

# Chronic Kidney Disease Prediction Using Machine Learning

O.Nikhilesh Reddy<sup>1</sup>, K Sai Gowtham<sup>2</sup>, Shaik Abdul Sami<sup>3</sup>, Dvm Karthik<sup>4</sup>

Integrated Mtech Software Engineering, Vit

## Abstract:

In today's era everyone is trying to be conscious about health although due to workload and busy schedule one gives attention to health when it shows any symptoms of some kind. However, CKD is a disease that either exhibits no symptoms at all or exhibits no signs that are particular to the condition, making it difficult to forecast, identify, and prevent such a disease and this could be led to permanently health damage, but machine learning can be hope in this problem it is best in prediction and analysis. We will employ several machine learning approaches, such as Decision Tree, KNN, Random Forest, SVM, Naive Bayes, using data that contains 24 health related attributes like age, blood pressure, sugar, glucose, taken in 2-month period of 400 patients in which 11 numeric and 14 nominal attributes in which it consists of class label named 'Class' which classifies patients having disease and not present. To build a model with maximum accuracy of predicting whether CKD or not and if yes then its Severity.

**Keywords:** Chronic, preprocessing, Accuracy, Feature selection ,Decision Tree, KNN, SVM, Naïve-Bayes, Random Forest, Cross validation, Early Detection, Model Evaluation.

## I. INTRODUCTION

Chronic kidney disease is a dangerous disease of the kidney which produces gradual loss in kidney functionality. In chronic kidney disease (CKD), kidney function gradually declines over several years. If CKD is not detected and cured in early stage then patient can show following Symptoms: Blood Pressure, anemia, weak bones, poor nutrition health and nerve damage, decreased immune response because at advanced stages dangerous levels of fluids, electrolytes, and wastes can build up in your blood and body. Therefore, it is crucial to identify CKD at an early stage. Machine learning accomplishes this by training a predictive model using historical CKD patient data. It can be calculated from the results of your blood creatinine, age, race, gender, and other factors. The earlier a disease is detected the better the chance of showing or stopping its progression.

## EXISTING SYSTEM:

The existing systems used algorithms like logistic regression, KNN, random forest, decision tree, SVM, neural networks. Some of the existing systems were developed based on the ant colony optimization. ACO is used for feature selection and some existing systems use gradient descent for the optimization purposes. The existing systems takes metrics like precision, recall, F1-Score. Based on these metrics the existing systems are evaluated. All existing systems use the dataset and pre-process the dataset to fill the missing values and remove noisy data. Then dataset is split into parts for training and testing. Then the dataset is fed into the algorithms to train. Later the algorithm performance is tested and evaluated.

**GAPS IDENTIFIED:**

- **High computational cost:** Computational cost refers to the amount of resources the neural network uses in training or inference. The models are using too much of resources. The resources usage needs to be minimized.
- **Small dataset:** The dataset used is too small. Small dataset can lead to many issues like outliers, overfitting and model may be biased.
- **Overfitting:** this can lead our model to give accurate predictions for training data but not for new data.
- **No Optimization:** Getting the better results is important.
- **Missing values:** Missing values can reduce the accuracy of the model if they present in the data.

**PROPOSED METHOD:**

In our project we are going to apply machine learning techniques on chronic kidney disease. The machine learning techniques we are going to apply are Decision Tree, KNN, Random Forest, SVM, Naive Bayes. In first step we will collect the data and then the pre-processing will take place to remove noisy data. Later we will create each model mentioned above. For optimization we will use keras tuner. Keras tuner will help to increase the accuracy. Main aim of introducing kerasTuner is to find the best suitable hyperparameter values. Here the hyperparameters are no. of neurons in each layer, learning rate and number of clusters. It will go through different values of hyperparameters and finds the best suitable hyperparameters for our model by using different search algorithms. After optimizing we will evaluate to find the best model. For implementation we will use jupyter notebook software tool and coded in python language.

**II.LITERATURE SURVEY AND EXISTING METHODOLOGIES**

1. This paper presents the idea of detecting the presence of kidney disease through machine learning based classification modelling by processing the patient's ECG signal. In this model, digitized ECG data is collected from the Physio net database ([www.physionet.org](http://www.physionet.org)) from open access databases such as PTB (for kidney patients) and Fantasia (for healthy people), and the model will be the same later. Validated using various data from online-database. The model succeeded in classifying users as healthy or kidney disease, and the validation process yielded satisfactory results. In this study, they found an accuracy of 97.6%. This was best when using both the QT and RR interval functions compared to using either function.
2. In this paper the researchers discussed about how to apply machine learning algorithms to predict the chronic kidney disease. They tried to predict this using the deep artificial neural networks. In this study they have used the dataset from UCI machine learning repository. The dataset they have used is collected from the apollo hospital Tamilnadu for nearly two months. In their model the sequence of layers are as follows at first input layer is present and next the output layer is present between both the layer there are hidden layers. For optimization purposes they used the stochastic gradient descent. They have utilised adam (adaptive moment estimation), which determines the weight that needs to be adjusted for various factors or independent adaptive learning rates. Finally, they used K-fold cross-validation. Finally researchers concluded that the network they have built gave the highest precision, recall, true positive compared to the other implementations.
3. The main aim of this paper is to examine the accuracy and the performance of the algorithms like naïve bayes, KNN, random forest. The dataset used in this paper is KFT dataset, four hundred instances and

