Prediction Of Chronic Kidney Disease Using Different Classification Algorithms: A Comparative Study.

Maria Yousef

Department of Computer Science, AL-al Bayt University, Mafraq, Jordan.

***Abstract* - Kidney disease (KD) is a serious public health concern is defined as a disorder that disrupts the kidneys' normal function. As the rate of people influenced by CKD continues to rise, efficient prediction techniques should be considered. Therefore, the goal of this work is to help in the prevention of Chronic Kidney Disease (CKD) applying machine learning approaches to detect CKD early on and keep chronic kidney disease from getting worse. In this paper, we focus on applying various classification algorithms perform on a dataset of 400 patients with 24 variables linked to chronic kidney disease diagnosis in aims to compare the performance of these algorithms in making predictions and justify the algorithm that achieves more correct prediction. The classification techniques used in this study include Random Forest (RF), Decision Tree (DT), and Naive Bayes (NB). To perform experiments, all missing and outlier values in the dataset were replaced by the mean of the corresponding attributes. Next, in the feature selection step, we optimizing the prediction model and obtaining the optimal subset of features from the dataset. In the proposed approach, correlation coefficient, and recursive feature elimination have been used as feature selection methods to decrease the number of features by eliminating irrelevant and useless features to maintain a good analytical result. Finally, different assessment parameters are used to evaluate the classifiers' performance such as Accuracy, Precision, Recall, and F1-score. The empirical results from the experiments serve that RF performed better than DT, and NB with accuracies of 100% in CKD predicting.**

Keywords - Kidney disease; prediction systems; Random Forest; Decision Tree; Naive Bayes.

I. INTRODUCTION

Kidneys are a couple of bean-shaped organs located near the lower back of the body. The major functionality of the Kidney is filtration for blood and to eliminate toxins from the body, where the kidney moves the toxins to the bladder then is later removed from the body during urination. Every 24 hours, the kidneys filter and return around 200 quarts of fluid to the bloodstream, which keeps the body alive. If kidneys cannot perform their normal job then the body becomes overloaded with toxins. This can induce kidney failure, which can lead to death. Moreover, there are two types of Kidney problems either acute or chronic.

Chronic kidney disease (CKD), also known as chronic renal disease, is one of the Kidney diseases in the medical field. It is characterized by abnormal kidney function or gradual renal failure over months. CKD is frequently discovered as a result of screening of persons who are known to be at risk for kidney difficulties, such as those with high blood pressure or diabetes, or those who have a blood relative who has the condition. Moreover, diabetes, high blood pressure, and unhealthy lifestyles have also increased the number of patients with CKD. According to the Saudi Center of Organ Transplantation Registry's most recent data, 10203 people with renal disease take hemodialysis every day [1].

CKD patients suffer from a variety of adverse impact. These complications involve infliction to the nervous and immunological systems that interrupt everyday activities [2]. Thus, realizing kidney failure in the first stage is extremely important to preserve people's lives. Moreover, early detection allows each kidney's action to be controlled, reducing the danger of irreparable damage. For this reason, routine check-ups and early diagnosis are critical to the patients, for they can prevent essential risks of renal failure and related diseases, which creates (n4, n5) and deletes (n3, n4) concurrently. A path starting from host n1 to host n5 may erroneously be returned, even though no such path ever existed.

To assist in the prevention of CKD, the CKD diagnoses system can assist medical experts in performing efficient and more accurate diagnoses. The idea of medical diagnostics system is to extraction of useful information from the very massive amount of eHealth data collected and stored in a medical organization's databases or repositories which typically consist of frequently accumulating medical records [3]. Early detection of CKD can be achieved using machine learning approaches. Machine Learning is a rapidly data with various variables. It is developed from the study of computational learning theory and pattern recognition in the artificial intelligence domain and includes computational strategies, algorithms, and analysis techniques.

In this paper, we present a comparative study to evaluate the performance of different Machine Learning algorithms to predict CKD disease. In this work, we used three classification algorithms (Decision Tree DT, Random Forest RF, Naïve Bayes NB) to train the model in making predictions and justify the algorithm that achieves more correct predictions. These

algorithms have been applied on a recently collected CKD dataset downloaded from the UCI repository with 400 data records and 25 attributes. Moreover, different measurement approaches, such as accuracy, precision, and recall, were used to assess the performance of different classification algorithms.

The rest of this paper is organized as follows: Section 2 provided a review of some of the work related to the proposed approach. While Section 3 explained the strategy and methodology used in building the proposed system. Furthermore, Findings and discussions are presented in Section 4. Section 5 concludes this study.

