International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Applying Data Mining Approaches for Chronic Kidney Disease Diagnosis

Sorayya Rezayi [1], Keivan Maghooli [2], Soheila Saeedi [3] *

Submitted: 10/08/2021 Accepted : 12/11/2021

Abstract: Kidney disease is one of the most common problems today that many people in the world deal with it. Therefore, in this study, our main objective is to use several computational-based algorithms to classify and diagnose Chronic Kidney Disease. The applied data in our study were publicly available data on chronic kidney disease. Eight classifiers were used to classify chronic kidney disease into two groups (patient or not). We used the Windows 10 operating system and RapidMiner Studio 9.8 version. The confusion matrix provides us the TP, FP, FN, and TN values; some performance measures were calculated to evaluate the used techniques. Evaluation of data mining algorithms revealed that Random Forest (with 100 trees), Deep Learning network (with five hidden layers), and Neural Network (with 0.02 training rate and 100 cycles) reached the highest accuracy rates with 99.09%, 98.04%, and 96.52% respectively. However, it is notable that Random Forest, Support Vector Machine, and Deep Learning network achieved 1 for AUC. Data mining on health-related issues can be considered one of the most useful data analysis tools. These classification methods are beneficial for specialists in the medical diagnosis process, and by using these techniques, hidden patterns are extracted from the raw data.

Keywords: *Data Mining; Chronic* Kidney *Disease; Supervised Technique; Machine Learning*

1. Introduction

Chronic Kidney Disease (CKD) is a public health problem that can have adverse consequences, including renal failure, dialysis, kidney transplantation, cardiovascular disease, and premature death (1). The incidence and mortality rate of this disease is rising day to day. According to the Global Burden of Disease (GBD) study results, from 1990 to 2016, the prevalence of this disease has increased by 87%, and the mortality rate of CKD is also increased by 98% to 1,186,561 (2). The CKD burden is higher in low- and middle-income countries, and disability-adjusted-life-years (DALYs) raised by 62% to 35,032,384 (2). Managing chronic diseases is often challenging; people with CKD encounter many problems throughout their lives. They must follow their treatment plans throughout their life, learn dialysis techniques, adapt to the disease's complications and various treatments. These patients will face psychological problems, which will significantly impact the affected people's quality of life (3).

If CKD is not treated in the early stages, the disease will progress. End-Stage-Renal-Disease (ESRD) has many complications, including renal replacement therapy, dialysis, and kidney transplantation. Besides, at this stage of the disease, dialysis and kidney transplantation have many complications too and the mortality rate increases. With early detection of kidney disease, the necessary interventions and treatments can be performed to prevent the disease's progression. One of the approaches that can help in the rapid diagnosis of kidney disease in the early stages is using data mining techniques. Data mining means extracting hidden patterns from a large dataset (4). Data related to health and medicine have great potential to be used to extract valuable knowledge. The healthcare industry can be considered as an entity full of rich data. A massive amount of data is produced daily in healthcare organizations, and this data is generated by electronic or paper-based health records and administrative reports. However, it is observed that this volume of rich data is not commonly used, and physicians rely on their knowledge (5, 6). Patterns and knowledge extracted from medical data can be utilized in the diagnosis and prognosis of the disease.

In recent years, many studies have employed various data mining methods to diagnose diseases, determine the stage of diseases, provide treatment solutions, predict the prevalence and mortality of diseases (7-10). Therefore, it can be concluded that various data mining techniques in health can be used to transform health data into knowledge. Data mining techniques in kidney disease can lead to the extraction of hidden patterns from patient data to diagnose the disease and help physicians greatly.

In a study conducted by Rady and Anwar to predict kidney disease stages, Probabilistic Neural Networks (PNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Radial Basis Function (RBF) techniques were used. The applied data in this study consisted of 361 CKD Indian patients and contained 25 variables. The evaluation of these techniques has shown that the PNN algorithm with 96.7% accuracy has the best result. The MLP algorithm with 51.5% accuracy has the weakest performance compared to other techniques and algorithms (4).

In one study, Naïve Bayes and Support Vector Machine classification algorithms were used to predict kidney disease. The

[1] *Health Information Management and Medical Informatics Department, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran, ORCID ID: 0000-0001-7423-8853*
[2] *Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran, ORCID ID: 0000-0003-0980-0154*
[3] *Health Information Management and Medical Informatics Department, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran, ORCID ID: 0000-0003-1315-794X*
** Corresponding Author Email: saeedi_s@razi.tums.ac.ir*

utilized data in this study was the kidney function test (KFT) dataset, which consisted of 584 samples and six attributes. The evaluation results showed that the SVM algorithm was more successful in predicting kidney disease than Naïve Bayes. However, in general, these two algorithms were able to diagnose about 70% of patients correctly (11).