- 25 entities. The researchers developed the algorithms in c sharp language. The metrics used by the researchers to compare are accuracy, precision, recall and f-measure. The precision for the naïve bayes is 1. Random Forest finished better in phrases of accuracy, and f degree over distinctive datasets, whereas Naive Bayes, shows better Precision. Finally, researchers concluded that the random forest is better than naïve bayes and KNN.
4. In this paper the researchers suggested different algorithms like probabilistic neural network, multilayer perceptron, support vector machine and radial basis function. These are the methods the researchers had implemented in the paper. The researchers also implemented the data mining techniques in this paper which helped them to find the interesting patterns in the data. The dataset used is UCI dataset with 361 CKD Indian patients data. By using the algorithms mentioned the researchers found that the PNN gave highest accuracy. The algorithm is much rapid than multilayer perceptron networks. It can be more exact than multilayer perceptron networks. The algorithm moderated than multilayer perceptron networks at classifying new cases. PNN needs a higher memory space to store the data.
  5. The main motivation for this paper is to determine the presence of chronic kidney disease by introducing various classification algorithms into the patient's medical records. This study focuses primarily on finding the most appropriate classification algorithm that can be used to diagnose CKD, based on classification reports and performance factors. Empirical work is performed on a variety of algorithms such as support vector machines, random forests, Boost, logistic regression, neural networks, and naive Bayes classifiers. Experimental results show that Random Forest and XGBoost perform better than other classification algorithms and produce an accuracy of 99.29.
  6. This paper tries to predict the chronic kidney disease by using ML procedures to analyse CKD at a beginning phase. Kidney infections are messes that upset the typical capability of the kidney. As the level of patients impacted by CKD is fundamentally expanding, successful expectation methodology ought to be thought of. In this paper, researchers applied various machine learning algorithms on a dataset of 400 patients and 24 attributes connected with determination of kidney disease. The machine learning algorithms used in this paper are ANN and SVM. tuning of the parameters also takes place to improve the accuracy. The exact outcomes from the trials showed that ANN performed better compared to SVM, with accuracies of 99.75% and 97.75%, respectively.
  7. The accuracy of the classification algorithm depends on using the correct feature selection algorithm to reduce the dimensions of the dataset. In this study, a support vector machine classification algorithm was used to diagnose chronic kidney disease. To diagnose chronic kidney disease, researchers chose two major types of feature selection methods: wrapper and filter approaches to reduce the dimension of the chronic kidney disease dataset. The wrapper approach used a classifier subset evaluator with a greedy step-by-step search engine and a wrapper subset evaluator with a best-first search engine. The filtering approach used a subset evaluator to select correlation features using a greedy step-by-step search engine, and a filtered subset evaluator using a best-first search engine. The results show that the support vector machine classifier, which uses a subset evaluator filtered by the search engine's best-first feature selection method, has a higher accuracy rate in diagnosing chronic kidney disease compared to other selected methods.
  8. This research aims at how well machine learning algorithms can identify CKD while taking into account the fewest tests or features possible. The dataset is based on CKD patients obtained from Apollo Hospital in India in 2015 over a two-month period. Overall Researchers have found that just

three indicators are sufficient to identify CKD. Using random forest (RF) and Gradient boosting (GB) models, we also discovered that haemoglobin contributes most to the detection of CKD while albumin contributes least. They came to this conclusion by using 10-fold cross-validation, the classifiers have been trained, examined, and validated. The gradient boosting approach improved performance in terms of F1-measure (99.1%), sensitivity (98.8%), and specificity (99.3%). These results are among the best when compared to earlier studies, although with fewer features reached thus far.