II. LITRETURE REVIEW

Several researchers have applied Machine Learning and data mining-based algorithms to solve problems in the field of medicine. In this part, we present a summary of the previous researches that introduced different models for predicting the problem of renal illness.

In a study performed by Ref [4], the authors compared two data mining techniques in predicting CKD by using Support Vector Machine (SVM) and Artificial Neural Network (ANN) classifier on 400-instances CKD dataset. The experimental analysis performed by the WEKA tool showed that the ANN classifier has more success than the SVM classifier in correct CKD predicting with an accuracy of 99.75%.

In the paper presented by [5], the authors attempted to evaluate the ability of machine learning algorithms in knowledge discovery from large databases in the healthcare field. In this study, the Support vector machine (SVM) and K-Nearest Neighbor (KNN) classifier are employed to predict patients with chronic kidney failure and normal people. In this work, the authors used the Kidney Function Test (KFT) dataset to study kidney disease. This dataset contains four hundred instances and twenty-five features. Finally, the performance of the Support vector machine (SVM) and K-Nearest Neighbor (KNN) classifier have been compared based on its accuracy, precision, and execution time for CKD prediction. The experimental results revealed that the performance of the KNN classifier is better than SVM its achieved 78% of accuracy.

Prediction is one of life's most fascinating and difficult activities. Moreover, data mining plays a primary role in the prediction of a medical dataset. Many different types of artificial neural networks have been discussed by many authors to be very effective in disease prediction. This paper [6], proposed a new chronic kidney disease decision support system by using three classifiers such as radial basis function network, multilayer perceptron, and logistic regression. The CKD dataset from the UCI Machine Learning Repository (University of California, Irvine) was used in this system's data analysis. Accuracy of the three classifiers is evaluated and analyzed, the evaluation parameters indicate that the multilayer perceptron gives a better performance than other neural networks, it achieved 99.7% of accuracy, while the radial basis and logistic regression achieved 98.5%, 97% of accuracy respectively.

Another effort to present a comparative study of different machine learning algorithms is performed by [7], the authors had used WEKA (Waikato Environment for Knowledge Analysis) data mining tool for predicting the early detection of chronic kidney disease for diabetic patients with the help of machine learning methods. 600 records collected by the authors from a leading Chennai-based diabetes research center were used as a training set to perform the prediction. To enhance the predicting performance level this model has been designing with 10-Cross validation. This validation is done by dividing the data into 10 components each one is consisting of 90% of the original data. Cross-validation will be very much useful to create a set of training data with validation folds. The authors have tested the dataset for classification using Naïve Bayes and the Decision tree method. The result shows that the Decision tree classification gives better accuracy more than Naive Byes it has achieved 91% of accuracy, while the Naive Byes achieved 86%.

Pasadana et al. [8] discussed the use of several decision tree algorithms such as HoeffdingTree, DecisionStump, CTC, LMT, J48, J48graft, NBTree, RandomTree, REPTree, SimpleCart, and RandomForest to assess their effectiveness in classifying kidney disease. These algorithms have been applied and tested using collected data at UCI's CKD dataset which is including features like age, blood pressure, specific gravity, albumin, sugar, red blood cells, plus cell, pus cell clumps, bacteria, blood glucose random, and blood urea. From the analysis, the experimental results showed that the RBF gains the most powerful accuracy than other algorithms incorrect classification. According to their observation, the authors point that the application of a decision tree in predicting CKD will benefit in maintaining health.

III. RESEARCH METHODOLGY

The major objectives in this paper are to (1) proposed a CKD prediction model to classify whether the patient has CKD or not; (2) comparing different classification algorithms in their predicting ability and (3) finding the most effective algorithm for predicting CKD.

This proposed work employs three classification methods to predict the presence of chronic kidney disease in humans. The classifiers used are the Random Forest, Naive Bayes, and Decision Tree classifier. The UCI chronic kidney disease dataset was collected and some statistical methods were applied to it to extract the optimized features from it and then fed to each classifier to make a prediction process. Also, the classifier's performance was assessed using the assessment parameters. Figure 1 depicts the main phases of the proposed prediction model for diagnosing CKD illness.