One of the algorithms that can be used to diagnose diseases is Decision Tree. A study by Chaurasia et al. developed the Decision Tree algorithm to diagnose kidney disease. In this study, the dataset from UCI Repository was applied, which included 400 samples, 24 attributes, and one class as output, of which 11 attributes were numerical, and 14 attributes were nominal. The results showed that the "Hemoglobin" attribute acted as a critical element in diagnosing the disease (12).

Given that the prevalence of chronic kidney disease is increasing in societies, the disease progressively affects people's quality of life. It leads to shorter life expectancy and violation of their health consequences. Therefore, in this study, due to the high prevalence of chronic kidney disease, the main goal is to diagnose the condition without wasting time and energy using data mining algorithms. Consequently, this study intends to classify Chronic Kidney Disease patients and apply computational-based data mining techniques on a public dataset. We developed eight well-known classification techniques. The difference between this paper and previous studies is that we used eight common and complex algorithms to classify people into two categories: healthy and patient. In this study, most of the performance evaluation indicators of classifiers were calculated, and the obtained results were compared.

2. Materials and Methods

The data mining process is shown in Figure 1. To produce the model, in the first step, data preparation was done. In the next step, the data were first divided into two parts: training and testing. Training is the part where the dataset is trained by input with determined output. Testing is a part where the evaluates the model's performance and determines the label related to the instances.

2.1. Dataset

The applied data in this study were publicly available data on Chronic Kidney Disease, which is obtained from Kaggle website. Class Distribution of this dataset contains 250 samples for patient records and 150 samples for healthy records too; there are 24 attributes (14 numeric,10 nominal), and the selector is a class label used for dividing into two groups (CKD patient or not patient, class = 25). This dataset has some missing values; we have replaced them with the average for numerical features. Missing nominal data was also replaced by their mode. The list of features for each data is provided in Table 1.

Table 1. The list of attributes of chronic kidney patients

#	***Attributes***	***Attribute Information***	***Type***	***Range***
1	Age	Age of the patient	Integer	2-90
2	BP	Blood pressure	Integer	50-180
3	SG	Specific gravity	Real	1.005-1.025
4	AL	Albumin	Integer	0-5
5	SU	Sugar	Integer	0-5
6	RBC	Red blood cells	Nominal	Abnormal-Normal
7	PC	Pus cell	Nominal	Abnormal-Normal
8	PCC	Pus cell clumps	Nominal	Present-Not present
9	BA	Bacteria	Nominal	Present-Not present
10	BGR	Blood glucose random	Integer	22-490
11	BU	Blood urea	Integer	10-391
12	SC	Serum creatinine	Real	0.400-76
13	SOD	Sodium	Integer	5-163
14	POT	Potassium	Real	2.500-47
15	HEMO	Hemoglobin	Real	3.100-17.800
16	PCV	Packed cell volume	Integer	9-54
17	WC	White blood cell count	Integer	2200-26400
18	RC	Red blood cell count	Real	2.100-8
19	HTN	Hypertension	Nominal	Yes-No
20	DM	Diabetes mellitus	Nominal	Yes-No
21	CAD	Coronary artery disease	Nominal	Yes-No
22	APPET	Appetite	Nominal	Good-Poor
23	PE	Pedal edema	Nominal	Yes-No
24	ANE	Anemia	Nominal	Yes-No
25	Class	Labels	Nominal	CKD-Not CKD

2.2. Machine learning algorithms

In this study, eight algorithms including Random Forest, SVM, K-NN, Deep learning, Auto-MLP, Naïve Bayes, Neural Net and Decision Tree were applied to classify the chronic kidney disease into two groups (patient or not). These utilized techniques on our chosen dataset were as follows: Decision Tree (Criterion= gain-ratio, Maximal depth= 10, Minimal depth= 0.01, Confidence= 0.1). Random Forest (criterion=Gini-index, maximal depth=10, and number of trees=100). Auto-MPL (training cycle=10 and number of generations=10). Deep Learning Net (Activation=rectifier, epochs=10, five hidden layers with 80 neurons.) Neural Net (training cycles= 100, training rate=0.02, and momentum=0.9 and hidden layer=1). Naïve Bayes. Support Vector Machine (kernel cache=200 and maximum iteration=200000) and K-Nearest Neighbors (k=5, Measure types= Mixed Measures).

