Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Diagnose of Chronic Kidney Diseases by Using Naive Bayes Algorithm

Noor S. Abd [a] , *Dr.prof. Dhahir A . Abdullah* [b]

[a] Department of Computer Science, College of Science, University of Diyala, Diyala, Iraq, 32001.
E_ mail : scicompms17@uodiyala.edu.iq.

[b] Department of Computer Science, College of Science, University of Diyala, Diyala, Iraq, 32001.
E_ mail : Dhahair@yahoo.com.

ARTICLE INFO

Article history:
Received: 02 /06/2021
Rrevised form: 12 /06/2021
Accepted : 07 /07/2021
Available online: 09 /07/2021

Keyword:

Chronic Kidney disease,
Naïve Bayes Algorithm.

ABSTRACT

Chronic kidney disease (CKD) develops gradually, usually after months or years when the kidneys lose function. In general, it may not be detected before it loses 25% of its functionality. Patients may begin to not recognize kidney failure because kidney failure may not give any symptoms at first. Treatment for kidney failure aims to control the causes and slow the progression of kidney failure. If the treatments are insufficient, the patient is in the end stage of kidney failure and the last treatment is dialysis or a kidney transplant. at this time. Therefore, it is necessary to make an early diagnosis to avoid reaching the stage of kidney failure. We conclude in this paper that the Naive Bayes algorithm is one of the best algorithms for diagnosing diseases with high accuracy of 99.24% and time of 0.003 seconds approximately because it is suitable for this kind of dataset.

MSC. 41A25; 41A35; 41A36

DOI : https://doi.org/10.29304/jqcm.2021.13.2.819

1. Introduction

The kidneys are a vital organ for the proper functioning of the human body. Its function is to filter blood, remove waste products, and control fluid balance in the body and urine formation. Chronic Kidney Failure (CKD) is a condition in which kidney function is altered, and the ability to function properly decreases, leading to an increase in the amount of waste products in the blood that makes the human body sick in the long term [1]. People with high blood pressure and diabetes and those who have family members with chronic kidney disease are at greater risk of

∗Corresponding author: Noor S. Abd

Email addresses: scicompms17@uodiyala.edu.iq

Communicated by: *Dr. Rana Jumaa Surayh aljanabi.*

developing kidney disease. The purpose of medical diagnosis is to mine useful information from the massive medical datasets which are accumulated frequently [2].

Data mining and machine learning can be used as an informative tool to extract useful information which helps pathologists and doctors import decisions making [3]. Machine learning researchers create algorithms that can improve a solution to a problem that contains huge data such as medical data. Moreover, the amount of relevant problem-related data available improves the accuracy of the solution [4]. Data mining, which is a branch of machine learning and artificial intelligence, has evolved to such an extent that it can now be used in a variety of areas, including risk assessment, industrial process control, healthcare, insurance, financial reporting, and forecasting of expense payments. business among many other fields [5].

Advantages of this Model (Naïve Bayes) "It is a relatively simple algorithm to understand and build". And "It is faster to predict classes using this algorithm than many other classification algorithms". In addition to "it can be easily trained using a small dataset suitable for this Kind of dataset which used in this paper" [19].

NOMENCLATURE

Aradius of

Bposition of

Cfurther nomenclature continues down the page inside the text box

1. Related Work

This section reviews some previous studies and explains the different techniques used to diagnose chronic kidney disease.

