2024 3rd International Conference for Advancement in Technology (ICONAT) | 979-8-3503-5417-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICONAT61936.2024.10774782

Enhancing Chronic Kidney Disease Prediction through Comparative Analysis of Decision Tree and Random Forest Algorithms

Arpanpreet Kaur[1]
Chitkara University Institute of Engineering and Technology,
Chitkara University,
Punjab, India
arpanpreet.kaur@chitkara.edu.in

Kanwarpartap Singh Gill[2]
Chitkara University Institute of Engineering and Technology,
Chitkara University,
Punjab, India
kanwarpartap.gill@chitkara.edu.in

Sonal Malhotra[3]
Computer Science & Engineering,
Graphic Era Hill University,
Dehradun, Uttarakhand, India, 248002
sonalm.cse@gmail.com

Swati Devliyal [4]
Computer Science & Engineering,
Graphic Era Deemed to be University,
Dehradun, Uttarakhand, India, 248002
swatidevliyal@gmail.com

Abstract— **A major public health concern, chronic kidney disease (CKD) is marked by a high incidence and serious consequences of a progressive loss of kidney function over time. The basis of effective treatment and care is a precise early diagnosis. This study is justified by the growing prevalence of CKD and the need for accurate predictive models for early detection. Application of Decision Tree (DT) and Random Forest (RF) algorithms to CKD prediction is investigated in this work. DT, a simple yet powerful classification technique, creates a model that predicts the value of a target variable based on several input variables. RF, an ensemble learning method, constructs multiple decision trees and merges their outcomes to improve prediction accuracy and control overfitting. Furthermore, the dataset includes patient medical records with demographic, clinical, and laboratory data such as RBC's, WBC's, etc. in the study. Moreover, the methodology consists of preprocessing of the data, feature selection, and Decision tree and Random Forest algorithm implementation. Furthermore, results show how well both models function; RF shows better accuracy of 93% whereas DT shows 79% respectively. The work shows, in summary, that machine learning methods—RF in particular—can greatly raise the accuracy of CKD prediction. Such models might be included into healthcare systems to help with early detection and improved control of chronic kidney disease, which would eventually lead to better patient outcomes. These models will be improved and their application in various patient populations will be investigated in future studies.**

Keywords— ***Machine learning, Ensemble learning, random forest, decision tree, health, diagnosis, medical, patient***

I. INTRODUCTION

Chronic kidney disease (CKD) is a serious public health problem characterized by a high incidence and disastrous consequences caused by the progressive decrease of renal function with time. This disease often leads to end-stage renal disease, which is very bad for patients' quality of life and healthcare systems and needs dialysis or transplantation. Early diagnosis and treatment reduce CKD's unfavourable effects. The disease's gradual onset and growth make diagnosis difficult. This work is driven by the rising prevalence of chronic kidney disease (CKD) and the need for accurate prediction models for early diagnosis and treatment.. In this paper, application of Decision Tree (DT) and Random Forest (RF) algorithms to CKD prediction is investigated. Simple but effective categorization approaches allow decision trees anticipate target variable values from multiple input variables.

Its interpretability and simplicity are prized. Random Forest prevents overfitting and improves prediction accuracy by combining decision trees. This dual method compares ensemble and individual decision tree CKD prediction methods considering their benefits. The dataset employed in this paper consists of red blood cells (RBCs), white blood cells (WBCs), and other relevant indicators together with patient medical records including demographic, clinical, and laboratory information. The different parameters give out the target from 400 rows by performing the various functions in terms of ckd and not ckd. The process has many steps: Outliers and missing data are removed during feature selection after data pretreatment to uncover CKD predictors. Then, Random Forest and Decision Tree are evaluated for accuracy and other criteria. Random Forest predicts 93% accurately, while Decision Tree predicts 79%. Ensemble approaches predict complex medical data well, as seen by this intriguing distinction. Averaging many decision trees to reduce volatility improves Random Forest forecasts. This research improves chronic kidney disease forecasts with Random Forest. Implementing such models in healthcare systems could improve early identification and treatment of chronic kidney disease and patient outcomes. These prediction algorithms help doctors identify at-risk patients early for early treatments, tailored treatment regimens, and better resource use. Future study will focus on improving these models, applying them to more patient populations, and incorporating them into therapy. Moreover, it will improve the models' interpretability so doctors may use their predictions in clinical decision-making. The goal is to provide a complete, accurate, and easy-to-use predictive tool for CKD patients' prognosis and proactive healthcare management.