We used the Windows 10 operating system and RapidMiner Studio 9.8 version.

2.3. Model performance evaluation

Our primary effort was comparing the various performance metrics for each technique. There are many indicators like accuracy, precision, sensitivity, specificity, and mean absolute error (MAE) and AUC (Area under the ROC-Curve) to evaluate the performance of employed algorithms. The confusion matrix provides us the TP, FP, FN, and TN values. The equations of the mentioned indicators are given in the following:

$$\text{Sensitivity} = {TP}/{TP + FN} \quad (1)$$

$$\text{Specificity} = {\text{TN}}/{\text{FP} + \text{TN}} \quad (2)$$

$$\text{Accuracy} = {\text{TP} + \text{TN}}/{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

$$\text{Precision} = {\text{TP}}/{\text{TP} + \text{FP}} \quad (4)$$

$$\text{MAE} = {\sum_{i=1}^{n} |yi - x\ i|}/{n} \quad (5)$$

(y_i =prediction x_i=true valuen=total number of data points)

TP: True Positive, TN: True Negative, FP: False Positive and FN: False Negative (13).

After importing the dataset in RapidMiner, we used 10-fold cross-validation, then we got evaluation results. The selected dataset was divided into the train set and the test set; training set was applied to build various classifiers, and test set was applied to validate the performance of models. We performed the training and testing steps with split size (70% training and 30% testing in each cross-validation) or a 0.7 sample ratio; we used stratified sampling type. data sample.

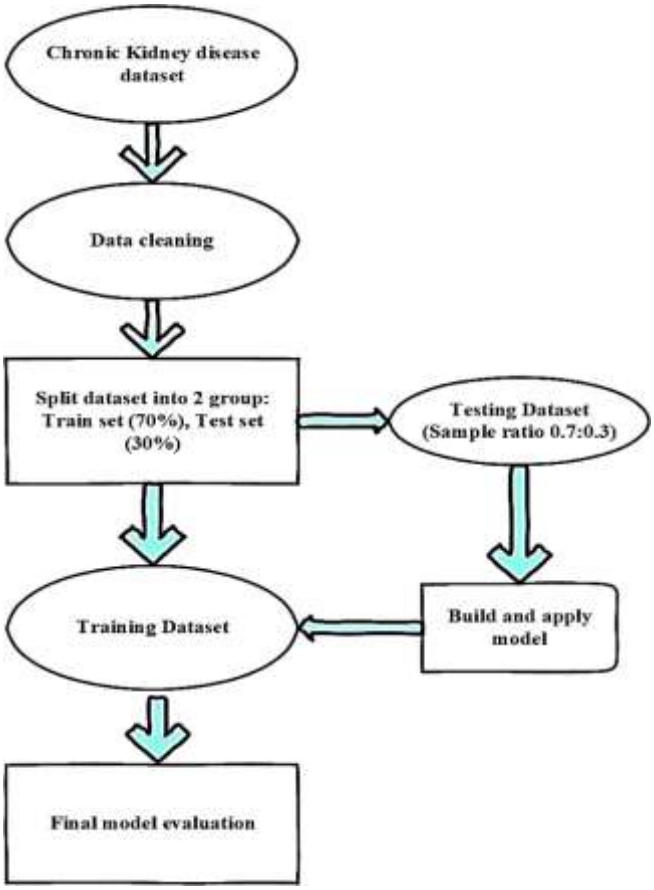

Figure 1. Data mining Processing Diagram

3. Results

In this study, eight algorithms including Random Forest, SVM, K-NN, Deep learning, Auto-MLP, Naïve Bayes, Neural Net and Decision Tree were applied to classify the chronic kidney disease into two groups (patient or not). The distribution of individuals based on age and existence of diseases (patient or healthy person) is shown in Figure 2. Table 2 provided the applied eight model performance in 10-cross-validation folds; several essential performance analysis metrics were calculated. In Figure 3, Area under Receiver Operating Characteristic curves (AUC) based on chronic kidney disease attributes were presented for four applied algorithms. Random forest, Deep Learning, and SVM achieved the maximum Area Under the Curve (AUC) score, i.e., 1. As it is notable, K-NN has the minimum value for AUC with 0.80. The comparison of accuracy, sensitivity, AUC, and precision for applied techniques is given in Figure 4. Random forest has the maximum value for accuracy of 99.09% and 99.45% of sensitivity; SVM has reached 100% precision.