- **Polat, H., et al, 2017, [3]** In this study, encapsulation and filtering methods were used CKD data set. And achieved 89% in Naïve Bayes Algorithm. Naive Bayes and ResolutionTree-J48, were selected subsets and compare performance workbooks. The results showed that Naïve Bayes on Nimda data set reduced by Classifier Highest F scale and rated Decision Tree-J48 when downgraded Code Red I dataset by WrapperSubsetEval got the extension Highest measure F. Classifier SVM on Slammer Reduced the dataset by WrapperSubsetEval achieved the highest performance measurements.
- **Alassaf, R. A., et al, 2018, [4].** In this study, Saudi medical records were investigated for the first time in the process of diagnosing CKD using machine learning techniques. The authors used correlation coefficient and recursive feature elimination for feature selection. Then, four classification algorithms were explored, namely: ANN, SVM, Naïve Bayes, and k-NN. The performance of each of these classifiers was examined by the classification accuracy, precision, recall, and f-measure achieved by the classifier. ANN, SVM, and NB all achieved an accuracy of 98 % while k-NN achieved an accuracy of 93.9%.
- **Padmanaban, K. A., et al, 2020 [5]** Chronic kidney disease has been analyzed and predicted for different classifiers: Naïve Bayes, SVM, KNN, and Decision tree. To compare the performance of these classifier algorithms, the WEKA tool has been used. From the performance result, it is observed that the decision tree algorithm provides the highest accuracy of 98%. The second most accurate classifier is SVM with an accuracy of 97.75%. It is also observed that the implementation of the ranking algorithm increases the performance for predicting CKD but a correct number of attributes must be selected. Some major factors like age, RBC, blood pressure, etc. have been considered for classification. Other parameters like nutrition, accommodation status, clean water availability, surroundings can be considered for the detection of CKD. In the future, the performance of other classifiers like ANN, Fuzzy logic can be compared using the WEKA tool for a similar situation and dataset.
- **Deepika, B., et al, 2020 [6]** This project is a medical sector application that helps medical practitioners in predicting CKD disease based on CKD parameters. It is automation to predict CKD and it identifies the disease and its stages effectively and economically method. It is accomplished through the KNN algorithms with 97% accuracy. And Naive Bayes classification algorithms 91% accuracy. This classification technique

comes under the data mining technique. This algorithm takes CKD parameters as input and predicts disease based on old CKD patient data.

3.Material and Method

3.1 Data Mining Applications in Healthcare

The huge amount of data available in the health sector and the need to extract knowledge from these is huge Data makes data mining techniques the most effective solution for processing such a quantity of data and extract knowledge [3]. Data mining is a process analyze and summarize the data into useful information that can be used to increase revenue, reducing costs, or both. It's the process to find relationships or patterns between tens of fields in large relational data. mining data consists of main components: information extraction, data storage and management, access provision, and analyze data and present the data in a useful format [8].

3.2 Naïve Bayes

Naive Bayes classifier is a powerful algorithm for the classification task. Even with working on a data set with millions of records with some attributes, the Naïve Bayes approach is best to use [3]. In Naïve Bayes, the probability of its being a target class is calculated in which the instance is classified as belonging to the target class of highest probabilities [8].

Bayes' theory uses the posterior probability and the previous probability. It represents the pre-probability of an event or hypothesis of the original probability where it was obtained before obtaining any additional information. The revised probability of the event through the use of additional information or evidence that was obtained is known as the posterior probability [3].

The theory is written as equation [9]:

$$P(C_i|A) = \frac{P(A|C_i)P(C_i)}{P(A)}$$

Where :
The prior probability of A is P(A)
The prior probability of Ci is P(Ci)
The posterior probability of A given Ci as P(A| Ci)
The posterior probability of Ci given A as P(Ci |A)

The classifier of Naive Bayes is a probabilistic simple and convenient classifier that depends on the application of the Bayes theorem. Naive Bayes regards each component of the attributes as an independent variable [10].

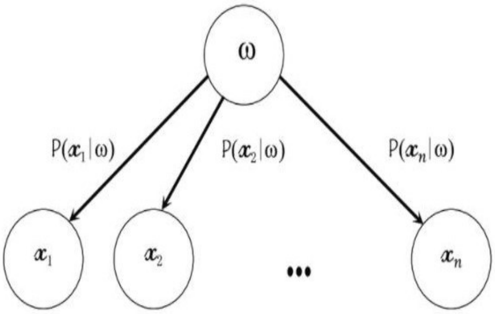


Figure (1): Naïve Bayes Model With the Assumption of Conditional Independence.

1. Proposed System Framework

In general, the proposed system involves four main phases: preprocessing, statistical analysis, feature selection, and classification. Each stage includes a set of sub-steps as figure (2) shows:

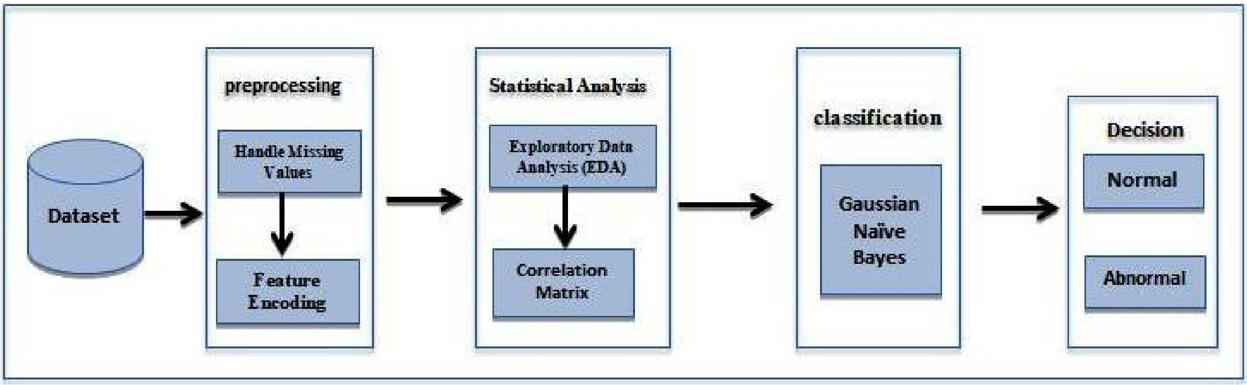


Figure (2): Proposed System Framework

The following sections provide a detailed explanation for each phase of the proposed and framework system:

4.1 Preprocessing Stage:

Pretreatment is one of the most important issues in the disease diagnosis and classification system. A set of data was adopted in the application of the proposed system containing two types of numerical and nominal data. When preparing the data, initially, the identifier drop column is executed. Because the identity of the patient is taken randomly and independently in the useful classification used in the proposed system, it does not contain any information in the classification. Then, the missing data is processed as will be explained later, and then a scan of all the features in the data set is done to see which features are similar and which are duplicates for deletion. This process is achieved depending on the Gaussian distribution.

A- Handle Missing Values

The data set contains two types of data: numerical and nominal, if the missing values are numerical, they are replaced by using the average value of the column, while in the second type the missing values are replaced by taking the adjacent value.

The data set used in this paper contains two types of data: 10 numerical and 13 nominal attributes.

Table 1 - The features before and after handling the missing values of the table shows that have no missing values after preprocessing it.

	The features	Number of missing value before handle missing values	Number of missing value after handle missing values
1	Age	9	0
2	bp	12	0
3	sg	47	0
4	al	46	0
5	su	49	0
6	rbc	152	0
7	pc	65	0
8	pcc	4	0
9	ba	4	0
10	bgr	44	0
11	bu	19	0
12	sc	17	0
13	sod	87	0
14	pot	88	0
15	hemo	52	0
16	pcv	71	0
17	wbcc	106	0
18	rbcc	131	0
19	htn	2	0
20	dm	3	0

21	cad	2	0
22	appet	1	0
23	pe	1	0
24	ane	1	0
25	class	0	0

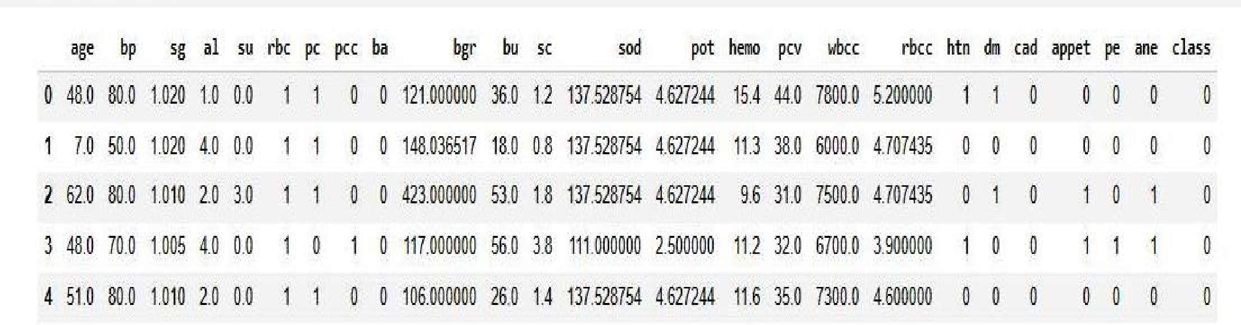
B. Encoding Categorical Features

In this step, converts each categorical value under a specified feature to a numerical value. To be dealt with in mathematical operations, as many machine learning algorithms cannot work with categorical data directly. It data must be digital.

this step will convert each nominal attribute to 1 or 0 value to use it in arithmetic operations of train and test, as shown in table (2).

