International Journal of Multidisciplinary Research and Modern Education (IJMRME)
International Peer Reviewed - Refereed Research Journal, Website: www.rdmodernresearch.com
Impact Factor: 7.315, ISSN (Online): 2454 - 6119, Volume 10, Issue 1, January - June, 2024

MACHINE LEARNING MODELS FOR PREDICTION OF CHRONIC RENAL DISEASE

C. Syamsundar Reddy*, G. Anjan Babu & Anbu Malar*****

* Research Scholar, Department of Computer Science, SVU College of CM&CS, Sri Venkateswara University, Tirupati. Andhra Pradesh
** Professor, Department of Computer Science, SVU College of CM&CS, Sri Venkateswara University, Tirupati. Andhra Pradesh
*** Faculty, National Institute of Electronics and Information Technology, Chennai. Tamil Nadu

Cite This Article: C. Syamsundar Reddy, G. Anjan Babu & Anbu Malar, "Machine Learning Models for Prediction of Chronic Renal Disease", International Journal of Multidisciplinary Research and Modern Education, International Peer Reviewed - Refereed Research Journal, Volume 10, Issue 1, January - June, Page Number 76-83, 2024.

Abstract:

Chronic kidney disease (CKD) is a long-term condition where the kidneys do not work as well as they should. It's a common condition often associated with getting older. It can affect anyone, but it's more common in people who are black or of south Asian origin. Early CRD diagnosis enables individuals to receive prompt therapy to slow the course of their illness. Because it quickly and accurately identifies potentials, model-based machine learning can help doctors accomplish this aim. In this project, we presented an ML strategy for CRD forecasting. The UCI machine learning repository provided the CRD dataset, which has a high number of uncertain data values. To address unknown values or unknown data, KNN estimation was employed, which picks multiple instances that contain the most comparable measures to manage the undetermined data value for every insufficient sample. Missing data is prevalent in real therapy situations; therefore, individuals may skip particular measurements for a variety of reasons. The technique is generated using five machine learning algorithms (SVM, Logistic Regression, k-nearest neighbor, Random Forest, and Neural Network). Random forest outscored the other models of machine learning regarding accuracy.

Key Words: Decision Tree, KNN, Logistic Regression, Neural Network, Random Forest, Renal, SVM

1. Introduction:

Chronic kidney disease (CKD) represents a heavy burden on the healthcare system because of the increasing number of patients, high risk of progression to end-stage renal disease, and poor prognosis of morbidity and mortality. The aim of this study is to develop a machine-learning model that uses the comorbidity and medication data obtained from Taiwan's National Health Insurance Research Database to forecast the occurrence of CKD within the next 6 or 12 months before its onset, and hence its prevalence in the population. A total of 18,000 people with CKD and 72,000 people without CKD diagnosis were selected using propensity score matching. Their demographic, medication and comorbidity data from their respective two-year observation period were used to build a predictive model. Among the approaches investigated, the Convolutional Neural Networks (CNN) model performed best with a test set AUROC of 0.957 and 0.954 for the 6-month and 12-month predictions, respectively. The most prominent predictors in the tree-based models were identified, including diabetes mellitus, age, gout, and medications such as sulfonamides and angiotensins. The model proposed in this study could be a useful tool for policymakers in predicting the trends of CKD in the population.

