International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 07 | July 2022 www.irjet.net p-ISSN: 2395-0072

Diagnosis Of Chronic Kidney Disease Using Machine Learning

Madhuri[1], Nandini S[2], Mrs.Mona[3]

[1,2] *Student Dept. of Information Science and Engineering, BNMIT Institute of Technology, Karnataka, India*
[3]*Assistant Professor, Dept. of Information Science and Engineering, BNM Institute of Technology, Karnataka, India*

Abstract - *About 10% of the adult population worldwide is affected by chronic kidney disease (CKD), one of the top 20 killers globally. CKD is a condition that impairs healthy kidney function. Effective preventative methods for the early detection of CKD are needed due to the rise in CKD cases. This study is new in that it has created a mechanism for diagnosing chronic renal disorders. With the aid of machine learning techniques, this study aids specialists in investigating CKD prevention strategies through early detection. This study was concerned with analysing a dataset made up of 400 patients and 24 features. To replace the missing numerical and nominal values, the mean and mode statistical analysis methods were used. Recursive Feature Elimination (RFE) was used to select the most important features. In this study, four classification algorithms were used: support vector machine (SVM), k-nearest neighbours (KNN), decision tree, and random forest. All of the classification algorithms performed well. The random forest algorithm outperformed all other applied algorithms, achieving 100 percent accuracy, precision, for all measures. CKD is a life-threatening disease with high morbidity and mortality rates. As a result, artificial intelligence techniques are critical in the early detection of CKD. These techniques aid experts and doctors in making early diagnoses in order to avoid kidney failure.*

Key Words: Recursive Feature Elimination, Nominal values, Chronic kidney disease, Classification, K-nearest neighbours, Support vector machice.

1. INTRODUCTION

In terms of the current state of society's health, chronic kidney disease (CKD) is regarded as a serious hazard. Regular laboratory testing can identify chronic kidney disease, and there are treatments available to stop the illness from progressing, lessen the problems of reduced Glomerular Filtration Rate (GFR), lowers the risk of cardiovascular disease, and enhance quality of life and survival. Lack of water intake, smoking, a poor diet, lack of sleep, and numerous other factors can lead to CKD. Globally, this illness impacted 753 million people in 2016, 417 million of them were female and 336 million male. The majority of the time, the illness is discovered at its advanced form, which can occasionally result in renal failure. The current diagnostic method relies on the analysis of urine with the aid of serum creatinine levels. This is accomplished using a variety of medical techniques, including ultrasonography and screening. Patients who have hypertension, a history of cardiovascular disease, a current illness, or who have had renal disease in a family member are all screened during the screening process. This method involves measuring the albumin-to-creatinine ratio (ACR) in a first-morning urine sample as well as estimating GFR from the serum creatinine level. This research focuses on machine learning approaches such as ACO and SVM for improving prediction accuracy by decreasing features and picking the best features.

An abnormal functioning of the kidneys or a lack of renal function that progresses over months or years is known as chronic kidney disease, also popular as chronic renal disease. People who are known to be at risk for kidney issues, such as those with high blood pressure, diabetes, or who have a blood family with chronic kidney disease (CKD), are typically screened for the condition. Therefore, early diagnosis and effective therapy are essential to battling the disease. In this work, the Ant Colony Optimization (ACO) method and the Support Vector Machine (SVM) classifier are suggested as machine learning strategies for CKD. Using the fewest number of features possible, the final product can determine if a person has CKD or not.

2. RELATED WORK

Chronic Kidney disease (CKD) disease that can be treated early on yet results in kidney failed at the very end. Chronic renal disease in 2016, 753 million people died as a result of it worldwide, where the Male fatalities total of 336 million, whereas female fatalities 417 million females died [1]. Due to its high mortality rate, chronic kidney disease (CKD) has attracted a lot of interest. According to the World Health Organization (WHO), chronic diseases have become a major hazard to emerging countries [2]. The reason kidney illness is referred to as a "chronic" condition is because it develops gradually over time and has an impact on how the urine system works. Other health issues that are brought on by the build up of waste products in the blood include diabetes, high and low blood pressure, bone and nerve damage, and cardiovascular disease. These issues are all accompanied by a variety of symptoms. Diabetes, hypertension, and cardiovascular disease (CVD) are risk factors for those with CKD [3]. Patients with CKD experience side effects, particularly in the late stages, which weaken the immunological and nervous systems. Patients may be in advanced stages in the developing nations, necessitating dialysis or kidney transplants. Glomerular filtration rate (GFR), a measure of kidney function, is used by medical professionals to identify renal illness. GFR is determined by factors including the patient's age, blood test results, gender,

