Int. J. Pure Appl. Sci. 9(2);274-281 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1382974

Evaluation of Prostate Cancer via Machine Learning

Fatma Söğüt[1]*, Evrim Ersin Kangal[2]
[1]Mersin University, Vocational School of Medical Services, Department of First Aid and Emergency, Mersin, Türkiye
[2]Mersin University, School of Applied Technology and Management of Erdemli, Computer Technology and Information Systems, Mersin, Türkiye
**fatmasogut@mersin.edu.tr, evrim.kangal@mersin.edu.tr*

Received date:29.10.2023, Accepted date: 01.12.2023

Abstract

By training computers with machine learning technique, patients can be prevented from being exposed to unnecessarily difficult examinations. In recent years, machine learning-based disease assessment approach has gained importance in terms of the benefits it provides to clinical methods. There is a remarkable increase in studies in this direction. There are a limited number of clinical guiding parameters in predicting some types of cancer, and this limitation pushes the patients under treatment to a very frustrating process. For this reason, apart from ordinary procedure of the traditional medicine, an alternative approach to predict the any type of cancer is making a computer-based evaluation that has become a highly studied method in recent years. In this study, a machine learning (ML) approach will be used to evaluate prostate cancer, which is the second most common cancer-related death in men worldwide. For this purpose, the K-Nearest Neighbor (kNN) algorithm based on ML will be used with feature selection, which is a dimension reduction technique. An open source data science platform, Kaggle, was used for the evaluation. The accuracy value of the used algorithm was obtained as 88%. Thus, the performance of the algorithm developed for predicting prostate cancer has an effective value. This result is above the acceptable accuracy limits of the approach followed in predicting the disease, based on studies in the literature. Additionally, in our ML approach, identifying and excluding trivial parameters used in diagnosis also increased the efficiency of our prediction model.

Keywords: Prostate cancer, machine learning, feature selection kNN.

Prostat Kanserinin Makine Öğrenimi Yoluyla Değerlendirilmesi

Öz

Bilgisayarların makine öğrenimi tekniği ile eğitilmesi ile hastaların gereksiz yere zor tetkiklere maruz kalması engellenebilir. Son yıllarda makine öğrenimi tabanlı hastalık değerlendirme yaklaşımı, klinik yöntemlere sağladığı faydalar açısından önem kazanmıştır. Bu yönde yapılan çalışmalarda dikkat çekici bir artış vardır. Bazı kanser türlerini öngörmede sınırlı sayıda klinik yol gösterici parametre vardır ve bu kısıtlılık tedavi gören hastaları oldukça yıpratıcı bir sürece itmektedir. Bu nedenle, geleneksel tıbbın alışılagelmiş prosedürlerinden farklı olarak, herhangi bir kanser türünü tahmin etmede alternatif bir yaklaşım, son yıllarda üzerinde çok çalışılan bir yöntem haline gelen bilgisayar tabanlı değerlendirme yapmaktır. Bu çalışmada, dünya çapında erkeklerde ikinci en yaygın kansere bağlı ölüm olan prostat kanserini değerlendirmek için bir makine öğrenimi (ML) yaklaşımı kullanılacaktır. Bu amaçla bir boyut küçültme tekniği olan öznitelik seçimi ile ML için K-En Yakın Komşu (kNN) algoritması kullanılacaktır. Değerlendirme için açık kaynaklı bir veri bilimi platformu olan Kaggle kullanılmıştır. Kullanılan algoritmanın doğruluk değeri %88 olarak elde edilmiştir. Böylece prostat kanserini tahmin etme için geliştirilen algoritmanın performansı etkili bir değere sahiptir. Bu sonuç, hastalığı tahmin etmede izlenen yaklaşımın, literatüreki çalışmalar esas alındığında kabul edilebilir doğruluk sınırlarının üzerinde bir değerdir. Ayrıca ML yaklaşımımızda, tanıda kullanılan önemsiz parametrelerin belirlenerek dışlanması öngörme modelimizin verimini de arttırmıştır.