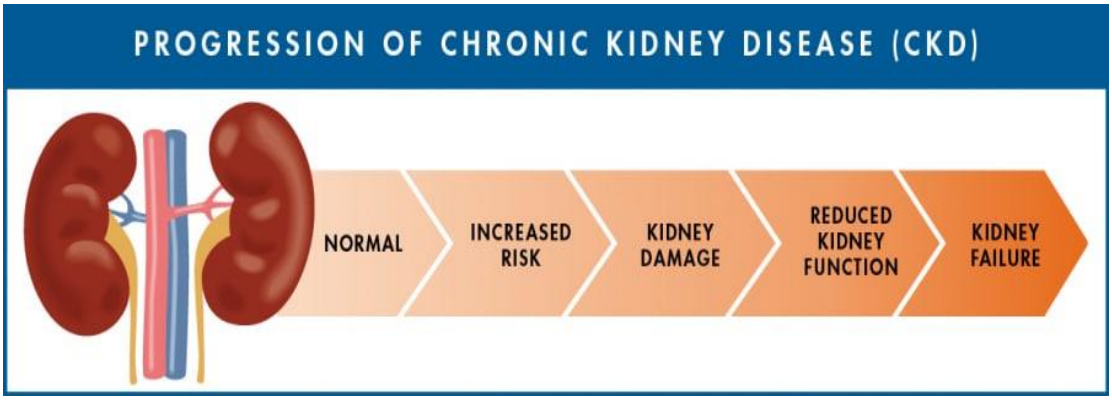

Figure 1: Progression of Chronic Renal Disease (CRD)

The models can allow close monitoring of people at risk, early detection of CKD, better allocation of resources, and patient-centric management (L. Zhang et al., 2012). The world notices 2.4 million deaths every

year and is currently the sixth most noteworthy developing reason for death. Also, these are the numbers when the cases in India, which is home to world's 17% populace, remain to a great extent undocumented and unregistered. Kidneys are one of the vital organs in the lower back, one kidney arranged on one or the other side of the spine. The kidneys main function is to purify blood and remove the waste from the body in the form of urine (Singh et. al, 2015). Renal failure is another name for chronic kidney disease. In the entire world, kidney illness affects one in 10 persons. According to the National Kidney Foundation, 10% of the world's population has chronic kidney disease (CKD), which affects one in five men and one in four women between the ages of 60 and 75. The major goal of this study is to create a machine learning-based model that can correctly forecast the onset of chronic renal disease. Chronic renal disease is a long-term medical condition that affects the kidneys' ability to filter waste and excess fluids from the body. The disease is often asymptomatic in the early stages, making it difficult to diagnose and treat. In this paper section-2 discusses the previous studies to identify and diagnosis of CKD,

2. Literature Review:

Chronic renal failure develops as a result of progressive, frequently permanent kidney function impairment. This takes place gradually over the course of months or years. Therefore, to treat chronic kidney disease, we must first determine if the patient is experiencing chronic kidney damage or not. The following list of information discusses the many approaches and procedures that have enabled us to anticipate chronic kidney disease from last decade.

Each bimonthly issue of *Advances in Kidney Disease and Health* presents focused review articles devoted to a single topic of current importance in clinical nephrology and related fields. There was convincing evidence that a healthy dietary pattern may lower CKD risk. Plant-based foods, coffee, and dairy may be beneficial. Unhealthy diets and their components, such as red (processed) meat and sugar-sweetened beverages, may promote kidney function loss (van Westing et al., 2020). A Comparative examination of temporal electronic health record (EHR) data proposed to analyze the usefulness of hierarchical risk prediction systems in forecasting the decline of renal activity. The study, conducted in February 2015, aimed to evaluate and compare different approaches for predicting the risk of renal activity deterioration using EHR (Singh et al. 2015). 2-Type fuzzy classifiers proposed to recognize CKD. This study uses FuRES and FOAM fuzzy classifiers to categorize people with renal disease. Utilizing kidney disease data from the UCI Machine-based Learning network, their viability and robustness were evaluated (Z.Chen et al. 2016). A novel technique was developed to determine severity of CKD in patients, where many classification models: K-Nearest Neighbor algorithms and Naïve Bayes classifiers are employed to accomplish this prediction (G.S. Drall et. Al, 2018). A long-term and practical model using machine learning algorithms was proposed to detect the various phases of kidney disease using machine learning algorithms using a dataset gathered from afflicted people's health records and employed the Random Forest and J48 algorithms. It shows high accuracy in prediction of various stages of kidney disease (Hamid aIlyas et.al, 2021). It is evident that Multi classification work was very important to know the stages of the disease and suggest needed treatments for the patients in order to save their lives (Dibaba Adeba et.al 2022). A predictive model to forecast Chronic Kidney diseases, assessing the risk of kidney failure and the need for dialysis or transplant. Three machine learning classifiers (Decision Tree, Logistic Regression and Support Vector Machine) were used to evaluate the model, with the decision tree classifier outperforming the others. The addition of a bagging ensemble method enhanced accuracy up to 97.23%, offering potential benefits for early disease diagnosis and treatment (Saurabh Pal, 2023). The in-depth scholarly review articles explore the care and management of persons with early kidney disease and kidney failure, as well as those at risk for kidney disease. Emphasis is on articles related to the early identification of kidney disease; prevention or delay in progression of kidney disease.