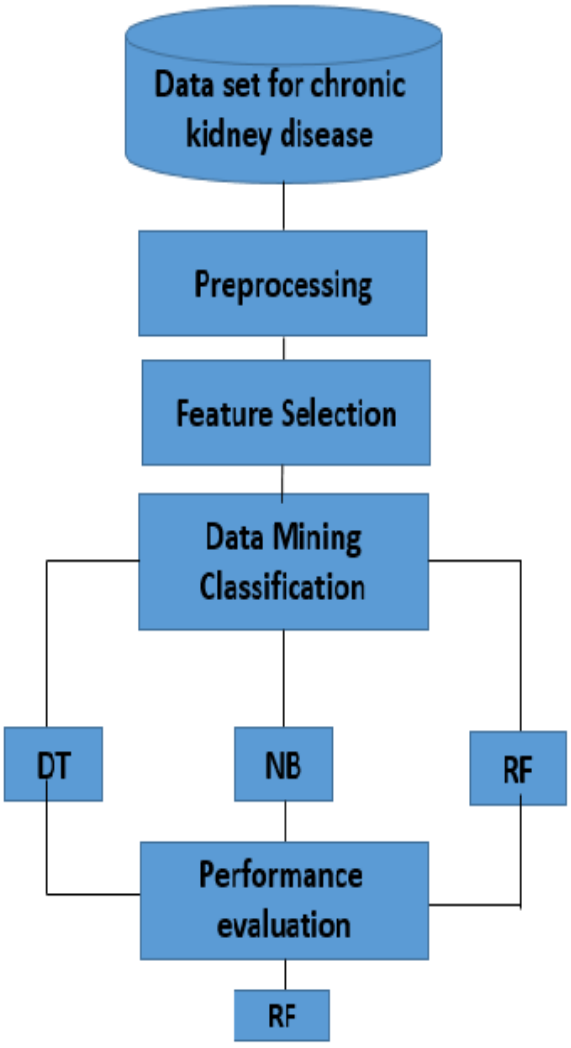

Figure 1. CKD Prediction Model Framework

A. Chronic Kidney Disease Dataset

In this study, machine-learning algorithms were trained by using the CKD dataset obtained from the UCI machine learning repository [9]. The Apollo Hospitals in Tamilnadu, India, published the CKD dataset in July 2015. This dataset has 24 features, including 11 numerical and 13 nominal features, that were gathered over two months from 400 instances. Moreover, there are 250 CKD patients and 150 non-CKD patients among the 400 instances. As a result, this dataset is used to classify patients as having or not having chronic kidney disease. Table 1 summarizes the description of CKD dataset attributes. Also, Figure 2 shows a portion of the CKD database that was used.

Table 1. CKD Attributes Description

Attribute Name	Attribute Code	Attribute Type	Attribute Value
Age	age	numeric	years
Blood Pressure	bp	numeric	mm/Hg
Specific Gravity	sg	numeric	1.005, 1.010, 1.015, 1.020, 1.025
Albumin	al	numeric	0, 1, 2, 3, 4, 5
Sugar	su	numeric	0, 1, 2, 3, 4, 5
Red Blood Cells	rbc	nominal	normal, abnormal
Plus Cell	pc	nominal	normal, abnormal
Plus Cell Clumps	pcc	nominal	present, notpresent
Bacteria	ba	nominal	present, notpresent
Blood Glucosa Random	bgr	numeric	mgs/dl
Serum Creatine	sc	numeric	mgs/dl
Sodium	sod	numeric	mEq/l
Potassium	pot	numeric	mEq/l
Hemoglobin	hemo	numeric	gms
Packed Cell Volume	pcv	numeric	-
White Blood Cell Count	wc	numeric	cells/cumm
Red Blood Cell Count	rc	numeric	millions/cumm
Hypertension	htn	numeric	yes, no
Diabetes Mellitus	dm	numeric	yes, no
Coronary Artery Disease	cad	nominal	yes, no
Appetite	appet	nominal	good, poor
Pedal Edema	pe	nominal	yes, no
Anemia	ane	nominal	yes, no
Class	class	nominal	ckd, notckd