9. The paper "Chronic kidney disease prediction using data mining" by Snegha, et al. (2020) discusses the use of data mining techniques for the prediction of chronic kidney disease. The authors used a dataset of patient information and applied various data mining techniques such as decision tree, random forest, and neural networks to predict the likelihood of a patient having chronic kidney disease. The dataset used in this paper is from Kaggle with 24 attributes and 400 records. They found that back propagation neural network had the highest accuracy in predicting the disease.
10. This article has the use of mechanical learning (ML) prediction models in diagnosis of chronic diseases. This analysis included various articles published between 2015 and 2019. Finally, 22 studies were selected to present all modeling methods in a concise manner that explain CD diagnosis and usage models of individual medical conditions, along with related strengths and limitations. Because each method has its strengths and weaknesses, their results suggest that there is no standard way to determine the best approach in real-time clinical practice. Of the methods considered, support vector machine (SVM), logistic regression (LR), and clustering were the most commonly used. These models are highly applicable for CD classification and diagnosis and are expected to become more important in the medical field in the near future.
11. The novelty of the current research work lies in the assumption that demonstrates the no crisp Rough K Means (RKM) clustering for figuring out the ambiguity in chronic disease datasets to improve the performance of the system. The RKM algorithm has clustered data into two sets, namely, the upper approximation and the lower approximation. The objects belonging to the upper approximation are favorable objects, whereas the ones belonging to the lower approximation are excluded and identified as ambiguous. These objects have been excluded to improve the algorithm. The machine learning algorithms, namely, naïve Bayes (NB), support vector machine (SVM), K-nearest neighbors (KNN), and random forest tree, are presented and compared. Chronic disease data is obtained from machine learning repositories and Kaggle to test and evaluate the proposed model. Test results show that the proposed system has been successfully used in the diagnosis of chronic diseases. The proposed model achieved the best results using Naive Bayes with AFM in the classification of diabetic disease (80.55%).
12. This paper discussed about the chronic kidney disease prediction using machine learning algorithms. In this paper the algorithms they used are SVM and ant colony optimization. The authors even tried to minimize the features and selecting best features to improve the accuracy of prediction. The dataset used in this paper is the dataset which is available from the UCI repository. This dataset contains 400 samples of two different classes. ACO is used in this paper to select the features. ACO suggests 12 best attributes for prediction. This helps to improve the accuracy. The metrics used to evaluate the paper are precision, recall, f1-score, support. The authors obtained 96 percent of accuracy.
13. Ekanayake and Herath proposed a workflow to predict CKD. They dealt with aspects of data collection and highlighted the significance of domain knowledge while dealing with machine learning. In this paper they trained 11 models to find out the best performed. The algorithms include decision tree

- classifier, random forest classifier, xgb classifier, extra tree classifier, ada boost classifier, knn, classical neural network, svc linear, logistic regression, svc, gaussian nb. They used dataset from the UCI repository which is considered as a good dataset. They found out that the best performing algorithm based on accuracy was extra tree classifier and random forest classifier.
14. In research paper, Authors presents the development of an ensemble approach to diagnose chronic kidney disease (CKD) using machine learning techniques. The ensemble approach combines multiple machine learning models to improve the diagnostic performance. The dataset used in this ensemble approach is a Chronic Kidney Disease dataset Collected from UCI repository. The researchers collected data from 100 patients with CKD and 100 healthy controls to train and test the ensemble model. They used a combination of four models: Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), and Random Forest (RF). The ensemble approach achieved an accuracy of 96.5%, sensitivity of 96.0% and specificity of 97.0%. Finally Researchers concluded that the ensemble approach is a promising method for the diagnosis of CKD and it has the potential to improve the diagnostic performance when compared to individual models.
  15. This paper tried to implement the four reliable approaches, the four approaches are support vector machine, AdaBoost, linear discriminant analysis and gradient boosting. They implemented this algorithms to predict the chronic kidney disease which is a slow and lately diagnosed disease. The algorithms are trained and tested using the online dataset gathered from the UCI machine learning repository. To make this study more accurate the researchers used multiple metrics like false negative rate, accuracy, precision, negative predicted value, f1 score, false discovery rate and etc. Missing value issues have been rectified using the imputation method of KNN and standard scaler technique have been used to make the values range between 0 to 1. This imputation have helped to gain more accurate results. Finally researchers conclude that GB classifier is the one which gives better results.
  16. Researchers primary goal is to compare the effectiveness of various machine learning algorithms based on their accuracy. To compare their performances, they idolised Rcode in this work. This study's primary goal is to classify cases of CKD and non-CKD using an analysis of the chronic kidney disease dataset. The methods used in this paper are KNN, SVM, Logistic regression. The morality of every algorithm is inspected. They divided the dataset in 30% and 70%. Results from each classifier were assessed using several assessment criteria, and 10-fold cross-validation was used to check for overfitting. Nested cross validation is a method that has been useful in optimising the model's parameters. Finally, the conclusion obtained is within the constrained parameters of this medical scenario, the Support Vector Machine algorithm diagnoses chronic kidney disease more accurately than Logistic Regression and K-Nearest Neighbours.
  17. In this study, researchers predicted the risk factors for chronic kidney disease (CKD) and the progression of CKD. The data set used in the research is a CKD dataset collected from the Nephrology Department of a hospital in Bangladesh. They have applied six machine learning algorithms. The dataset is processed through six different algorithms, and the best results were obtained through the classification of risk factors. Out of those six They are getting high accuracy (98.8858%) with the Random Forest method and it is best fit for the dataset. In comparison to previous models, Researchers stated that their approach model is better at identifying the CKD risk factors that are most significant and non-significant.
  18. The literature review in this paper focuses on the use of machine learning approaches for the performance analysis of Chronic Kidney Disease (CKD). The authors cite previous research that has

employed various machine learning techniques, such as decision tree, artificial neural network, and support vector machine, for the prediction and diagnosis of CKD. They also note that some studies have used ensemble methods to improve the accuracy of these models. The authors also mention that the use of feature selection techniques such as correlation-based feature selection and principal component analysis have been used to improve the performance of the machine learning models. The paper aims to analyze the performance of different machine learning models for CKD diagnosis and prediction.