3. Problem Definition:

The full spectrum of basic science through clinical care is covered in these reviews. Clinical care issues stress the multidisciplinary team approach to the care of kidney patients. Chronic renal failure develops as a result of progressive, frequently permanent kidney function impairment. Early detection of CKD plays vital role in providing proper treatment to the patients. With advantage of digital technology, so many methods are proposed using AI/ML techniques, but they too have certain limitations in accurately predicting the CKD.

3.1. Existing System:

Chronic kidney disease (CKD) is defined as kidney damage or an estimated glomerular filtration rate (eGFR) less than 60 ml/min/1.73 $m^{2,}$ persisting for three months or more, irrespective of the cause. The CKD has different stages, i.e., Stage 1 represents normal kidney function, Stage 2 represents mildly reduced kidney function, Stage 3A represents moderately reduced kidney function, Stage 3B represents moderately reduced kidney function, Stage 4 represents severely. CKD is usually asymptomatic till stages IV and V. The main draw backs in present study are prediction of CKD with medical records is high complex and time consuming process.

4. Proposed System:

Machine learning in health care relies on the collection of patient data. Using systems and tools

designed to sort and categorize data, machine learning algorithms can discover patterns in datasets that allow medical professionals to identify new diseases and predict treatment outcomes. Various machine learning algorithms such as Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Preceptron Neural Network (ANN) were used in this proposed study.

5. Data Set:

We utilize the kidney disease dataset for machine-learning predictions of chronic renal disease. This dataset is made taken from the UCI network through the UC Irvine Machine Learning storehouse. There are 400 occurrences in this dataset with 24 features (ID, AGE, BP, SG, AL, SU, RBC, PC, PCC, BA, BGR, BU, SC, SOD, POT, HEMO, PCV, WC, RC, HTN, CAD, APPET, PE, ANE), 1 target or class variable (Classification). The sample data in chosen dataset shown in following table.

id	age	bp	sg	al	su	rbc	pc	pcc
24	42	100	1.015	4	0	Normal	Abnormal	Not Present
125	72	90	NaN	NaN	NaN	NaN	NaN	Not Present
155	50	70	1.02	3	0	Abnormal	Normal	Present
159	59	80	1.01	1	0	Abnormal	Normal	Not Present
43	35	80	1.01	1	0	Abnormal	NaN	Not Present
...	...	...	...	...	...	...	...	...
193	32	90	1.025	1	0	Abnormal	Abnormal	Not Present
90	63	100	1.01	2	2	Normal	Normal	Not Present
374	79	80	1.025	0	0	Normal	Normal	Not Present

...	rc	htn	cad	appet	pe	ane	Classification
...	4.6	Yes	No	Poor	No	No	ckd
...	NaN	Yes	No	Poor	No	No	ckd
...	NaN	No	No	Good	No	No	ckd
...	4.3	No	No	Poor	No	No	ckd
...	3.1	No	No	Good	No	No	ckd
...	...	...	...	...	...	...	...
...	2.8	Yes	No	Poor	Yes	Yes	ckd
...	4.2	Yes	Yes	Good	No	No	ckd
...	6.4	No	No	Good	No	No	notckd

Table 1: Sample Data Elements in Chosen Dataset

6. Methodology:

The proposed methodology where we are going to use different machine learning algorithms to predict the CKD based on the chosen dataset was explained as follows. The architecture of the