age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	class
48	80	1.02	1	0		normal	notpresent	notpresent	121	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
7	50	1.02	4	0		normal	notpresent	notpresent	148	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
62	80	1.01	2	3	normal	normal	notpresent	notpresent	423	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
48	70	1.005	4	0	normal	abnormal	present	notpresent	117	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
51	80	1.01	2	0	normal	normal	notpresent	notpresent	106	1.4			11.6	35	7300	4.6	no	no	no	good	no	no	ckd
60	90	1.015	3	0			notpresent	notpresent	74	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
68	70	1.01	0	0		normal	notpresent	notpresent	100	24	104	4	12.4	36			no	no	no	good	no	no	ckd
24		1.015	2	4	normal	abnormal	notpresent	notpresent	410	1.1			12.4	44	6900	5	no	yes	no	good	yes	no	ckd
52	100	1.015	3	0	normal	abnormal	present	notpresent	138	1.9			10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
53	90	1.02	2	0	abnormal	abnormal	present	notpresent	70	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
50	60	1.01	2	4		abnormal	present	notpresent	490	4			9.4	28			yes	yes	no	good	no	yes	ckd
63	70	1.01	3	0	abnormal	abnormal	present	notpresent	380	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
68	70	1.015	3	1		normal	present	notpresent	208	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
68	70						notpresent	notpresent	98	4.6	135	3.4	9.8				yes	yes	yes	poor	yes	no	ckd
68	80	1.01	3	2	normal	abnormal	present	present	157	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd
40	80	1.015	3	0		normal	notpresent	notpresent	76	9.6	141	4.9	7.6	24	3800	2.8	yes	no	no	good	no	yes	ckd
47	70	1.015	2	0		normal	notpresent	notpresent	99	2.2	138	4.1	12.6				no	no	no	good	no	no	ckd

Figure 2. Part Of CKD Database

B. Statistical Study Of The CKD Dataset

To more understand and visualize the data, Table 2 shows the results of the statistical analysis of the data, where the numerical properties in the dataset such as mean, median, standard deviation, minimum, and maximum are introduced.

Table 2. Statistical analysis of the dataset

Attribute	Mean	Standard Deviation	Maximum	Minimum
Age	51.483	17.17	90	2
Blood Pressure	76.469	13.684	180	50
Blood Glucose Random	148.037	79.282	490	22
Serum Creatine	3.072	5.741	76	0.4
Sodium	137.529	10.409	163	4.5
Potassium	4.627	3.194	47	2.5
Hemoglobin	12.526	2.913	17.8	3.1
Packed Cell Volume	38.884	8.151	54	9
White Blood Cell Coun	8406.12	2523.22	26400	2200
Red Blood Cell Count	4.707	0.84	8	2.1

C. Preprocessing Step

Today's real-world database, particularly clinical database, are prone to noisy, missing, redundant, and incompatible data. When you work with low quality data, you'll get low quality outcomes. Therefore, in each machine learning application, the first step is to study and comprehend the dataset and its properties in order to prepare it in a more useful and suitable format for the modeling process. [10]. This manner is generally known as data pre-processing.

1) Outlier Detecting

Outlier is defined as an observation that is inconsistent with the remains of the set of data values [11]. In other words, outliers are data points that are significantly different from the rest of the data also it's located far away from the feature central tendency. Invalid outliers in the dataset usually occur due to errors in measurement and data entry, which are pointed to as noise in the data.

When dealing with outliers, medical data cannot be treated the same as other data since the outliers could be valid or significant. For this reason, each outlier found in the CKD dataset is checked to identify if it is realistic or not. In this study, the extreme data points that go above the medically permissible range have been considered as missing data and then changed and modified as described in the missing data phase.

In this study, we use the Box plots tool to detect outliers in the CKD dataset, as shown in Figure 3, there were a few outliers discovered for blood glucose random features that arrived at 500 mg/dl. However, according to [12], the highest blood glucose level recorded for a living patient in 2008 was 2,656 mg/dl. As a result, these outliers are legitimate and should not be changed.

In contrast, three other unacceptable data points were detected in potassium and sodium features. In the potassium feature, two outliers data point have been concentration found with values 39 and 47 as shown in Figure 4. While the highest potassium level observed was 7.6 mEq/L [13]. This suggests that a potassium level of 39 and 47 is impossible and usually due to a mistake. Similarly, with sodium, as Figure 5 displays, one outlier data point was identified, which is 4.5. Normally, sodium levels should be within 135 and 145 mEq/L, and if it is less than 135, then the patient suffers from hyponatremia [14]. As a result, a score of 4.5 is either impossible or unachievable.

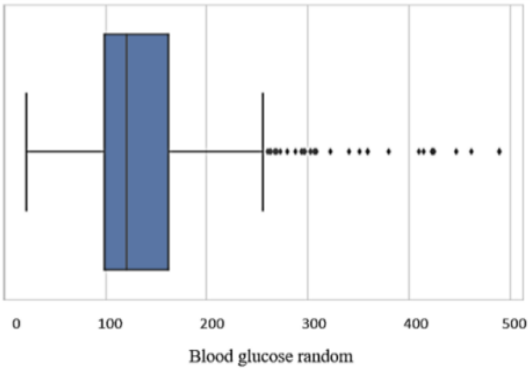

Figure. 3. Outlier Random Blood Glucose values.