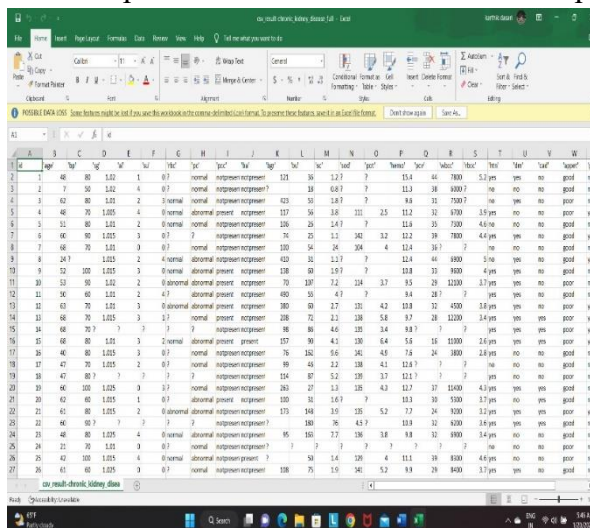
19. In this paper the authors used ML approaches to predict and diagnose chronic kidney disease. They have used a dataset from the University of California, Irvine (UCI) repository that contained information on 400 people for this. To maximise the number of features affecting chronic kidney disease, several feature selection strategies have been employed. These desirable traits are then picked and applied to several ML models, and their accuracy and sensitivity are contrasted. Logistic regression, Naive Bayes, KNN, SVM, Decision Trees, Random Forest Classifier, and Extra Trees Classifier are just a few of the machine learning algorithms that have been investigated. The Extra Trees Classifier model, with its best accuracy of 99.36% and one of the quickest execution durations, was found to produce six ideal features from Decision Trees using information gain. In the future, this suggested model may also be used to study several other illnesses. The illness detection method will become less invasive and more economical as a result.
20. This study focused on the evaluation of records collected by 400 patients, including 24 functions. Medical analysis methods and modes were used to replace missing numerical values and nominal values. Recursive function removal (RFE) was used to select the most important functionality. In this study, the use algorithm for four classification algorithms was a support vector machine (SVM), hybrid (KNN), decision tree and random forest. All classification algorithms have achieved promising performance. The Random Forest algorithm outperformed all other algorithms applied and achieved 100% accuracy, precision, recall, and F1 score on all measurements.
21. The study provides a methodology for predicting CKD status using clinical data that includes feature extraction, data aggregation, data preprocessing, and a strategy for handling missing values. The dataset used in this paper is CKD dataset from Kaggle . They have used three machine learning techniques Decision Tree (DT), Logistic Regression(LR) and K-Nearest Neighbor (KNN) with 96.25%, 97% and 73% accuracies respectively. The accuracy percentage of the models employed in this research is significantly higher than that of earlier studies, showing that the models used in this study are more trustworthy than those used in earlier studies.
22. The paper "Implementation of Machine Learning Models for the Prevention of Kidney Diseases (CKD) or Their Derivatives" by Twarish Alhamazani et al. (2021) reviews the use of machine learning models in predicting and preventing kidney diseases (CKD) or their derivatives. In this paper the researchers took the crisp dm model as a reference. The data was pre-processed in the azure cloud platform. The data was balanced using the smote technique. The machine learning algorithms those were implemented are logistic regression, decision forest, neural network and jungle of decisions. The dataset is gathered from the clinic in Iraq. No significant data sample due to the restrictions of medical data and its legal effects. Finally researchers concluded that with a score of 92%, the decision forest outperformed the other machine learning models.
23. Here In this paper the authors have proposed a method for predicting kidney disease in real time. The objective is to identify the most effective machine learning (ML) application that can accurately

identify and forecast the state of chronic renal disease. They made advantage of the machine learning repository data from UCI. In this study, five key machine learning classification techniques—KNN, Logistic Regression, Random Forest Classifier, SVM, and Decision Tree Classifier—were considered for the prediction of chronic kidney disease. The data has been split into two parts for this operation. A part with a trained dataset and another with a test dataset were examined. The analysis's findings demonstrate that Decision Tree Classifier and Logistic Regression algorithms outperformed the competition, with an accuracy of 98.75%.