and other circumstances [4]. The process of categorising legitimate, innovative, possibly helpful, and eventually intelligible patterns in data is known as data mining. In simple terms, it's the process of extracting data from a large database. There are many uses for data mining, including those in business, education, government, the health care sector, and science and engineering. Data mining is primarily utilised for disease prediction in the healthcare sector. There are a vast array of data mining techniques, including classification, clustering, association rules, summarizations, regression, and others, that can be used to forecast diseases. Using classification techniques like Naive Bayes and Support Vector Machine, the primary goal of this research is to predict kidney disorders. The major goal of this research was to identify the classification algorithm with the highest classification accuracy and execution time performance. According to the experimental findings, the SVM performs better than the Naive Bayes classifier algorithm [5]. A key aspect of healthcare informatics is the prediction of chronic diseases. Early disease detection is critical for successful treatment. In order to diagnose and predict chronic diseases, this study gives a survey on the application of feature selection and classification algorithms. In order to improve the accuracy of classification systems, proper feature selection is crucial[6].

According to the current state of study, the prevalence of chronic kidney disease (CKD) rises yearly. The ability of machine learning algorithms to classify data with high accuracy makes them more crucial in medical diagnosis and a source for future therapy in CKD prognosis. In the past, the accuracy of classification algorithms was determined by how well feature selection algorithms were used to reduce data size. On the Internet of Medical Things (IoMT) platform, the Heterogeneous Modified Artifical Neural Network (HMANN) has been suggested for the early detection, segmentation, and diagnosis of chronic renal failure [7]. Given that Chronic Kidney Disease (CKD) is one of the diseases that can be fatal, early detection and appropriate care of CKD are encouraged to increase survivability. Age, blood pressure, specific grativity, albumin, sugar, red blood cells, plus cells, pus cell clumping, bacteria, blood glucose random, and blood urea are some of the characteristics included in the UCI's CKD dataset that was chosen for this study. This work's primary goal is to compute and compare the effectiveness of various decision tree algorithms. Decision Stump, Hoeffding Tree, J48, CTC, J48graft, LMT, NB Tree, Random Forest, Random Tree, REP Tree, and Simple Cart are some of the decision tree methods employed in this study. As a result, the findings demonstrate that Random Forest provides the highest accuracy in CKD identification [8].

3.PROPOSED METHODOLOGY

The system is developed using a variety of design principles; the design specification outlines the features of the system, its competitors or constituent parts, and how they will appear to end users. Machines were used to carry out a number of experiments. SVM, KNN, decision trees, and random forests are the learning algorithms used to assess the CKD dataset. The general organisation of the CKD diagnosis in this paper is shown in Figure. The missing nominal values were computed using the mode technique during preprocessing, and the missing numerical values were computed using the mean method. The RFE algorithm was used to choose the features of relevance linked to the features of importance for CKD diagnosis. In order to diagnose diseases, these chosen traits were fed into classifiers. SVM, KNN, decision trees, and random forests were the four classifiers used in this work to diagnose CKD. For classifying a dataset into CKD or a normal kidney, all classifiers demonstrated good results.

To determine whether kidney illness exists or not, the system will first do a pre-process, then it will extract the key information and use a variety of classification algorithms. In addition to the class features, such as "ckd" and "notckd" for classification, the dataset has 24 features total, separated into 11 numerical features and 13 categorical features. Age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clusters, bacteria, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edoema, and anaemia are among the characteristics. ckd and notckd are the two values in the diagnostic class.

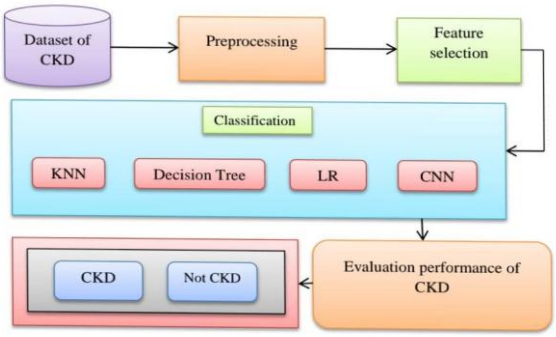

Fig. 1. System Architecture

The dataset needed to be cleaned up in a pre-processing stage because it had outliers and noise. Pre-processing tasks included checking for uneven data, normalising, estimating missing values, and removing noise like outliers. Selection of features. It is necessary to find the significant aspects that have a strong and positive connection with features of importance for disease diagnosis after computing the missing values. A robust diagnostic model cannot be built since the vector characteristics must be extracted to exclude features that are irrelevant and unhelpful for prediction.

	sg	al	sc	hemo	pcv	htn	classification
0	1.020	1.0	10.1	5.0	29	1	0
1	1.020	4.0	10.2	5.1	23	0	0
2	1.010	2.0	10.3	5.2	16	0	0
3	1.005	4.0	10.4	5.3	17	1	0
4	1.010	2.0	10.5	5.4	20	0	0

Fig. 2. Pre-processed Data

After computing the missing values and figuring out the vital capabilities having a robust and fine correlation with capabilities of significance for disease prognosis. Extracting the vector capabilities removes vain capabilities for prediction and features which are irrelevant and prevents the development of a robust diagnostic model. In this study, we used the RFE technique to extract the maximum vital capabilities of a prediction. The Recursive Feature Elimination (RFE) algorithm may be very famous because of its ease of use and configurations and its effectiveness in choosing capabilities in education datasets applicable to predicting goal variables and doing away with weak capabilities.