Table (2)

Sample of encoding categorical features



	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wbcc	rbcc	htn	dm	cad	appet	pe	ane	class
0	48.0	80.0	1.020	1.0	0.0	1	1	0	0	121.000000	36.0	1.2	137.528754	4.627244	15.4	44.0	7800.0	5.200000	1	1	0	0	0	0	0
1	7.0	50.0	1.020	4.0	0.0	1	1	0	0	148.036517	18.0	0.8	137.528754	4.627244	11.3	38.0	6000.0	4.707435	0	0	0	0	0	0	0
2	62.0	80.0	1.010	2.0	3.0	1	1	0	0	423.000000	53.0	1.8	137.528754	4.627244	9.6	31.0	7500.0	4.707435	0	1	0	1	0	1	0
3	48.0	70.0	1.005	4.0	0.0	1	0	1	0	117.000000	56.0	3.8	111.000000	2.500000	11.2	32.0	6700.0	3.900000	1	0	0	1	1	1	0
4	51.0	80.0	1.010	2.0	0.0	1	1	0	0	106.000000	26.0	1.4	137.528754	4.627244	11.6	35.0	7300.0	4.600000	0	0	0	0	0	0	0

4.2 Statistical Analysis

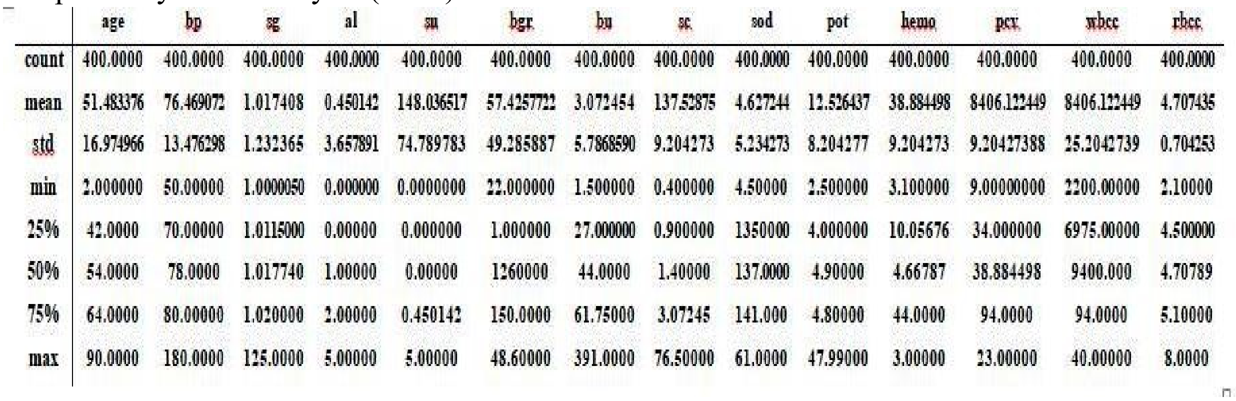
Here will show the result of the Exploratory Data Analysis and Correlation Matrix**.**

A. Exploratory Data Analysis (EDA)

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

Table (3)

Exploratory Data Analysis (EDA)



	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo	pcv	wbcc	rbcc
count	400.0000	400.0000	400.0000	400.0000	400.0000	400.0000	400.0000	400.0000	400.0000	400.0000	400.0000	400.0000	400.0000	400.0000
mean	51.483376	76.469072	1.017408	0.450142	148.036517	57.4257722	3.072454	137.52875	4.627244	12.526437	38.884498	8406.122449	8406.122449	4.707435
std	16.974966	13.476298	1.232365	3.657891	74.789783	49.285887	5.7868590	9.204273	5.234273	8.204277	9.204273	9.20427388	25.2042739	0.704253
min	2.000000	50.00000	1.0000050	0.000000	0.0000000	22.000000	1.500000	0.400000	4.50000	2.500000	3.100000	9.00000000	2200.00000	2.10000
25%	42.0000	70.00000	1.0115000	0.00000	0.000000	1.000000	27.000000	0.900000	1350000	4.000000	10.05676	34.000000	6975.00000	4.500000
50%	54.0000	78.0000	1.017740	1.00000	0.00000	1260000	44.0000	1.40000	137.0000	4.90000	4.66787	38.884498	9400.000	4.70789
75%	64.0000	80.00000	1.020000	2.00000	0.450142	150.0000	61.75000	3.07245	141.000	4.80000	44.0000	94.0000	94.0000	5.10000
max	90.0000	180.0000	125.0000	5.00000	5.00000	48.60000	391.0000	76.50000	61.0000	47.99000	3.00000	23.00000	40.00000	8.0000

