu.tr).

[2] Department of Biostatistics, and Medical Informatics, Faculty of Medicine, Inonu University, Malatya 44210, Turkey (e-mail: hilal.yagin@inonu.edu.tr).

ARTICLE INFO

Received: Aug.,03.2022
Revised: Oct, 11.2022
Accepted: Dec.,21.2022

Keywords:
Machine learning
Classification
Prostate cancer
Random forest

Corresponding author: Onural Özhan
✉ onural.ozhan@inonu.edu.tr
☎ +90 535 932 33 44

ISSN: 2548-0650

DOI:
https://doi.org/10.52876/jcs.1221425

ABSTRACT

In this study, it is aimed to classify cancer based on machine learning (ML) and to determine the most important risk factors by using risk factors for prostate cancer patients. Clinical data of 100 patients with prostate cancer were used. A prediction model was created with the random forest (RF) algorithm to classify prostate cancer. The performance of the model was obtained by Monte-Carlo cross validation (MCCV) using balanced subsampling. In each MCCV, two-thirds (2/3) of the samples were used to assess the significance of the feature. In order to evaluate the performance of the model, graph, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-score and Area under the ROC Curve (AUC) criteria including prediction class probabilities and confusion matrix were calculated. When the results were examined, the sensitivity, specificity, positive predictive value, negative predictive value, accuracy, F1-score, and AUC values obtained from the RF model were 0.89, 0.84, 0.77, 0.93, 0.86, 0.83, and 0.88, respectively. Area, perimeter, and texture were the three most important risk factors for differentiating prostate cancer. In conclusion, when the RF algorithm can be successfully predicted prostate cancer. The important risk factors determined by the RF model may contribute to diagnosis, follow-up and treatment researches in prostate cancer patients.

1. INTRODUCTION

THE abnormal division of cells in the prostate gland is one of the characteristics that define the type of cancer known as prostate cancer [1]. Research indicates that prostate cancer is the second most common form of cancer in men and the fifth leading cause of death on a global scale. On the other hand, it is the type of cancer that is diagnosed in more men over the age of middle age than any other type, in both developed and developing nations. If a man is between the ages of 40 and 59, his risk of developing prostate cancer is 2.58%, but between the ages of 60 and 79, his risk increases to 14.7%. The probability of developing prostate cancer in a man between the ages of 0 and 39 is only 0.01%. In addition, the likelihood of a man developing prostate cancer during the course of his lifetime is approximately 17.8% [2,3].

The predisposition of one's family to develop prostate cancer, as well as advanced age, race, genetics, diet, environmental factors, and hormonal factors, are all considered to be risk factors [4]. The correct management of the treatment, diagnosis, and follow-up process of prostate cancer is important not only for the patient and the doctor, but also for national health policies [5,6].
For this reason, the need for methods that can detect prostate cancer rapidly and accurately is increasing.

Machine learning, also known as ML, is a subset of artificial intelligence that identifies patterns in unprocessed data through the application of a specific algorithm or method. The primary objective of machine learning is to make it possible for computer systems to learn from experience on their own, without the need for explicit programming or intervention from humans. ML methods are frequently used in different areas of medicine and are less costly, more accurate and faster results in the diagnosis of different diseases. ML methods increase the predictive power thanks to their ability to combine data from various sources and manage large amounts of data [7-11].

Classification is one of the important tasks of ML. Classification includes approaches used to estimate the output variable when the output variable is categorical. The model obtained by using classification algorithms is used to predict

the unknown output variable when new data is obtained [12-14].

In this study, it is aimed to classify cancer based on ML and to determine the most important risk factors by using risk factors for prostate cancer patients.

2. MATERIAL AND METHODS

2.1. Data

The dataset to classify and predict prostate cancer in the study was obtained from https://www.kaggle.com/sajidsaifi/prostate-cancer. Of the patients in the data set, 62 (62%) were diagnosed as malignant and 38 (38%) as benign. Variables used to predict prostate cancer in the dataset: radius (mean distances from the center to perimeter points), texture (the standard deviation of grayscale values), perimeter (mean size of the core tumor), area, smoothness (mean of local variation in radius lengths), compactness ((mean of perimeter)2 / (area - 1)), symmetry and fractal dimension (mean for "coastline approximation").

2.2. Methods

2.2.1. Machine Learning Approach

The Random forest (RF) algorithm was used to predict prostate cancer in the study. The RF algorithm is among the ensemble classification methods created by Leo Breiman. Ensemble classification techniques are learning algorithms that produce more than one classifier instead of just one classifier and classify new data with the votes obtained as a result of the predictions of the classifiers produced [15,16].

