A Random Forest Classifier Approach for Predicting Chronic Kidney Disease

Dr. M. Ganesan[1], D. Haritha[2], V. Ragapriya[3], S. Vidya[4]

Assistant Professor[1], B. Tech Students[2,3,4]
Department of Computer Science and Engineering
Sri Manakula Vinayagar Engineering College, Puducherry

Abstract: Chronic Kidney Disease (CKD) is an international fitness trouble with excessive morbidity, mortality, and different illnesses. Kidney ailment impacts 1 in 10 human beings worldwide. As the quantity of CKD sufferers increases, powerful predictors for early detection of CKD are needed. Therefore, a higher prognosis of persistent kidney ailment is wanted to save you endured progression. Machine learning assist clinicians gain this intention with their rapid and correct prediction capabilities. In this project, we've the usage of Random Forest Algorithm to decide if a person has CKD or Not. The CKD dataset has been taken from the Kaggle repository. This helps in the early detection of persistent kidney ailment.

Keywords: Chronic Kidney Disease, Prediction, Random Forest, CKD, Machine Learning

I. INTRODUCTION

Both in the industrialised and developing worlds, chronic renal illness is becoming an increasingly serious issue. In developing nations, people are adopting unhealthy lifestyles that encourage diabetes and high blood pressure, the major causes of chronic renal illness, while 10–20% of persons with diabetes pass away from kidney disease, according to a CNN study. On the other hand, 1 in 10 adults in affluent nations like the USA, or 26 million Americans, have CKD, and millions more are at higher risk. An eight-point risk factor checklist for predicting chronic renal disease has been developed by US researchers. Older age, anaemia, female sex, hypertension, diabetes, peripheral vascular disease, and any prior congestive heart failure or cardiovascular disease history are all on the list. The most common causes of chronic renal failure are related to:

Poorly controlled diabetes, Kidney stones, Polycystic Kidney Disease and Prostate disease. Kidney failure may initially show no signs (not producing any symptoms). The symptoms of declining kidney function are caused by the body's inability to control water and electrolyte balances, remove waste items, and encourage the creation of red blood cells. Lethargy, a lack of energy, breathlessness, and swelling all over could happen. Life-threatening situations might arise if they go unnoticed or untreated. Anemia, or a low red blood cell count, can produce generalised weakness because erythropoietin levels that are too low do not sufficiently activate the bone marrow. The body fatigue easily because a reduction in red cells causes a reduction in the blood's ability to carry oxygen, which in turn affects how much oxygen is delivered to the cells for work. Additionally, when there is less oxygen available, cells are more likely to use anaerobic metabolism (an=without + aerobic=oxygen), which results in greater acid generation that the kidneys' already failing ability to filter out cannot handle. Kidneys are greatly concerned with three major functionalities: Pressure filtration, Tubular secretion, Urine formation.

A. Analysis of Chronic Kidney Disease (CKD)

Kidneys are one of the most important organs that filter all waste and water from the human body to make urine. Chronic kidney disease (CKD), also commonly known as chronic kidney disease or chronic renal failure, is a life-threatening condition that results from the inability of the kidneys to perform their normal functions. Continuously declining glomerular filtration rate (GFR) for more than 3 months is a universal health problem. The dataset is pulled from Kaggle's Machine Learning repository. Preprocessing is considered by replacing or splitting the dataset to eliminate missing values. Random Forest, a classification algorithm, is used to create the model. The dataset contains 400 illustrations, each with 26 features. The class target variable contains the value 'CKD' or 'NOT CKD'. "CKD" indicates a positive test for chronic kidney disease, "NOT CKD" indicates a negative test. There are 250 cases in the "CKD" class and 150 cases in the "not CKD" class. 0 means negative, meaning no CKD, 1 means positive, meaning CKD.