24. In order to predict chronic kidney disease (CKD), this study examines various machine learning techniques, in particular classification and association algorithms. The study examines the outcomes of combining feature selection and classification algorithms. Weka's classification methods are utilised to benchmark the CKD dataset. Tenfold cross-validation with and without the feature selection technique is used to calculate the results. Instances that were successfully classified, the kappa statistic, and the mean absolute value with and without the feature selection technique are compared as outcomes. The Apriori association technique is used in the preparation of the benchmark dataset. The WEKA models ZeroR, OneR, J48, IBk, and naive Bayes are further tested using the recovered data. The results show that IBk with the Apriori associative algorithm may produce the best outcomes with 99% accuracy.
25. This research provides hybrid machine learning strategies for detecting chronic kidney disease that incorporate feature selection methods and machine learning classification algorithms based on big data platforms (Apache Spark) (CKD). The Relief-F and chi-squared feature selection techniques were employed to identify the key features. They have used the CKD dataset from the UCI machine learning repository, and dataset was separated by 80% of the training set and 20% of the testing set. In this study, Using PySpark six machine learning classification techniques were used. With the chosen features, the Support Vector Machine (SVM), Decision tree (DT), and Gradient Boosted Trees (GBT Classifier) had the best performance at 100% accuracy, and Relief-F feature selection performs better than chisquare feature selection in terms of performance.
26. This study can help in our understanding of the reasons behind the variations in kidney disease around the world, as well as what we can do to address them and work together to attain equity in kidney health globally. A good feature-based prediction model for identifying renal disease is provided by this study. On a publicly accessible dataset of healthy and kidney disease patients, various prediction models were built and analysed using a variety of machine learning algorithms, such as the k-nearest neighbours' algorithm (KNN), artificial neural networks (ANN), support vector machines (SVM), naive bayes (NB), and others. RFE and Chi-Square test feature-selection techniques were also used. According to the experiments, a prediction model based on logistic regression that used the Chi-Square approach to choose the best features achieved a 98.75 percent accuracy rate. [27] This study sought to determine whether machine learning (ML) could accurately predict the likelihood that patients with chronic renal disease would develop end-stage kidney disease (CKD). Five ML algorithms were trained and evaluated using fivefold cross-validation: logistic regression, naive Bayes, random forest, decision tree, and Knearest neighbours. Each model's performance was evaluated against that of the Kidney Failure Risk Equation (KFRE). 748 people with CKD were included in the sample. In comparison to the KFRE, three ML models—logistic regression, naive Bayes, and random forest—showed equal predictability and higher sensitivity. This research showed the efficacy of using ML to assess the prediction of CKD based on conveniently located features. A potential utility for patient

screens is suggested by three ML models with sufficient performance and sensitivity ratings. The study's findings point to the viability of ML models in carrying out this crucial clinical task as well as their potential.

27. The purpose of this study was to solve two issues: first to choose the most informative features to help with the accurate identification of CKD. The second objective was to create an efficient, cost-sensitive AdaBoost classifier that correctly identified samples in the minority class. They have used CKD dataset prepared in 2015 by Apollo Hospitals, Tamil Nadu, India. In order to improve the identification of CKD, Researchers proposed a method that integrates information-gain based feature selection with a cost-sensitive AdaBoost classifier. For the performance comparison, They have implemented six machine learning classifiers. Among these classifiers the proposed cost-sensitive AdaBoost had 99.8% accuracy, 100% sensitivity, and 99.8% specificity.
28. The paper "Machine learning techniques for chronic kidney disease risk prediction" by Dritsas and Trigka (2022) reviews various machine learning techniques for predicting the risk of chronic kidney disease (CKD). They discuss the use of linear and non-linear models, such as logistic regression and random forests, as well as deep learning techniques, such as convolutional neural networks. The authors evaluate the performance of these models using various metrics such as accuracy, precision, recall and F1 score, and find that the deep learning models outperform the traditional models. However, the authors also note that these models have a higher computational cost and may require more data for accurate prediction. They also mention that the interpretability of deep learning models is a concern and it's hard to know why the model is making certain predictions. They also mention that further research is needed to improve the performance of these models in real-world settings. The authors also mention that the proposed models can be used as a tool for early diagnosis and prevention of CKD.
29. The main aim of the paper is to identify the chronic kidney disease in the early stages. For this the researchers adopted the prediction of the chronic kidney disease using the machine learning techniques. The models they have selected are random forest, support vector machine and decision tree. The feature selection is applied to select the best features in the dataset this helps the model to predict accurately. The feature selection is applied by using the analysis of variance and recursive feature elimination using cross validation. Dataset consists of 400 instances represented by 13 input features and 1 for the target class. The experiment findings showed that RF based on recursive feature reduction with cross validation outperforms SVM and DT in terms of performance.