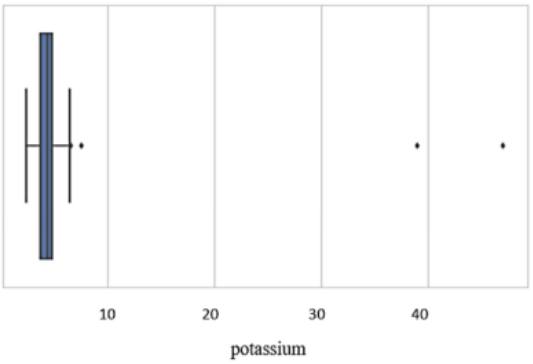

Figure. 4. Outlier Potassium values.

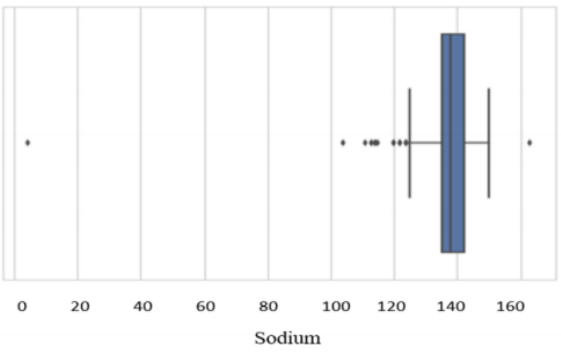

Figure. 5. Outlier sodium values.

2) Missing Values

Missing data is a widespread problem, particularly in the medical field. Almost, each patient's record and each feature has some missing values [15]. The CKD dataset contains a substantial occurrence of missing values in many features. Eliminating samples or features with missing values is simply not feasible because significant information may be lost in the process. To expel missing values from the dataset, a filter technique is utilized, which takes the mean from the training data and uses it to fill in the gaps.

3) Data Transformation

Data transformation is the process of converting data into a suitable format that may be used for mining. Normalization is one of the data scaling techniques, which is the process of reducing the values of the attributes to a limited range [16]. Because different scales of attributes complicate attribute comparison and degrade the ability of algorithms to learn, the normalization approach has been used before feature selection and modeling phases. Consequently, the Min-Max normalization technique was used on numeric data types in this investigation. So that the minimum value of each feature is converted to a 0, the highest value is converted to a 1, and all other values are converted to a decimal between 0 and 1 [17].

Label Encoding is another method for data transformation that has been done to convert categorical variables into specific numerical forms [18]. This is because some machine learning algorithms, particularly in prediction tasks, cannot handle categorical data.

D. Feature Selection

Usually, many mining algorithms don't perform well with large amounts of features or attributes. Therefore, a feature selection approach must be applied before any kind of mining algorithm is applied. The process of identifying the most discriminating features in a dataset is known as feature selection [19].

The main objectives of feature selection are to avoiding overfitting, enhancing the model's performance, and providing faster and more cost-effective models.

In this study, correlation coefficient, and recursive feature elimination have been used as feature selection methods. As indicated in Table 3, the correlation coefficient, which includes Pearson's correlation and Cramer's V, is used to rank the features having the strongest association with the outcome variable from highest to lowest correlation. Next, the recursive feature elimination method is applied to search for a best-performing features subset, by recursively removing the bottom half of the listed features until a single feature has remained. Each subset works all four classifiers with 10-fold cross-validation to decide the subset offering the most powerful performance.

Accordingly, Table 4 shows that the highest average accuracy was achieved using the top 6 features: Albumin, Hemoglobin, Packed cell volume, specific gravity, Red blood cell count, and Diabetes Mellitus.

class [20].

Table 3. The correlation confusion value of CKD features

Numerical Features	Correlation coefficient	Nominal Features	Correlation coefficient
Hemoglobin	0.729	Albumin	0.730
Packed cell volume	0.690	pacific gravity	0.687
Red blood cell count	0.590	Hypertension	0.590
Blood glucose random	0.401	Diabetes Mellitus	0.544
Sodium	0.342	Red blood cell	0.540
Blood pressure	0.294	Sugar level	0.432
Serum creatinine	0.294	Appetite	0.393
Age	0.227	Pedal edema	0.365
White blood cell count	0.205	Anemia	0.325
Potassium	0.076	Coronary artery disease	0.236
		Pus cell clumps	0.214
		Bacteria	0.120

Table 4. The recursive features parameters number and their accuracy result

Number of Attributes	Decision Tree	Random Forest	Naïve Bayes	AVG Accuracy
23	0.754	0.926	0.737	0.805
12	0.837	0.934	0.762	0.844
6	0.966	0.100	0.956	0.974
3	0.881	0.946	0.623	0.816
2	0.902	0.951	0.983	0.945
1	0.939	0.943	0.774	0.885

E. Description Of The Proposed Techniques

Three machine learning techniques are applied on the dataset during the modeling stage to assess their capability of detecting CKD. These algorithms are Decision Tree (DT), Naive Bayes (NB), and random forest (RF).