II. TECHNOLOGY USED IN PROPOSED SYSTEM

A. Machine Learning

The need for machine learning is increasing day by day. Machine learning is necessary because it can perform tasks too complex for humans to implement directly. Humans have some limitations due to their inability to manually access large amounts of data. To do this, you need some computer system. This is where machine learning comes in to make things easier for us. You can train machine learning algorithms by feeding them large amounts of data, letting them explore the data, build models, and automatically predict the desired output. The performance of machine learning algorithms depends on the amount of data and can be determined by a cost function. You can save both time and money with the help of machine learning. The importance of machine learning is easily understood through its use cases. Today, machine learning is being used in self-driving cars, cyber fraud detection, facial recognition, Facebook friend suggestions, and more. Various top companies such as Netflix and Amazon use large amounts of data to develop machine learning models that analyze user interests and recommend products accordingly.

B. Types of Machine Learning

1) Supervised Learning

Supervised learning is a type of machine learning in which you provide labeled sample data to train a machine learning system and predict an output based on it. The system uses the labeled data to build a model, understand the data set, and learn more about each data. Once trained and processed, test the model by providing sample data to see if it predicts accurate outputs. The goal of supervised learning is to map input data to output data. Supervised learning can be further classified into her two categories of algorithms: Classification and Regression.

2) Unsupervised Learning

Unsupervised learning is a learning method in which machines learn without supervision. Training is provided to the machine using a labeled, classified, or unclassified dataset, and the algorithm should operate on that data unsupervised. The goal of unsupervised learning is to reconstruct the input data into a set of objects with new features or similar patterns. In unsupervised learning there are no predetermined outcomes. It can be further classifieds into two categories of algorithms: Clustering and Association.

2) Reinforcement Learning

Reinforcement learning is a feedback-based learning method in which the learning agent receives rewards for all correct actions and penalties for all wrong actions. Agents automatically learn from this feedback and improve their performance. In reinforcement learning, agents interact and explore their environment. The agent's goal is to earn maximum reward points and improve their performance.

III. SYSTEM STUDY

A. Problem Definition

Early Chronic Kidney Disease (CKD) detection and therapy are highly preferred since they assist avoid negative outcomes. Recently, the use of machine learning techniques for the early detection and diagnosis of numerous diseases has become increasingly popular. The goal of this project is to use machine learning classification methods to predict the presence of CKD using datasets gathered from medical records of affected individuals. We specifically developed a sustainable and workable technique for identifying various phases of CKD utilising random forests, which has broad medical accuracy.

B. Existing System

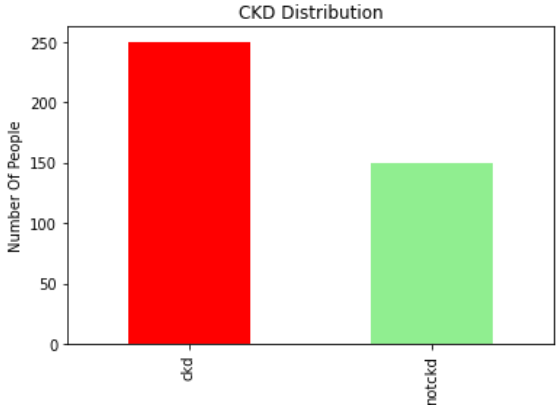
Everyone has experienced being a patient at some point, and we all desire receiving high-quality medical care. We assume that all physicians are board-certified medical specialists and that the data supporting their recommendations is reliable. But it isn't always the case. They can't memorise all the information needed for every situation, and they probably don't have it on hand either. Even if they had access to the enormous volumes of data necessary to compare treatment outcomes for all the diseases they see, it would still take them time and experience to analyse the data and integrate it with the patient's unique medical profile. A doctor, however, cannot be expected to do a comprehensive statistical analysis or study of this kind. They want a doctor who can talk to them, listen to them, and give them advice on how to restore and protect their health in the future. Takes precedence over desire. Systems in use today are built with complex architectures and limited use. Most of the developed systems do not have proper prototypes, making it difficult to implement previously created systems. Existing systems have used various methods to predict chronic kidney disease, but machine learning has not been used because it does not improve accuracy.