Anahtar Kelimeler: Prostat kanseri, makine öğrenimi, özellik seçimi, kNN.

INTRODUCTION

Prostate cancer (PCa) is the most common type of cancer in men and the second leading cause of cancer-related deaths. The prostate is a glandular organ in men that produces various secretions for reproductive activities. PCa occurs when some cells that make up the prostate tissue form abnormal tumor structures. While some responsible gene groups have been identified for the causes of PCa, obesity and tobacco use are also risk factors among the causes of PCa. Delay in the diagnosis and treatment of PCa may cause the cancer to spread to neighboring organs and tissues. Involvement may be seen in nearby lymph nodes, bones, and adjacent organs. Currently, there is no early diagnosis method that can reveal that a healthy man will have PCa. With the use of some advanced technological methods (such as Multiparametric magnetic resonance (MR), MR-Ultrasound Fusion Biopsy), it is often diagnosed with a PSA (Prostate Specific Antigen) test, which is examined with a blood sample, or by prostate biopsy (Mydlo *et al.,* 2016; Anand et al., 2023). However, the PSA test can cause pointless, expensive and painful needle biopsies. Multiparametric magnetic resonance imagining (MRI), which is used in radiological settings and depends on diffusion-weighted magnetic resonance imaging (DWI), has become a standard technique that is increasingly applied in the diagnosis of PCa, but still creates various problems due to the differences in the variables seen among patients (Yoo et al., 2019).

The lack of clear rules in the evaluation of PCa symptoms and the inadequacy of current diagnostic methods to predict the disease made us think that ML methods may be effective in predicting the disease in addition to medical methods. The ML is a field of artificial intelligence (AI) based on the idea that a computer system uses data from a large pool of data to learn a model and make decisions or predictions on new data. Conventional feature engineering relies on extracting countable imaging features such as shape, odor, color, volume, density from a raw data from imaging data and then using an ML classifier such as Support Vector Machines (SVM), Adaboost, and Decision Trees (DT). In recent years, deep learning methods have yielded important results for various medical imaging tasks, including histopathological image processing in computer-based vision with various stages such as segmentation, classification, and object detection (Coiture et al., 2018; Qaiser et al., 2019). Computer-based deep convolutional neural networks (CNNs) (Bektaş, 2022) have achieved significant success in the processes such as object detection and segmentation, as well as imaging and evaluation of objects with a computer, making them one of the indispensable methods in the detection of difficult diseases such as cancer. For this reason, more attention has been given to the development of computer-based evaluation algorithms for medical imaging in recent decades and there are increasingly studied different CNN architectures in this direction (Yoo et al., 2019).

It is thought that a system in which the computer can make decisions based on various data obtained from the patient may be more successful in predicting cancer. From this point of view, the aim of this study is to use the kNN method, which is a single-supervised machine learning algorithm, to predict PCa. Using data from Kaggle, he will be able to predict the diagnosis of the disease by learning the relationship between PCa-specific parameters of his system. In this way, there will be no need for some wearing processes that the patient will be exposed to during the diagnosis process.

The paper is organized as follows. In section 2, we review the related works that have tried to estimate PCa. Section 3 explains the methodology and selected data pool. In section 4, we give the results and related remarks. In section 5, we discuss the obtained results and compare them with the literature. Finally, we give conclusion of the study in section 6.