6.1 Architecture of Proposed Study:

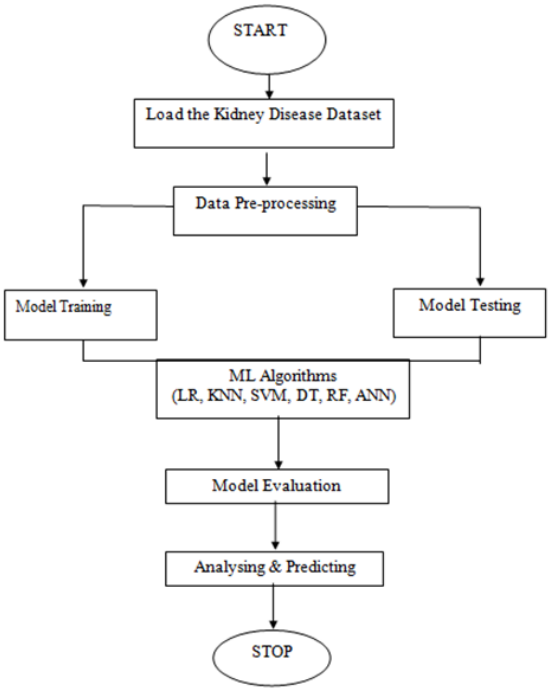

Figure 2: Architecture of the proposed study

6.2. Steps Involved in Proposed Study:

- Step 1: Data collection and preprocessing: Gather a relevant dataset that contains features (such as age, sex, blood pressure, serum creatinine levels, etc.) that may be indicative of chronic kidney disease.
- Step 2: Feature selection: Select a subset of features that are most relevant for predicting chronic kidney disease.
- Step 3: Split the data into training and testing sets: Randomly split the data into two sets: a training set, which will be used to train the machine learning models, and a testing set, which will be used to evaluate the performance of the models.
- Step 4: Train the machine learning models: Use various machine learning algorithms such as Logistic Regression, KNN, SVM, Decision Tree, Random Forest, and Perceptron Neural Networks to train models on the training set.
- Step 5: Evaluate model performance: Use the testing set to evaluate the performance of each model. Metrics such as accuracy, precision, recall, F1-score, and AUC-ROC can be used to evaluate model performance.
- Step 6: Select the best model: Compare the performance of each model and select the one with the highest performance.
- Step 7: Deploy the model: Once the best model has been selected, it can be deployed and used to make predictions on new data.
- Once the best model has been selected, it can be deployed and used to make predictions on new data.

6.3. Machine Learning Models used in Proposed Study:

In this proposed study, we used five machine learning algorithms to predict the Chronic Renal Disease for the given medical record(s) of a respective person. The algorithms are

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machines
- Decision Tree
- Random Forest
- Neural Network

Logistic Regression:

Logistic regression is a popular statistical model used for binary classification problems, where the goal is to predict the probability of an event occurring based on input variables. It is a popular and widely used algorithm in machine learning and statistics. to transform a linear combination of input features into a probability value between 0 and 1.

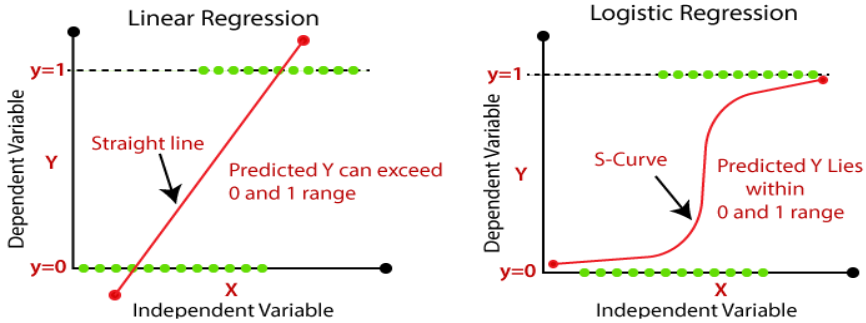

Figure 3: Regression Models

K-Nearest Neighbors:

KNN algorithm is a supervised learning algorithm. It can use for both classification and regression techniques to assign weights.

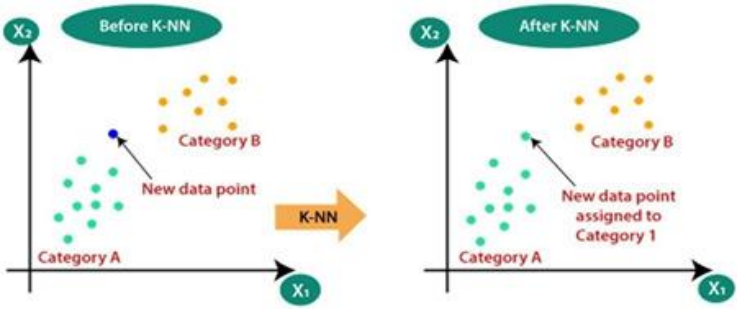

Figure 4: KNN Model

Here, we can see that in this fig the blue point is a new data point. Category-A and category-B are neighbors. This algorithm calculates the nearest distance the new data point is assigned to the category-A, because the distance is very less comparing to category-B.