- **Decision Tree:** It's a tree-structured prediction strategy for creating classification or regression models. It is the most intuitive classifier for any classification task. This method classifies the training dataset into branch-like segments by constructing a binary tree with a root node, internal nodes, and leaf nodes that describe the target

A root node, also called a decision node, indicates a choice that will result in the subdivision of all records into two or more mutually incompatible subsets. Whereas, internal nodes, also known as chance nodes reflect one of the options accessible at a certain position in the tree structure. Additionally, Leaf nodes also called end nodes, describe the final result of a sequence of decisions or events. Where the most crucial steps in developing a decision tree model are splitting, stopping, and pruning. Once the tree is constructed, all the records in the database are classified based on the tree. The decision tree is a non-parametric algorithm and can efficiently handle large and complex data sets.

- **Naïve Bayes (NB):** The Naïve Bayes classifier is a simple probabilistic algorithm that is based on Bayes theorem assumptions to deal with simple conditional probabilities [21]. It's a supervised learning algorithm that uses the maximum likelihood method to classify data. Naïve Bayes classifiers assume that the influence of a variable value on a particular class is independent of the values of other variables. This assumption is called class conditional independence. It's designed to make computations easier, therefore it's referred to as "Nave" [22]. One of the advantages of Nave Bayes is that it just requires a few amounts of training data to determine the classification parameters. To apply NB in any dataset, let $x = (x_1, x_2, \dots, x_n)$ denote to a feature records and $c = (c_1, c_2, \dots, c_n)$ indicate all possible categories that the record may belong to. The principle of a Naive Bayes classifier is to calculate the probabilities $p_1, p_2, \dots, p_n$ for x where p_1 is the probability that x belongs to category c_1. By determining the value of max$(p_1, p_2, \dots, p_n)$, we will identify which category the feature record x belongs to. Consequently, the classification problem can be considered as deciding the maximum value of the following Eq. (1):

$$P(c_j|x_1, x_2, .., x_n) = \frac{P(x_1, x_2, .., x_n|c_j)\, P(c_j)}{P(c_1, c_2, .., c_n)} \quad \dots\dots\dots\dots(1)$$

- **Random Forest (RF):** The Random Forest is made up of several decision trees, each tree represents a different decision by selecting a random number of attributes for splitting [23]. After producing a huge number of trees, they vote for the most common class for the given set of inputs. The more trees in the classification process, the more accurate the result will be. The random forest algorithm will do the overall estimate, and it has the final decision through voiding techniques. Moreover, the random forest technique can handle unbalanced data, is resistant to overfitting, and has significantly shorter runtimes [24].

IV. PERFORMANCE ANALYSIS AND RESULT

The performance of the prediction system was evaluated using the following metrics.

A. Classification Evaluation Measures

Different assessment parameters are used to evaluate the classifiers' performance.

In our study, the confusion matrix is used as a performance measurement of the classification models as shown in Figure 6. The confusion matrix is a summary of prediction results on a classification problem that is used to shows the miss-classifications (the false positive and false negatives) and corrects classification (the true positive and true negative) of the (DT, RF, and NB) algorithms on the classification of the Kidney diseases [25]. Also, we can calculate many performance criteria depending on the confusion matrix such as Accuracy, precision, recall, and f-Score.

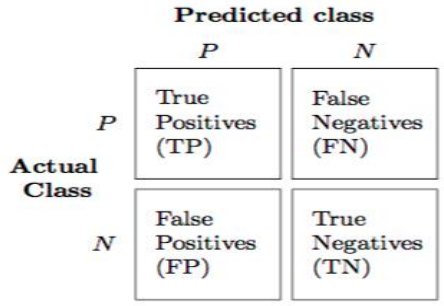

Figure. 6. The confusion matrix for two-class classification problem.

Where:

True positive (TP): The amount of positively identified classes that have been classed as positive "properly predicts positive class".

True Negative (TN): The number of negatively labeled classes that were classed as negative "predicts the Negative class appropriately".

False positive (FP): The number of wrongly identified as Positive negatively labeled classes "incorrectly predicts the Positive class".

False Negative (FN): The number of correctly categorized classes that have been wrongly classed as Negative "incorrectly forecasts the Negative class".