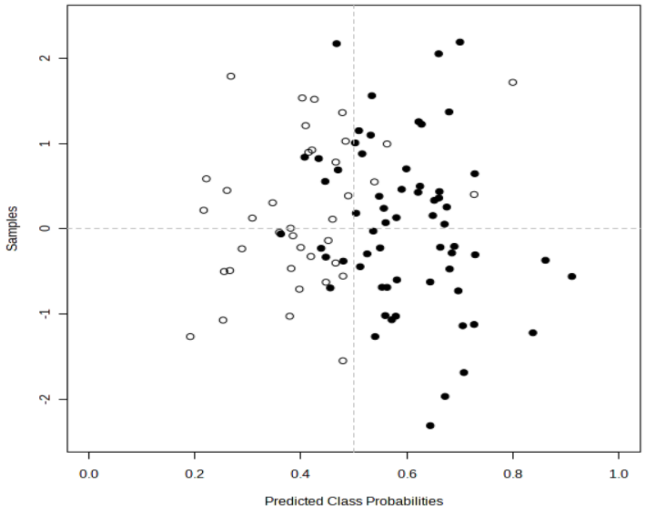
In the RF classification method, as in other ensemble learning methods, the performance values of weak learners (single decision tree, single sensor, etc.) are increased by a voting scheme. The classification method with the RF algorithm is based on the decision tree model. One of the advantages of the random forest algorithm is that it can use both continuous and discrete variables together. It can also be used in large or small size data sets [17,18].

The performance of the model was obtained by Monte-Carlo cross validation (MCCV) using balanced subsampling [19]. In each MCCV, two-thirds (2/3) of the samples were used to assess the significance of the feature. In order to evaluate the performance of the model, graph, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-score and Area under the ROC Curve (AUC) criteria including prediction class probabilities and confusion matrix were calculated.

3. RESULTS

In order to examine the performance of the RF model, confusion matrix with class probabilities is given in Figure 1. According to Figure 1, the model correctly classified (predicted) 52 of the 62 malignant patients and misclassified 10 patients.

Table 1 shows the results of the criteria related to the performance of the model, and Figure 2 shows the ROC curve. When the results were examined, the sensitivity, specificity, positive predictive value, negative predictive value, accuracy, F1 score, and AUC values obtained from the RF model were 0.89, 0.84, 0.77, 0.93, 0.86, 0.83, and 0.88, respectively



Confusion Matrix
(Cross-Validation)

	0	1
0	34	10
1	4	52

Fig. 1. Class probabilities and confusion matrix for model estimation (0: benign; 1 = malignant)

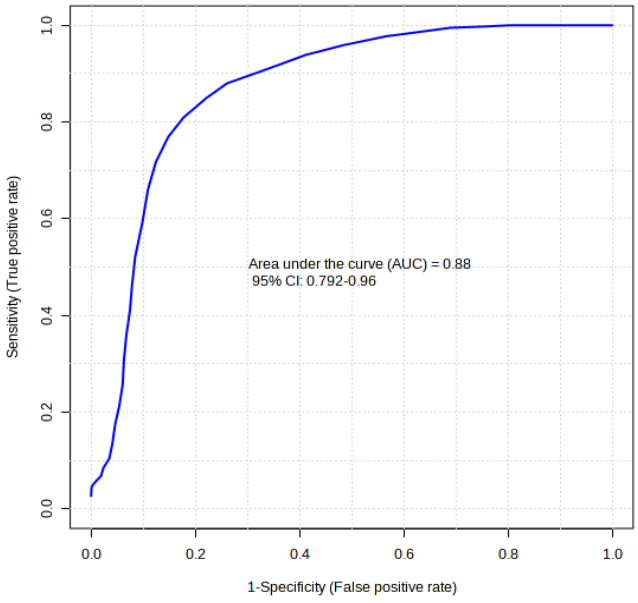


Fig. 2. Modelin performansına ilişkin ROC eğrisi

In Figure 3, the importance plot of the variables according to their contribution to the RF model created to predict prostate cancer is given. According to Figure 3, area, perimeter, and texture were the three most important factors to differentiate prostate cancer. In particular, the importance level of area was high compared to other risk factors.

