PREDICTION OF RENAL DISEASE USING MACHINE LEARNING

B Padmaja [1], S Padmaja [2], S Sneha Sruthi [3], S Sai Varnika[4] and P Roja[5]
[1,2,3,4,5] Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India.

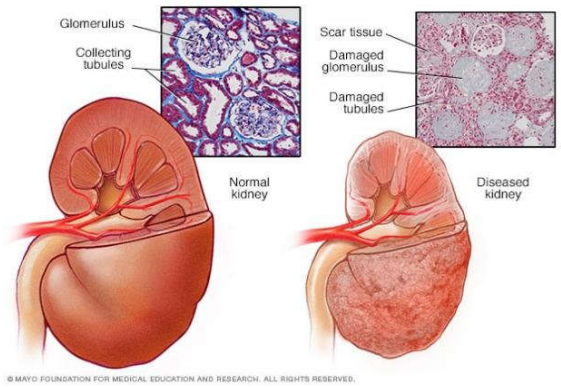
ABSTRACT

Many current medical expert systems find it challenging to diagnose diseases. In a medical diagnosis, there are various degrees of uncertainty. By identifying the contributing elements to the condition, a proper diagnosis can be made. Chronic kidney failure, another name for renal disease, refers to the loss of kidney function. The most popular area of research for diagnosing medical conditions uses machine learning to generate descriptive and predictive knowledge. Based on the patient's symptoms and other information provided as input, this system forecasts an illness. The linked factors are discovered using the Random Forest algorithm. Our analysis demonstrates that the Random Forest Algorithm outperforms all other machine learning classification algorithms with 99% accuracy. The chronic kidney disease dataset comprising the properties of blood pressure, sugar, random blood glucose, haemoglobin, red blood cells, white blood cells, and anaemia was subjected to analysis. The random forest algorithm is used to obtain an accurate assessment of chronic kidney disease (CKD) and predicts whether a patient has CKD or not. This algorithm can assist doctors in effectively managing patients and in making diagnoses more quickly. The study of the results demonstrates the value of employing the random forest classifier for clinical decision making. It can help in the early detection of CKD and its associated stages, which slows the course of renal damage.

KEY WORDS Chronic kidney Disease, Random Forest Algorithm

I. INTRODUCTION

Osmo regulation and excretion are the main duties of the kidney, a vital organ in the human body. Filtering by the kidney eliminates waste from the blood. Protein can leak into urine and waste materials can stay in blood if this filtration system is compromised. Eventually, the kidney's filtering capacity is lost. Chronic kidney disease (CKD), also known as Chronic Renal Disease, is the term used to describe this kidney failure. When a person is ill, it is possible for the kidney to become injured due to sudden changes in the cause or by taking specific medications; this condition is known as acute kidney injury. People often experience this disease in accordance with their age, but as of recent years, children and youth as young as 5 years old are also experiencing CKD disease. Around 10% of the world's population has chronic kidney disease (CKD), which causes millions of fatalities each year. As the main cause of death globally in 1990, chronic renal disease came in at number 27. It comes in at number 18 in the year 2010. Muscle cramps, nausea and vomiting, appetite loss, swelling in your feet and ankles, too much or too little urine, difficulties breathing, trouble sleeping, fever, and vomiting are some signs that your kidneys are starting to fail. Diabetes, smoking, insufficient sleep, hypertension, an unsuitable diet, and others are risk factors for CKD. The most serious of these is diabetes. The patient must undergo kidney transplantation or dialysis at the end of the process.



The early stages of CKD can be divided into five phases:

1. At the initial stage, a person may change while having their kidneys function normally and may even encounter mild complications.
2. During the second phase, a person may experience one to three minor setbacks with their kidney function.
3. Phase three also misrepresents, with a person experiencing mild to serious renal work trouble.
4. In the fourth phase, renal function will suffer a severe setback.
5. A person will get complete renal failure in Phase 5.

