IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Chronic Kidney Disease (CKD) Prediction Using Machine Learning Algorithms

Ms. G.OBULAMMA [#1], Mr. L.D.RAMAYYA [#2], Mr. D.D.D.SURIBABU [#3]

[#1]M.Tech Student, [#2]Associate Professor, [#3]Associate Professor & HOD
Department of Software Engineering
International School of Technology and Sciences (Women)
(Affiliated to JNTUK),East Gonagudem, Rajanagaram, Rajamahendravaram,
East Godavari District, Andhra Pradesh, India– 533294.

ABSTRACT

With a high rate of morbidity and mortality as well as the ability to spread other diseases, chronic kidney disease (CKD) is a major worldwide health concern. Patients frequently overlook the disease in the early stages of CKD since there are no evident symptoms. Early diagnosis of CKD enables patients to receive effective treatment in time to slow the disease's progression. Due to their quick and precise detection capabilities, machine learning models can help therapists accomplish this goal efficiently. In this research, we suggest a machine learning approach to CKD diagnosis. The CKD data set, which contains a significant number of missing values, was downloaded from the website KAGGLE. For object data types (strings), we replaced the missing values with the most frequent object using the mean value as a filler (string).Since patients may overlook particular measurements for a variety of reasons, missing values are typically observed in real-world medical scenarios. Four machine learning algorithms—Logistic Regression, SVM, Random Forest Classifier, and Decision Tree Classifier—were applied to create models after successfully completing the incomplete data set. Random Forest has the highest accuracy of these machine learning models.

KEYWORDS:

Logistic Regression, SVM, Random Forest Classifier, Decision Tree Classifier, Chronic Kidney Disease

1. INTRODUCTION

The adverse effects of chronic kidney disease (CKD), which include renal failure, cardiovascular disease, and early death, are a serious public health concern worldwide [1]. Chronic kidney disease (CKD), which rose from 27th place in 1990 to 18th place in 2010 according to the Global Burden of Disease Study (GBDS), is one of the top causes of death worldwide [2]. Over 500 million individuals globally suffer from chronic renal disease [3, 4], with South Asia and sub-Saharan Africa bearing a disproportionately high burden [5]. In high-income countries, there were 110 million people with CKD (men 48.3 million, women 61.7 million), but in low- and middle-income nations, there were 387.5 million [6].

In Bangladesh, a heavily populated developing nation in Southeast Asia, chronic kidney disease rates continue to climb. In a global survey of six countries, including Bangladesh, the prevalence of CKD was estimated to be 14% [7]. A

26% prevalence of chronic renal disease was found in a different study among urban Dhaka residents over the age of 30 [8], and a 13% prevalence was found in a different study among urban Dhaka inhabitants over the age of 15 [9]. One-third of rural individuals in Bangladesh were at risk of developing CKD, which was frequently misdiagnosed at the time, according to a community-based prevalence survey conducted there in 2013 [10].

On the other hand, the observed variations in CKD prevalence between Bangladeshi categories could be attributed to a variety of variables, including the study period, the cross-sectional research design, and the geographic distribution of urban and rural areas. CKD prevalence varies by age group, gender, socioeconomic level, and geographic region, claims one study. [7].

Patients with end-stage renal disease (ESRD), which necessitates costly treatment options like dialysis and kidney transplantation [11], are more likely to have chronic kidney disease (CKD), which increases their risk of developing ESRD and increases their risk of long-term medical and psychological problems [12, 13]. Additionally, uncontrolled diabetes and high blood pressure are two risk factors that have an impact on the occurrence of CKD globally. In order for decision-makers (such as ministries, insurers, hospital administrators, and so on) to prevent an increase in the number of patients, it is critical from the standpoint of public health to be able to estimate CKD occurrence trends. As it has been shown that lifestyle changes (weight loss, improved diet, increased physical activity, decreased alcohol consumption, avoided smoking, early referral to nephrologists, appropriate medication use, and treatment options to manage other risk factors) are the most effective, rising population screening for CKD-related risks and awareness programmes are examples of such mitigation strategies. Creating suitable hemodialysis facilities and providing personnel with the necessary training are additional mitigating techniques.

2. EXISTING SYSTEM AND ITS LIMITATIONS

In the existing system there was no proper method to identify the chronic kidney disease prediction using data mining algorithms. The following are the main limitations in the existing system.

LIMITATION OF PRIMITIVE SYSTEM

1. More Time Delay in finding the route cause of kidney diseases
2. There is no prevention technique due to late prediction.
3. There is no early prediction of chronic kidney disease.
4. There is no method to identify the kidney diseases using ML algorithms

3. PROPOSED SYSTEM AND ITS ADVANTAGES

In proposed system we are applying different Machine Learning Algorithms to detect Chronic Kidney disease. In this Project we used the features like age, sugar(su),blood pressure(BP),hyper tension(htn),pus cell(pu), RBC, WBC, Coronary Artery Disease (cad) , etc,to classify CKD.After data pre-processing a very well cleaned data is inputed to these algorithms(Logistic regression, Support Vector Machine, Random forest Classifier, Decision Tree Classifier).So each algorithm has shown 100% Accuracy then we considered Time Complexity of each algorithm. After comparing each algorithm's Time complexity Decision Tree has shown optimum time. So we predict decision tree as best algorithm.