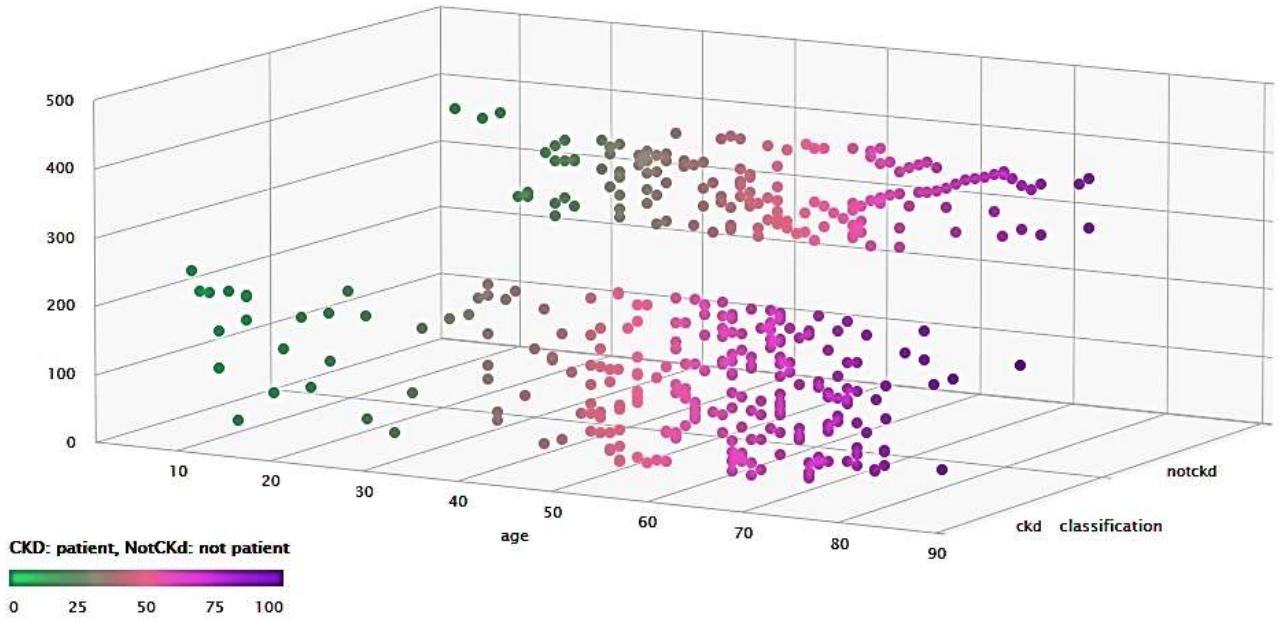

Figure 2. Distribution of instances based on age and existence of CKD

Table 2. Comparison of model performance and specific metrics of eight numerous applied classifiers

Indicators	Random Forest	Naïve Bayes	SVM
Accuracy	99.09% +/- 0.66%	96.14% +/- 1.29%	93.69% +/- 1.78%
Recall	99.45% +/- 0.53%	94.27% +/- 1.26%	89.91% +/- 2.87%
Precision	99.10% +/- 1.02%	99.59% +/- 1.05%	100.00% +/- 0.00%
AUC	1.000 +/- 0.001	0.994 +/- 0.017	1.000 +/- 0.000
Sensitivity	99.45% +/- 0.53%	94.27% +/- 1.26%	89.91% +/- 2.87%
Specificity	98.57% +/- 1.45%	99.18% +/- 2.14%	100.00% +/- 0.00%
MAE	0.91% +/- 0.66%	3.86% +/- 1.29%	6.31% +/- 1.78%
-	**Neural Net**	**Deep Learning**	**KNN**
Accuracy	96.52% +/- 1.06%	98.04% +/- 0.93%	71.56% +/- 2.66%
Recall	95.29% +/- 1.28%	97.80% +/- 1.58%	70.07% +/- 5.05%
Precision	99.07% +/- 1.22%	99.08% +/- 1.20%	82.44% +/- 6.31%
AUC	0.994 +/- 0.006	1.000 +/- 0.000	0.805 +/- 0.026
Sensitivity	95.29% +/- 1.28%	97.80% +/- 1.58%	70.07% +/- 5.05%
Specificity	98.55% +/- 1.92%	98.57% +/- 1.90%	74.86% +/- 8.81%
MAE	3.48% +/- 1.06%	1.96% +/- 0.93%	28.44% +/- 2.66%
-	**Decision Tree**	**Auto-MLP**	**-**
Accuracy	95.42% +/- 2.35%	96.22% +/- 1.56%	-
Recall	94.95% +/- 3.81%	94.00% +/- 2.45%	-
Precision	97.63% +/- 1.33%	99.91% +/- 0.28%	-
AUC	0.952 +/- 0.024	0.996 +/- 0.004	-
Sensitivity	94.95% +/- 3.81%	94.00% +/- 2.45%	-
Specificity	96.21% +/- 1.96%	99.87% +/- 0.41%	-
MAE	4.58% +/- 2.35%	3.78% +/- 1.56%	-