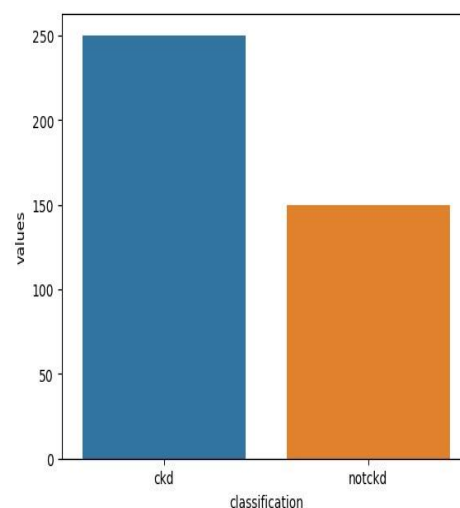


|    | A  | B    | C     | D     | E   | F      | G          | H          | I          | J          | K    | L    | M    | N     | O     | P    | Q     | R     | S    | T    | U    | V    | W    |    |  |  |
|----|----|------|-------|-------|-----|--------|------------|------------|------------|------------|------|------|------|-------|-------|------|-------|-------|------|------|------|------|------|----|--|--|
| 1  | 48 | 80   | 1.02  | 1     | 0.7 | normal | reciprocal | reciprocal | 121        | 36         | 1.27 | 7    | 15.4 | 44    | 7800  | 5.3  | yes   | no    | good | no   |      |      |      |    |  |  |
| 2  | 1  | 7    | 56    | 1.02  | 4   | 0.7    | normal     | reciprocal | reciprocal | 75         | 68   | 7    | 11.3 | 26    | 10077 | 16   | no    | no    | good | no   |      |      |      |    |  |  |
| 3  | 43 | 80   | 1.02  | 2     | 0.7 | normal | reciprocal | reciprocal | 423        | 30         | 1.67 | 7    | 8.8  | 31    | 7007  | 7    | yes   | no    | poor | no   |      |      |      |    |  |  |
| 4  | 4  | 48   | 70    | 1.025 | 4   | 0.7    | normal     | abnormal   | present    | reciprocal | 117  | 56   | 1.38 | 111   | 25    | 11.2 | 32    | 1700  | 3.9  | yes  | no   | poor | yes  |    |  |  |
| 5  | 5  | 51   | 85    | 1.01  | 2   | 0.7    | normal     | reciprocal | reciprocal | 106        | 25   | 1.47 | 7    | 11.8  | 25    | 7300 | 4.6   | no    | no   | good | no   |      |      |    |  |  |
| 7  | 8  | 80   | 80    | 1.025 | 3   | 0.7    | 7          | reciprocal | reciprocal | 76         | 25   | 1.5  | 142  | 32    | 11.2  | 38   | 7800  | 4.4   | yes  | no   | good | yes  |      |    |  |  |
| 8  | 7  | 48   | 70    | 1.01  | 3   | 0.7    | normal     | reciprocal | reciprocal | 100        | 34   | 1.6  | 26   | 4     | 10.4  | 18.7 | 7     | no    | no   | good | no   |      |      |    |  |  |
| 9  | 8  | 24.7 | 1.015 | 2     | 4   | normal | abnormal   | reciprocal | reciprocal | 423        | 31   | 1.27 | 7    | 12.4  | 44    | 6900 | 5     | no    | yes  | no   | good | yes  |      |    |  |  |
| 13 | 9  | 52   | 100   | 1.025 | 3   | 0.7    | normal     | abnormal   | present    | reciprocal | 138  | 68   | 1.67 | 7     | 13.8  | 33   | 9000  | 4.9   | yes  | no   | good | no   |      |    |  |  |
| 11 | 30 | 52   | 80    | 1.02  | 2   | 0.7    | abnormal   | abnormal   | present    | reciprocal | 75   | 107  | 7.2  | 124   | 3.7   | 5.5  | 19    | 12100 | 11.7 | yes  | no   | poor | no   |    |  |  |
| 12 | 11 | 30   | 80    | 1.01  | 2   | 0.7    | abnormal   | present    | reciprocal | 480        | 56   | 4.7  | 3    | 9.4   | 28.7  | 7    | yes   | no    | good | no   |      |      |      |    |  |  |
| 13 | 11 | 63   | 70    | 1.01  | 3   | 0.7    | abnormal   | abnormal   | present    | reciprocal | 380  | 68   | 2.7  | 131   | 4.2   | 10.8 | 32    | 4500  | 10.8 | yes  | no   | poor | yes  |    |  |  |
| 14 | 11 | 68   | 70    | 1.015 | 3   | 0.7    | normal     | present    | reciprocal | 208        | 72   | 2.3  | 138  | 5.8   | 9.7   | 18   | 12100 | 14.4  | yes  | yes  | poor | yes  |      |    |  |  |
| 15 | 14 | 68   | 70.7  | 7     | 7   | 7      | reciprocal | reciprocal | 95         | 86         | 4.6  | 105  | 2.4  | 8.8   | 7     | 7    | yes   | yes   | poor | no   |      |      |      |    |  |  |
| 16 | 15 | 68   | 80    | 1.01  | 3   | 2      | normal     | abnormal   | present    | present    | 157  | 90   | 4.1  | 130   | 6.4   | 5.5  | 15    | 10100 | 1.6  | yes  | yes  | poor | yes  |    |  |  |
| 17 | 16 | 40   | 80    | 1.015 | 3   | 0.7    | normal     | reciprocal | reciprocal | 76         | 162  | 5.6  | 141  | 4.5   | 7.8   | 24   | 3800  | 2.8   | yes  | no   | no   | good | no   |    |  |  |
| 18 | 17 | 47   | 70    | 1.015 | 2   | 0.7    | normal     | reciprocal | reciprocal | 99         | 46   | 2.2  | 138  | 4.1   | 118.7 | 7    | 7     | no    | no   | no   | good | no   |      |    |  |  |
| 19 | 18 | 47   | 80.9  | 7     | 7   | 7      | reciprocal | reciprocal | 144        | 87         | 5.2  | 136  | 17   | 112.7 | 7     | 7    | yes   | no    | poor | no   |      |      |      |    |  |  |
| 20 | 19 | 40   | 100   | 1.025 | 3   | 0.7    | normal     | reciprocal | reciprocal | 263        | 27   | 1.3  | 135  | 4.3   | 12.7  | 17   | 15400 | 4.1   | yes  | yes  | yes  | good | no   |    |  |  |
| 21 | 20 | 42   | 85    | 1.015 | 1   | 0.7    | abnormal   | present    | reciprocal | 100        | 31   | 1.67 | 7    | 18.3  | 10    | 1300 | 1.7   | yes   | no   | good | no   |      |      |    |  |  |
| 22 | 21 | 42   | 80    | 1.015 | 2   | 0.7    | abnormal   | abnormal   | reciprocal | 172        | 140  | 1.5  | 105  | 5.2   | 7.7   | 24   | 3200  | 11.7  | yes  | yes  | poor | no   |      |    |  |  |
| 23 | 22 | 40   | 60.7  | 7     | 7   | 7      | reciprocal | reciprocal | 180        | 26         | 4.5  | 7    | 10.9 | 12    | 4200  | 14.4 | yes   | no    | poor | no   |      |      |      |    |  |  |
| 24 | 23 | 48   | 80    | 1.025 | 4   | 0.7    | normal     | abnormal   | reciprocal | reciprocal | 95   | 161  | 7.7  | 136   | 3.8   | 8.8  | 32    | 6900  | 14.4 | yes  | no   | no   | good | no |  |  |
| 25 | 24 | 25   | 70    | 1.01  | 3   | 0.7    | normal     | reciprocal | reciprocal | 7          | 7    | 7    | 7    | 7     | 7     | 7    | 7     | no    | no   | no   | poor | no   |      |    |  |  |
| 26 | 25 | 42   | 100   | 1.025 | 4   | 0.7    | normal     | abnormal   | reciprocal | present    | 7    | 38   | 1.4  | 130   | 4     | 11.2 | 18    | 8200  | 14.4 | yes  | no   | poor | no   |    |  |  |
| 27 | 26 | 63   | 60    | 1.025 | 3   | 0.7    | normal     | reciprocal | reciprocal | 108        | 75   | 1.5  | 141  | 5.2   | 8.9   | 28   | 9400  | 11.7  | yes  | no   | good | no   |      |    |  |  |



