

Kidney disease prediction using Random forest classifier and Logistic regression

¹ Perumal Sivaraman & ² Dr.Javalkar Dinesh Kumar

Research scholar, Lingayas Vidyapeeth, Haryana, India

Professor, Department of CSE,Lingayas vidyapeeth Haryana, india

Abstract

Sustainable Development Goals (SDG) includes Good health and well-being. Wellness of kidney plays an important role. Soundness of kidney affects inadvertently due to lack of exercise, outside food and adulteration in the food items. The other causes are due to increased sugar level and eating too much of salt. The other medical issues that leads to kidney diseases includes blood pressure, specific gravity, blood glucose random(BGR), blood urea(BU), serum creatine(SC),pedal edema(PEE),anemia (ANE).wholeness of kidney is affected by more than 18 % and keeps on increasing. Blood test can be done to detect the ferocity of the diseases. Used a novel technique for prediction of kidney diseases using Random forest.Accuracy,Recall, F1-Score and Precision are in the range of 96.15, 93, 96 and 1.

Keywords—kidney diseases(KID), random forest, machine learning

I. INTRODUCTION

The main function of the kidney is to filter the blood. Kidney not working properly leads to many complications such as inability to filter blood and excessive wastes from the human body. This leads to swelling of the legs, unable to walk, feeling tired and making the normal work schedule very complicated. Sleep apnea also indication of KID. If left without treating it may lead to Ischemic Attack, stage called coma. Retaining to normal position takes time.

KID in initial stage can be treated effectively using tablets. Next level of treatment requires dialysis. It is a cooperative treatment until a kidney transplant is done. Dialysis takes 3 to 4 hours and cleaned blood is again passed to the body. with the correct support and care they can have a peaceful life, connect with friends, family and colleagues. Their economic status can be improved if found and detected early.

KID occurs due to processed food, intake of tea and coffee in high level. High blood level and high salt intake also leads to KID. Regular exercise, walking, kinesthetics and free from stress and depression also avoids KID. Initial stage it's not easy to find but advanced level leads to renal failure. Nephrologists or urologists are in shortage that makes patients to wait for longer period of time. Commutation also difficult for patients to move to hospitals located far away.

The finding of KID in patients is essential as it solves the KID in the early stage. Machine learning algorithms are well suited to predict KID.

II. Related Work

B. Ragu ,Dr. S. Perumal et al [1] Machine learning benefits many industries, including healthcare. Its alert and decision-making technologies improve patient safety and healthcare quality. This manuscript performs a detailed analysis of human disease prediction models based on Deep Learning and Machine Learning algorithms to determine the pros and cons of the existing methodologies

T. Kalpana¹,Dr.R.Thamilselvan et al [2] Five supervised learning algorithms (Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, K-Nearest Neighbour's (KNN)) were evaluated in terms of performance accuracy, precision, recall, F1-score, and ROCAUCOverall, Random Forest classifier was found to provide the best performance with all around better accuracy and better generalization than all other models.

Manpreet Kaur Grewal,Baibaswata Mohapatra et al [3] This study presents a methodology which applies different of clinical information to predict the status of CKID. This work implements four ML Models namely Decision Tree which is DT, SVM, XGBoost and Logistic Regression LR. The experimental results show that the Decision Tree (DT) model outperforms with score rate of 98.75%.

K. M. T. Jawad et al. [4] XAI algorithms would allow the patients or subjects understand the causes behind their disease at early stages so that they can be cautious about the

progression of the disease to later stages.

Dr. N. Srihari Rao, Priyanka S Marellavar et al [5] A very user- friendly frontend interface was designed to enter patient details and quickly see outputs from prediction, while a backend API for model inference is accessible. This project demonstrates the utility of synthetic data generation with machine learning to provide usable diagnostic tools that will be useful in healthcare.

Mohammad Anisur Rahman, Waleed Alsabhan et al [6] The most important characteristics in the early detection of CKID were hemoglobin, serum creatinine, and red blood cell count. The findings show the possibility of AI-based approaches to enhance clinical decisions and enable prompt treatment of chronic conditions. S. Poongodi, K. Sasirekha et al The Adaptive Neural Optimized Model is used for the classification process, which leverages the Extreme Learning Model's power for classification. The particle swarm optimization method is used for feature selection and optimization. The algorithm's output is evaluated using a number of performance criteria, including accuracy, precision, and recall. With the highest accuracy of 98%, the

suggested model performs better than the current models.

Gaurav Dhuriya, Aradhna Saini et al [7] Overall, because patients may miss a few estimations for a variety of reasons, missing qualities are usually seen in clinical settings. Once the inadequate informational index was adequately rounded off, models were constructed using six AI computations: arbitrary woodlands, support vector machine, closest neighbor, innocent Bayes classifier, decision tree, and strategic relapse. The choice tree and irregular forests AI models outperformed the others with a 99.00% determination precision. Therefore, it is estimated that this procedure may apply to more complicated clinical data for sickness diagnosis by examining the misconceptions caused by the setup models.

Polinati Vinod Babu, Bellamgubba Anoch et al [8] The accuracy is 98.36%, and the AUC is 0.992. Serum creatinine, hemoglobin, blood urea, blood pressure, and albumin are the most significant factors, according to the SHAP analysis. This improves the model's outputs' dependability by being in line with the body of clinical literature already in existence. The findings

demonstrate the recommended approach is a useful and simple way for medical professionals to make prompt decisions regarding CKID in a clinical context.

Karthikadevi M 1 and Sneha et al [9] Comparative performance evaluation against other machine learning models like K-Nearest Neighbors (KNN) and Decision Trees establishes the advantage of LightGBM. The model reaches 96% accuracy, 93% recall, and an F1-score of 89%, demonstrating the reliability of the model for CKID prediction. The research also pinpoints significant risk factors affecting CKID, providing important information for future research and clinical use. The results highlight the capability of machine learning to improve early detection of CKID and better patient outcomes.

Narander Kumar, Sabita Khatri et al [10] compared J48, Naïve Bayes, Random Forest, SVM and k-NN classifiers using performance measures like ROC, kappa statistics, RMSE and MAE using WEKA tool. Authors have also compared these classifiers on various accuracy measures like TP rate, FP rate, precision, recall and f-measure by implementing on WEKA. Experimental result shows that random forest classifier has better classification accuracy over others for chronic kidney disease dataset.

James H. W. Li et al [11] study demonstrates the feasibility of a lowcost, portable EIT system combined with machine learning for point-of-care CKID detection and eGFR estimation. The system has the potential to enable widespread, non-invasive, and self administrable CKID screening at a community level and telemedicine settings, reducing healthcare costs and improving early detection and management of CKID.

Arpit Pathak, Om prakash singh et al [12] The random forest (RF), gradient boosting (GB), Adaboost (AdaB), CatBoost (CatB) models achieve the highest overall performance, with an accuracy of 100% and the naive bayes (NB), and KNN approaching 97%, the findings demonstrated remarkable precision.

This study demonstrates how ML can improve the early detection of CKID and promote more effective healthcare practices.

III Methodology

For the prediction of chronic kidney disease using Random forest classifier. The steps involved is shown in the following diagram.

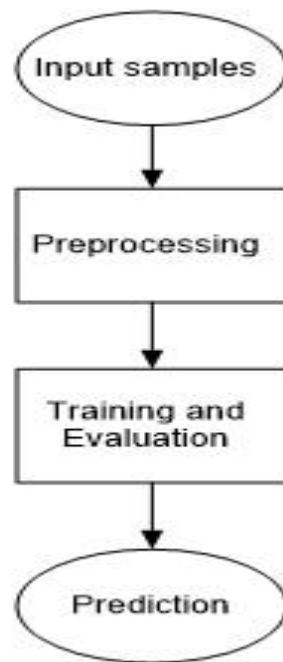


Fig1. Proposed Model with Random Forest

A. Preprocessing

Preprocessing is major role in preparing the dataset for Random forest classifier. First it cleans to address missing values. Feature selection or extraction techniques are implemented to identify attributes for predicting kidney disease. Improves accuracy of the model. Prevents misleading of results and improves training time

B. Random forest classifier

It uses many decision trees to make better predictions. The results are combined by voting for classification or averaging for regression. It selects most important feature for prediction. It works with big and complex data.

Preprocess the data by cleaning, organizing and transforming it. This involves missing data, outliers, encoding variable scaling numerical features and splitting the dataset for training and testing. This will evaluate early detection of disease. It works good with large datasets .works both with numerical and categorical data.

C. Experimental setup

Diagnosing kidney diseases done using Random forest classifier. This outperformed logistic regression classifier. There are 25 columns and 400 entries in the dataset.

D. Performance Measure

Performance metrics is calculated with accuracy, sensitivity, Recall and F1-score.

E. System design

Seaborn : used for drawing and statistical Graphs.

Matplotlib : used to create static , interactive and animated visualizations.

Pandas: Tool for manipulation and analysis.

Numpy: Working with arrays. used in domains of linear algebra ,fourier transforms.

Sklearn: simple and efficient tool for predictive data analysis.

F. Data Exploration

Bp-blood pressure

Pcc-pus cell clumps

Ba-Bacteria presence

Appet-Appetite status

S.No	Parameter	Random Forest
1	Accuracy	96.15
2	Recall	93
3	F1-Score	96
4	Precision	1
5	Macro Avg F1-score	96
6	Weighted Avg F1-score	96

Pe-pedal edema

Ane-Anemia status

Grf-Glomerular filtration rate

Ag-age

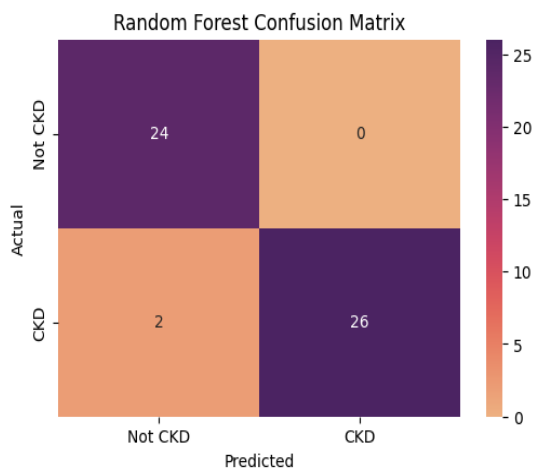
Hae-Haemoglobin status

Wbcc-white blood cell count

Rbcc-red blood cell count

These are some of the nominal and numeric Datas.

G. Feature Selection



IV RESULT AND DISCUSSION

A survey is conducted using age, blood pressure ,sodium, potassium, packed cell volume ,hypertension , pedal edema are the samples collected from datasets. Mortality rate reduction is the objective of this model. We got the accuracy results and are in the range of 96.15, 93, 96 and 1 for Accuracy, Recall, F1-Score and Precision.

TABLE II PERFORMANCE OF RANDOM CLASSIFIER

TABLE II COMPARISON OF PERFORMANCE

Classifiers	Accuracy	Recall	F1-Score	Precision	Macro Avg F1-score	Weighted Avg F1-score
Rafo	96.15	93	96	1	96	96
Logistic Regression	94.23	0.89	94	1	94	94

V CONCLUSION

we developed novel model for kidney diseases in the early stage by taking blood pressure,specific gravity,blood glucose random(BGR),blood urea(BU),serum creatine(SC),pedal edema(PEE),anemia (ANE).People who are having diabetes are more likely to have KID.

References

- 2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech) | 979-8-3315-2076-2/25/\$31.00 ©2025 IEEE | DOI: 10.1109/IEMENTECH65115.2025.10959624-Arpit Pathak,Om prakash singh
- 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3315-8618-8/25/\$31.00 ©2025 IEEE | DOI: 10.1109/EMBC58623.2025.11253076-James H. W. Li1,2, Runze Liu1,3, Michael Lawson1,4, Eddie C. Wong1,4,Iris Y. Zhou5,6, Desmond Y.H. Yap7,8, and Russell W. Chan1,4,*
- 3rd IEEE International Conference on "Computational Intelligence and Communication Technology" (IEEE-CICT 2017)-Narander Kumar,Sabita Khatri
- IEEE Xplore Part Number: CFP258A9-ART; ISBN: 979-8-3315-0574-5/25/\$31.00 ©2025 IEEE 1512025 International Conference on Machine Learning and Autonomous Systems (ICMLAS) | 979-8-3315-0574-5/25/\$31.00 ©2025

- IEEE | DOI: 10.1109/ICMLAS64557.2025.10967970-
Karthikadevi M 1 and Sneha N 2 Department of
Computer Application, Karpagam Academy Higher
Education 1,2 Coimbatore,TamilNadu, India,
karthikadevi056@gmail.com 1
snehanatarajan26@gmail.com 2
5. 2025 International Conference on Sustainable
Communication Networks and Application (ICSCN) |
979-8-3315-9420-6/25/\$31.00 ©2025 IEEE | DOI:
10.1109/ICSCN67106.2025.11308705-Polinati
Vinod Babu,Bellamgubba Anoch
6. 2025 3rd International Conference on Disruptive
Technologies (ICDT)2025 3rd International
Conference on Disruptive Technologies (ICDT) |
979-8-3315-1958-2/25/\$31.00 ©2025 IEEE | DOI:
10.1109/ICDT63985.2025.10986563-Gaurav
Dhuriya,Aradhna Saini
7. 2025 International Conference on Inventive
Computation Technologies (ICICT) | 979-8-3315-
1224-8/25/\$31.00 ©2025 IEEE | DOI:
10.1109/ICICT64420.2025.11004674-
S.Poongodi,K.Sasirekha
8. 2025 1st International Conference on Artificial
Intelligence Technology (ICoAIT) | 979-8-3315-
8679-9/25/\$31.00 ©2025 IEEE | DOI:
10.1109/ICoAIT67446.2025.11309087-Mohammad
Anisur Rahman, Waleed Alsabhan
9. 2025 9th International Conference on Inventive
Systems and Control (ICISC) | 979-8-3315-1247-
7/25/\$31.00 ©2025 IEEE | DOI:
10.1109/ICISC65841.2025.11187733-Dr. N. Srihari
Rao,Priyanka S Marellavar
10. K. M. T. Jawad et al.: Study on the Application of
Explainable AI on Ensemble Models-2024
11. 2025 3rd International Conference on
Communication, Security, and Artificial Intelligence
(ICCSAI) | 979-8-3315-3607-7/25/\$31.00 ©2025
IEEE | DOI: 10.1109/ICCSAI64074.2025.11064007 -
Manpreet Kaur Grewal,Baibaswata Mohapatra
12. 2025 3rd International Conference on Intelligent
Cyber Physical Systems and Internet of Things
(ICoICI) |978-1-6654-5753-8/25/\$31.00 ©2025 IEEE
|DOI: 10.1109/ICoICI65217.2025.11254615-
T.Kalpana1 ,Dr.R.Thamilselvan2,Dr.P.Vijayakumar
3,Anie H Syndhiya4 ,Gobinath M4 ,Madhubalan C
N4
13. 2025 3rd International Conference on Intelligent
Cyber Physical Systems and Internet of Things
(ICoICI) |978-1-6654-5753-8/25/\$31.00 ©2025 IEEE
|DOI: 10.1109/ICoICI65217.2025.11254615
14. 2025 3rd International Conference on Intelligent
Cyber Physical Systems and Internet of Things
(ICoICI) |978-1-6654-5753-8/25/\$31.00 ©2025 IEEE
|DOI: 10.1109/ICoICI65217.2025.11254615,B.
Ragu,Dr. S. Perumal
15. 2025 8th International Conference on Electronics,
Materials Engineering & Nano-Technology
(IEMENTech) | 979-8-3315-2076-2/25/\$31.00
©2025 IEEE |DOI:
10.1109/IEMENTECH65115.2025.10959447Optimiz
ing Diabetic Kidney Disease Diagnosis with Machine
Learning: A Novel Approach.
16. International Journal of Innovative Technology and
Exploring Engineering (IJITEE) ISSN: 2278-3075
(Online), Volume-14 Issue-6, May 2025 -
The Natural Language Processing Axioms in Classical
Tamil for Zonal Dialects using Machine Learning -
Perumal Sivarman, Prabakaran G, Senthil Kumar R
17. Journal of Harbin Engineering University - ISSN:
1006-7043 Integrative Feature Selection and
Enhanced Classification of Chronic Kidney Disease
through Novel Weight