C. Proposed System

The aim of this project is to detect primarily chronic kidney disease at an early stage using random forests. Existing systems are not much accurate in their predictions. Using Random Forest in Machine Learning allows a user to predict her CKD and provide results with maximum accuracy. As a result, we have ensured that the use of the random forest algorithm is highly accurate compared to all other existing systems. Existing systems could not provide better accuracy. These were the main drawbacks overcome in the proposed system. Given the data results and analysis ideas, Random Forest gives the best results.

1) Data Description

The dataset is pulled from Kaggle's Machine Learning repository. Preprocessing is considered by replacing or splitting the dataset to eliminate missing values. RandomForest, a classification algorithm, is used to build the model. The dataset contains 400 illustrations, each with 25 features.Blood pressure, potassium, anemia, albumin, sugar, pus cells, appetite, pus cell mass, bacteria, hypoglycemia, hypertension, specific gravity, sodium, hemoglobin, age, packed cell volume, red blood cells, white blood cell count, red blood cell count, blood urea, diabetes mellitus, serum creatinine, coronary artery disease, leg edema and class. The class target variable contains the value 'CKD' or 'NOT CKD'. On the other hand, for chronic kidney disease, "CKD" indicates a positive test and "NOT CKD" indicates a negative test. There are 250 cases in the "CKD" class and 150 cases in the "not CKD" class. Distribution of CKD patients - We built a model to predict CKD, but the data set was slightly imbalanced with approximately 400 classes. 0 means negative, meaning no CKD, 1 means positive, meaning CKD.



2) Data Pre-processing

The most essential phase is data pre-processing. Most data pertaining to healthcare has missing values and other contaminants that can affect the efficacy of the data. Data pre-processing is done to increase the quality and efficacy of the results from the mining process. This procedure is crucial for accurate results and good prediction when applying machine learning techniques to the dataset.

a) Missing Values removal

There will be missing values when the data is real-world data. The accuracy of the predictions is further shifted as a result. A good way to deal with missing values is to use the mean or average of the observed attribute or value. We now have more accurate data and better prediction results as a result.

b) Data Transformation

In this step, we convert the actual data given to us into the required format. The downloaded data consists of nominal, real and fractional values. This step converts nominal data to numeric data in formats 0 and 1 (yes/1 or no/0). The resulting CSV file contains all integer and fractional values for various CKD-related attributes.

c) Splitting of data

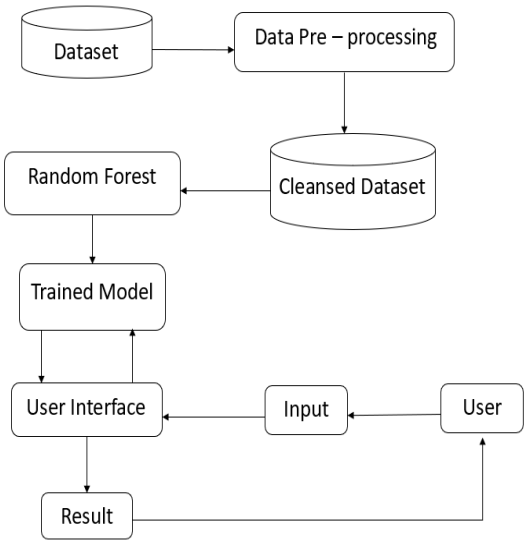
So, after cleansing the data, we normalize the data to train and test the model. When splitting the data, train the algorithm on a training data set, but a separate test data set. This training process generates a training model based on the logic, algorithms, and feature values of your training data. The goal of normalization is to put all attributes on the same scale.