PREVIOUS STUDIES

ML prediction has been applied to classify PCa using different medical datasets (Goldenberg et al., 2019). In a study performed by Wang et al. (Wang et al., 2022), they compared the "diagnostic performance" of their ML studies with MRI to construct a ML prediction model for PCa using transrectal ultrasound video clips of the prostate gland. They selected 14 features and used SVM and random forest (RF) algorithms to generate radiomic models based on these features. In addition, they proposed a machine learning model assisted diagnosis algorithm (MLAD) consisting of SVM, RF and MRI-based diagnosis of radiologists to evaluate the performance of ML models in computer-assisted diagnosis (CAD) and evaluated MRI-based radiologist diagnosis. Srivenkatesh (Srivenkatesh , 2020) used ML to predict PCa using the Kaggle dataset. In his study, he proposed a model to predict PCa. He compared the accuracy of the algorithms of SVM, Random forest, Naive Bayes (NB) classifier and logistic regression (LR) on the dataset from Kaggle. Yoo et al. were used (Yoo et al., 2019) diffusion-weighted data from 427 patients for the detection of clinically significant PCa. They developed an automated CNN-based pipeline using magnetic resonance imaging (DWI) images. The performance of their method when applied to an independent test data outperformed other studies that introduced CAD tools using CNNs for the detection of PCa. In a study performed by Valero et al., (Valero et al., 2022) they present an automatic system for evaluating multi-parametric MRI images by using Deep Learning. The system performs localization, segmentation and GGG estimation of PCa lesions from prostate multi-parametric MRIs. They use a specific instance detection and segmentation network called as the 3D Retina U-Net27 for this study.

Zhang et al. developed a new approach for diagnosing the PCa through magnetic resonance imaging (Zhang et al., 2021). They studied with advenced GrowCut algorithm for segmentation of the questionable cancer area and the combined ML algorithms like kNN, SVM and MLP (Multilayer perceptron) in the combined learning system to determine PCa. They have obtained the accuracy of ensembled methods has improved nearly 20% compared to traditional methods.

Araujo et al. proposed a new method to identify patients at risk of PCa. The method, which is developed by using various medical records of the patient, was tested via machine learning algorithms such as SVM, NB, kNN, DT and ANN and the presence of PCa were evaluated based on the available parameters and determined that the best performance was achieved through the linear SVM model (Araujo et al., 2023).

Urologists use Partin tables based on LR to evaluate PCa. For this purpose, the pathological stage estimates given in these tables are used. Coudert et al adapted Partin tables to patients in the UK. They also compared the performance of various classifiers with each other using the same data (Coudert, 2012).

MATERIAL AND METHODS

We present an algorithm to predict the PCa from the data obtained at Kaggle that includes well documented and good quality datasets. In line with the results obtained from the correlation table, interrelated features were removed from the dataset. Then, the kNN classifier working with the wrapper framework is run for this reduced dataset. The flow diagram of a such procedure can be summarized by Fig. 1.

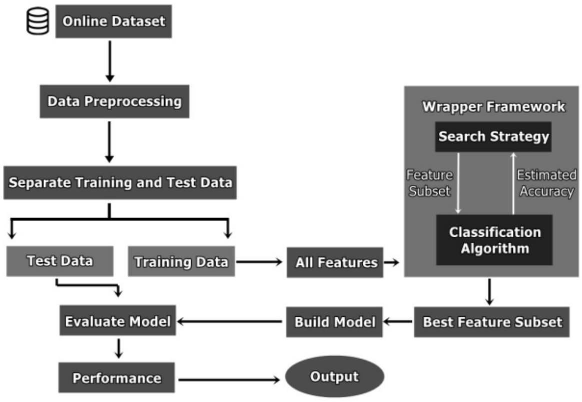


Figure 1. Schematic diagram of the classification process of PCa based on ML.

Dataset

In the process of the classification of dataset to diagnose of the PCa, a dataset is retrieved from the Kaggle (Sajid, 2018).

Separating an image into smaller parts that form it, as close to its true value as possible, is one of the most important stages of the process. The fact that the cells that make up the data are heterogeneous and that their size, format and image density properties vary greatly makes this process even more difficult. One of the approach is a quasi-automatic segmentation procedure that is known as snakes. Starting with an approximate boundary as an initialization, the snake points the actual boundary of the cell nucleus. Contrary to other segmentation methods, snakes can approach boundaries even in areas of little or no gray-scale contrast.