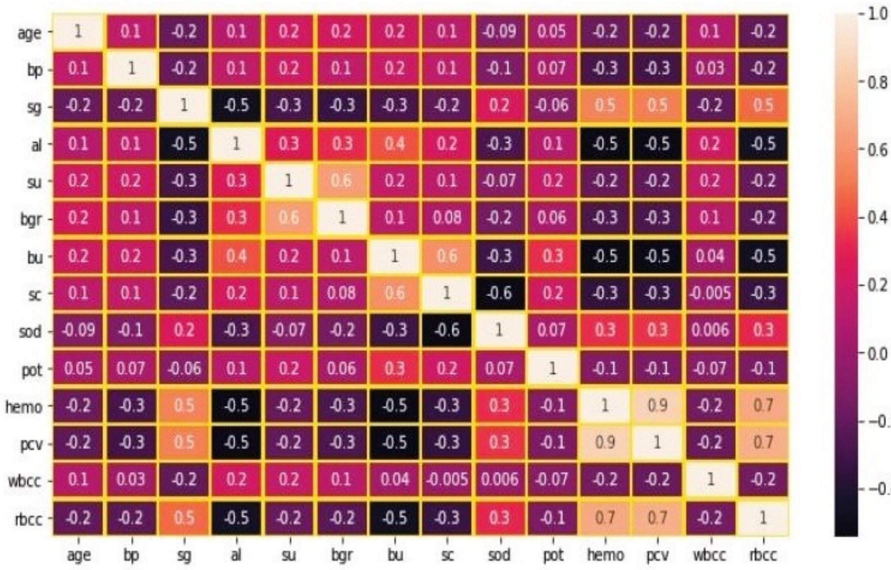
B. Correlation Matrix

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables and is diagnostic for advanced analysis**.**

Pearson correlation measures the linear association between two variables. It has a value between -1 and 1 where:

- -1 indicates a perfect negative linear correlation between two variables.

- 0 indicates no linear correlation between two variables.
- 1 indicates a perfect positive linear correlation between two variables.



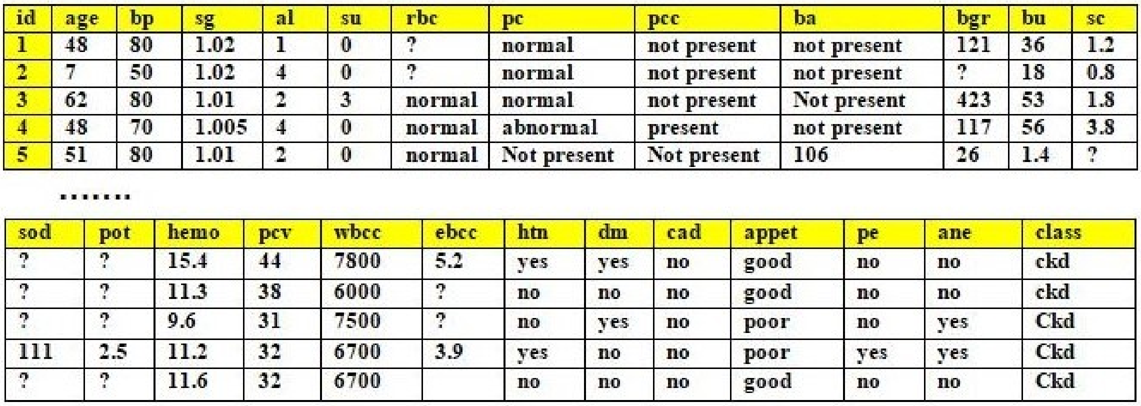
Figurer (3) Correlation Matrix

5.Dataset Description

This experiment was performed using the CKD dataset available in the UCI machine learning, The CKD dataset was released in July 2015 by Apollo Hospitals, The dataset consists of 25 attributes, including 11 numerical and 13 nominal attributes, These 400 instances consist of 250 CKD patients and 150 non-CKD patients. The data set is divided into two parts, 70% of the first part is CKD data for training, and the second part uses about 30% of the tests to test the classification states. As a result, this dataset is used for binary classification of either CKD or non-CKD **[19].**

Table (4)

Sample of kidney Dataset.



id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc
1	48	80	1.02	1	0	?	normal	not present	not present	121	36	1.2
2	7	50	1.02	4	0	?	normal	not present	not present	?	18	0.8
3	62	80	1.01	2	3	normal	normal	not present	Not present	423	53	1.8
4	48	70	1.005	4	0	normal	abnormal	present	not present	117	56	3.8
5	51	80	1.01	2	0	normal	Not present	Not present	106	26	1.4	?

