

Towards a better diagnosis of prostate cancer: Application of machine learning algorithms

Soheila Saeedi¹, Keivan Maghooli², Shahrzad Amirazodi³, Sorayya Rezayi^{1*}

¹Department of Health Information Management, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran

²Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

³Department of Management and Health Information Technology, School of Management and Medical Information, Isfahan University of Medical Sciences, Isfahan, Iran

Article Info

Article type:
Research

Article History:

Received: 2022-05-27

Accepted: 2022-07-11

Published: 2022-07-15

* Corresponding author:

Sorayya Rezayi

Department of Health Information
Management, School of Allied
Medical Sciences, Tehran University
of Medical Sciences, Tehran, Iran

Email: s_rezayi@razi.tums.ac.ir

Keywords:

Prostate Cancer
Data Mining
Machine Learning
Diagnose
Neural Network
Deep Learning

ABSTRACT

Introduction: Prostate cancer is one of the leading causes of death in men, and the early detection of this disease can be a significant factor in controlling and managing it. Applying data mining techniques can lead to the extraction of hidden knowledge from a huge amount of data and can help diagnose this disease by physicians. This study aims to determine the algorithm with the best performance to diagnose prostate cancer.

Material and Methods: In this study, nine data mining techniques, including Support Vector Machine, Decision Tree, Naive Bayes, K-Nearest Neighbors, Neural Network, Random Forest, Deep Learning, Auto-MLP, and Rule Induction algorithms, were used to extract hidden patterns from prostate cancer data. In this study, the data of 100 patients, which included eight characteristics, were used, and the RapidMiner Studio environment was employed for modeling. To compare the performance of the mentioned approaches used in this study to diagnose prostate cancer, accuracy, recall, precision, AUC, sensitivity, and specificity were calculated and reported for all techniques.

Results: The results of this study showed that the accuracy of the applied algorithms was between 77% and 84%. Using different criteria to evaluate the techniques used showed that the two algorithms K-Nearest Neighbors and Neural Network, had better performance and accuracy (84%) than other methods. The sensitivity in these two algorithms was 80% for Neural Networks and 85% for K-Nearest Neighbors, respectively.

Conclusion: The usage of different data mining techniques can lead to the discovery of hidden patterns among an enormous amount of data related to prostate cancer, and as a result, it leads to the early diagnosis of this disease and saves the subsequent costs.

Cite this paper as:

Saeedi S, Maghooli K, Amirazodi S, Rezayi S. Towards a better diagnosis of prostate cancer: Application of machine learning algorithms. *Front Health Inform.* 2022; 11: 116. DOI: [10.30699/fhi.v11i1.382](https://doi.org/10.30699/fhi.v11i1.382)

INTRODUCTION

Prostate cancer is the type of cancer where an uncontrolled extension of cancer cells appears in prostate tissue. This cancer is the second most common cancer among men after lung cancer, ranking first in developed countries [1]. Various prostate tumors spread gently and are confined to the prostate gland, where they may not create dangerous harm. The expansion of prostate cancer is correlated with ethnicity, family history of cancer, age, and a high-fat diet. Men have a 1 in 6 chance of developing prostate cancer in their lifetime, and 1 in 32 people die from this type of cancer [2]. According to

GLOBOCAN 2018 statistics, there are 1.3 million prostate cancer cases or around 7.1% of all carcinoma cases globally. Men aged from 70 to 79 years are the group who undergo the most from this disease [3].

Early prostate carcinoma regularly possesses no symptoms; still, it was observed that, when detected early, many prostate tumors can be cured and treated; when prostate cancer is diagnosed following symptoms have emerged, metastases are commonly present [4]. Prostate cancer is divided into different types depending on the starting point. In more than 95% of cases, cancer starts in cells inside the gland

called adenocarcinoma. Accordingly, we mean the same type [5].

Prostate cancer has a long prognosis. Approximately 80 to 85% of prostate cancers are diagnosed when the cancer cell is localized or regional. Nearly 100% of people whose disease is diagnosed and treated at this stage are recognized as disease-free after five years [6]. Prostate cancer is a life-taking disease, and early detection can decrease the incidence of mortality. An investigation of the numerous recent data has revealed that the survival rate is 98% after five years of diagnosis and 99% after ten years of diagnosis. Subsequent prostate cancer diagnosis, staging provides essential information about the degree of cancer in the body and foreseen response to therapy [7].