Numerous studies and reports Using machine learning algorithms, various research investigations have been carried out to create predictive models for chronic kidney disease (CKD), a significant health concern that affects people all over the world. Several machine learning methods were employed in the study by Ekanayake and Hearth (2020) to predict CKD. Several early-stage CKD prediction models were contrasted by Rubini and Eswaran (2015). For CKD prediction, Gopika and Vanitha (2017) employed clustering approaches. ECG signals were used by Rahman et al. (2019) to identify CKD in its early stages. An intelligent diagnostic prediction and classification system for CKD was created by Elhoseny et al. in 2019. For CKD prediction, Raju et al. (2019) used data science. Machine learning models' use to the detection of chronic diseases was studied by Battineni et al. in 2020. Senan et al. (2021) employed recursive feature removal methods and classification algorithms for diagnosis of CKD. Soft clustering was employed by Al Dhyani et al. (2020) to improve the diagnosis of chronic disorders. Eventually, ensemble classifiers were used by Basar and Akan (2017) to detect CKD. Overall, these studies show that machine learning has a promise for predicting and diagnosing CKD, and other strategies are being investigated for increased performance and accuracy.

### III.PROBLEM STATEMENT



The number of patients with kidney disease continues to increase due to the consumption of junk food and lack of water in our body and many other reasons. Diagnosis of kidney infections can be very costly and dangerous if kidney tests are done frequently. For these reasons, many patients are neglecting treatment. Kidney disease is a major chronic disease associated with high blood pressure, diabetes and aging. The main function of the kidneys is to remove waste products and excess water from our body. In our project we are going to apply machine learning techniques on chronic kidney disease. Nowadays it leads to major issues with a steady growth rate. Machine learning will help to identify chronic kidney diseases in the early stage itself without leading to critical problems.

#### **Datasets Description and sample data:**

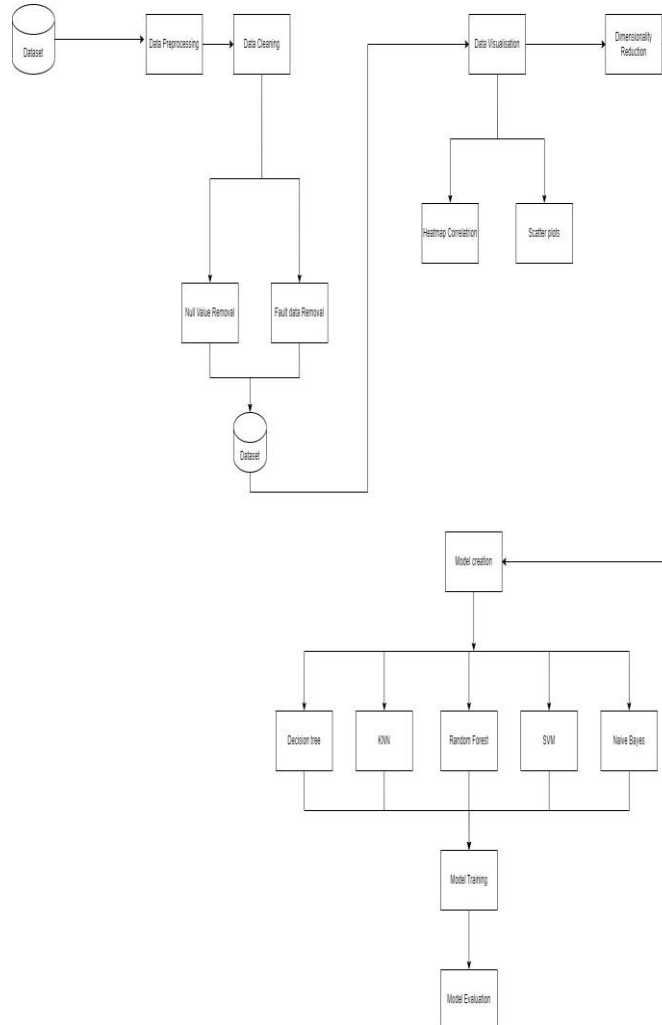
The Dataset we are using is from UCI machine learning repository. The Dataset contains 24 health related attributes like age, blood pressure, sugar, glucose, taken in 2-month period of 400 patients in which 11