TABLE I
The Performance of the Models

Metric	Value
Sensitivity	0.89
Specificity	0.84
Positive Predictive Value	0.77
Negative Predictive Value	0.93
Accuracy	0.86
F1 score	0.83
AUC	0.88

AUC: Area under the ROC Curve

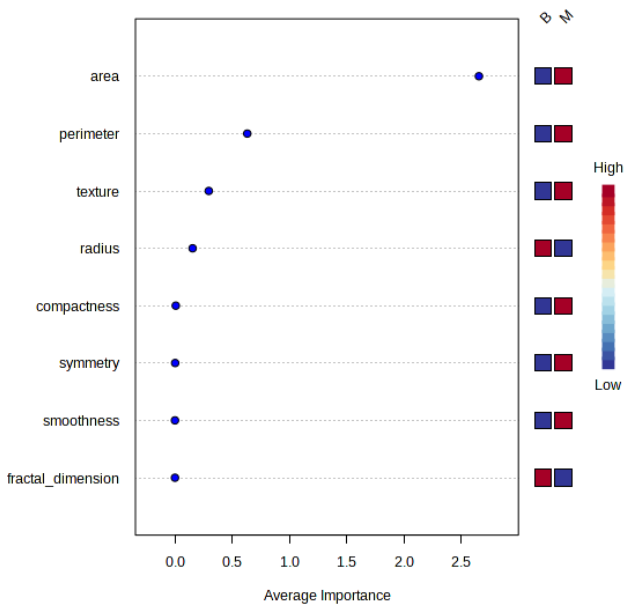


Fig. 3. Importance plot for clinical factors

4. DISCUSSION

The prostate cancer is a disease that can begin in any part of the prostate gland, progress slowly for the first five to ten years, and then rapidly grow after that and can also spread to other organs. Cancer of the prostate is a significant contributor to male health problems and deaths. Initiating factors for prostate cancer, such as genetic factors, chronic inflammation and infection, high-fat diet, smoking, alcohol use, and obesity, are not fully understood at this time. Prostate cancer is caused by a combination of factors. In terms of both incidence and mortality, prostate cancer ranks among the top five most common cancers in the world. Therefore, early detection of prostate cancer allows for the possibility of preventing the progression of the disease as well as the application of alternative treatment protocols [20,21].

ML methods have been used frequently for cancer detection and classification in recent years. Clinical decision support systems developed based on ML can help clinicians in the pre-diagnosis, follow-up and treatment of diseases [22,23].

This research focused on the prediction of prostate cancer with ML methods, which is one of the most common causes of cancer-related death in men and shows symptoms similar to benign enlargement. Diagnosing diseases is one of the most challenging aspects of the medical field. The fact that there are no established guidelines for evaluating prostate cancer symptoms and that the diagnostic methods that are currently available have poor predictive rates makes this study extremely valuable. In situations where there are no hard-and-fast rules to follow but where the factors that will influence an event can be anticipated, such as in the case of prostate cancer, we believe that methods of machine learning may be useful in making accurate predictions.

Based on this, the RF algorithm, which is one of the supervised machine learning methods and contributed to the creation of high-performance models, was used in this research to evaluate the accuracy of the prediction of prostate cancer.

In the model, patients were assigned to one of the output classes based on the class probabilities, thus obtaining the confusion matrix. The sensitivity, specificity, positive predictive value, negative predictive value, accuracy, F1 score, and AUC values obtained from the RF model were 0.89, 0.84, 0.77, 0.93, 0.86, 0.83, and 0.88, respectively. Our results showed that the RF model could successfully predict prostate cancer. In addition, the importance of the clinical features examined in order to distinguish prostate cancer was examined in the study. Our results showed that area, perimeter, and texture are the most important features in differentiating prostate cancer.

In a study using the same data set in the literature and comparing the results using various ML approaches, the highest classification accuracy was obtained with the k-nearest neighbor and naive bayes methods [24]. The classification rate accuracy obtained with the optimal model of this study was found to be 0.80. Another study using the same dataset compared the performance of some popular ML methods to predict prostate cancer. The authors achieved the best performance with the Recurrent Neural Network (RNN) model with an accuracy rate of 0.813 [25]. In our study, we classified prostate cancer with an accuracy of 0.86, and the RF model had the ability to discriminate quite well.

As a result, the proposed RF model can successfully classify prostate cancer and the model can help clinicians to pre-diagnose prostate cancer.

5. CONCLUSIONS

In conclusion, prostate cancer risk can be successfully predicted with methodology combined using clinical information and RF algorithm. Furthurmore, the model created with the RF algorithm can help clinicians in the diagnosis, follow-up and treatment of patients with prostate cancer.

REFERENCES

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6), 394-424.

[2] Jemal, A. (2005). murray t, Ward e, samuels A, tiwari RC, Ghafoor A, Feuer eJ, thun mJ. Cancer statistics, 10-30.