As a result, it is expressly important in the early management, supervision, and diagnosis of the illness. Due to its active and covert nature in the early stages, as well as patient heterogeneity, it is crucial to accurately and reliably detect CKD disease. Consequently, accurate diagnosis and timely therapy may be able to halt or delay the progression of this chronic illness.

The main factors specified as a reason for the disease is

- Diabetes
- High blood pressure
- Heart (cardiovascular) disease
- Smoking
- Obesity
- Being Black, Native American or Asian American
- Family history of kidney disease
- Abnormal kidney structure
- Older age
- Frequent use of medications that can damage the kidneys.

Early kidney disease detection protects the patient from life-threatening consequences. The factors that cause renal illnesses must be properly examined in order to forecast them. All of these variables can be converted into information to forecast renal illness and offer a treatment plan to enhance the patient's health. Continuity, multiple attributions, incompleteness, and 1 2 2 2 12 time-related qualities are all present in medical knowledge. For the healthcare sector, the issue of effectively employing massive amounts of data is growing.

Now a days machine learning plays a vital role in health care domain. ML already makes disease diagnosis simpler and aids with more accurate treatment plan creation for doctors. Its algorithms analyse enormous volumes of patient data to draw valid findings more quickly and precisely than the human brain. The healthcare sector has a wide range of applications those are diagnosis and disease prediction, predictive medicine, medical care personalization, administrative workflow, medical image analysis, drug discovery, medical documentation flow and so on.

.

II. LITERATURE REVIEW

Chilakamarthi Prem Kashyap; Gollapudi Sai Dayakar Reddy; M Balamurugan proposed a model using Four machine learning algorithms Support vector machine classifier (SVM), K-Nearest Neighbour (KNN), Random Forest, and decision tree were used to propose a model. Data pre-processing techniques are used on the dataset to improve the accuracy of these algorithms' CKD prediction after they have been trained.

*Sweety Kumari; Sunil Kumar Singh*in this model in orderto predict the cases of CKD and non-CKD, they used machine, deep, and ensemble learning-based models. They also used a stacking and voting ensemble comprising SVM, RF, Adaboost, LDA, and MLP models to accurately and swiftly predict the CKD and non-CKD cases.

N. Mohana Suganthi; Jemin V.M; P. Rama; E. Chandralekha proposed a model using Logistic Regression (LR) and its ensembles, KNN, Decision Trees (DT), Random Forest (RF), Naive Bayes (NB), and others. KNN, Decision Tree, Naive Bayes, Random Forest, and Logistic Regression were used to compare the outcomes. According to experimental findings, the AdaBoost-Random Forest ensemble is 96% accurate at spotting renal disease early on.

Arif-Ul-Islam; Shamim H Ripon proposed a model which evaluates the effectiveness of boosting algorithms for CKD detection and derives rules showing relationships between the CKD characteristics they employed. Here, the performance of categorization is compared using AdaBoost and LogitBoost. Another data mining algorithm that uses the Ant Colony Optimization method is called Ant-Miner. In this model, rules are derived using Ant-Miner and Decision Tree.

*Helmie Arif Wibawa; Indra Malik; Nurdin Bahtiar*proposed a model whichIt evaluates the effectiveness of boosting algorithms for CKD detection and derives rules showing relationships between the CKD characteristics they employed. Here, the performance of categorization is compared using AdaBoost and LogitBoost. Another data mining algorithm that uses the Ant Colony Optimization method is called Ant-Miner. In this model, rules are derived using Ant-Miner and Decision Tree.

*Uma N Dulhare; Mohammad Ayesha*proposed a model which is not only extracting action rules based on stages but also predicting CKD by using naïve bayes with OneR attribute selector which helps to prevent the advancing of chronic renal disease to further stages.

*A. Vijayalakshmi; V. Sumalatha*When it is possible to anticipate the patient's CKD and non-CKD status utilising different categorization techniques. The survey has covered numerous ML algorithms that are used to identify renal illness, and the key concerns are briefly highlighted. Therefore, the analysis of recent research on ML applications in kidney illness is highly valued by physicians and will significantly improve clinical practise in the future.