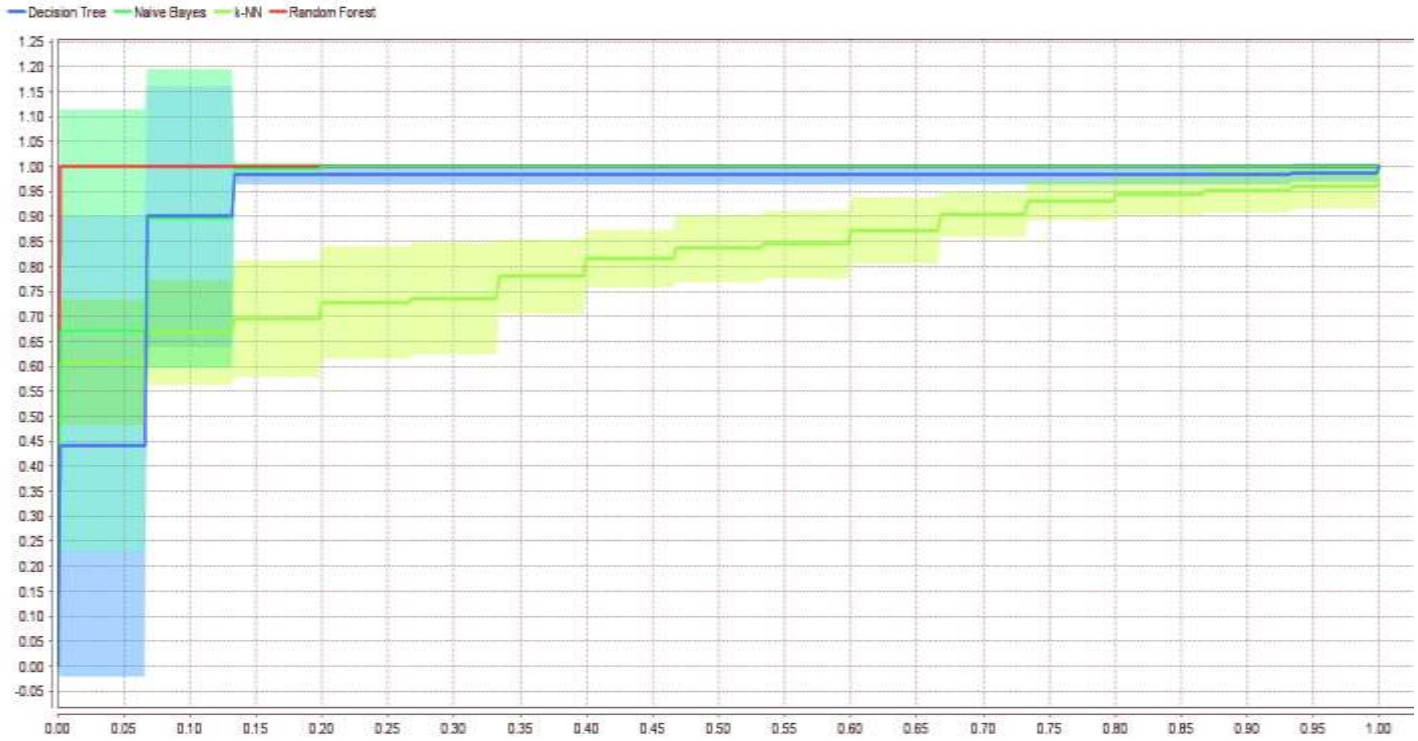

Figure 3. ROC for DT, K-NN, NB and RF

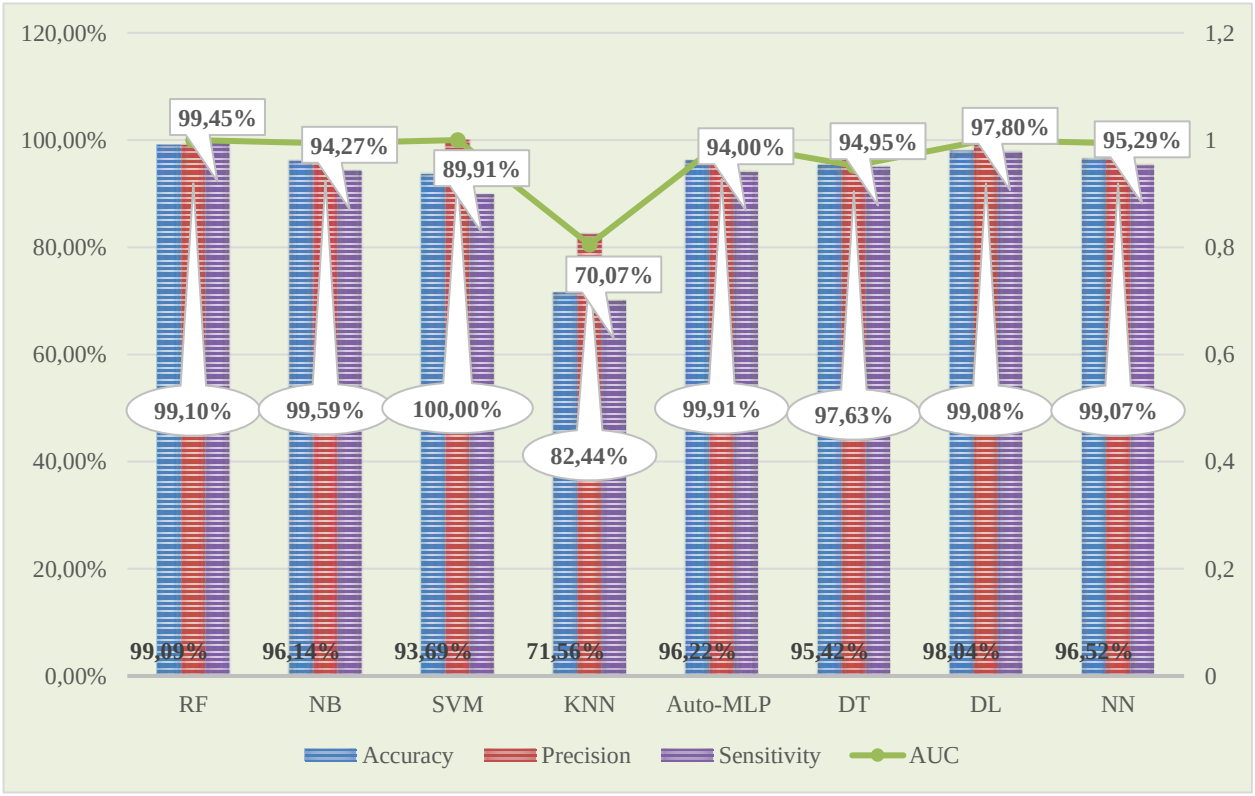

Figure 4. Comparison of main indicators

4. Discussion

The dramatic growth of chronic kidney disease, its effects, complications, and the high costs it incurs on society has led the medical community to seek programs for further investigation, prevention, early detection, and effective treatment. Using data mining and knowledge discovery in the system of medical centers, valuable knowledge can be created, which can improve the quality of service provided by the center's managers and can also be used by physicians to determine the future behavior of kidney patients. Predict the given history and diagnose kidney disease from various features and symptoms, evaluate risk factors are the most important applications of data mining and knowledge discovery in kidney patients' system.

The main objective of this study was to apply various powerful machine learning algorithms to classify and rapidly diagnose patients with CKD. The used dataset in this paper is related to CKD from Kaggle repository. Replacing missing values, using eight robust computational-based techniques like 1-RF, 2-NB, 3- ANN, 4- DNN, 5- SVM, 6-AutoMLP 7-K-NN, and 8- DT with optimal and high performances were the critical differences of our work with others. What is remarkable in this study is the use of eight powerful data mining algorithms to classify people into two classes of healthy and patient.