....

sod	pot	hemo	pcv	wbcc	ebcc	htn	dm	cad	appet	pe	ane	class
?	?	15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
?	?	11.3	38	6000	?	no	no	no	good	no	no	ckd
?	?	9.6	31	7500	?	no	yes	no	poor	no	yes	Ckd
111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	Ckd
?	?	11.6	32	6700		no	no	no	good	no	no	Ckd

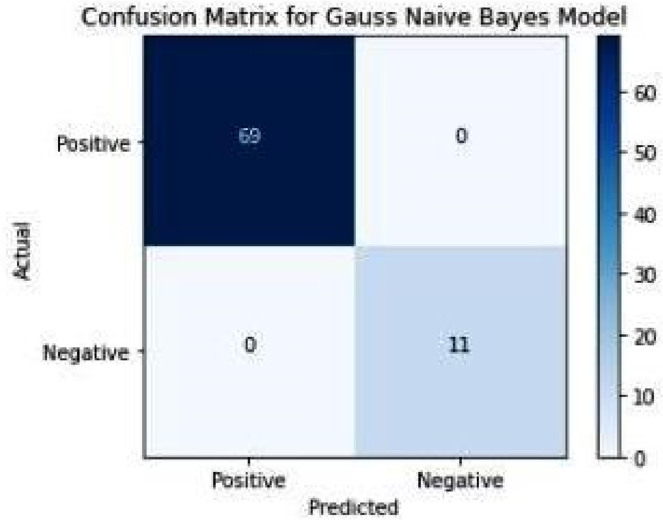
Where :
age - age
bp - blood pressure
sg - specific gravity
al - albumin

su - sugar
rbc - red blood cells
pc - pus cell
pcc - pus cell clumps
ba - bacteria
bgr - blood glucose random
bu - blood urea
sc - serum creatinine
sod - sodium
pot - potassium
hemo - hemoglobin
pcv - packed cell volume
wc - white blood cell count
rc - red blood cell count
htn - hypertension
dm - diabetes mellitus
cad - coronary artery disease
appet - appetite
pe - pedal edema
ane - anemia
class – class

6. The performance of Naive Bayes Algorithm

The GNB classifier in the proposal model divides the data set into an 70% training set that is used to train the classifier, while a 30% test set is used to evaluate the performance accuracy of the classification model. Training phase for CKD classification To be the required classification on the input data set, a GNB classifier with no error rate was implemented.

Confusion matrix of Naïve Bayes shows as figure (5) :



Figurer (4) Confusion Matrix

$$\text{Accuracy} = \left(\frac{\text{Tp+TN}}{\text{Tp+TN+FP+FN}} \right) * 100 \qquad \textbf{(2)}$$

$$\text{Sensitivity} = (\frac{Tp}{Tp+ FN}) * 100 \qquad \textbf{(3)}$$

$$\text{Specificity} = (\frac{TN}{TN+FP}) * 100 \qquad \textbf{(4)}$$

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

Table (5)

The performance of Naïve Bayes :

1	Accuracy	0.99
2	Sensitivity or Recall	0.94
3	Specificity	1.0
4	Precision	1.0

8. Comparison between other Existing Works and the this Proposed Work.

Data Set	Researchers	Methodology	Accuracy
CKD	Polat, H., et al, 2017, [3]	In this study, encapsulation and filtering methods were used CKD data set.	89 %
CKD	Padmanaban, K. A., et al, 2020 [5]	Chronic kidney disease has been analyzed and predicted for different classifiers: Naïve Bayes	98 %
CKD	Deepika, B., et al, 2020 [6]	It is a mechanism for predicting CKD that identifies the disease and its stages efficiently and economically using the Naive Bayes classification algorithm	91 %.
CKD	This Proposed Work	Naïve Bayes	99.24 %

Conclusion

This paper is a medical sector application that helps medical practitioners in predicting the CKD disease based on the CKD parameters. We conclude that the Naive Bayes algorithm is one of the best algorithms for classification and diagnosis in the medical fields. We recommend the use of other algorithms with the same level of accuracy in diagnosis and speed in time.