The most wide spread capabilities consistent with RFE; it's far mentioned that albumin function has maximum correction (17.99%), featured via way of means of 14.34%, then the packed cell volume feature via way of means of 12.91%, and the serum creatinine function via way of means of 12.09%. RFECV plots the range of capabilities withinside the dataset at the side of a cross-established rating and visualizes the chosen capabilities is presented in Figure 2.

A. KNN:

K-nearest neighbour (KNN), which employs the supervised learning methodology, is one of the simplest machine learning algorithms. Depending on how much a new instance resembles existing categories, it is categorised accordingly. The KNN approach is what is used in this. By using the KNN approach, all of your data can be stored, and new data may be categorised according to how similar it is to the old. This implies that the KNN approach can categorise new data into predefined categories quickly.

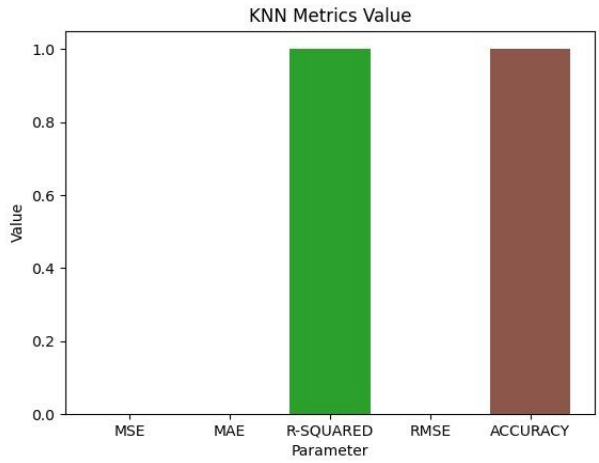

Fig.3. KNN Metrics Value

The KNN approach can be applied to regression even though it is frequently employed for classification problems. The KNN approach, which is nonparametric and also popularly called as a "lazy learner algorithm," does not instantly learn from the training set, but instead stores and organises the data for later analysis. The KNN places new data it receives into a category that is quite similar to the new data it saved during training.

B. LR:

The widely used statistical technique of logistic regression is used to model binary outcomes. In statistical research, logistic regression is carried out using various learning techniques. The LR algorithm was developed using a different neural network technique. Although this approach is easier to set up and utilise, it shares many similarities with neural networks.

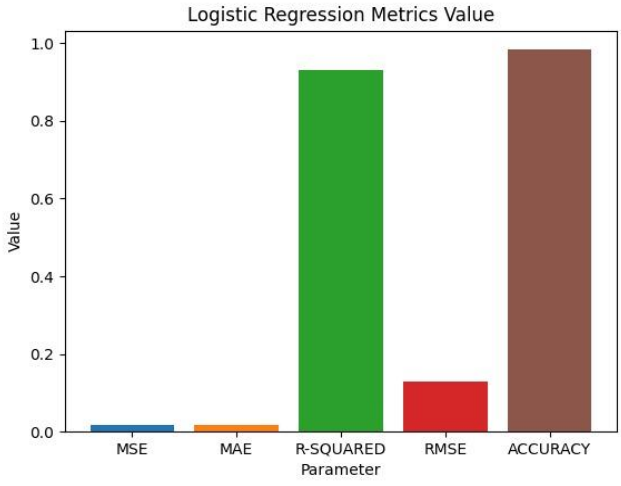

Fig. 4. Logistic Regression Metrics Value.

C. CNN:

We are constructing the CNN algorithm to forecast the kidney disease CNN model comprises following layers . A feed-forward neural network known as a convolutional network analyses data by processing it in a grid-like architecture. It is also referred to as a ConvNet. Data detection and classification are done using a convolutional neural network.