II. LITERATURE

A large corpus of literature indicates the current substantial interest in the use of machine learning (ML) for the diagnosis and prediction of chronic kidney disease (CKD). An extensive review of several machine learning models is given by Rahman, Al-Amin, and Hossain, who also show how well they diagnose and forecast chronic kidney disease. Their work emphasizes the crucial role machine learning plays in improving early CKD detection and diagnostic accuracy, two

979-8-3503-5417-1/24/$31.00 ©2024 IEEE

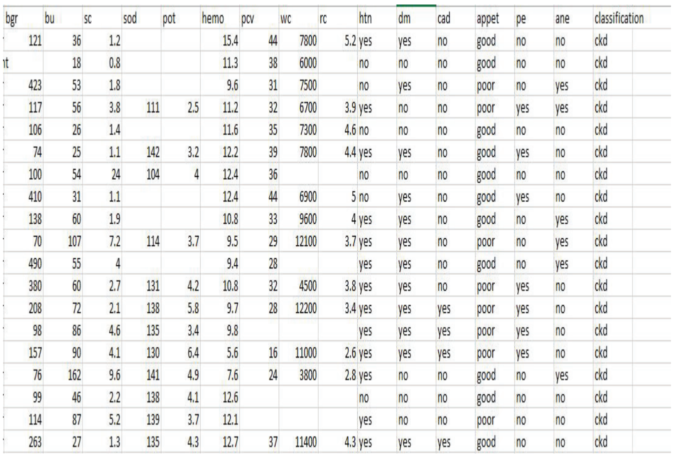
requirements for prompt intervention and treatment [1]. Similar in approach, Arora, Sehgal, and Agarwal evaluated many machine learning techniques for CKD prediction. Their findings suggest choosing models based on clinical needs and dataset features. presenting at the 2024 International Conference on Data Science & Engineering, Cloud Computing [2]. Khalid and colleagues investigated a hybrid machine learning model to combine the benefits of multiple approaches for better CKD prediction.. Published in "Computational Intelligence and Neuroscience," their work shows that hybrid models can overcome the shortcomings of each technique to outperform single approaches in prediction accuracy[3]. In a different study, Liu and associates used the Random Forest approach to assess risk factors associated to chronic kidney disease. Their observational study, which was written up in the "Asian/Pacific Island Nursing Journal," shows how machine learning (ML) may be used to identify and measure risk factors, providing a practical approach for CKD risk assessment[4]. Islam, Majumder, and Hussein reported their chronic renal disease prediction machine learning findings in the "Journal of Pathology Informatics." The findings show that numerous machine learning techniques are needed to diagnose chronic kidney disease (CKD) early to improve patient outcomes and reduce disease progression [5].In "Sustainability," Iftikhar and colleagues evaluated many machine learning methods. A model performance comparison shows the need for reliable and long-lasting machine learning solutions in healthcare, especially for chronic diseases like renal disease [6]. The 2023 International Conference on Electrical Engineering, Computer Science, and Informatics featured ensemble machine learning models by Haque and colleagues. Ensemble models predict renal failure better than single models[7]. Reddy and colleagues tested explainable chronic kidney disease machine learning models on a small pathology dataset. Their medRxiv work underlines the need for explainability in machine learning models to promote physician acceptance and trust by offering explicit decision-making procedures[9]. Shanmugarajeshwari and Ilayaraja used intelligent decision support systems to diagnose chronic renal disease stages.. Published in the "International Journal of Intelligent Information Technologies," their study demonstrates how machine learning algorithms may help physicians make well-informed choices on the treatment of chronic kidney disease[11]. Finally, in the "Journal of Computer Science and Technology Studies," Rahat et al. compared early-stage CKD detection machine learning techniques. Early chronic kidney disease diagnosis is crucial, and machine learning can enhance diagnostics [12]. The literature generally shows great agreement on how machine learning (ML) might completely transform CKD diagnosis and prediction. Every study adds to our knowledge of the advantages and disadvantages of certain machine learning techniques and provides insightful information about their useful applications in the medical field.The following points cover the proposed methodology of the research work:

- The researchers focus on chronic kidney disease prediction and use models that accurately diagnose the disease.
- The work is done on annotated data and shows the potential use of such an approach in practice.
- All this work tries to enhance the accuracy and efficiency of chronic kidney disease and shows rather good results.
- Therefore, the implementation of machine learning will help identify this issue faster and more accurately, which is especially relevant for disease prediction.

III. INPUT DATASET

The work uses the dataset which includes 25 features like red and white blood cell counts, age, blood pressure, and more. Predict chronic kidney disease. It uses decision tree and random forest algorithms. Decision trees use simple but effective classification to predict target variable values from several input variables.. Its interpretability and simplicity is really appreciated. Conversely, an ensemble learning technique called Random Forest constructs several decision trees and combines their results to improve prediction accuracy and reduce overfitting. The table 1 below provides the input dataset in CSV format, which is extracted from Kaggle open-source platform.

TABLE I. INPUT DATASET



bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
121	36	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
nt	18	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
423	53	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
106	26	1.4			11.6	35	7300	4.6	no	no	no	good	no	no	ckd
74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
100	54	24	104	4	12.4	36			no	no	no	good	no	no	ckd
410	31	1.1			12.4	44	6900	5	no	yes	no	good	yes	no	ckd
138	60	1.9			10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
490	55	4			9.4	28			yes	yes	no	good	no	yes	ckd
380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
98	86	4.6	135	3.4	9.8				yes	yes	yes	poor	yes	no	ckd
157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd
76	162	9.6	141	4.9	7.6	24	3800	2.8	yes	no	no	good	no	yes	ckd
99	46	2.2	138	4.1	12.6				no	no	no	good	no	no	ckd
114	87	5.2	139	3.7	12.1				yes	no	no	poor	no	no	ckd
263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes	yes	good	no	no	ckd

The different parameters give out the target from 400 rows by performing the various functions in terms of ckd and not ckd. The dataset format ensures an exacting evaluation of the proposed methodology. This disease often leads to end-stage renal disease, which is very bad for patients' quality of life and healthcare systems and needs dialysis or transplantation. CKD's deleterious effects are mitigated with early diagnosis and treatment. Slow disease onset and progression complicate diagnosis. This project is driven by the rising prevalence of chronic kidney disease (CKD) and the need for accurate prediction models for early detection and treatment Two accurate models: RF 93%, DT 79%. RF-ML improves CKD prediction. These models could detect and treat chronic kidney disease early, improving patient outcomes.

IV. PROPOSED METHADOLOGY

The paper shows that using machine learning models, Decision tree and Random Forest, may reliably and accurately identify chronic kidney disease. The method was developed with a methodical guidance provided by the figure 1 below, which shows dataset data pretreatment and methodology evaluation in addition to data collecting. Data collection is obtained from the chronic kidney disease dataset on the Kaggle open source platform, which contained dataset with different parameters such as red blood cell count, white blood cell count, age, blood pressure, age, and many more. Data cleaning was done which included the transfer of columns name and analysing categorical columns. Further NaN values replacement and exploratory data analysis.

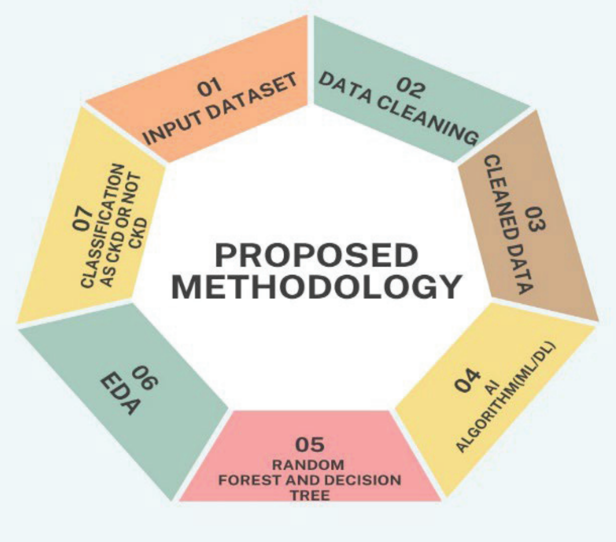


Fig. 1. Proposed Methadology for Chronic Kidney Disease Detection using decision tree and random forest

Then two machine learning methods, Decision Tree (DT) and Random Forest (RF), are put into practice. As a baseline, the well-known and readily understood DT model is used; the ensemble method RF constructs many decision trees to improve prediction accuracy and reduce overfitting. Moreover, accuracy, precision, recall, and F1-score are used as evaluation criteria for the variance inflammatory factor check, train and split the data, and the models. Supporting the results, which revealed an RF reaching an astounding accuracy of 93% compared to DT's 79%, interpretation and debate were conducted, pointing up model advantages and disadvantages and making recommendations for future study.

V. RESULTS

A. Results for Decision Tree

1) Confusion Matrix

The confusion matrix shown in Figure 2 depicts the performance of the decision tree for chronic kidney disease prediction. The matrix illustrates that nine were wrongfully identified as non-CKD, whereas thirty were correctly predicted as CKD. The matrix predicts that the model achieved 79% accuracy. Although the Decision Tree model could improve, notably in reducing false negatives, which are critical in medical diagnosis, this study highlights its good CKD prediction ability. Future study may refine the model or investigate ensemble techniques like Random Forest to improve forecast accuracy and reliability.

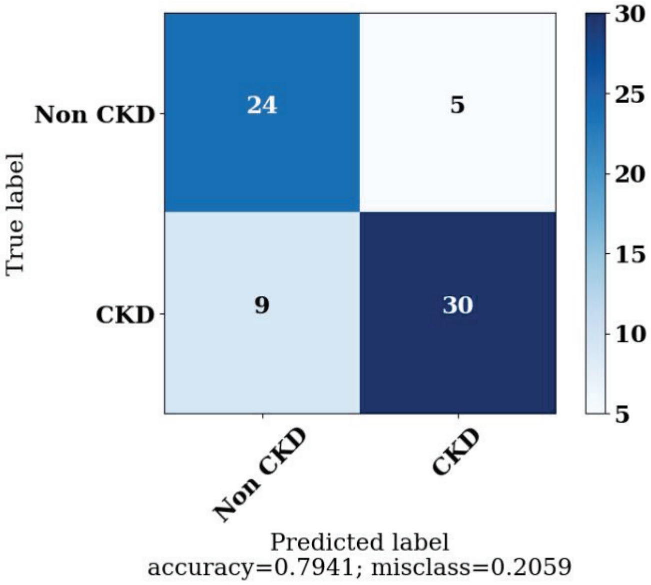


Fig. 2. Confusion Matrix Analysis

2) Configuration Report

The classification report shown in Table 2 below is the performance matrix of a machine learning classifier, perhaps for the diagnosis of chronic kidney disease. The report provides support, a f1-score, recall, and precision for each class. Class 0 (non-CKD) patients produce 0.77 as the f1-score, 0.83 as the recall, and 0.73 as the precision. Precision, recall, f1-score, and support all score 0.86 for Class 1 (CKD). Overall accuracy of the classifier is 79%. For precision, the measured unweighted average is 0.80; for recall, it is 0.79. For f1-score, 0.79 is the macro average, and for precision, 0.80. The decision tree can do better at lowering the number of CKD cases overlooked (recall), even though it is very strong at precisely detecting CKD cases (high precision).

TABLE II. CONFIGURATION REPORT ANALYSIS

	precision	recall	F1-score	Support
0	0.73	0.83	0.77	29
1	0.86	0.77	0.81	39
Accuracy			0.79	68
Macro avg	0.79	0.80	0.79	68
Weighted avg	0.80	0.79	0.80	68

B. Results for Random Forest

1) Confusion Matrix

When compared to the earlier decision tree classifier, the random forest classifier's confusion matrix shown in figure 3 demonstrates a notable performance gain. Random Forest has correctly predicted all non CKD cases and most of the CKD cases correctly hence, predicting strong performance. The model's better sensitivity is shown by the fact that it missed just 5 CKD cases, a considerable decrease over the decision tree classifier. Furthermore, the classifier has shown overall 93% of accuracy and still shows a remarkable ability to correctly classify both CKD and non CKD cases.

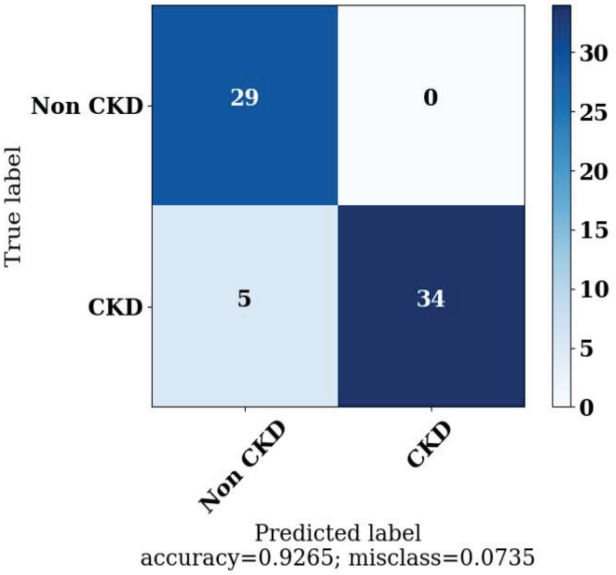


Fig. 3. Confusion Matrix Analysis

2) *Configuration Report*

The classification report shown in the table 3 below gives overall statistics and a thorough analysis of the random forest classifier's performance, decomposing the metrics for each class (CKD and Non CKD). For each class there is support,f1-score,precision and recall. Class 0(non- CKD) provides 85% of precision, 92% of f1-score. 1% precision for class 1(CKD)patients, 87% of recall and 93% of accuracy. Further, for precision and f1-score macro average is 93% and 87% for recall. Overall it means that the model is very good at correctly classifying both positive and negative cases.

TABLE III. CONFIGURATION REPORT ANALYSIS

	precision	**recall**	**F1-score**	**Support**
0	0.85	1.00	0.92	29
1	1.00	0.87	0.93	39
Accuracy			0.93	68
Macro avg	0.93	0.94	0.93	68
Weighted avg	0.94	0.93	0.93	68

C. *Result comparison of both Decision Tree and Random Forest*

Within the realm of chronic renal illness prediction, it has been discovered that the decision tree classifier performs significantly less well than the random forest classifier. In addition to achieving a higher overall accuracy (92.65% versus 79%), the random forest model also achieves higher precision, recall, and F1-scores for both the Chronic Kidney Disease (CKD) and Non-CKD classes. This model also demonstrates a better balance between precision and recall. The random forest is a more reliable solution for the medical diagnosis of chronic kidney disease (CKD) due to its superior performance, particularly in terms of reducing the number of false negatives and eliminating false positives. For the purpose of enhancing the accuracy and dependability of forecasts in medical applications, the primary objective of future study should be to consider the implementation of ensemble methods such as random forest.