The K-NN working can be explained on the basis of the below algorithm:

- Step-1: Select the number K of the neighbors
- Step-2: Calculate the Euclidean distance of K number of neighbors
- Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
- Step-4: Among these k neighbors, count the number of the data points in each category.
- Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
- Step-6: Our model is ready

Support Vector Machines:

As Chronic Kidney Disease progresses slowly, early detection and effective treatment are the only cure to reduce the mortality rate. Machine learning techniques are gaining significance in medical diagnosis because of their classification ability with high accuracy rates. The accuracy of classification algorithms depend on the use of correct feature selection algorithms to reduce the dimension of datasets. In this study, Support Vector Machine classification algorithm was used to diagnose Chronic Kidney Disease. This hyper plane margin is Soft Margin.

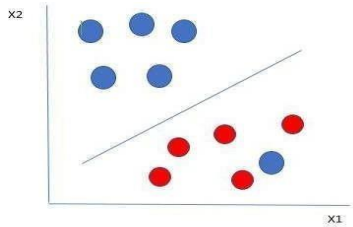

Figure 5: SVM Model

The SVM model working can be explained on the basis of the below algorithm:

- Step-1: AttributeSupportVector(ASV) = {Closest Attribute Pair from Opposite Classes}
- Step-2: while margin constraint violating points exist do
- Step-3: Find the Violator
- Step-4: ASV = ASV U Violater
- Step-5: if any $a_p < 0$ because of addition of c to S then
- Step-6: ASV = (ASV/p)
- Step-7: Repeat all the violating points are pruned
- Step-8: end if
- Step-9: end while

Decision Tree:

A decision tree is a decision support hierarchical model that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

- Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
- Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM)
- Step-3: Divide the S into subsets that contains possible values for the best attributes.
- Step-4: Generate the decision tree node, which contains the best attribute.
- Step-5: Recursively make new decision trees using the subsets of the dataset created in
- Step-6: Continue this process until a stage is reached where you cannot further classify the node / nodes and called the final node as a leaf node.

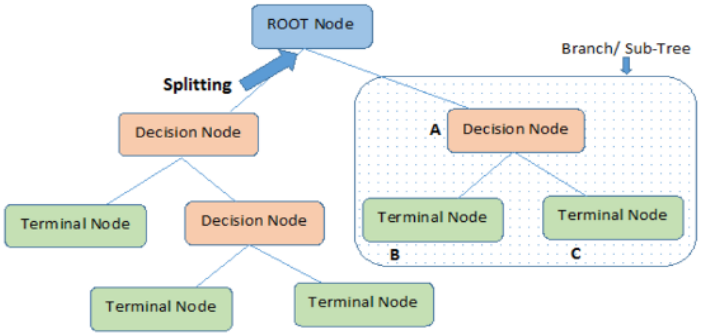

Figure 6: Decision Tree

Random Forest:

Random Forest works in two-phase. *First is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.*

- **Bagging:** It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.
- **Boosting:** It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

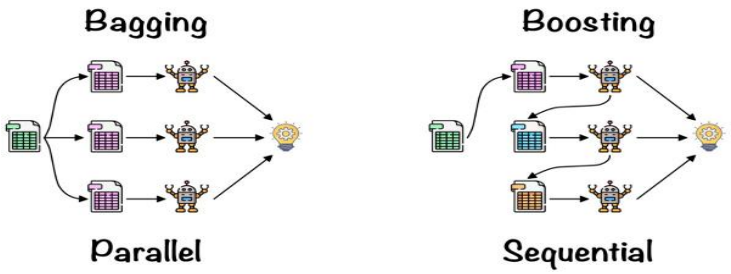

Figure 7: Random Forest

The Working process can be explained in the below steps and diagram:

- Step-1: Select random K-data points from the training set.
- Step-2: Build the decision trees associated with the selected data points (Subsets).
- Step-3: Choose the number N for the decision trees that we want to build.
- Step-4: RepeatSteps1&2.
- Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