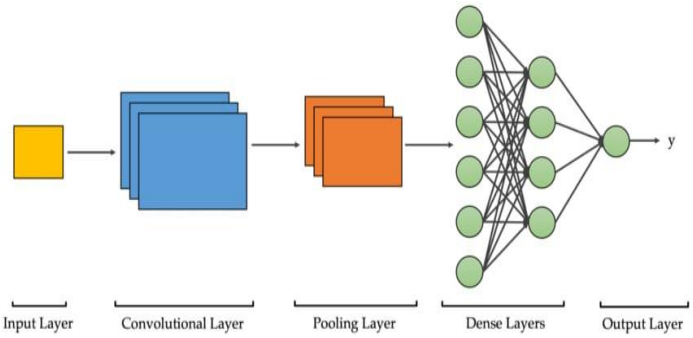

Fig. 5. Convolutional neural network

4. RESULTS AND DISCUSSION

Several prior related studies were used to evaluate the performance of the proposed systems. The accuracy ranges of the previous research are between 96.8 percent and 66.3 percent, it should be emphasised, but the suggested system has achieved accuracy of 100 percent using the Convolutional neural network method. When compared to current systems, it is seen that the proposed has the best outcomes. The dataset is randomly split into 25% for testing and validation and 75% for training. To choose the irrelevant subset characteristics, the Recursive Feature Elimination approach was presented. Then, classifiers were used to process the chosen features in order to diagnose CKD. It should be highlighted that the suggested system has produced encouraging outcomes. In order to prioritise the features and assign a percentage to each feature based on the correlation with the target feature, we utilised the RFE algorithm to determine the best associations between each feature and the target features.

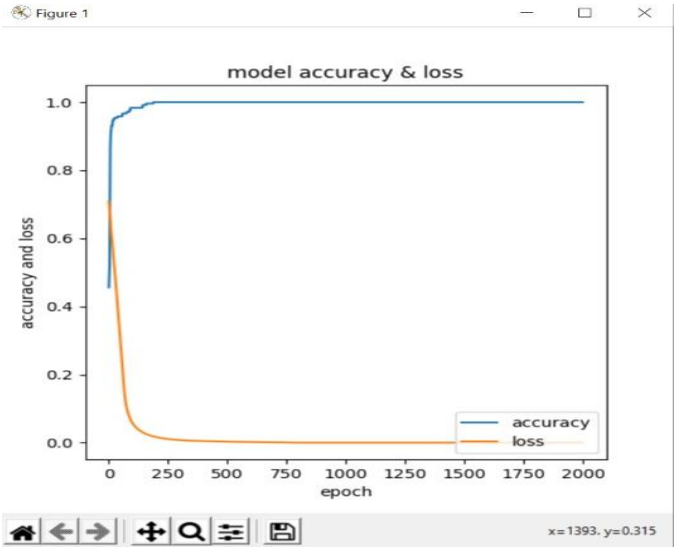

Fig.6. Model Accuracy and Loss

$$\text{Accuracy} = (TN + TP/TN + TP + FN + FP) * 100\%$$

where TN stands for True Negative,

TP stands for True Positive,

FN stands for False Negative, and

FP stands for False Positive.

5. CONCLUSIONS

This study shed light on the diagnosis of CKD patients to address their condition and obtain therapy in the early stages of the disease. The data was gathered from 24 characteristics were found in 400 patients. 25% of the dataset was used for testing and validation, and the remaining 75% was used for training. In order to replace missing numerical and nominal values and remove outliers from the dataset, mean and mode statistical measures were used, respectively. The most strongly representative CKD characteristics were chosen using the RFE method. The classification algorithms SVM, KNN, decision tree, and random forest were fed with specific features. All classifiers' parameters were adjusted for the best classification performance, and the results from all methods were positive.

REFERENCES

[1] Gunarathne W.H.S.D,Perera K.D.M, Kahandawaarachchi K.A.D.C.P, “Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)”,2017 IEEE 17th International Conference on Bioinformatics and Bioengineering.

[2] S.Ramya, Dr. N.Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative Research in Computer and Communication Engineering,Vol. 4, Issue 1, January 2016.

[3] S.Dilli Arasu and Dr. R.Thirumalaiselvi, "Review of Chronic Kidney Disease based on Data Mining Techniques",International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13498-13505

[4] L. Rubini, "Early stage of chronic kidney disease UCI machine learning repository,"2015. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[5] S. A. Shinde and P. R. Rajeswari, "Intelligent health risk prediction systems using machine learning : a review," IJET, vol. 7, no. 3, pp. 1019– 1023, 2018. M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available:http://www.ctan.org/texarchive/macros/latex/contrib/supported/IEEEtran/

[6] Himanshu Sharma,M A Rizvi,"Prediction of Heart Disease using Machine Learning Algorithms: A Survey",International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169,Volume: 5 Issue: 8

[7] Asif Salekin, John Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes," Proc. IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Oct. 2016, doi:10.1109/ICHI.2016.36. A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

[8] Pinar Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction," Proc. 41st IEEE International Conference on Computer Software and Applications (COMPSAC), IEEE, Jul. 2017, doi: 10.1109/COMPSAC.2017.84 *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.