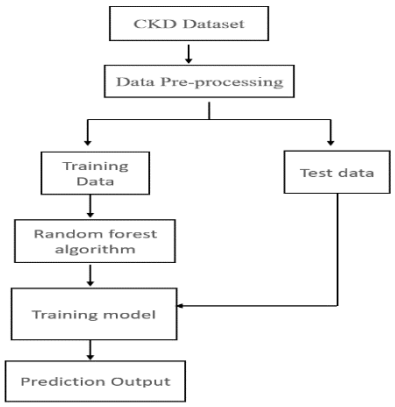
*Nosin Ibna Mahbub; Md. Imran Hasan; Md. Martuza Ahamad; Sakifa Aktar; Mohammad Ali Moni*proposed a modelto categorise the most significant factors that lead to CKD and to anticipate the diagnosis of CKD based on symptoms or features observed in a specific instance, regardless of the stage being acute or chronic. In order to predict the development of CKD, they used seven machine learning techniques, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree, Random Forest, Naive Bayes, and XGBboost classifiers, as well as two statistical tests, Student's T-test and chi-squared test, to determine the most important features. With an accuracy of 94%, the results demonstrated that the tree-based classifier performed well in the renal diagnosis method.

*Zhen-Yi Tang; Yen-Chung Lin; Che-Chou Shen*proposed dual-path convolutional neural network (DPCNN) In order to concurrently extract and integrate high-level and low-level information from the US image for

IFTA prediction The proposed DPCNN for binary IFTA classification of non-diabetic nephropathy obtains accuracy of 0.856 (0.818-0.876), recall of 0.761 (0.715-0.817), specificity of 0.927 (0.862-0.952), precision of 0.887 (0.804-0.920), F1 score of 0.819 (0.776-0.846), and area under the receiver operating characteristic curve (AUC) of 0.922 with five-fold cross-validation (0.893-0.944).

III. PROPOSED SYSTEM

The data collecting, pre-processing, feature engineering with machine learning, classification, and assessment processes make up the methodology of our work. The picture below shows a graphic representation of these computing processes.



A.DATA SET

The UCI Repository was used to obtain the Training data set for kidney disease. The information was gathered from various hospitals, and although it has 24 qualities in total, only 19 of them are crucial for establishing that association, according to pre-processing. There are 400 samples in all. This will enable accurate early identification of chronic kidney disease using the forecasts.

ID	Attribute Symbols	Attribute Description	Attribute Type
1	Age	Numerical	Years
2	Blood pressure	Numerical	Mm/Hg
3	Specific gravity	Nominal	0.005,1.010,1.015,1.020,1.025
4	Albumin	Nominal	0.1.2.3.4.5
5	Sugar	Nominal	0.1.2.3.4.5
6	Red blood cells	Nominal	Normal, abnormal
7	Pus cells	Nominal	Normal, abnormal
8	Pus cell clumps	Nominal	Present, not present
9	Bacteria	Nominal	Present, not present
10	Blood glucose	Random Numerical	Mgs/dl
11	Blood urea	Numerical	Mgs/dl
12	Serum creatinine	Numerical	Mgs/dl
13	Sodium	Numerical	mEq/L
14	Potassium	Numerical	mEq/L
15	Hemoglobin	Numerical	Gms
16	Packed cell volume	Numerical	Gms
17	White blood cell count	Numerical	Cells/cum
18	Red blood cell count	Numerical	Millions/cm
19	Hypertension	Nominal	Yes, no
20	Diabetes mellitus	Nominal	Yes, no
21	Coronary artery disease	Nominal	Yes, no
22	Appetite	Nominal	Good, poor
23	Pedal edema	Nominal	Yes, no
24	Anemia	Nominal	Yes, no
25	Class	Nominal	CKD, NOCKD