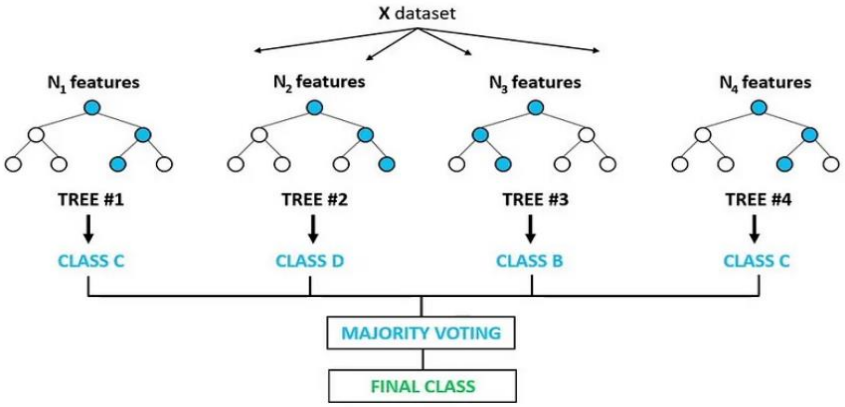

Figure 8: Random Forest

Neural Network:

Artificial intelligence includes artificial neural networks. This kind of machine learning is supervised. Similar to the human brain, it has the same structure. Like human neurons, ANN neurons are linked to other ANN neurons and layers of the network, exactly as human neurons are. ANNs are capable of solving issues that, by human or statistical criteria, have never been achievable. ANNs are easy to handle, produces accurate results in with minimal time complexity.

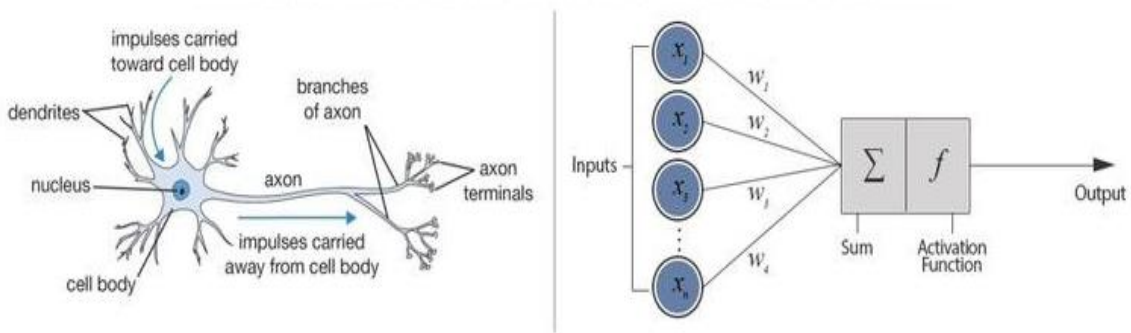

Figure 9: Biological Neuron and Artificial Neural Network

Algorithm 1 Perceptron Algorithm

Initialize $\theta = 0$.
while there exist a misclassified feature vector **do**
  for $i = 1, \ldots, n$ **do**
    if $y_i \times (\theta \cdot x_i) \leq 0$ **then** ▷ This is true when x_i is misclassified by θ
      $\theta = \theta + y_i \times x_i$ ▷ θ moves in the "right direction"
    end if
  end for
end while

System Requirements:

- Hardware: To implement the proposed, we used a windows operating system based desktop with RAM:8GB, HDD:500GB, Processor: Intel Core i5-2400S 2.5 GHz.
- Software: Anaconda - Jupyter notebook with Python 3.9 or higher and other required machine learning packages in python.

7. Results and Discussion:

7.1 Presentation of Results:

Our chosen dataset in this proposed study was using most of the researchers to implement their machine learning classification models or predictive models. To assess the predictive capabilities of the chosen machine learning techniques in proposed study, we split the dataset into a training set (70%) and testing dataset (30%).

7.2 Evaluation of Proposed Methodology:

To evaluate our proposed methodology in prediction of CKD for given input diagnostics data, we use different performance measures like accuracy, precision, recall, F1-score, specificity and Area Under Curve (AUC) by the confusion matrix.