Acknowledgments

The authors would like to thank Diyala University and Al_ Qadisiyah University - Iraq for their support in the present work.

References

[1] Khan, B., Naseem, R., Muhammad, F., Abbas, G., & Kim, S. (2020). An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy. IEEE Access, 8, 55012-55022.

[2] Singh, V. P., Srivastava, S., & Srivastava, R. (2017). Effective mammogram classification based on center symmetric-LBP features in wavelet domain using random forests. Technology and Health Care, 25(4), 709-727.

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

[4] Alassaf, R. A., Alsulaim, K. A., Alroomi, N. Y., Alsharif, N. S., Aljubeir, M. F., Olatunji, S. O., ... & Alturayeif, N. S. (2018, November). Preemptive diagnosis of chronic kidney disease using machine learning techniques. In 2018 international conference on innovations in information technology (IIT) (pp. 99-104). IEEE.

[5] Padmanaban, K. A., & Parthiban, G. (2016). Applying machine learning techniques for predicting the risk of chronic kidney disease. Indian Journal of Science and Technology, 9(29), 1-6.

[6] Devika, R., Avilala, S. V., & Subramaniyaswamy, V. (2019, March). Comparative study of classifier for chronic kidney disease prediction using naive Bayes, KNN, and random forest. In 2019 3rd International conference on computing methodologies and communication (ICCMC) (pp. 679-684). IEEE.

[7] Deepika, B., Rao, V. K. R., Rampure, D. N., Prajwal, P., & Gowda, D. G. (2020). Early prediction of chronic kidney disease by using machine learning techniques. Amer. J. Comput. Sci. Eng. The survey, 8(2), 7.

[8] Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2019). A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. IEEE Access, 8, 20991-21002.

[9] Tazin, N., Sabab, S. A., & Chowdhury, M. T. (2016, December). Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique. In 2016 International Conference on Medical Engineering, Health Informatics and Technology (Meditec) (pp. 1-6). IEEE.

[10] Kaviani, P., & Dhotre, S. (2017). Short survey on naive Bayes algorithm. International Journal of Advanced Engineering and Research Development, 4(11), 607-611.

[11] Sankara Subbu, R. (2017). Brief Study of Classification Algorithms in Machine Learning.

[12] Balakrishnan, S. (2020). Feature Selection Using Improved Teaching Learning Based Algorithm on Chronic Kidney Disease Dataset. Procedia Computer Science, 171, 1660-1669.

[13] Matter Nassir, L. (2017). A comparison among methods for estimation of the parameter of the Maxwell-Boltzmann distribution using simulation. Journal of Al-Qadisiyah for Computer Science and Mathematics, 6(2), 186-200. Retrieved from https://qu.edu.iq/journalcm/index.php/journalcm/article/view/118.

[14] Abduljabbar Saad, I. (2018). An Efficient Classification Algorithms for Image Retrieval Based Color and Texture Features. Journal of Al-Qadisiyah for Computer Science and Mathematics, 10(1), Comp Page 42 - 53. https://doi.org/10.29304/jqcm.2018.10.1.350.

[15] Hssein, A. (2021). Estimate Reliability Function of Inverse Lindley for Strength – Stress Models. Journal of Al-Qadisiyah for Computer Science and Mathematics, 12(4), Math Page 61-. https://doi.org/10.29304/jqcm.2020.12.4.725.

[16] El-Husseini, Z. alabdin, & Flaih, A. (2019). Estimating Mediation and direct effects of the multiple model with application Kidney. Journal of Al-Qadisiyah for Computer Science and Mathematics, 11(4), Stat Page 1-20. https://doi.org/10.29304/jqcm.2019.11.4.617.

[17] Neamah Hussein, K. (2018). Video Frames Edge Detection of Red Blood Cells: A Performance Evaluation. Journal of Al-Qadisiyah for Computer Science and Mathematics, 10(1), Comp Page 16 - 27. https://doi.org/10.29304/jqcm.2018.10.1.347.

[18] A. Mohammed, H. (2018). Design and Implementation of Smart Dust Sensing System for Baghdad City. Journal of Al-Qadisiyah for Computer Science and Mathematics, 10(1), Math Page 80 - 87. https://doi.org/10.29304/jqcm.2018.10.1.353 .

[19] https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease