Kidney Disease Prediction Using Machine Learning

CH. PAPA RAO[1], SYED AMJAD[2], M. RUPA VISWANATH[3], G. VASAVI[4], SK. MANSOOR[5]
[1] *Assoc Prof. in CSE Dept, GVR & S College of Engineering, Guntur, A. P, INDIA.*
[2, 3, 4, 5] *B. Tech Student, Department of CSE, GVR & S College of Engineering, Guntur, A.P, INDIA.*

Abstract*— A serious condition that can last a lifetime, chronic kidney disease (CKD) is brought on by either impaired kidney function or kidney cancer. It is possible to stop or limit the advancement of this chronic illness to the point when a patient's sole options for survival are dialysis or surgery. An earlier diagnosis and the right treatment can make this more likely to occur. The potential of various distinct machine learning techniques for offering an early diagnosis of chronic kidney disease (CKD) has been examined throughout this study. On this subject, a substantial quantity of research has been done. Nevertheless, by utilizing predictive modelling, we are strengthening our strategy. As such, in our methodology, we explore the relationship between data elements and target class features. Because predictive modelling allows for the introduction of improved attribute measures, we may use machine learning and predictive analytics to build a collection of prediction models.

Index Terms*— Chronic Kidney Disease, Predictive Modelling, Data Elements, Target Class Features.

I. INTRODUCTION

The illness known as chronic kidney disease, or CKD, occurs when the kidneys are so severely damaged that they are unable to filter blood as effectively as they should. The primary function of the kidneys is to eliminate excess water and waste from the circulation. Urine is created in this manner. Waste accumulation in the body is indicated by CKD. The reason this ailment is considered chronic is that the damage develops gradually over an extended period of time. People are affected by this disease anywhere in the world. As a result of CKD, you may encounter a number of health issues. Yes, a variety of symptoms, including back pain, stomach ache, diarrhoea, fever, nosebleeds, rash, and vomiting, may be present. The two most prevalent conditions that could harm the body over time are CKD can result from a wide range of illnesses, diabetes, high blood pressure, and heart disease being only three of terms.

Who develops chronic kidney disease (CKD) is influenced by age and gender in addition to these grave health issues. If you have diabetes and one or both of your kidneys aren't functioning, Machine learning for renal disease prediction is an exciting and important field of research that uses sophisticated computational methods to evaluate data and predict the probability of people getting kidney problems. In order to develop prediction models that can help healthcare practitioners with early detection and prevention methods, this field merges the domains of medicine, data science, and artificial intelligence. Historical patient data, including as test results, medical records, demographics, lifestyle characteristics, and, if accessible, genetic information, is used to train machine learning algorithms. Then, using these algorithms, it will be possible to find trends, connections, and risk factors related to kidney disorders, including acute kidney injury (AKI) and chronic kidney disease (CKD).

Stages of KD

1.Early stages of KD

Early on in its course, CKD usually shows no symptoms. This is because the human body can usually adapt to a significant decline in kidney function. Until a normal test for another condition, like a blood or urine test, finds a possible problem, kidney disease is frequently not detected until this point. Early detection and treatment with medicine, along with regular testing, may help stop it from developing into a more serious condition.

2. KD in its advanced stages

Kidney disease can present with several symptoms if it is not detected early or if it worsens despite treatment. Chronic kidney disease ends in kidney failure. It is also known as established renal failure or end-stage renal disease. It is probable that at some point a kidney transplant or dialysis will be required.

3. Tests for KD

When a disease or condition interferes with the kidneys' ability to function, the damage to the kidneys worsens over time and is referred to as chronic renal disease. When the kidneys are impacted by another illness or condition, this can happen.

4. Urine test

The Doctor also asks for a urine sample to assess kidney function. The kidneys generate urine. Blood and protein in your urine are signs that either or both of your kidneys.

5. Blood pressure

Your blood pressure is taken by the doctor because it indicates how well your heart is pumping blood within a given range. The patient has reached the terminal stage of renal disease if their GFR result is less than 15. Renal failure now has just two treatments available: kidney transplantation and dialysis. Age, gender, the frequency and duration of dialysis treatments, the degree of physical mobility the patient has, and their mental condition are all factors that affect how long the patient will live following dialysis. If dialysis proves to be unsuccessful, the doctor's sole remaining choice is kidney transplantation. However, the cost is really excessively.