In order to evaluate the cancer existence, pathologists try to identify the size, shape and texture of the cell nuclei. There are ten variables computed for the dataset (radius, texture, perimeter, area, smoothness,compactness, symmetry, fractal dimension, diagnosis_result and ID) (Street et al., 1993). The dataset belongs to 100 patients and data-

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

types of attributes are “numerical” except the “diagnosis result” which is categorical.

Framework of the analysis

The class asymmetry is a case which requires to be handled carefully in data mining studies. Commonly, the classification algorithms achieve investigation by accepting that the raw data has been tuned, but this is not the usual case. While some of the classes in the data set have enough elements, some may have very few elements. In these cases, incorrect predictions may be made as a result of the algorithm because the classification algorithm will be inadequately trained for elements for which it has limited characteristic information. The aim of classification algorithms is to boost the useful estimation rate. For this reason, we need to tune the distribution of aspects in a convenient way. Therefore, in Fig. 2, we discuss distribution of the results given in the repository of Kaggle and see that the number of “Benign (B) ” cases is much higher than the “Malignity (M)” ones. This status can affect adversely the performance of approach; thence we need to generate new “Malignity” situations via a satisfactory Python 3.6 package to cancel this issue.

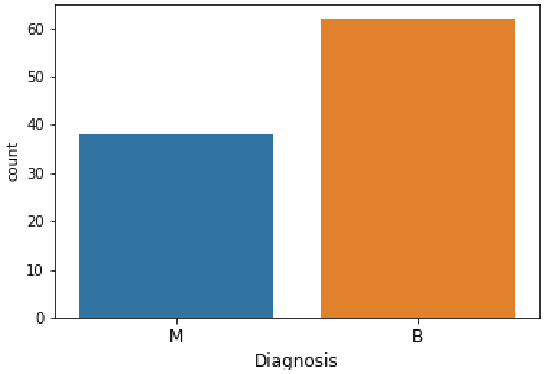


Figure 2. The number of “Benign” and “Malignity” cases.

The most commonly used data classification algorithms based on the ML can be sorted as the Decision Trees (DTs), Linear Regression (LR), Logistic Regression (LoR), Support Vector Machines (SVM), Naive Bayes Classifier (NBC), Neural Networks (NN), Instance-based Learning (IBL) and the k-Nearest Neighbors (kNN). Among them, a widely known method of machine learning algorithm to predict treatment approaches and outcomes for disease using various medical data is the kNN method and can be used to classify various data. The “k-means” clustering approach is an unsupervised algorithm that iteratively divides data into “k-numbers”. Here k is the number of repetitions. Each point in the feature space is given an identifier corresponding to the k-center number that is spatially closest to it. In the next stage, a new mean is determined for each cluster and the locations of the centroids of each cluster are changed to be consistent with the new mean. The process of assigning and renewing centers continues until no changes occur. Thus, the number of classes forming a cluster is denoted by the letter k. The most important thing to keep in mind when using the kNN method is that it always produces the same results, even when using the same learning data (Anand, 2023).

The large number of variables to be used in the analysis process negatively affects the algorithm in the decision-making process and also increases the compilation time of the process. One method to address this issue is the feature selection (FS) approach. In this way, the negative features that most affect the decision-support mechanism can be determined and removed from the dataset. For this purpose, we will use the "wrapper based supervised FS (WBS-FS)" method in this study. In this method, "feature subsets" are created in the first stage, then the results of each selected subset are compared with random selections from the feature space (Elkhani, 2017). The analysis will be performed by Sequential Forward Selection (SFS) wrapper in which a universal attribute pool is generated by adding the attributes one by one to the dataset.

Visualization of Data

As mentioned above, an unbalanced data set creates serious weaknesses in the performance of the algorithm being studied at the decision-support stage. There are two ways to fix this problem: under-sampling and over-sampling. For this purpose, we can use the SMOTE (Synthetic Minority Oversampling TEchnique) approach (Torgo,2013).

Depending on the case presented in Fig. 2, we create sufficient synthetic “Benign” cases for the attribute “Diagnosis” with the help of the SMOTE. After this task, we plot Fig. 3, which shows that number of “Benign” and “Malignity” cases are now balanced.

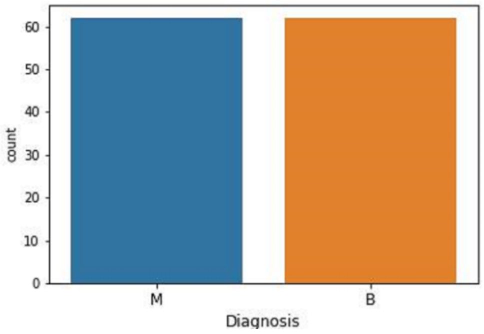


Figure 3. The number of “Benign” and “Malignity” cases after the SMOTE.

Correlation between features is also one of the factors that reduce the performance of the analysis. The correlation table consists of rows and columns in which the variables of the data set are represented. If the correlation between any two features is in the range of [-1,-0.5] then it is labeled as negative correlation. Also, positive correlation between any two attributes requires the correlation coefficient in the range of [0.5, 1]. Correlation coefficients have the value 1 on the main diagonal. In this case, it is understood that each variable always has a perfect correlation with itself, but the other correlation values are not correlated. Fig. 4 is provided to show the positive and negative correlations between the features of our dataset. The performance of the algorithm still remains in the desired state, since the attributes are unrelated.

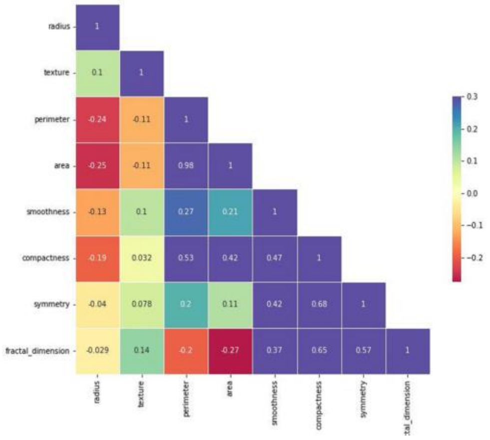


Figure 4. The correlation coefficients for all features.

By using WBS-FS approach, the features negatively affecting the analysis were removed from the data set for cases where the positive correlation was greater than (0.5) and the negative correlation was less than (-0.5). In this case, the correlation table of the remaining features is given by Fig. 5.

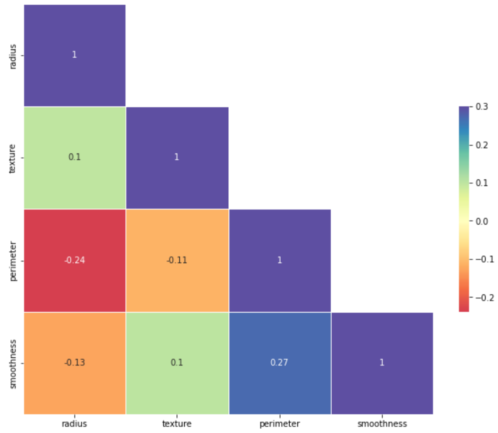


Figure 5. The correlation coefficients for important features.

To see how the averages of the remaining parameters change relative to each other depending on whether the cancer is benign or malignant, the value ranges of these parameters are given in Fig. 6, along with their standard deviations.

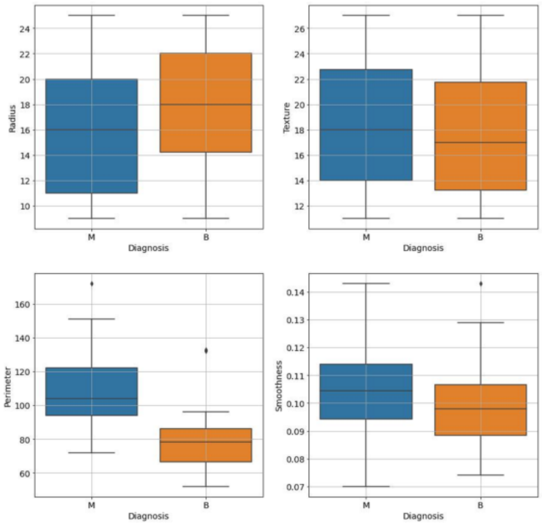


Figure 6. The dynamic of remaining features versus diagnosis.

When the four situations given in the graph are examined, it is clearly seen that the averages of each feature are different from each other for benign or malignant tumors. The algorithm uses the difference when evaluating.

RESULTS AND DISCUSSION

Results

When we started the analysis, we had Kaggle dataset for 100 patients. This number of data was increased to 124 after the SMOTE process. Therefore, analysis was carried out with a total of 124 data. Of the 124 data of patients counted in this study, 30% data of the patients (which belongs to 38 data) was used randomly for testing and 70% (belongs to 86 data) for training to verify classification accuracy and performance.

The confusion matrix (CM) is a measure for the performance evaluations of the methods used for binary and multiclass classification (Kulkarni, 2020). It gives counts from predicted and actual values. The popular metric criteria for performing classification is accuracy. The accuracy of a model (through a CM) is obtained by using the following formula:

$$Accuracy = \frac{TN+TP}{TN+FP+FN+TP} \tag{1}$$

The abbrevations TN means True Negative (gives the number of negative examples classified accurately), TP is for True Positive (and presents the number of positive examples classified accurately), FP stands for False Positive value (the number of actual negative examples classified as positive) and FN means a False Negative value (the number of actual positive examples classified as negative).

As seen in the “Accuracy” plot from Fig. 7., the algorithm shows a success rate of detecting malignant tumors at a level of 88% .

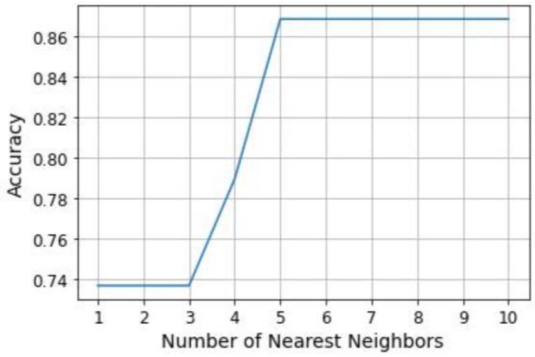
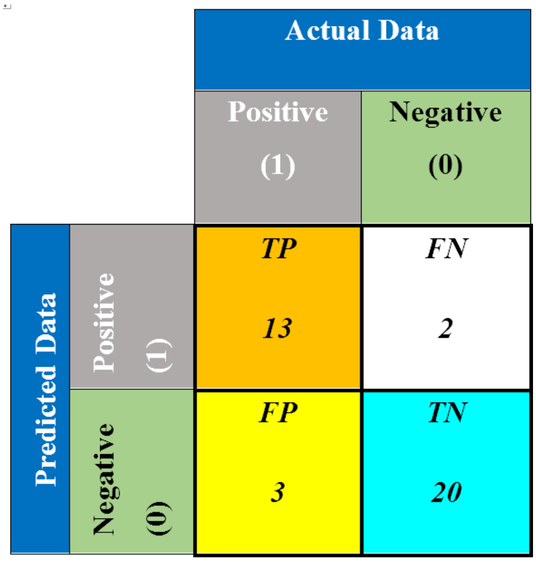


Figure 7. Number of nearest neighbors used in the classification.

The algorithm is run for the test data by selecting the optimized number of nearest neighbors as five, and the resulting CM matrix is given as Fig. 8.



		Actual Data	
		Positive (1)	Negative (0)
Predicted Data	Positive (1)	*TP* *13*	*FN* *2*
	Negative (0)	*FP* *3*	*TN* *20*

Figure 8. For the binary classification, the framework of the CM is presented.

Discussion

The public Kaggle PCa dataset is available online since 2018 and has been used by some studies in the literature (Erdem, 2020). We also used this data for a ML investigation of PCa prediction. The dataset belogs to 100 patients. Since the small number of data may reduce the prediction performance of the algorithm we aim to develop, the number of data was increased with the help of SMOTE before creating our model. SMOTE can balance the class ratio by creating new samples from existing data.

Moreover, in our study, we evaluated which clinical features used to differentiate PCa were more meaningful. According to the data we obtained as a result of the examination, radius, perimeter, smoothness and texture were the most important features in the evaluation of PCa.

After obtaining the convenient data rate, the classification is performed by using kNN algorithm. We obtained the accuracy rate for trained data as %88 and %87 (according to usage of CM table in Eq. 1) for the test data. According to this result, it can be thought that if the computer is trained based on the information obtained from the patient, the developed algorithm will be useful to clinicians in predicting cancer. With the help of this method, patients will not need to undergo some tiring medical procedures.

After obtaining the cross validation (CV) reports of the remaining 4 features, the average accuracy score is 0.826 only for perimeter feature. Even if radius and smoothness are added to this situation, the average accuracy value remains the same. When the texture feature is included in these features, the success rate drops to 0.802. Therefore, it is clear that the perimeter feature alone is sufficient for decision making. The reason for this is that the shape of the benign tumor is symmetrical, while the shape of the malignant tumor is asymmetrical.

The Kaggle dataset we used in this study was also used in another study (Erdem, 2020). However, in that study, the unbalanced data set was not converted to balanced data using SMOTE. Additionally, the feature selection method was not used and they have used 10 parameters in the analysis. However, as we stated in this study, 6 of these 10 parameters negatively affect accuracy performance. The authors ignored this situation. Additionally, when we look at the CM produced in this study, it is understood that we worked with 200 data in total. However, it was stated that 100 data were taken from Kaggle in the study. While an 83% accuracy rate was achieved, confusion arose about how much data was used in total. As a result of SMOTE and WBS-FS processes, the classification rate accuracy value for our study was obtained as 88%, while in that study this value was obtained as 83%. The important thing to consider here is the numbers in the data set. In that study, the authors also used 10 ML algorithms. In some of the algorithms, they achieved very high values. However, as we mentioned above, when SMOTE and WBS-FS are applied to other methods other than kNN, different and meaningful values will be obtained for these accuracy values.

Finally, we can say that if the obtained 4 features with the correlation approach are used as input data in the developed algorithm, an online PCa risk assessment can be made with this model. An online risk calculation model has also been developed by Wang et al. (Wang et al., 2022).

CONCLUSION

PCa is one of the most common causes of death in men in later ages. When cancer is diagnosed early, starting the treatment process as soon as possible will increase the patient's chance of survival. Medical technologies that have developed in recent years are widely used to detect the disease in the light of expert evaluations. However, when traditional approaches are used, the objective evaluation of the process creates difficulties in some areas, as the medical data evaluated can vary widely between patients. For this reason, in recent years, computer-aided diagnosis attempts have been made to evaluate complex data in the medical field. Therefore, developing a model to predict PCa with the help of ML based on clinical data is very important to generate low-cost medical expenses and increase early diagnosis rates.

In this work we put our attention to develope a ML model for predicting the PCa by using the Kaggle dataset. We used the kNN method to obtain the accuracy performance of the model. With the algorithm developed in this study, a PCa prediction rate of 88% accuracy was achieved.

CONFLICT OF INTEREST

The Authors report no conflict of interest relevant to this article.

RESEARCH AND PUBLICATION ETHICS STATEMENT

The authors declare that this study complies with research and publication ethics.

REFERENCES

Anand L., et al. (2023). Diagnosis of Prostate Cancer Using GLCM Enabled KNN Technique by Analyzing MRI Images, BioMed Research International, Article ID 3913351.

Araujo W. B. D., et al. (2023). Method to aid the diagnosis of PCa using machine learning and clinical data, PREPRINT (Version 1), Research Square.

Bektaş J. et al. (2022). Integrating a novel SRCRN network for segmentation with representative batch-mode experiments for detecting melanoma. Biomedical Signal Processing and Control 71(B): 103218.

Coudert O. R. et al. (2012). Machine learning for improved pathological staging of prostate cancer: A performance comparison on a range of classifiers, Artificial Intelligence in Medicine, Volume 55, Issue 1, 25-35.

Couture H. D. et al. (2018). Image analysis with deep learning to predict breast cancer grade, er status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer 4(1):30

Elkhani N. and Muniyandi R.C. (2017). Intell. Data Anal. 21, S137.

Erdem E., and Bozkurt F. (2021). A Comparison of Various Supervised Machine Learning Techniques for Prostate Cancer Prediction, Eur. J. Sci. Tech., 21, 610-620.

Goldenberg, S. L., Nir, G. and Salcudean, S. E. (2019). A new era: artificial intelligence and machine learning in prostate cancer. Nature Reviews Urology, 16(7), 391-403.

Kulkarni A., Chong D. and Batarseh F. A. (2020). 5 Foundations of data imbalance and solutions for a data democracy, Editor(s): Feras A. Batarseh, Ruixin Yang, Data Democracy, Academic Press, 83-106.

Mydlo J. H. And Godec C. J. (2016). Prostate Cancer, Academic Press.

Pellicer-Valero, O.J. et al. (2022). Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of PCa in multiparametric magnetic resonance images. Sci Rep 12, 2975.

Qaiser T, Rajpoot N. M. (2019). Learning where to see: a novel attention model for automated immunohistochemical scoring. IEEE Trans Med Imag 38(11), 2620–2631.

Regnier-Coudert O., et al. (2012). Machine learning for improved pathological staging of prostate cancer: A performance comparison on a range of classifiers, Artificial Intelligence in Medicine, 55(1), 25–35.

Sajid S. (2018). Prostate cancer dataset, [Online]. Available: https://www.kaggle.com/sajidsaifi/prostate-cancer

Srivenkatesh, M. (2020). Prediction of Prostate Cancer using Machine Learning Algorithms, International Journal of Recent Technology and Engineering. 8 (5).

Street W.N., Wolberg W.H. and Mangasarian O.L (1993). Nuclear feature extraction for breast tumor diagnosis. International Symposium Electronic Imaging: Science and Technology; 1-4; San Jose, CA, USA. vol. 1905, 861-870.

Torgo L., et al. (2013). SMOTE for Regression, Progress in Artificial Intelligence, Lec-ture Notes in Computer Science, vol 8154, Springer, Berlin, Heidelberg.

Valero P. et al. (2022). Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. Scientific reports, 12(1), 2975.

Wang K. et al. (2022). Machine learning prediction of PCa from transrectal ultrasound video clips. Front. Oncol. 12, 948662.

Wang C. et al. (2022). PCa Risk Prediction and Online Calculation Based on Machine Learning Algorithm, Chin Med Sci J., 37(3), 210-217.

Yoo, S. et al. (2019). PCa Detection using Deep Convolutional Neural Networks. Sci Rep 9, 19518.

Zhang L. et al. (2021). A new approach to diagnosing PCa through magnetic resonance imaging, Alexandria Engineering Journal, 60(1), 897-904.