3) Apply Machine Learning

Once the data is ready, apply machine learning techniques. A random forest algorithm classification technique was used to predict chronic kidney disease (CKD).

a) *Random Forest*

Classification and regression problems are typically addressed using Random Forest, a supervised machine learning technique. Create decision trees using various samples, and then use the majority of votes for classification and average when performing regression. The random forest algorithm's capacity to handle datasets with continuous variables, as in regression, and categorical variables, as in classification, is one of its most crucial aspects. In categorization issues, it produces superior outcomes. It is built on the idea of ensemble learning, which brings together various classifiers to address difficult issues and enhance model performance. A random forest is a classifier that uses an average of several decision trees for various subsets of a given dataset to increase the dataset's predictive accuracy. A random forest uses predictions from each tree to forecast the ultimate result based on the predictions' majority vote rather than depending on decision trees. The accuracy and reduction of overfitting issues improve with increasing forest tree density.



V. WORK MODULES

A. Exploratory Data Analysis

Exploratory data analysis primarily relies on graphical representation and visualization of data. Statistical modeling provides a simple, low-dimensional representation of relationships between variables, but usually requires a thorough understanding of statistical methods and mathematical concepts. Visualizations and charts are often easier to create and much better to interpret, allowing you to quickly explore different characteristics of your dataset. The ultimate goal is to produce a concise summary of information supporting the request. This isn't the final step in the data science process, but it's still important. The final graphics will differ from the graphics produced in EDA. In the course of exploring your dataset, you will probably create dozens, if not hundreds, of exploration charts. We encourage you to post one or two of these charts in their final form. His one goal of EDA is to help develop one's own knowledge of data, so all code and graphics should aim at this goal. Exploratory graphics do not have to include any significant information that you might include when you publish the graphic2. Exploratory data analysis is a type of data analysis that examines data distributions to identify inherent laws. EDA (exploratory data analysis) uses visual techniques to identify the inner structure of data. Visual data analysis techniques can be traced back centuries, as the human eye and brain play a very important role in data exploration and have a long history in various fields. Exploratory data analysis also

has the ability to uncover unexpected differences not possible with traditional models. A key element of exploratory data analysis is flexibility, both in how it applies to data structures and how later analytical steps respond to the disclosed modes of analysis.

B. Knowledge Discovery and Analysis

Process data using a variety of feature extraction approaches such as statistical description, association, and pattern analysis. Data reduction: This phase aims to reduce the modified data using the key features of mining technology. After thoroughly analyzing the data, we remove irrelevant information and deduce and consider patterns that will help the training process.

C. Model Training and Prediction

Training a machine language model is the process of feeding an ML algorithm with data that helps it identify and learn good values for all relevant attributes. There are many types of machine learning models, but the most common are supervised and unsupervised models. This project uses supervised learning. Supervised learning is possible when the training data contains both input and output values. Each record containing inputs and expected outputs is called a supervisory signal. Training is based on the deviation of processed results from documented results when inputs are given to the model. "Prediction" refers to the output of an algorithm after it has been trained on historical data sets and applied to new data to predict the likelihood of a particular outcome. Predictions should be based on prior knowledge, experience, observations and research.

IV. EXPERIMENTAL RESULTS

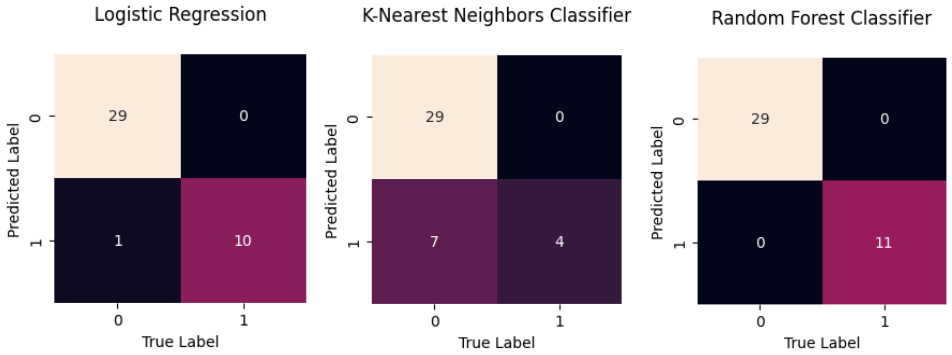
A classification procedure is implemented with the following results: The TP rate is associated with true positives, which are correctly classified cases in the dataset, and the FP rate is associated with false positives, which are misclassified occurrences in the class.