Accuracy: the measurement used to determine the percentage of correct predictions for the test data (rate of correct classifications) [26]. It is calculated by Eq. (2).

$$\frac{(TP + TN)}{(TP + TN + FP + FN)} \text{ ……..(2)}$$

Precision: The ability of the algorithm to determine how close prediction results are to each other, regardless of whether those predictions are accurate or not [27]. It is calculated by Eq. (3).

$$\frac{TP}{FP + TP} \text{ ……..(3)}$$

Recall: Also called Sensitivity or the True Positive Rate. It is the sick patient's categorization probability, which refers to a test's capacity to correctly identify people who have the condition. It is calculated by the Eq. (4).

$$\frac{TP}{(TP + FN)} \text{ ……..(4)}$$

F1-Score: combines the classifier's precision and recall into a single metric by taking their harmonic mean. It's mostly used to compare the results of two different classifiers [28]. It is calculated by the Eq. (5).

$$2*((precision*recall)/(precision + recall) \text{ ……..(5)}$$

B. Experimental Result

This section shows the experimental findings and analysis performed by this study. In this study, the experiments are conducted using Python 3.9.3 programming language through the Jupyter Notebook web application. Also, several libraries from Sciket-learn [29] have been used, which is a free software for the machine learning library in Python. Three classifiers, including RF, DT, and NB, are used in this study. Where the database is divided into two sections trainset and testset (65% and 35% respectively), the training set is used to build the classifier and the testing set is used to evaluate it [30]. The experimental results in the prediction of kidney problems of each model in terms of accuracy, F1-score, Recall, and Precision are demonstrated in Table 5.

Table 5. Performance Comparison Of Classification Techniques

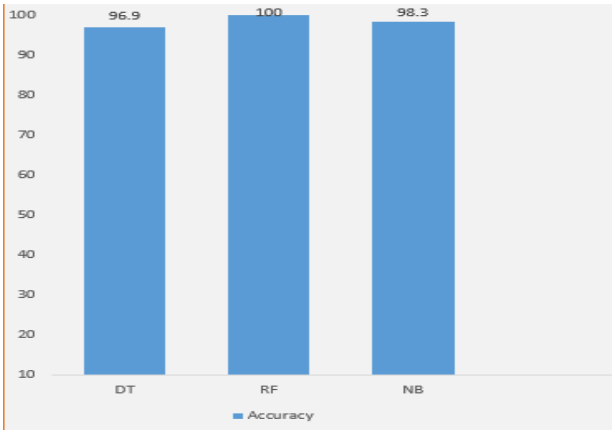

Figure. 7. Accuracy Graph.

By applying different data miming learning techniques and conducting experiments on the given dataset, we conclude that the RF method is relatively higher than the NB, and DT methods. According to Table 5, and Figure 7. Also, the RF shows the highest performance in terms of all measures (accuracy, recall, precision, and f-measure). So, The RF method can be adopted since it has an accuracy of 100% in the prediction of chronic kidney disease.

V. CONCLUSION

Early identification of CKD aids in the timely treatment of patients with the condition as well as preventing the disease from worsening. The medical sector has to be able to predict diseases early and treat them quickly. In this research, Chronic Kidney Disease is predicted using three distinct algorithms and and their performance is compared. Furthermore, this work highlights the importance of data preprocessing and feature selection when analyzing clinical data related to CKD. In this study, correlation coefficient, and recursive feature elimination have been used as feature selection methods to identify features needed for the prediction algorithm, and practically it decreases the number of medical examinations to be taken. From the experiments, we found that out of three classifiers NB, DT, and RF, the RF classifier outperforms the others in terms of precision F-measure, and accuracy

Criteria	DT	RF	NB
True Positive	77	80	78
True Negative	39	40	40
False Positive	1	0	0
False Negative	3	0	2
Accuracy	96.6%	100%	98.3%
precision	98%	100%	100%
Recall	96%	100%	97%
F1-score	96%	100%	98%

REFRENCES

[1] A. A. Al-Sayyari and F. A Shaheen, "End stage chronic kidney disease in Saudi Arabia", *Saudi Medical Journal*, vol. 32, no.4, pp. 339–346, April 2011.

[2] P. Tikariha and P. Richhariya, "Comparative Study of Chronic Kidney Disease Prediction Using Different Classification Techniques", Lecture Notes in Networks and Systems. Springer Singapore, 2018.

[3] C. E. Nulsen, A. M. Fox, and G. R. Hammond, "Adult Chronic Kidney Disease: Neurocognition in Chronic Renal Failure", Neuropsychology Review, vol. 20, no. 1, pp. 33–51, 2010.