Confusion Matrix Description:

A confusion matrix is a performance measurement tool in machine learning. It's a 2X2 matrix that allows visualization of the performance of an algorithm, typically a classifier. It used to understand how well the model is performing in terms of correctly and incorrectly classified instances. Performance matrix consists of four sections: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy:

Accuracy is the ratio of correctly predicted observations to the total observations. It measures the overall correctness of the model and is calculated as:

Accuracy = Number of Correct Predictions / Total Number of Predictions

Precision:

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It measures the correctness of positive predictions and is calculated as:

Precision = True Positives / (True Positives + False Positives)

Recall (Sensitivity): Recall, also known as Sensitivity, is the ratio of correctly predicted positive observations to the all observations in actual class. It measures the ability of the model to capture all the positive instances and is calculated as:

Recall = True Positives / (True Positives + False Negatives)

F1 Score:

The F1 Score is the harmonic mean of Precision and Recall. It combines both Precision and Recall into a single metric and is calculated as:

F1 Score = 2 × (Precision × Recall) / (Precision + Recall)

Specificity:

It also known as True Negative Rate, is a performance metric in binary classification tasks that measures the proportion of actual negative instances that are correctly identified as negative by the model. It is calculated as:

Specificity = True Negatives / (True Negatives + False Positives)

AUC (Area Under the ROC Curve):

AUC represents the area under the receiver operating characteristic (ROC) curve. It quantifies the ability of the model to discriminate between positive and negative classes across various threshold settings. A higher AUC indicates better model performance, with 1 being the perfect classifier and 0.5 representing a random classifier.

7.3 Comparative Analysis:

On our examining the performance of the considered machine learning algorithms, it produced the following results,

	Accuracy	FI-Score	Precision	Recall	Specificity	AUC
LR	93	95	99	92	97	96
KNN	90	92	100	86	100	98
SVM	97	98	100	96	100	98
DT	98	99	99	99	97	98
RF	100	100	100	100	100	100
ANN	73	81	80	82	54	100

Table 2: Comparison of results of ML algorithms in proposed study

With comparative analysis of the above performance measures of each machine learning model/ algorithm, it is found that, Radom Forrest shown high score for all considered metrics and is best for prediction of chronic renal disease.

8. Conclusion:

For early detection of CKD, proposed a study with six different machine learning algorithms. After analyzing the results of proposed study, it was found that Random Forest is the most accurate in predicting chronic renal disease among all other chosen machine learning algorithms. Based on the outcomes of this study, it is possible to infer that each machine learning algorithm utilized in this study has its own unique set of strengths and shortcomings in predicting chronic renal disease.

9. References:

1. L. Zhang et al., “Prevalence of chronic kidney disease in China: a cross-sectional survey,” Lancet, vol. 379, pp. 815-822, Aug. 2012, https://pubmed.ncbi.nlm.nih.gov/22386035/
2. Singh et al., “Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration,” J. Biomed. Inform, vol.53, pp. 220-228, Feb. 2015. https://pubmed.ncbi.nlm.nih.gov/25460205/
3. Westing, A. & Küpers, L. & Geleijnse, J. Diet and Kidney Function: a Literature Review. Current Hypertension Reports (2020) 22. https://doi.org/10.1007/s11906-020-1020-1
4. Z. Chen et al., “Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers,” Chemometrics. Intell. Lab., vol. 153, pp. 140-145, Apr.2016.
5. Sujata Drall et al., Chronic Kidney Disease Prediction Using Machine Learning: A New Approach, Volume 8, Issue, V, May 2018.
6. Ilyas, Hamida & Ali, Sajid & Ponum, Mahvish & Hasan, Osman & Mahmood, Muhammad & Iftikhar, Mehwish & Malik, Mubasher. (2021). Chronic kidney disease diagnosis using decision tree algorithms. BMC Nephrology. 22. 10.1186/s12882-021-02474-z.
7. Dibaba Adeba Debal1 and Tilahun Melak Sito, et.al Chronic kidney disease prediction using machine learning techniques, https://doi.org/10.1186/s40537-022-00657-5
8. Saurabh Pal, et.al Chronic Kidney Disease Prediction Using Machine Learning Techniques, Biomedical Materials & Devices (2023) 1:534–540, https://doi.org/10.1007/s44174-022-00027-y