II. LITERATURE REVIEW

Determining which features or variables are most relevant for predicting kidney disease is crucial. Researchers need to explore various clinical, demographic, laboratory, and genetic factors to identify the most informative features while avoiding redundant or noisy data. Ensuring the quality and reliability of input data is essential for building robust machine learning models. This includes handling missing values, outliers, and inconsistencies in the data through appropriate prepossessing techniques such as imputation, normalization, and feature scaling. In healthcare datasets, the classes (e.g., patients with kidney disease vs. without kidney disease) are often imbalanced, with fewer instances of positive cases. Dealing with class imbalance requires techniques such as oversampling, under sampling, or using advanced algorithms like SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples. While achieving high prediction accuracy is important

III. EXISTED METHOD

The proposed hybrid model is implemented in Python with pandas, Sk-learn, Matplotlib, NumPy, and other essential libraries. We have downloaded the CKD datasets from the UCI repository. The datasets contains two groups (CKD represented by 1 and non-CKD represented by of chronic kidney disease in the downloaded information. The machine learning algorithm that has best accuracy is selected for analysis and implementation so that repeated results are produced. We have also developed a hybrid model based on knowledge that we gained during the analysis and implementation. The hybrid model consists of Gaussian.

Naive Bayes, gradient boosting, and decision tree as base classifiers and random forest as a meta classifier. We have selected the tree-based machine learning algorithms for achieving the highest accuracy, while at the same time, it can handle the over-fitting.

IV. DISADVANTAGES OF EXISTING METHOD

1. Data Availabilty and Quality: Accessing comprehensive,High_quality medical data is challenging due to availability,consistency and biases
2. Feature Engineering Complexity: Extracting meaningful features from raw medical data requies domain expertise and careful consideration, impacting model effectiveness.
3. model Selection and Turning: Choosing suitable algorithms and hyperparameters requires extensive experimentation and validation,which can be computationally insentive.

V. PROPOSED METHOD

Data Collection: Gather relevant medical data related to kidney disease. This can include patient demographics, medical history, laboratory test results (e.g., creatinine levels, urine protein levels), imaging data (like ultrasounds or CT scans), and any other pertinent information.

Data Preprocessing: Clean the data to handle missing values, outliers, and inconsistencies. Perform feature engineering to extract meaningful features from raw

data that can help in predicting kidney disease outcomes.

Feature Selection/Extraction: Use techniques such as correlation analysis, feature importance ranking, or dimensionality reduction (like PCA) to select the most relevant features or create new features that capture important information from the data

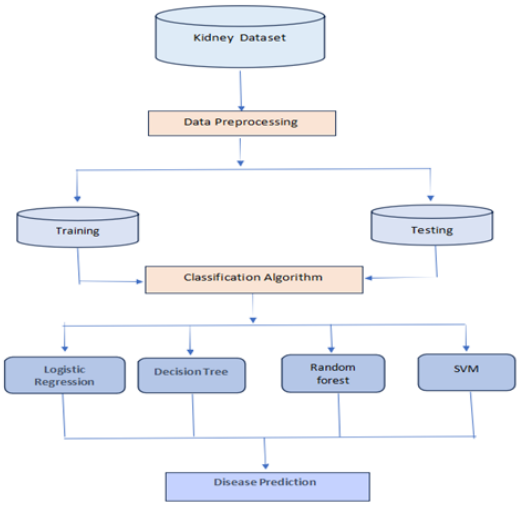
Model Selection: Choose appropriate machine learning models based on the nature of the problem. For binary classification tasks (e.g., presence or absence of kidney disease), models like logistic regression, support vector machines (SVMs), decision trees, random forests, or gradient boosting machines (GBMs) could be suitable. For regression tasks (e.g., predicting kidney function), linear regression, decision trees, or more advanced models like neural networks may be used.

Model Training: Split your data into training, validation, and test sets. Train your chosen models on the training data and tune hyper-parameters using the validation set to optimize model performance.

Model Evaluation: Evaluate the trained models using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, area under the ROC curve (AUC-ROC), or mean squared error (MSE) for regression tasks. Choose the metric(s) most relevant to your specific application.

Model Interpretation: For clinical applications, it's crucial to interpret the model predictions to understand which features contribute most to the predictions. Techniques like feature importance plots, SHAP values, or LIME (Local Interpretable Model- agnostic Explanations) can help in model interpretation.

Deployment and Monitoring: Once you have a well-performing model, deploy it in a real-world healthcare setting. Continuously monitor model performance and update the model as needed with new data or improved algorithm.



BLOCK DIAGRAM OF PROPOSED METHORD

5.1 ALGORITHMS

1. KNN

The KNN method is a popular and adaptive machine-learning technique that is widely utilized due to its ease of implementation. It is also an adaptable approach for a variety of datasets types in classification and regression applications due to its ability to handle both numerical and categorical data. It is a non-parametric forecasting technique that uses the degree of similarity between data points in a particular datasets. The KNN method calculates the K nearest neighbors to a given data point using a distance measure such as Euclidean distance. The majority vote or average of the K neighbors is then utilized to determine the data point's class or value. KNN has applications in recommendation engines, pattern recognition, and data preparation.

2. Decision Tree

Decision trees are intricate diagrams resembling botanical formations, depicting choices and outcomes. A sturdy trunk symbolizes the initial decision, with branches extending into paths, and nodes marking pivotal points. Each decision is strategic, splitting data into manageable subsets to maximize insights. Leaves represent final projections, offering clarity amid complexity. Data scientists, like arborists, prune and refine these trees to unveil hidden patterns. Dynamic and structured, decision trees capture choice, uncertainty, and information gain.

3.Random Forest

Random forests epitomize ensemble learning, blending diverse decision trees' perspectives for refined forecasts in regression and classification. Versatile in handling large datasets, they navigate complex landscapes adeptly. Resilient against overfitting, they balance bias and variance, ensuring accuracy even with noisy data. Collaboration enhances their predictive power, with each tree contributing a unique perspective. Unified forecasts transcend individual limitations, showcasing the power of collective intelligence. In supervised learning, random forests demonstrate the whole exceeding the sum of its parts, extracting insights from complex datasets effectively.

6. Ada Boost

AdaBoost, a stalwart in ensemble learning, amalgamates weak learners' insights iteratively for robust predictions. Like a conductor, it adjusts focus dynamically, emphasizing misclassified instances to refine accuracy. Initially egalitarian, it assigns greater importance to inaccurately classified instances over iterations. AdaBoost orchestrates diverse insights into a precise model, transcending individual limitations. Its adaptability and efficacy make it essential for tackling complex tasks, empowering practitioners with valuable insights.

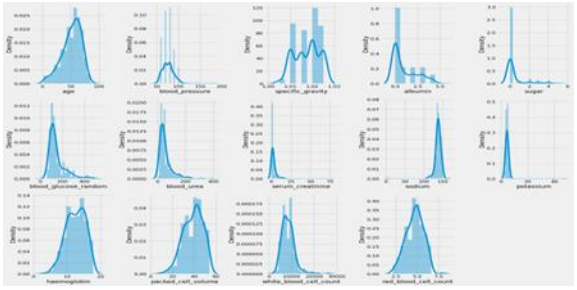
7. Gradient Boost

Gradient Boosting, a formidable boosting technique, iteratively enhances weak learners' predictions to create a robust model. Sequentially trained models rectify errors of predecessors, guided by minimizing a chosen loss function through gradient descent. Each iteration acts like a meticulous sculptor, refining predictions with precision. Initial patterns are captured by the base model, iteratively fine-tuned with subsequent learners. The ensemble's predictive power grows with each addition, capturing complex patterns adeptly. Gradient Boosting transforms weak learners into a formidable ensemble, outperforming individual models. Its iterative refinement and optimization make it a cornerstone in machine learning, offering precision and efficacy in diverse prediction tasks.

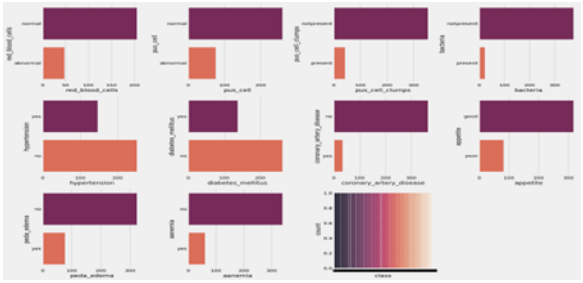
6 XG Boost

XGBoost, a trailblazer in machine learning, balances efficiency, scalability, and predictive power. Leveraging distributed gradient boosting, it orchestrates multiple weak models for robust forecasts, handling massive datasets seamlessly. Its adept management of missing values streamlines preprocessing, expediting model development. Parallel processing capabilities accelerate training, reducing computational overhead with large datasets. XGBoost's versatility spans regression to classification tasks, empowering practitioners with unparalleled speed and efficacy in extracting insights

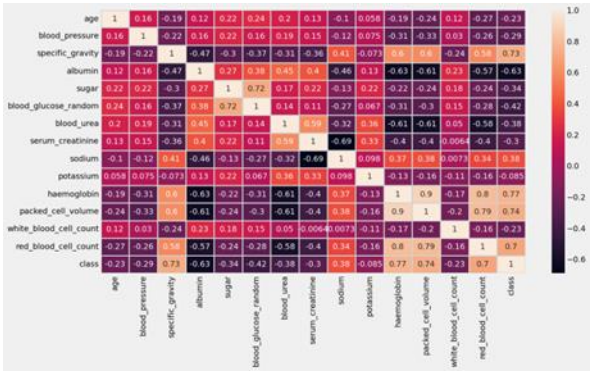
IV. RESULTS



1. Numerical feature distribution



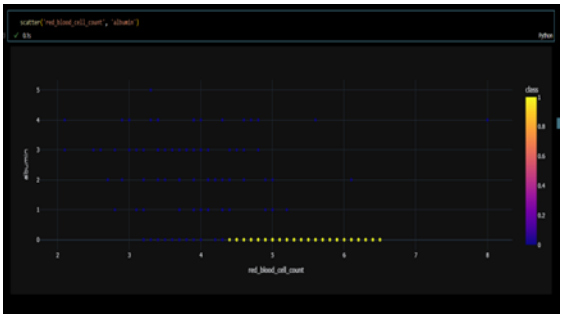
2. Categorical Column



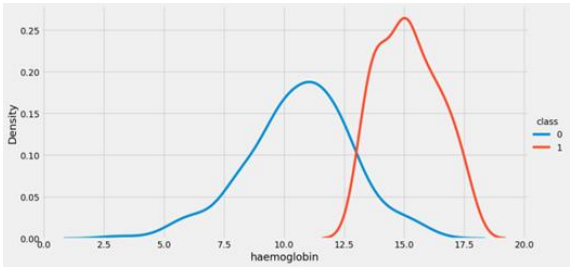
3. Numerical column

```
Index(['age', 'blood_pressure', 'specific_gravity', 'albumin', 'sugar',
       'red_blood_cells', 'pus_cell', 'pus_cell_clumps', 'bacteria',
       'blood_glucose_random', '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',
       'peda_edema', 'aanemia', 'class'],
      dtype='object')
```

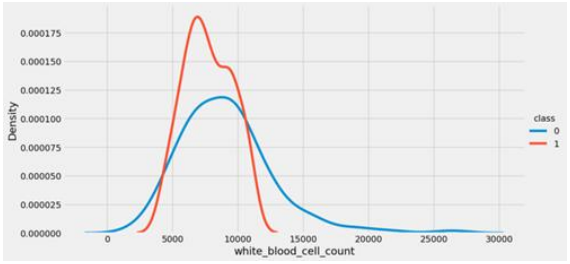
4. Exploratory data analysis (EDA)



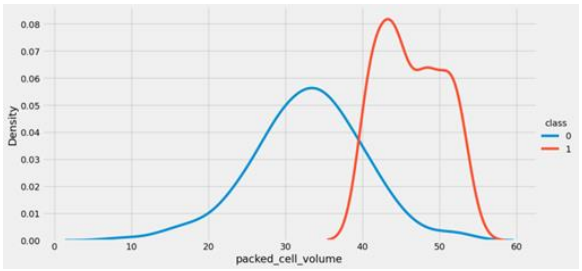
5. Red blood cell count



6. Haemoglobin



7. White blood cell count



8. Packed cell volume

CONCLUSION

In conclusion, this project investigated the efficacy of various machine learning algorithms for kidney disease prediction based on a datasets of blood pressure, specific gravity ext. We explored K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, AdaBoost, Random Forests, and XGBoost classifiers. All models achieved acceptable accuracy, with AdaBoost achieving the highest accuracy of 97.31%. This project demonstrates the potential of machine learning for early kidney disease using sugar blood urea, sodium. Future work could involve incorporating additional feature, such as demographic information or other physiological measurements, to improve model performance. Additionally, this approach could be validated on a larger and more datasets.