VI. CONCLUSION

In its conclusion, the study recommends decision tree (DT) and random forest (RF) algorithms for chronic renal disease detection. This research aim was to investigate if these algorithms help make predictions using their data. Random Forest enhanced CKD prediction, and the decision tree classifier had good precision and recall. The Random Forest classifier has 93% accuracy, high precision, and recall for CKD and non-CKD classifications, whereas the Decision Tree had 79%. Random Forest classifier was accurate too. This key discovery shows that ensemble methods like Random Forest may accurately predict complex medical data. Healthcare system models may help manage chronic renal illness. When it comes to clinical decision-making processes, it is projected that future research will concentrate on further refining these models, broadening their applicability to a wider range of patient groups, and improving their interpretability more than ever before.

REFERENCES

[1] Rahman, M.M., Al-Amin, M. and Hossain, J., 2024. Machine learning models for chronic kidney disease diagnosis and prediction. *Biomedical Signal Processing and Control*, *87*, p.105368.

[2] Arora, A., Sehgal, C. and Agarwal, N., 2024, January. An Analysis of Machine Learning Algorithms for Chronic Kidney Disease Prediction. In *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 581-586). IEEE.

[3] Khalid, H., Khan, A., Zahid Khan, M., Mehmood, G. and Shuaib Qureshi, M., 2023. Machine learning hybrid model for the prediction of chronic kidney disease. *Computational Intelligence and Neuroscience*, *2023*(1), p.9266889.

[4] Liu, P., Liu, Y., Liu, H., Xiong, L., Mei, C. and Yuan, L., 2024. A Random Forest Algorithm for Assessing Risk Factors Associated With Chronic Kidney Disease: Observational Study. *Asian/Pacific Island Nursing Journal*, *8*, p.e48378.

[5] Islam, M.A., Majumder, M.Z.H. and Hussein, M.A., 2023. Chronic kidney disease prediction based on machine l earning algorithms. *Journal of pathology informatics*, *14*, p.100189.

[6] Iftikhar, H., Khan, M., Khan, Z., Khan, F., Alshanbari, H.M. and Ahmad, Z., 2023. A comparative analysis of machine learning models: a case study in predicting chronic kidney disease. *Sustainability*, *15*(3), p.2754.

[7] Haque, M.S., Amin, M.S., Ahmad, S., Sayed, M.A., Raihan, A. and Hossain, M.A., 2023, September. Predicting Kidney Failure using an Ensemble Machine Learning Model: A Comparative Study. In *2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)* (pp. 31-37). IEEE.

[8] Gill, K.S. and Gupta, R., 2023, May. Chronic Kidney Disease Detection Using GridSearchCV Cross Validation Method. In *2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON)* (pp. 318-322). IEEE.

[9] Reddy, S., Roy, S., Choy, K.W., Sharma, S., Dwyer, K.M., Manapragada, C. and Nakisa, B., 2024. Predicting chronic kidney disease progression using small pathology datasets and explainable machine learning models. *medRxiv*, pp.2024-04.

[10] Gill, K.S., Anand, V., Chauhan, R., Rawat, R.S. and Gupta, R., 2023, November. Kidney Disease X-Ray Image Classification Using a ResNet50V2 Model based Machine Learning Approach. In *2023 2nd International Conference on Futuristic Technologies (INCOFT)* (pp. 1-5). IEEE.

[11] Shanmugarajeshwari, V. and Ilayaraja, M., 2024. Intelligent Decision Support for Identifying Chronic Kidney Disease Stages: Machine Learning Algorithms. *International Journal of Intelligent Information Technologies (IJIIT)*, *20*(1), pp.1-22.

[12] Rahat, M.A.R., Islam, M.T., Cao, D.M., Tayaba, M., Ghosh, B.P., Ayon, E.H., Nobe, N., Akter, T., Rahman, M. and Bhuiyan, M.S., 2024. Comparing Machine Learning Techniques for Detecting Chronic Kidney Disease in Early Stage. *Journal of Computer Science and Technology Studies*, *6*(1), pp.20-32.