In a study conducted by Huseyin Polat, et al. SVM technique was applied with feature selection methods to reduce the dimension of inputs and diagnose chronic kidney disease. Based on the achievements of mentioned paper, after feature selecting by wrapping and filtering, SVM reached 0.96, 0, 1, 0.96, 0.98 of TP, FP, Precision, Recall and AUC respectively; the accuracy rate of SVM by 13 attributes was highest value (95.5%). Similar to our work, SVM had good potential for classification and predicting CKD patients from healthy persons (14). However, it can improve the generalization performance by managing non-linear classification rules with mapping the data into high feature space. The main difference between this article and our work is the use of feature selection methods. Using feature selection can have many advantages, such as eliminating noise and duplicate data, removing irrelevant data, and reducing learning time. However, using this method may also have disadvantages, such as being more prone to over-fitting and requiring expensive computation (15). In our study, we preferred to use all the features for classification so that no data is lost and the feature space is not limited (16). It should be noted that the dataset used in this study was the same as our study, and there was no difference in the specifications of the model used in these two studies, and the only difference was the use of feature selection in the mentioned study.

In (17), five classifiers like MLP, SVM, K-NN, C4.5, and RF were employed to diagnose CKD accurately. In line with our study, the CKD dataset was utilized. In this study, the default parameters were used for C4.5 and Random Forest Classifier. For K-NN, k was 1; for ANN, the learning rate was 0.3, and momentum was 0.2 with ten nodes in the hidden layer; for SVM, C=100 and L=0.001 with poly kernel. Based on results, RF has maximum values for performance metrics; all calculated measures achieved 100% by this technique. Considering, in our work, random forest has optimal maximum accuracy, recall, precision, AUC. It is remarkable that RF technique is less prone to overfitting and may change by a small change in the source data and time needed for training and testing is low. So, its computation goes far more complicated than other machine learning approaches (18).

In another work which was conducted by Veenita Kunwar et al., Chronic Kidney Disease analysis was performed by various data mining classification techniques. SVM, ANN, NB, and DT were developed to categorize the patients and healthy persons with high accuracy and low cost. Similar to our work, the clinical data of 400 records have been taken from Kaggle public repository, and RapidMiner software was used for implementation. However, we applied eight classification techniques for analyzing the dataset. Our finding indicated that after RF, Naïve Bayes, ANN, and DNN produced more accurate results than K-NN and SVM. Given that K-NN does not work well with large and high dimensional datasets and requires done feature scaling or normalizing process, so according to our results, this algorithm did not achieve optimal performance compared to other ones. Besides, in our paper, the performance indicators of SVM classifier (with the linear kernel) have been assessed to take accurate scores. So, our experimental results presented that SVM can be appropriately implemented with an overall performance of 93.69%, like the beforementioned paper achievement for SVM performance (94.60%). According to the survey results, this classifier can perform more effectively in high-dimension spaces like our chosen dataset and is relatively memory efficient (19).

In (20), SVM predictor was used for CKD dataset; comparing the performance of the mentioned technique based on its accuracy, precision, execution time was the main target. Reported evaluation metric was 94.6 for SVM. The dataset used in this study was the same, and the reported accuracy for SVM in these two studies was very close to each other. Since the parameters used for this classifier are not mentioned, no further comparisons can be made.

In this study, we used the CKD dataset from the UCI Machine Learning Repository, which included 400 samples with 24 features. Dataset characteristics can affect algorithm performance, both in terms of accuracy and elapsed time. The characteristics of the dataset that affect the algorithm's performance include the following: Sample size: insufficient sample size can lead to a decrease in classifier performance and ultimately lead to an increase in classification error, which in this study was 400, which has a sufficient number of samples. Class type: whether the class type is binary, multiple, or categorical can affect the method's performance. In the present study, the output class had two groups of patients and healthy, reducing the complexity of the model. Unlabeled class: in the present study, there was no class without labels, which can be effective in the high accuracy of classifiers. Missing values: if the missing value is not managed in the preprocessing stage, it can reduce the performance of the employed algorithms. In this study, the missing data in the preprocessing step were considered. Class dimensionality: the larger dimensions can lead to more complexity of the model. In this study, the number of features is 24, which in subsequent studies can be reduced with the help of dimensionality reduction methods such as PCA or FuzzyToolkit (21).

In this study, important and powerful classification algorithms for disease identification were developed, and by using specific indicators comparing was given. The results showed that Random Forest, Deep Learning Network, Artificial Neural Net, and Naïve Bayes methods had the best classification potential between healthy persons and patients. One of the limitations of our work was the existence of missing values in the dataset, which the researchers decided to identify and fill in with acceptable figures. In future directions, researchers want to use a local dataset to multiply the value and credibility of the work and compare the results with current work.