ADVANTAGES OF THE PROPOSED SYSTEM

1) By using data mining techniques it takes less time for the prediction of the disease with more accuracy.

2) In this paper we survey different papers in which one or more algorithms of data mining used for the prediction of CKD disease.

3) Applying data mining techniques to CKD disease treatment data can provide as reliable performance as that achieved in diagnosing kidney disease.

4. IMPLEMENTATION PHASE

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment. The front end of the application takes Google Collaboratory and as a Back-End Data base we took UCI Chronic kidney Patients Records as dataset. Here we are using Python as Programming Language to implement the current application. The application is divided mainly into following 5 modules. They are as follows:

1. Import Necessary Libraries
2. Load Dataset Module
3. Data Pre-Processing
4. Train the Model Using Several ML Algorithms
5. Find the Performance of ML Algorithms

1) IMPORT LIBRARY MODULE

In this module initially we need to import all the necessary libraries which are required for building the model. Here we try to use all the libraries which are used to convert the data into meaningful manner. Here the data is divided into numerical values which are easily identified by the system, hence we try to import numpy module and for plotting the data in graphs and charts we used matplot library.

2) LOAD DATASET MODULE

In this module the we try to load the dataset which is downloaded or collected from Kaggle repository. Here we store the dataset names as `ckdisease.zip` file and this dataset contains the following information such as :

```
df=pd.read_csv('kidney_disease.csv')
df.head()
```

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd

5 rows × 26 columns

Each and every attribute contains some information which are tested and collected based on individual patient id.

3) DATA PRE-PROCESSING MODULE

Here in this section we try to pre-process the input dataset and find out if there are any missing values or in-complete data present in the dataset. If there is any such data present in the dataset, the application will ignore those values and load only valid rows which have all the valid inputs.

4) TRAIN THE MODEL

Here we try to train the current model on given dataset using several ML classification algorithms and then try to find out which algorithms suits best in order to identify and classify the input dataset accurately and efficiently. Here we try to use following algorithms on input dataset such as:

1. Logistic Regression
2. Support Vector Machine
3. Decision Tree
4. Random Forest

5) PERFORMANCE ANALYSIS MODULE

Here in this module we try to compare each and every classification algorithm on given input dataset and then try to find out which one suits best for finding the accurate results. Finally we will identify the best algorithm which gives accurate results in very less time. Here we can see **Random Forest** gives more accurate result compared with other ML Algorithms.

5. EXPERIMENTAL RESULTS

In this section we try to design our current model using Python as programming language and we used Google Collab as working environment for executing the application. Now we can check the performance of our proposed application as follows:

IMPORT LIBRARIES

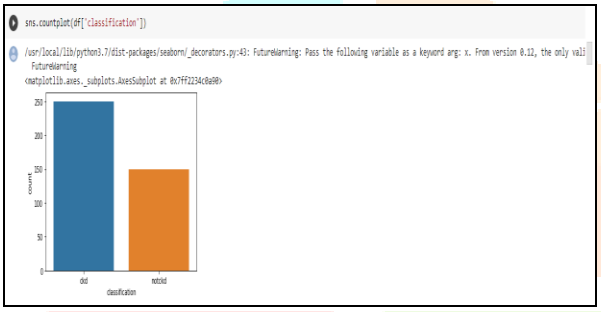
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import time

df=pd.read_csv('kidney_disease.csv')
df.head()
```

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd

The above window clearly represents the list of several modules used in our application.

REMOVE DUPLIATE DATA

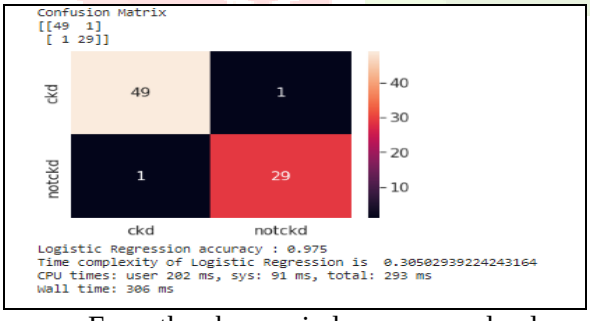


```
sns.countplot(df['classification'])
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only vali
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7ff2234c8a90>
```


From the above window we can see duplicate data is removed.

CONFUSION MATRIX



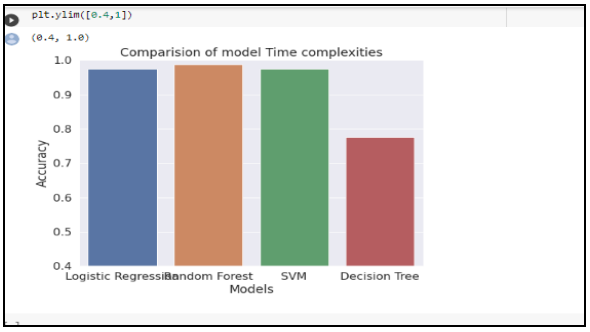
```
Confusion Matrix
[[49  1]
 [ 1 29]]
```