In current years, there is a growing interest in applying data-driven approaches to the early detection of various cancer to help improve accuracy; these computational Intelligence approaches covered data mining methods [8]. The field of medicine and health is one of the important sectors in industrial societies. Data mining means extracting latent information, recognizing hidden relationships and patterns, and generally discovering useful knowledge from high-volume data [9].

Extraction of classification rules is a type of data mining in which knowledge is discovered in several comprehensible and straightforward rules of data and used in the future for decision making and prediction [10]. By using data mining algorithms, intelligent systems can be developed that can automatically understand and interpret the medical properties of individuals without the need for physician supervision or discover useful information that helps experts make sound judgments [11]. Extracting knowledge from the vast amount of data related to individuals' medical histories using the data mining process can lead to identifying the laws of disease development and growth [12].

These techniques provide valuable information to health professionals and practitioners to identify the causes of disease occurrence, diagnosis, prognosis, and treatment of diseases according to the prevailing environmental factors. Several studies have centered on the confirmed diagnosis of prostate carcinoma employing data mining methods. Certain studies have applied diverse data mining approaches and have performed various results. Details of several studies are provided in this part. This paper aims to classify prostate patients and apply computational-based data mining techniques on public datasets.

Maliha et al. [13] conducted a study to predict various cancer diseases using Naïve Bayes, K-nearest neighbor, and J48 techniques. Datasets of patients were collected from different sources; physicians and experts helped researchers gather reliable data

elements. One of the nine cancers that the researcher team predict was prostate neoplasm. For performing the train and test process of the three before mentioned algorithms, Weka 3.6 was used. Numerous metrics like accuracy, sensitivity, specificity, error rate, and F-score were reported for each cancer type. In another work [14], two data mining techniques - Support Vector Machine-Recursive Features Elimination (SVM-RFE) and One Dimensional-Naïve Bayes- were used to feature selection and classification methods. Applied datasets in this study were publicly available for prostate cancer and breast neoplasm. Analysis of classifications and results of breast cancer and prostate cancer were generated by Python 3.6.

Accuracy, precision, and recall were reported for each classifier; for prostate cancer, 1-DBC has the highest accuracy, i.e., 85%, and for breast cancer SVM-RFE achieved an accuracy of 95.65%. Wang et al. [2] performed a study on early detection of prostate cancer. This research proposed a stacking-based decision tree ensemble technique that can produce effective diagnostic rules and reliable detection. The prediction model dataset for detecting prostate cancer is gathered from patients' demographic information and all kinds of examination results, including blood routine examinations, urine routine examinations, ultrasound scans, and Prostate-Specific Antigen (PSA) testing. Accuracy, sensitivity, and specificity of the proposed method were calculated, respectively, 82%, 73%, and 85%.

MATERIAL AND METHODS

Dataset

The data used in this study are prostate cancer data taken from the Kaggle data repository. This data includes 100 samples and eight features (Table 1). These eight features determine whether prostate cancer is benign or malignant. Thirty-eight tuples have been labeled benign, and 62 tuples have been labeled malignant.

Table1: Features of prostate cancer data set

Features	
Radius (16.85;4.88)	Fractal Dimension (0.06;0.008)
Texture (18.23;5.19)	Symmetry (0.19;0.03)
Area (702.88;319.71)	Compactness (0.12;0.06)
Perimeter (96.78;23.67)	Smoothness (0.1;0.01)

Classification

Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes, K-Nearest Neighbors (k-NN), Neural Network (NN), Random Forest (RF), Deep Learning (DL), AutoMLP, and Rule Induction algorithms were used to classify patients with prostate cancer into benign and malignant groups. RapidMiner Studio

environment was used for modeling.

Naive Bayes is a probabilistic classifier based on the Bayesian theory suitable for working with huge data sets. This classifier assumes that the variables are not independent of each other [15]. One of the most widely used algorithms is Decision Tree, which can be used to solve regression and classification problems. Decision Tree is a supervised algorithm [16]. Support Vector Machine is a machine learning method that classifies training data based on hyperplane by creating a discrete hyperplane in descriptor space [17]. K-NN Classifier, called the non-parametric lazy algorithm, performs classifications based on the closest examples in the feature space [18]. Neural Network is a computational classification model based on a biological neural network and includes three layers: input, hidden, and output [19]. Deep learning uses a backpropagation algorithm to discover the complex structures in huge data and allows computational models to learn with multiple layers [20]. Random Forest Algorithm is an ensemble classifier that randomly selects training samples to produce multiple decision trees [21, 22]. The parameters related to each of the algorithms used are given below:

- 1) Rule Induction: criterion: information-gain, sample ratio: 0.9, pureness: 0.9, minimal prune benefit: 0.3
- 2) Random Forest: criterion: gain-ratio, maximal depth=10, and number of trees=110
- 3) Decision Tree: criterion: gain-ratio, maximal depth: 20, confidence: 0.25, minimal gain: 0.1, minimal leaf size: 2
- 4) AutoMLP
- 5) Neural Network: training cycles= 60, learning rate=0.01, momentum=0.9 and hidden layer=1
- 6) K-Nearest Neighbors: K: 5
- 7) Naïve Bayes
- 8) Support Vector Machine: kernel type: dot, C: 0.0, convergence epsilon: 0.001, L pos:1.0, L neg: 1.0, epsilon: 0.0, epsilon plus: 0.0, epsilon minus:0.0
- 9) Deep Learning: Activation: rectifier, epochs=10

Model performance evaluation

After training the model with various data mining techniques, the performance of the model should be tested. A 10-fold cross-validation technique was used to compare different data mining models' accuracy. The evaluation results of different models were expressed with accuracy, sensitivity, specificity, precision, and recall (equations 1-5).

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (1)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

RESULTS

Nine machine learning algorithms, including SVM, DT, Naive Bayes, k-NN, NN, RF, DL, AutoMLP, and Rule Induction, were used to diagnose malignant and benign prostate cancer. Comparing the performance of different algorithms based on different evaluation parameters, including accuracy, sensitivity, specificity, recall, precision, and area under the ROC curve (AUC), are shown in Table 2.

Fig 1 shows the accuracy of prostate cancer diagnosis by nine machine learning techniques. Of the nine machine learning algorithms, Random Forest with 77% accuracy had worse performance than other algorithms, and the Neural Network and K-NN classification techniques performed best with 84% accuracy. The highest sensitivity was related to Naïve Bayes algorithm with 90%, and the lowest was related to Decision Tree and Random Forest techniques.

Based on Decision Tree classifier's results, the two attributes of perimeter and texture have played an essential role in diagnosing people with the malignant type of prostate cancer and its benign type (Fig 2).

Table 2: Performance Evaluation of prostate cancer diagnosis with various classifiers.

Data Mining Technique	Decision Tree	Rule Induction	SVM
Accuracy	79.00% +/- 9.94%	83.00% +/- 10.59%	82.00% +/- 11.35%
Precision	77.50% +/- 20.81%	77.67% +/- 18.78%	77.50% +/- 17.15%
Recall	71.67% +/- 18.51%	85.00% +/- 17.48%	80.00% +/- 19.72%
Sensitivity	71.67% +/- 18.51%	85.00% +/- 17.48%	80.00% +/- 19.72%
Specificity	84.29% +/- 16.16%	82.62% +/- 15.19%	84.05% +/- 14.11%
AUC	0.777 +/- 0.151	0.853 +/- 0.097	0.904 +/- 0.121
Data Mining Technique	Deep Learning	Neural Network	Random Forest
Accuracy	79.00% +/- 11.97%	84.00% +/- 10.75%	77.00% +/- 8.23%
Precision	75.83% +/- 19.02%	79.83% +/- 15.02%	74.67% +/- 20.66%
Recall	75.00% +/- 26.35%	80.00% +/- 19.72%	71.67% +/- 21.94%
Sensitivity	75.00% +/- 26.35%	80.00% +/- 19.72%	71.67% +/- 21.94%
Specificity	82.62% +/- 15.19%	87.14% +/- 9.60%	81.19% +/- 15.53%
AUC	0.890 +/- 0.113	0.908 +/- 0.121	0.879 +/- 0.084
Data Mining Technique	Naïve Bayes	K-NN	AutoMLP
Accuracy	83.00% +/- 14.18%	84.00% +/- 11.74%	81.00% +/- 7.38%
Precision	74.67% +/- 18.98%	80.17% +/- 20.01%	80.33% +/- 14.92%
Recall	90.00% +/- 17.48%	85.00% +/- 17.48%	72.50% +/- 24.86%
Sensitivity	90.00% +/- 17.48%	85.00% +/- 17.48%	72.50% +/- 24.86%
Specificity	79.29% +/- 16.57%	84.05% +/- 16.15%	87.38% +/- 9.53%
AUC	0.908 +/- 0.116	0.838 +/- 0.144	0.895 +/- 0.123