5. Conclusion

In this study, different data mining techniques were used to diagnose unhealthy and healthy people. The evaluation results of the developed techniques showed the high accuracy of these techniques in the diagnosis of CKD. The use of these methods due to the high volume of data produced in health care can help in the better diagnosis of kidney disease. Using these techniques can help identify hidden patterns in the vast amount of data created and avoid physician confusion. Diagnosis of this disease in the early stages can lead to a lack of disease progression. Also, since many different factors must be considered to diagnose CKD, using these methods to diagnose the disease by primary caregivers can be very helpful.

References

[1] Levey AS, Eckardt K-U, Tsukamoto Y, Levin A, Coresh J, Rossert J, et al. Definition and classification of chronic kidney disease: a position statement from Kidney Disease: Improving Global Outcomes (KDIGO). Kidney international. 2005;67(6):2089-100.

[2] Xie Y, Bowe B, Mokdad AH, Xian H, Yan Y, Li T, et al. Analysis of the Global Burden of Disease study highlights the global, regional, and national trends of chronic kidney disease epidemiology from 1990 to 2016. Kidney international. 2018;94(3):567-81.

[3] Goh ZS, Griva K. Anxiety and depression in patients with end-stage renal disease: impact and management challenges–a narrative review. International journal of nephrology and renovascular disease. 2018;11:93.

[4] Rady E-HA, Anwar AS. Prediction of kidney disease stages using data mining algorithms. Informatics in Medicine Unlocked. 2019;15:100178.

[5] Jothi N, Husain W. Data mining in healthcare–a review. Procedia computer science. 2015;72:306-13.

[6] Yuliastuti GE, Alfiyatin AN, Rizki AM, Hamdianah A, Taufiq H, Mahmudy W. Performance Analysis of Data Mining Methods for Sexually Transmitted Disease Classification. International Journal of Electrical & Computer Engineering (2088-8708). 2018;8(5).

[7] Itani S, Lecron F, Fortemps P. Specifics of medical data mining for diagnosis aid: A survey. Expert systems with applications. 2019;118:300-14.

[8] Oskouei RJ, Kor NM, Maleki SA. Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges. American journal of cancer research. 2017;7(3):610.

[9] Shouman M, Turner T, Stocker R, editors. Using data mining techniques in heart disease diagnosis and treatment. 2012 Japan-Egypt Conference on Electronics, Communications and Computers; 2012: IEEE.

[10] Safdari R, Rezayi S, Saeedi S, Tanhapour M, Gholamzadeh M. Using data mining techniques to fight and control epidemics: A scoping review. Health and Technology. 2021:1-13.

[11] Vijayarani S, Dhayanand S. Data mining classification algorithms for kidney disease prediction. Int J Cybernetics Inform. 2015;4(4):13-25.

[12] Chaurasia V, Pal S, Tiwari B. Chronic kidney disease: a predictive model using decision tree. International Journal of engineering Research and technology. 2018.

[13] Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:201016061. 2020.

[14] Polat H, Mehr HD, Cetin A. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. Journal of medical systems. 2017;41(4):55.

[15] Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. Journal of Applied Science and Technology Trends. 2020;1(2):56-70.

[16] Karamizadeh S, Abdullah SM, Halimi M, Shayan J, javad Rajabi M, editors. Advantage and drawback of support vector machine functionality. 2014 International conference on computer, communications, and control technology (I4CT); 2014: IEEE.

[17] Ao Y, Li H, Zhu L, Ali S, Yang Z. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. Journal of Petroleum Science and Engineering. 2019;174:776-89.

[18] Azar AT, Elshazly HI, Hassanien AE, Elkorany AM. A random forest classifier for lymph diseases. Computer methods and programs in biomedicine. 2014;113(2):465-73.

[19] Raghavendra S, Santosh KJ. Performance evaluation of random forest with feature selection methods in prediction of diabetes. International Journal of Electrical and Computer Engineering. 2020;10(1):353.

[20] Amirgaliyev Y, Shamiluulu S, Serek A, editors. Analysis of chronic kidney disease dataset by applying machine learning methods. 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT); 2018: IEEE.

[21] Kwon O, Sim JM. Effects of data set features on the performances of classification algorithms. Expert Systems with Applications. 2013;40(5):1847-57.