A. Performance Evaluation of Classification

Classification performance is evaluated by calculating accuracy, sensitivity, specificity, F1-value, and confusion matrix using the appropriate mathematical relationships described below.

1) Confusion matrix

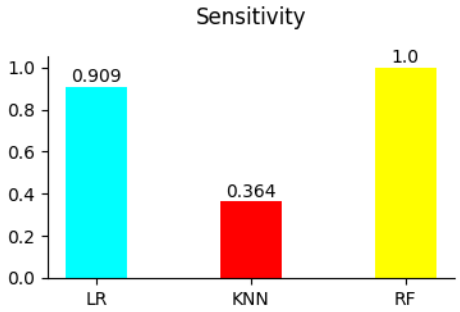
A confusion matrix, also called an error matrix, is a table commonly used to describe the performance of a classification model (or "classifier") on a set of test data whose true values are known. This allows you to visualize the performance of your algorithm. The key to the confusion matrix is not just the number of errors, but the number of correct and incorrect predictions, summarized in counts and sorted into each class.



2) Sensitivity

Also called true positive rate (TPR), hit rate, or recall. It represents the ratio of correctly classified positive instances to the total number of positive instances. The formula used in this study to calculate sensitivity is:

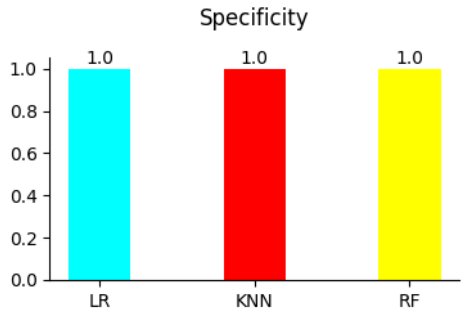
$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$



3) Specificity

This is also called true negative rate (TNR) or reverse recall. Measures the percentage of correctly classified negative instances out of the total number of negative instances. The formula used in this study to calculate specificity is:

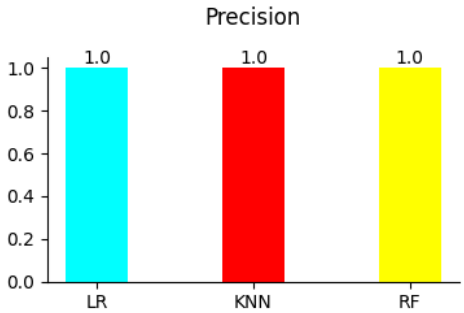
$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$



4)Precision

Precision is defined as the percentage of positive identifications that were correct. The formula used to calculate accuracy in this study is:

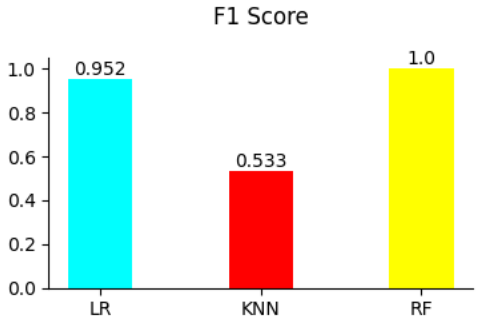
$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$



5) F1-measure

The F1 - measure is calculated by taking the weighted average of the sensitivity and precision values. The F1 - measure uses the domain of information retrieval to estimate classification performance. The formula used to calculate F1 - measure in this study is:

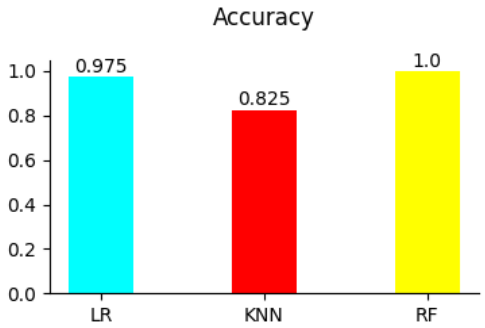
$$F1 - Measure = \frac{2 * \mathrm{sensitive} * \mathrm{precision}}{\mathrm{sensitive} + \mathrm{precision}}$$



6) Accuracy

One of the most commonly used classification performance metrics is accuracy. This is the ratio of correctly classified samples to the total number of samples. The formula used to calculate accuracy in this study is:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$



B. Tabulation of Results

Classifiers	Accuracy
Random Forest (proposed system)	100%
K – Nearest Neighors Classifier	82.5%
Logistic Regression	97.5%

VI. CONCLUSION

Chronic kidney infection is one health problem that requires better diagnosis. Prognostication of this disease in its early stages can halt its progression. Our system therefore aims to predict this at an early stage. Predictions are made using the random forest algorithm, a machine learning technique. A model obtained from a CKD patient is trained and validated using the previously mentioned input parameters. Also, 100% classification accuracy was achieved. The main purpose of the machine learning module is to label a patient's CKD status. In the future, a variety of more complex and specialized data can be collected to train the model and improve its generalization performance while determining disease severity. We believe this version will continue to improve as the length and data quality increase.

REFERENCES

1. Ganapathi Raju, N. V., Prasanna Lakshmi, K., Praharshitha, K. G., & Likhitha, C. (2019). Prediction of chronic kidney disease (CKD) using Data Science. 2019 International Conference on Intelligent Computing and Control Systems (ICCS).
2. M, M., & Balakrishnan, S. (2020). An Ensemble Feature Selection Method for Prediction of CKD. 2020 International Conference on Computer Communication and Informatics (ICCCI).
3. S. Smys et al. (2019). “Feature Selection and Ensemble Entropy Attribute Weighted Deep Neural Network (EEAw-DNN) for Chronic Kidney Disease (CKD) Prediction.” (Eds.): ICCVBIC 2019, AISC 1108, pp. 1232–1247, 2020
4. Elhoseny, M. et al. (2019). “Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease.” Scientific reports, 9(1): 9583.
5. Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., … Bolshev, V. (2021). Prediction of Chronic Kidney Disease - A Machine Learning Perspective. IEEE Access, 9, 17312–17334.
6. M. Almasoud and T. E. Ward, ‘‘Detection of chronic kidney disease using machine learning algorithms with least number of predictors,’’ Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 89–96, 2019.
7. P. G. Scholar, ‘‘Chronic kidney disease prediction using machine learning,’’ Int. J. Eng. Res. Technol., vol. 9, no. 7, pp. 137–140, 2020.
8. F. Ma, T. Sun, L. Liu, and H. Jing, ‘‘Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network,’’ Future Gener. Comput. Syst., vol. 111, pp. 17–26, Oct. 2020
9. G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, and J. A. Recio-Garcia, ‘‘Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning,’’ IEEE Access, vol. 7, pp. 152900–152910, 2019.
10. J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, ‘‘A machine learning methodology for diagnosing chronic kidney disease,’’ IEEE Access, vol. 8, pp. 20991–21002, 2020.
11. Gupta, R., Koli, N., Mahor, N., & Tejashri, N. (2020). Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease. 2020 International Conference for Emerging Technology (INCET).
12. Gudeti, B., Mishra, S., Malik, S., Fernandez, T. F., Tyagi, A. K., & Kumari, S. (2020). A Novel Approach to Predict Chronic Kidney Disease using Machine Learning Algorithms. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA).
13. J. Xiao et al, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of Translational Medicine, vol. 17, (1), pp. 119, 2019.