B. DATA PRE-PROCESSING

Pre-processing comes after the data has been analysed and visualised. Data cleansing and preparation for usage in machine learning algorithms are accomplished through the crucial step of data pre-processing. Pre-processing primarily focuses on addressing any missing values and outliers as well as inaccurate or outlier-containing data. There are two approaches to deal with missing data. The first approach is to just eliminate

the entire row that contains the inaccurate or missing value. Although this method is simple to apply, it is best to limit its application to huge datasets. This strategy can result in an excessive reduction in the size of tiny datasets, especially if there are many missing values.This could significantly reduce the accuracy of the outcome. We won't be employing this strategy because the dataset we have is relatively tiny. Alternatively, depending on the type of attribute, we would substitute the average or mode of the column for the missing values. If the property has a nominal value, we will use the average; otherwise, we will use the mode. We had to translate and encode the string-formatted values in the dataset we utilised into integer values before feeding them as input to the neural network. We first transformed the data into categorical pandas data and produced a separate data frame.

From the viewpoint of the characteristics, the missing data can be random, monotone, or univariate. It is univariate if all attributes have 0 missing values. The pattern is monotone if at least three attributes have missing values. While it is arbitrary if the missing values pertain to random properties.

The justification for missing value imputation is that there is no dataset reduction because the data imputation retains the complete sample size. Many imputation approaches with different features are available to impute the missing data.

C. RANDOM FOREST ALGORITHM

We carried out our research utilising machine learning-based architectures to train and evaluate the dataset. a random forest algorithm-based machine learning technique. The steps to implement the random forest algorithm I are as follows.

i.Decision trees creation:

Making decision trees is the initial stage in the random forest method. This is accomplished by dividing the data into subsets according to specific attributes. The decision tree that results from each subset is then utilised to make predictions.

The decision tree is built by analysing each attribute and choosing the one that divides the data most effectively. This is accomplished by figuring out which feature has the largest information gain for each feature. Up till the decision tree is finished, this process is repeated.

Gini Index= $1- \sum_j P_j^2$

Entropy = $- \sum p(x) \log p(x)$

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy (each feature)]

ii. Evaluating Predictions:

The decision trees' predictions are evaluated as the next stage in the random forest process. This is accomplished by calculating the prediction accuracy and deciding whether the forecasts are reliable enough to be used for CKD prediction.

By contrasting the forecasts with the actual values of the data, one can assess the predictions' accuracy. The decision trees are used to produce predictions if the predictions are accurate enough. If not, the decision trees are improved to increase the predictability.

iii. Refining Decision Trees:

The random forest algorithm's next step is to improve the decision trees. This is accomplished by determining which traits are crucial and which features are not crucial based on the outcomes of the evaluation of the forecasts.

Thereafter, the decision trees are changed to utilise only the crucial features.As a result, the model performs better and the forecasts are more accurate. Up till the predictions' accuracy is adequate, the process is repeated.

iv. Testing the model:

The random forest algorithm's subsequent phase involves comparing the model to the data. This is accomplished by making predictions on fresh data using the improved decision trees. To assess the forecasts' accuracy, the predictions are then contrasted with the data's actual values.

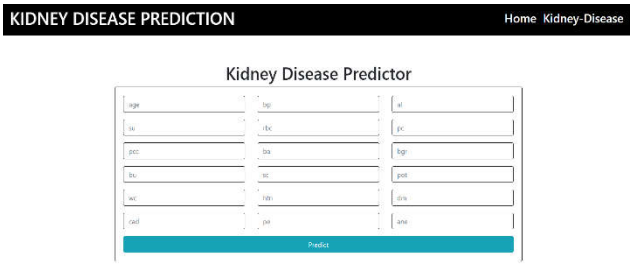
The model is deemed successful if the forecasts' accuracy is adequate. If not, the decision trees are improved and the model is put to another test. Until the predictions' accuracy is adequate, this process is repeated.

Algorithm:

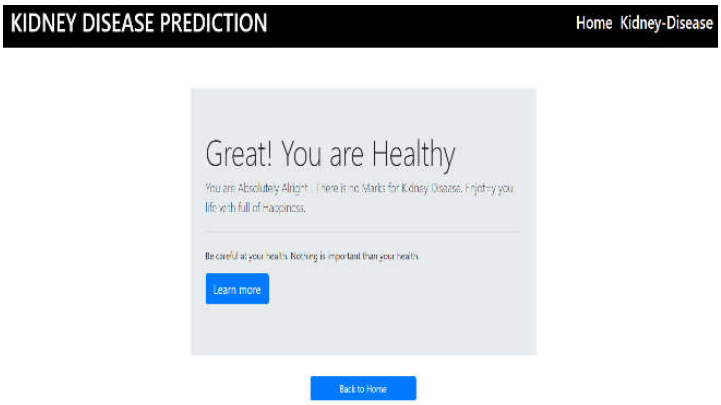
1. Randomly select “k” features from total “m” featureWhere k << m
2. Among the “k” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until “l” number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

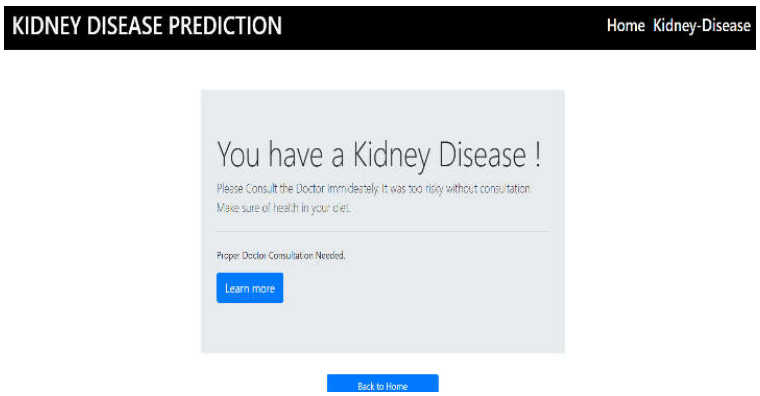
IV.RESULTS AND DISCUSSIONS

The outcomes for the suggested model during training and testing are shown in the pictures below..



Here the user need enter the values like age, bp, albumin(al), sugar(su),red blood cells count(rbc), pus cells(pc), pus cells clumps(pcc), bacteria(ba), blood glucose(bgr), blood urea(bu), serum creatinine(sc), potassium(pot), white blood cells(wc), hypertension(htn), diabetes mellitus(dm), coronary artery disease(cad), pedal edema(pe), anemia(ane)) before clicking on the "predict" button. The prediction will then determine whether the user has CKD or not.





V. CONCLUSION

In this study, a clever machine learning model that uses the random forest technique for classification and prediction of a dataset of chronic kidney disease is proposed. We pre-process the dataset using the Kaggle machine learning database to handle the missing values.

The suggested model's efficacy is evaluated, and its performance is compared to that of the existing models. The proposed model performs better and has a 98.52% accuracy compared to the existing models. The proposed model thus provides a reliable predictor and classifier for CKD and also determines whether an individual has the disorder.

VI. REFERENCES

1. Chilakamarthi Prem Kashyap; Gollapudi Sai Dayakar ReddyPrediction of Chronic disease in kidneys using machine learning classifiers Doi: 10.1109/ICCST55948.2022.10040329
2. Sweety Kumari; Sunil Kumar SinghAn ensemble learning-based model for effective chronic kidney disease prediction Doi:10.1109/ICCCIS56430.2022.10037698
3. Pinar YildirimChronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease PredictionDoi: 10.1109/COMPSAC.2017.84
4. N. Mohana Suganthi; Jemin V.M; P. Rama; E. Chandralekha Chronic kidney disease detection using AdaBoosting Ensemble method and K-Fold cross validation Doi: 10.1109/ICACRS55517.2022.10029047
5. Arif-Ul-Islam; Shamim H RiponRule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree Doi: 10.1109/ECACE.2019.8679388
6. Helmie Arif Wibawa; Indra Malik; Nurdin BahtiarEvaluation of Kernel-Based Extreme Learning Machine Performance for Prediction of Chronic Kidney Disease Doi: 10.1109/ICICOS.2018.8621762
7. Uma N Dulhare; Mohammad Ayesha Extraction of action rules for chronic kidney disease using Naïve bayes classifier Doi: 10.1109/ICCIC.2016.7919649
8. A. Vijayalakshmi; V. SumalathaSurvey on Diagnosis of Chronic Kidney Disease UsingMachine Learning Algorithms Doi: 10.1109/ICISS49785.2020.9315880
9. Nosin Ibna Mahbub; Md. Imran Hasan; Md. Martuza AhamadMachine Learning Approaches to Identify Significant Features for the Diagnosis and Prognosis of Chronic Kidney Disease Doi:10.1109/ICISET54810.2022.9775827