```
Logistic Regression accuracy : 0.975
Time complexity of Logistic Regression is  0.3050293922424316
CPU times: user 202 ms, sys: 91 ms, total: 293 ms
Wall time: 306 ms
```

From the above window we can clearly see confusion matrix is formed.

COMPARISION OF ML-ALGORITHMS



```
plt.ylim([0.4,1])
(0.4, 1.0)
```


From the above window we can clearly the comparison of several ML algorithms and we can see Random Forest has good accuracy.

6. CONCLUSION

The decision tree method and logistic regression can be used to forecast chronic kidney disease more precisely, the study's findings suggest. Their accuracy was 97 percent and precision was 96.25 percent, according to the study. The accuracy percentage of the models employed in this investigation is significantly higher than that of earlier studies, showing that the models used in this study are more trustworthy than those used in earlier studies. The LR technique outperforms the other methods when crossvalidation measurements are used to forecast chronic kidney disease. By creating a web application that uses these methods and a larger dataset than the one used in this study, future research may build on this work. This will enhance outcomes and increase the precision and effectiveness of healthcare professionals' ability to identify renal problems. This will improve both the framework's durability and its visual appeal. It is hoped that this will motivate people to make positive changes in their life and seek early treatment for chronic renal illness.

7. REFERENCES

1. A. S. Levey, R. Atkins, J. Coresh et al., "Chronic kidney disease as a global public health problem: approaches and initiatives–a position statement from kidney disease improving global outcomes," *Kidney International*, vol. 72, no. 3, pp. 247–259, 2007.View at: Publisher Site | Google Scholar

2. V. Jha, G. Garcia-Garcia, K. Iseki et al., "Chronic kidney disease: global dimension and perspectives," *The Lancet*, vol. 382, no. 9888, pp.

260–272, 2013.View at: Publisher Site | Google Scholar

3. N. R. Hill, S. T. Fatoba, J. L. Oke et al., "Global prevalence of chronic kidney disease – a systematic review and meta-analysis," *PLoS One*, vol. 11, no. 7, article e0158765, 2016.View at: Publisher Site | Google Scholar

4. H. Nasri, "World kidney day 2014; chronic kidney disease and aging: a global health alert," *Iranian Journal of Public Health*, vol. 43, no. 1, pp. 126-127, 2014.View at: Google Scholar

5. G. Abraham, S. Varughese, T. Thandavan et al., "Chronic kidney disease hotspots in developing countries in South Asia," *Clinical Kidney Journal*, vol. 9, no. 1, pp. 135–141, 2016.View at: Publisher Site | Google Scholar

6. K. T. Mills, T. Xu, W. Zhang et al., "A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010," *Kidney International*, vol. 88, no. 5, pp. 950–957, 2015.View at: Publisher Site | Google Scholar

7. B. Ene-Iordache, N. Perico, B. Bikbov et al., "Chronic kidney disease and cardiovascular risk in six regions of the world (ISN-KDDC): a cross-sectional study," *The Lancet Global Health*, vol. 4, no. 5, pp. e307–e319, 2016.View at: Publisher Site | Google Scholar

8. S. Anand, M. A. Khanam, J. Saquib et al., "High prevalence of chronic kidney disease in a community survey of urban Bangladeshis: a cross-sectional study," *Glob Health*, vol. 10, no. 1, p. 9, 2014.View at: Publisher Site | Google Scholar

9. L. Ali, K. Fatema, Z. Abedin et al., "Screening for chronic kidney diseases among an adult population," *Saudi Journal of Kidney Diseases and Transplantation*, vol. 24, no. 3, p. 534, 2013.View at: Publisher Site | Google Scholar

10. M. J. Hasan, M. A. Kashem, M. H. Rahman et al., "Prevalence of chronic kidney disease (CKD) and identification of associated risk factors among rural population by mass screening," *Community Based Medical Journal*, vol. 1, pp. 20–26, 2013.View at: Google Scholar

11. M. J. Lysaght, "Maintenance dialysis population dynamics: current trends and long-term implications," *Journal American Soietyc Nephrology*, vol. 13, suppl 1, pp. S37–S40, 2002.View at: Publisher Site | Google Scholar

12. M. Bakhshayeshkaram, J. Roozbeh, S. T. Heydari et al., "A population-based study on the prevalence and risk factors of chronic kidney disease in adult population of shiraz, southern Iran," *Galen Medical Journal*, vol. 8, no. 935, p. 935, 2019.View at: Publisher Site | Google Scholar

13. K. U. Eckardt, J. Coresh, O. Devuyst et al., "Evolving importance of kidney disease: from subspecialty to global health burden," *The Lancet*, vol. 382, no. 9887, pp. 158–169, 2013.View at: Publisher Site | Google Scholar