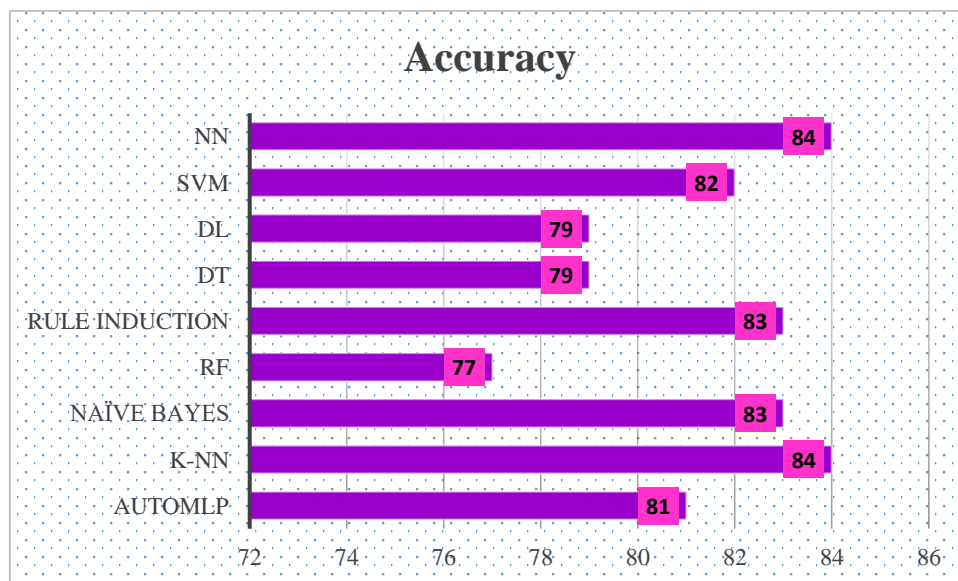


Fig 1: Accuracy of various data mining algorithms to diagnose prostate cancer

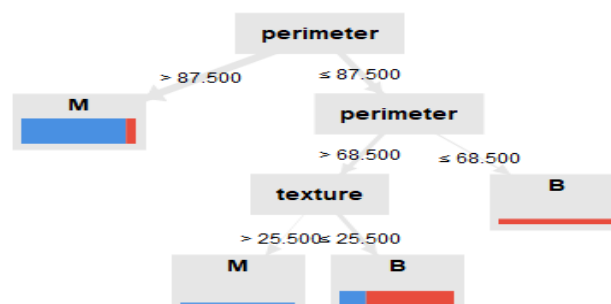


Fig 2: Decision tree for prostate cancer diagnosis

One of the nine algorithms used in this study was rule induction, which led to the development of rules for diagnosing cancer; some of the rules are listed below.

```

if area > 575 then M (51 / 4)
if perimeter ≤ 77.500 then B (1 / 17)
if perimeter ≤ 83.500 then B (3 / 9)
if radius ≤ 18.500 and radius > 15.500 then B (0 / 3)
    
```

The ROC curve for comparison of the nine algorithms used is shown in Fig 3 and 4. Neural Network, Naïve Bayes, and SVM had the maximum AUC scores. ROC curve for Naïve Bayes with AUC 0.908 had better performance in comparison to other algorithms.

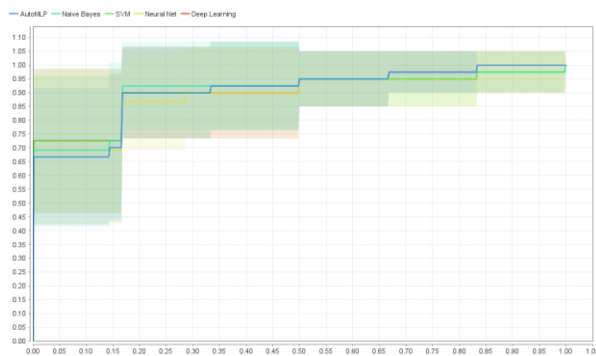


Fig 3: ROC curve for AutoMLP, Naïve Bayes, SVM, Neural Network, and deep learning

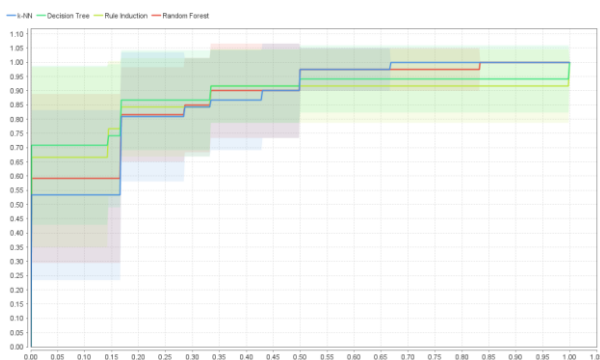


Fig 4: ROC curve for Random Forest, K-NN, Decision Tree, and Rule Induction

DISCUSSION

This study aimed to use different machine learning techniques to diagnose prostate cancer. SVM, NN, AutoMPL, K-NN, Naïve Bayes, RF, DL, Rule induction, and DT algorithms were used to diagnose prostate cancer. The evaluation results showed that the lowest accuracy was related to the Random Forest technique with 77%, and the highest accuracy was related to K-NN and Neural Network with 84%.

Kunwar et al. compared the diagnostic accuracy of Neural Network and Naive Bayes. Based on this study's results, Naive Bayes has been the most accurate classifier with 100% accuracy [15]. Jalali et al. in a study to diagnose malignant and benign prostate cancer used six data mining algorithms including DT, SVM, KNN, Naive Bayes, RF, and NN. the results showed that the detection accuracy of the different algorithms was between 77% and 93% and SVM had the best accuracy [23]. In another study, five different algorithms, including decision tree learner C4.5, MLP, Naïve Bayes classifier, Radial Basis Function (RBF) network, and K-nearest neighbor classifier used. The evaluation results show that the Naïve Bayes and K-NN performed better than other algorithms [24].

In evaluating the classifier's performance, each of the algorithms in different studies had different performances. In various studies, many factors can affect the accuracy of each algorithm. A decision support system was developed in a study by Sidiropoulos et al. to diagnose rare brain cancer cases. In this research, the Probabilistic Neural Network classifier was used to diagnose the diseases. This study has shown that the number of features can affect the accuracy of the designed system, and increasing the number of features can decrease the system's accuracy [25]. Ohmann et al. also used different data mining algorithms to diagnose acute abdominal pain and concluded that the sample size could significantly affect the performance of different algorithms, and a high sample size can improve the diagnostic system's accuracy [26]. Therefore, to obtain the best result, it is necessary to select the different algorithms according to the sample size, number of features, and each algorithm's unique features and characteristics.

CONCLUSION

As we have seen before, various applications of data mining and machine learning techniques in diagnosing diseases, using different data mining methods in diagnosing prostate cancer can help doctors extract valuable knowledge from a huge amount of data. This study uses various data mining techniques, including SVM, DT, Naive Bayes, k-NN, NN, RF, DL, AutoMLP, and Rule Induction. The results showed that out of nine classifiers, NN and k-NN Algorithms had the best performance and were more successful in diagnosing prostate cancer.

AUTHOR'S CONTRIBUTION

SS, SR, KM, and SHA designed the study. SS and SR conducted the analysis and interpretation. All authors contributed to drafting the manuscript, read

and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

REFERENCES

- Churilov L, Bagirov A, Schwartz D, Smith K, Dally M. Data mining with combined use of optimization techniques and self-organizing maps for improving risk grouping rules: Application to prostate cancer patients. *Journal of Management Information Systems*. 2005; 21(4): 85-100.
- Wang Y, Wang D, Geng N, Wang Y, Yin Y, Jin Y. Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Applied Soft Computing*. 2019; 77: 188-204.
- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018; 68(6): 394-424. PMID: 30207593 DOI: 10.3322/caac.21492 [[PubMed](#)]
- Mohler J, Bahnson RR, Boston B, Busby JE, D'Amico A, Eastham JA, et al. NCCN clinical practice guidelines in oncology: Prostate cancer. *J Natl Compr Canc Netw*. 2010; 8(2): 162-200. PMID: 20141676 DOI: 10.6004/jnccn.2010.0012 [[PubMed](#)]
- Grigore AD, Ben-Jacob E, Farach-Carson MC. Prostate cancer and neuroendocrine differentiation: More neuronal, less endocrine? *Front Oncol*. 2015; 5: 37. PMID: 25785244 DOI: 10.3389/fonc.2015.00037 [[PubMed](#)]
- Wu CH, Fang K, Chen TC. Applying data mining for prostate cancer. *International Conference on New Trends in Information and Service Science*. IEEE; 2009.
- Zhang YY, Li Q, Xin Y, Lv WQ. Differentiating prostate cancer from benign prostatic hyperplasia using PSAD based on machine learning: Single-center retrospective study in China. *IEEE/ACM Trans Comput Biol Bioinform*. 2018; 16(3): 936-41. PMID: 29993659 DOI: 10.1109/TCBB.2018.2822675 [[PubMed](#)]
- Dunning MJ, Vowler SL, Lalonde E, Ross-Adams H, Boutros P, Mills IG, et al. Mining human prostate cancer datasets: The "camcAPP" Shiny App. *EBioMedicine*. 2017; 17: 5-6. PMID: 28286059 DOI: 10.1016/j.ebiom.2017.02.022 [[PubMed](#)]
- Ngai EWT, Xiu L, Chau DCK. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*. 2009; 36(2): 2592-602.
- Freitas AA. A survey of evolutionary algorithms for data mining and knowledge discovery. In: Ghosh A, Tsutsui S (eds.). *Advances in evolutionary computing*. Springer; 2003.
- Alonso F, Martínez L, Pérez A, Valente JP. Cooperation between expert knowledge and data mining discovered knowledge: Lessons learned. *Expert Systems with Applications*. 2012; 39(8): 7524-35.
- Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. *Artif Intell Med*. 2005; 34(2): 113-27. PMID: 15894176 DOI: 10.1016/j.artmed.2004.07.002 [[PubMed](#)]
- Maliha SK, Ema RR, Ghosh SK, Ahmed H, Mollick MRJ, Islam T. Cancer disease prediction using naive bayes, K-nearest neighbor and J48 algorithm. *International Conference on Computing, Communication and Networking Technologies*. IEEE; 2019.
- Bustamam A, Bachtiar A, Sarwinda D. Selecting features subsets based on support vector machine-recursive features elimination and one dimensional Naive Bayes classifier using support vector machines for classification of prostate and breast cancer. *Procedia Computer Science*. 2019; 157: 450-8.
- Kunwar V, Chandel K, Sabitha AS, Bansal A. Chronic kidney disease analysis using data mining classification techniques. *International Conference of Cloud System and Big Data Engineering*. IEEE; 2016.
- Chaurasia V, Pal S, Tiwari B. Chronic kidney disease: A predictive model using decision tree. *International Journal of Engineering Research and Technology*. 2018; 11(11): 1781-94.
- Vijayarani S, Dhayanand S. Data mining classification algorithms for kidney disease prediction. *International Journal of Cybernetics & Informatics*. 2015; 4(4): 13-25.
- Bahrani B, Shirvani MH. Prediction and diagnosis of heart disease by data mining techniques. *Journal of Multidisciplinary Engineering Science and Technology*. 2015; 2(2): 164-8.
- Dangare CS, Apte SS. Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*. 2012; 47(10): 44-8.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015; 521(7553):436-44. PMID: 26017442 DOI: 10.1038/nature14539 [[PubMed](#)]
- Belgiu M, Drăguț L. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2016; 114: 24-31.
- Oshiro TM, Perez PS, Baranauskas JA. How many trees in a random forest? *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer; 2012.

FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

-
23. Jalali SMJ, Moro S, Mahmoudi MR, Ghaffary KA, Maleki M, Alidoostan A. A comparative analysis of classifiers in cancer prediction using multiple data mining techniques. *International Journal of Business Intelligence and Systems Engineering*. 2017; 1(2): 166-78.
 24. Mallios N, Papageorgiou E, Samarinas M. Comparison of machine learning techniques using the WEKA environment for prostate cancer therapy plan. *International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*. IEEE; 2011.
 25. Sidiropoulos K, Glotsos D, Kostopoulos S, Ravazoula P, Kalatzis I, Cavouras D, et al. Real time decision support system for diagnosis of rare cancers, trained in parallel, on a graphics processing unit. *Comput Biol Med*. 2012; 42(4): 376-86. PMID: 22197115 DOI: 10.1016/j.combiomed.2011.12.004 [[PubMed](#)]
 26. Ohmann C, Moustakis V, Yang Q, Lang K. Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain. *Artif Intell Med*. 1996; 8(1): 23-36. PMID: 8963379 DOI: 10.1016/0933-3657(95)00018-6 [[PubMed](#)]