[4] N. A Almansour *et al.*, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study", *Computers in Biology and Medicine*, vol. 109, pp. 101–111, April 2019.

[5] Parul Sinha and Poonam Sinha, "Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM", International Journal of Engineering Research and, vol. 4, no. 12, pp. 608–612, December 2015.

[6] Lj. Rubini and P. Eswaran, "Generating comparative analysis of early stage prediction of Chronic Kidney Disease", International Journal Of Modern Engineering Research (IJMER), vol. 5, no.49, pp. 49–55, 2015.

[7] K. Shankar et al., "Optimal Feature Selection for Chronic Kidney Disease Classification using Deep Learning Classifier", 2018 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2018. IEEE, pp. 1–5,2018.

[8] I. A. Pasadana et al., "Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques", Journal of Physics: Conference Series, vol. 1255, no.1, 2019.

[9] Machine Learning Repository, Center for Machine Learning and Intelligent Systems https://archive.ics.uci.edu/ml/index.php.

[10] S. Almuhaideb and M. E. B. Menai, "Impact of preprocessing on medical data classification", Frontiers of Computer Science, vol. 10, no.6, pp. 1082–1102, 2016.

[11] A. Christy, M. G. Gandhi, and S. Vaithyasubramanian, "Cluster based outlier detection algorithm for healthcare data", Procedia Computer Science, vol. 50, pp. 209–215, 2015.

[12] M. E. Bowen et al., "Random blood glucose: A robust risk factor for type 2 diabetes", Journal of Clinical Endocrinology and Metabolism, vol. 100, no. 4, pp. 1503–1510, 2015.

[13] G. Gheno et al., "Variations of serum potassium level and risk of hyperkalemia in inpatients receiving low-molecular-weight heparin," Eur. J. Clin. Pharmacol, vol. 59, no. (5-6), pp. 373-377, 2003.

[14] D. A. Henry, "In The Clinic: Hyponatremia," Ann. Intern. Med., vol. 163, no. 3, pp. ITC1-ITC19, 2015.

[15] Z. Hu *et al.*, "Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record", *Journal of Biomedical Informatics*, vol. 68, pp. 112–120, 2017.

[16] S. G. K. Patro and K. K. sahu, "Normalization: A Preprocessing Stage", *Iarjset*, pp. 20–22, 2015.

[17] C. Saranya and G. Manikandan, "A study on normalization techniques for privacy preserving data mining", *International*

Journal of Engineering and Technology, vol. 5, no.3, pp. 2701–2704, 2013.

[18] W. Bi and J. T. Kwok, "Efficient multi-label classification with many labels", *30th International Conference on Machine Learning, ICML 2013*, vol. 28, no. 2, pp. 1442–1450, 2013.

[19] R. Zebari *et al.*, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction", *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 56–70, 2020.

[20] B. R. Patel and K. K. Rana, "A Survey on Decision Tree Algorithm For Classification", Ijedr, vol. 2, no. 1, pp. 1–5. 2014.

[21] D. Berrar, "Bayes' theorem and naive bayes classifier", Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, pp. 403–412, January 2018.

[22] P. Chandrasekar and K. Qian, "The Impact of Data Preprocessing on the Performance of a Naïve Bayes Classifier", Proceedings - International Computer Software and Applications Conference, vol. 2, pp. 618–619, 2016.

[23] F. Livingston, '"Implementation of Breiman's Random Forest Machine Learning Algorithm", Machine Learning Journal Paper, pp. 1–1, 2005.

[24] N. M. Abdulkareem and A. M. Abdulazeez, "Machine learning classification based on Radom Forest Algorithm: A review", Journal of Science and Business, pp. 128–142, 2021.

[25] E. Beauxis-aussalet and L. Hardman, "Visualization of Confusion Matrix for Non-Expert Users", IEEE Information Visualization 2014.

[26] M. S. Esfahani and E. R. Dougherty, "Effect of separate sampling on classification accuracy", Bioinformatics, vol. 30, no. 2, pp. 242–250, 2014.

[27] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation", IEEE, pp. 37–63, 2020.

[28] D. K. Barupal and O. Fiehn, "Generating the blood exposome database using a comprehensive text mining and database fusion approach", Environmental Health Perspectives, vol. 127, no. 9, pp. 2825–2830. 2019.

[29] F. Pedregosaet l., " Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[30] C. C. Aggarwal et al., "Active learning: A survey", Data Classification: Algorithms and Applications, pp. 571–605. 2014.