FUTURE SCOPE

The future of predictive modeling in chronic kidney disease (CKD) holds tremendous promise, poised to revolutionize both diagnosis and treatment. Technological advancements, particularly the fusion of advanced imaging methods with artificial intelligence (AI) algorithms, are set to boost diagnostic precision and enable early detection of CKD-related issues. By harnessing machine learning models trained on diverse datasets encompassing kidney ultrasound, CT/MRI scans, and histopathological data, clinicians can uncover subtle structural changes indicative of early-stage CKD, paving the way for proactive intervention and personalized treatment plans.

Moreover, the integration of genetic and molecular markers into predictive models offers a path to personalized risk assessment and treatment

optimization for CKD patients. By incorporating genetic variations linked to kidney disease risk and biomarkers signaling kidney damage or dysfunction, predictive models can stratify patients based on their genetic predisposition and disease progression. This personalized approach empowers clinicians to tailor interventions to each patient's unique needs, leading to more effective disease management and better outcomes.

Additionally, the rise of real-time patient monitoring devices and wearable tech opens doors for continuous remote monitoring of CKD patients, enabling early detection of disease exacerbations and proactive intervention. By leveraging data from devices like smartwatches and fitness trackers, alongside predictive modeling techniques, healthcare providers can track vital signs and detect deviations from baseline, facilitating prompt adjustments to treatment plans and lifestyle interventions. This integration of predictive modeling with remote patient monitoring holds the potential to enhance patient engagement, improve healthcare delivery, and reshape the landscape of CKD management.

REFERENCES

[1] Y. Amirgaliyev, S. Shamiluulu and A. Serek, "Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods," 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan, 2018, pp. 1-4, doi: 10.1109/ICAICT.2018.8747140.

[2] N. K. Pareek, D. Soni and S. Degadwala, "Early Stage Chronic Kidney Disease Prediction using Convolution Neural Network," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 16-20, doi: 10.1109/ICAAIC56838.2023.10141322.

[3] Anurag, N. Vyas, V. Sharma and D. Balla, "Chronic Kidney Disease Prediction Using Robust Approach in Machine Learning," 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT), Dehradun, India, 2023, pp. 1-5, doi: 10.1109/CISCT57197.2023.10351277.

[4] H. H. Yördan, M. Karakoç, E. Çalğici, D. Kandaz and M. K. UÇar, "Hybrid AI-Based Chronic Kidney Disease Risk Prediction," 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), Sivas, Turkiye, 2023, pp. 1-4, doi: 10.1109/ASYU58738.2023.10296642.

[5] R. Al-Momani, G. Al-Mustafa, R. Zeidan, H. Alquran, W. A. Mustafa and A. Alkhayyat, "Chronic Kidney Disease Detection Using Machine Learning Technique," 2022 5th International Conference on Engineering Technology and its Applications (IICETA), Al-Najaf, Iraq, 2022, pp. 153-158, doi: 10.1109/IICETA54559.2022.9888564.

[6] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," 2016 Management and Innovation Technology International Conference (MITicon), Bang-San, Thailand, 2016, pp. MIT-80-MIT-83, doi: 10.1109/MITICON.2016.8025242

[7] L. A. Akinyemi, O. P. Oshinuga, S. O. Ekwe and S. O. Oladejo, "Enhancing Chronic Kidney Disease Prediction Through Data Preprocessing Optimization and Machine Learning Techniques," 2023 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, 2023, pp. 1-6, doi: 10.1109/ICECET58911.2023.10389513.

[8] A. Vijayalakshmi and V. Sumalatha, "Survey on Diagnosis of Chronic Kidney Disease UsingMachine Learning Algorithms," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 590-595, doi: 10.1109/ICISS49785.2020.9315880.

[9] S. S. Rasheed and I. H. Glob, "Classifying and Prediction for Patient Disease Using Machine Learning Algorithms," 2022 3rd Information Technology To Enhance e-learning and Other Application (IT-ELA), Baghdad, Iraq, 2022, pp. 196-200, doi: 10.1109/IT-ELA57378.2022.10107935.