[3] Rawla, P. (2019). Epidemiology of prostate cancer. World journal of oncology, 10(2), 63.

[4] Jemal, A., Thomas, A., Murray, T., & Thun, M. (2002). Cancer statistics, 2002. Ca-A Cancer Journal for Clinicians, 52(1), 23-47.

[5] Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. CA: a cancer journal for clinicians, 69(1), 7-34.

[6] Dimakakos, A., Armakolas, A., & Koutsilieris, M. (2014). Novel tools for prostate cancer prognosis, diagnosis, and follow-up. BioMed research international, 2014.

[7] Yağin, F. H., Yağin, B., Arslan, A. K., & Çolak, C. (2021). Comparison of Performances of Associative Classification Methods for Cervical

Cancer Prediction: Observational Study. Turkiye Klinikleri Journal of Biostatistics, 13(3).
[8] Deo RC. (2015). Machine learning in medicine: Circulation, 132(20), 1920-30.
[9] Sidey-Gibbons, J. A., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction: BMC medical research methodology, 19(1), 1-18.
[10] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction: Computational and structural biotechnology journal, 13, 8-17.
[11] Richter, A. N., & Khoshgoftaar, T. M. (2018). A review of statistical and machine learning methods for modeling cancer risk using structured clinical data: Artificial intelligence in medicine, 90, 1-14.
[12] Paksoy, N., & Yağin, F. H. (2022). Artificial Intelligence-based Colon Cancer Prediction by Identifying Genomic Biomarkers: Medical Records, 4(2), 196-202.
[13] Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques: Artificial Intelligence Review, 26(3), 159-190.
[14] Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: applications and issues: Journal of Basic & Applied Sciences, 13, 459-465.
[15] Yilmaz, R., & Yağin, F. H. (2022). Early detection of coronary heart disease based on machine learning methods: Medical Records, 4(1), 1-6.
[16] Khan, M. A., Memon, S. A., Farooq, F., Javed, M. F., Aslam, F., & Alyousef, R. (2021). Compressive strength of fly-ash-based geopolymer concrete by gene expression programming and random forest: Advances in Civil Engineering, 2021.
[17] Gupta, V. K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model: Big Data Mining and Analytics, 4(2), 116-123.
[18] Palimkar, P., Shaw, R. N., & Ghosh, A. (2022). Machine learning technique to prognosis diabetes disease: random forest classifier approach Advanced Computing and Intelligent Technologies: Springer, 219-244.
[19] Shan, G. (2022). Monte Carlo cross-validation for a study with binary outcome and limited sample size: BMC Medical Informatics and Decision Making, 22(1), 1-15.
[20] Gandaglia, G., Leni, R., Bray, F., Fleshner, N., Freedland, S. J., Kibel, A., . . . La Vecchia, C. (2021). Epidemiology and prevention of prostate cancer: European urology oncology.
[21] Habib, A., Jaffar, G., Khalid, M. S., Hussain, Z., Zainab, S. W., Ashraf, Z., . . . Habib, P. (2021). Risk Factors Associated with Prostate Cancer: Journal of Drug Delivery and Therapeutics, 11(2), 188-193.
[22] Yağin, F. H., Güldoğan, E., Ucuzal, H., & Çolak, C.(2021). A Computer-Assisted Diagnosis Tool for Classifying COVID-19 based on Chest X-Ray Images: Konuralp Medical Journal, 13(S1), 438-445.
[23] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine: New England Journal of Medicine, 380(14), 1347-1358.
[24] https://www.kaggle.com/alihantabak/prostate-cancer-predictions-with-ml-and-dl-methods.
[25] Laabidi, A., & Aissaoui, M. (2020). Performance analysis of Machine learning classifiers for predicting diabetes and prostate cancer: Paper presented at the 2020 1st international conference on innovative research in applied science, engineering and technology (IRASET).

BIOGRAPHIES

Onural Ozhan obtained his BSc. degree in pharmacy from Inonu University in 2009. He received MSc. degree in pharmacognosy from Ankara University in 2012. He received Ph.D. degree in medical pharmacology from Inonu University in 2019. In 2015, he joined the department of medical pharmacology at Inonu University as a research assistant. Currently, he continues to work as an Asst. Prof. at Inonu University. His research interests are the cardiovascular system, oxidative stress, and ischemia-reperfusion injury.

Fatma Hilal Yagin obtained her BSc. degree in Statistics from Gazi University in 2017. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2020. She currently continues Ph.D. education in biostatistics and medical informatics from the Inonu University. In 2019, 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, and image processing.