10. Zhen-Yi Tang; Yen-Chung Lin; Che-Chou Shen Dual – path Convolutional Neural Network for chronic kidney Disease Classification in Ultrasound Echography Doi: 10.1109/IUS54386.2022.9957954
11. C. S. S. Anupama, M. Sivaram, E. L. Lydia, D. Gupta, and K. Shankar, ''Synergic deep learning model–based automated detection and classification of brain intracranial hemorrhage images in wearable networks,'' Pers. Ubiquitous Compute., Nov. 2020, Doi: 10.1007/s00779-020-01492-2.
12. A. Khamparia, G. Saini, B. Pandey, S. Tiwari, D. Gupta, and A. Khanna, ''KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network,'' Multimedia Tools Appl., vol. 79, nos. 47–48, pp. 35425–35440, Dec. 2020, Doi: 10.1007/s11042-019-07839-z.
13. A. Khamparia, B. Pandey, S. Tiwari, D. Gupta, A. Khanna, and J. J. P. C. Rodrigues, ''An integrated hybrid CNN-RNN model for visual description and generation of captions,'' Circuits Syst., Signal Process., vol. 39, pp. 776–788, Nov. 2020, Doi: 10.1007/s00034-019-01306-8.
14. M. Sharma, B. Jain, C. Kargeti, V. Gupta, and D. Gupta, ''Detection and diagnosis of skin diseases using residual neural networks (RESNET),'' Int. J. Image Graph., Dec. 2020, Art. no. 2140002, Doi: 10.1142/S0219467821400027.
15. S.-H. Wang, S. Xie, X. Chen, D. S. Guttery, C. Tang, J. Sun, and Y.-D. Zhang, ''Alcoholism identification based on an AlexNet transfer learning model,'' Frontiers Psychiatry, vol. 10, Apr. 2019.
16. J. Zhao, S. Gu, and A. McDermaid, ''Predicting outcomes of chronic kidney disease from EMR data based on random forest regression,'' Math. Biosci., vol. 310, pp. 24–30, Apr. 2019.
17. A. Abdelaziz, A. S. Salama, A. M. Riad, and A. N. Mahmoud, ''A machine learning model for predicting of chronic kidney disease-based Internet of Things and cloud computing in smart cities,'' in Security in Smart Cities: Models, Applications, and Challenges (Lecture Notes in Intelligent Transportation and Infrastructure), 2019, pp. 93–114, Doi: 10.1007/ 978-3-030-01560-2_5.
18. S. M. K. Chaitanya and P. R. Kumar, ''Detection of chronic kidney disease by using artificial neural networks and gravitational search algorithm,'' in Innovations in Electronics and Communication Engineering (Lecture Notes in Networks and Systems), vol. 33, 2019.
19. Ankit, B. Besra, and B. Majhi, ''An analysis on chronic kidney disease prediction system: Cleaning, pre-processing, and effective classification of data,'' in Recent Findings in Intelligent Computing Techniques (Advances in Intelligent Systems and Computing), vol. 707, 2019.
20. S. Tekale, P. Shingavi, and S. Wandhekar, ''Prediction of chronic kidney disease using machine learning algorithm,'' Int. J. Adv. Res. Compute. Commune. Eng., vol. 7, no. 10, pp. 92–96, Oct. 2018