numeric and 14 nominal attributes in which it consists of class label named ‘Class’ which classifies patients having disease and not present.

[https://archive.ics.uci.edu/ml/datasets/chronic\\_kidney\\_disease](https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease)

There are 24 features + class = 25 attributes

### Architecture or Flowchart with explanation:



- Data Collection
- Data Pre processing
  - Null value removal
  - Fault data removal
- Data Visualization
  - Heat map Correlation
  - Scatter Plots
- Dimensionality reduction
- Model Creation
  - Decision Tree
  - KNN
  - Random Forest
  - SVM
  - Naïve Bayes
- Model Training
- Model Evaluation

**Data Collection:**

In this step, we will collect the data related to the kidney patients. This data includes 24 health related attributes like age, blood pressure, sugar, glucose. This data contains 11 numeric and 14 nominal attributes. A data with total 25 attributes is collected.

**Data Pre-processing:**

Data pre-processing is a crucial step in machine learning that involves cleaning and transforming raw data to prepare it for analysis. It includes tasks such as removing duplicates, dealing with missing values, scaling and normalizing features, and encoding categorical variables. We are going to perform null value removal and fault data removal on the data we have collected in the previous step. Proper data pre-processing can improve the accuracy and effectiveness of machine learning models.

**Null value removal:**

Null value removal is a common preprocessing step in machine learning that involves identifying and removing missing values from the dataset. This can be done by either deleting rows or columns with null values or imputing missing values with a suitable replacement value. In our project we are going to impute the values with the most repeated values.

**Fault Data removal:**

Fault data removal is a process in machine learning that involves identifying and removing erroneous or inconsistent data from a dataset. This can be due to data entry errors, measurement errors, or other types of inaccuracies. Faulty data can negatively impact the performance of machine learning models and lead to inaccurate predictions. Techniques for identifying and removing faulty data can include statistical analysis, visualization, and domain knowledge.

**Data Visualization:**

In this step the data visualization is done to visualize the data. Data visualization will help us to visualize the data. So that we can identify the relation between the attributes and identify the outliers in the data. Removal of the outliers will help the model to improve its accuracy.

**Dimensionality reduction:**

Dimensionality reduction is a technique in machine learning that involves reducing the number of features or variables in a dataset while preserving as much of the relevant information as possible. This can help in improving computation efficiency and reducing overfitting.

**Model creation:**

In this step the model is created and we will use over dataset and split the dataset into training and testing. Later the training dataset is used to train the model. In the project we are going to use Decision Tree, KNN, Random Forest, SVM, Naive Bayes.

**Model Evaluation:**

The model is evaluated using the testing data. In our project we are going to use different measures like f score, recall and other to evaluate the model.

**Proposed Algorithm (step by step):****Decision tree:**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

- > **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-3. Continue this process until a stage is reached where you can't further classify the nodes and called the final node as a leaf node

**KNN:**

The k-nearest neighbour (K-NN) classifier is considered as an example-based classifier, which means that the training documents are used for comparison instead of an exact class illustration, like the class profiles utilized by other classifiers. As such, there's no real training section. once a new document must be classified, the k most similar documents (neighbours) are found and if a large enough proportion of them are allotted to a precise class, the new document is also appointed to the present class, otherwise not. Additionally, finding the closest neighbours is quickened using traditional classification strategies.

- > **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.

**Random Forest:**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. 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).

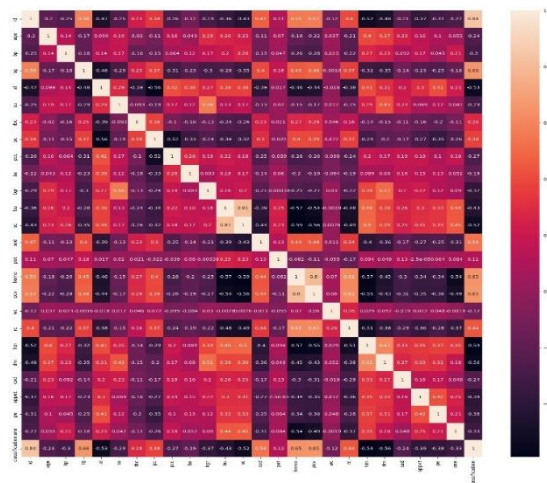
- Step-3: Choose the number N for decision trees that you want to build.
- Step-4: Repeat Step 1 & 2.

## SVM:

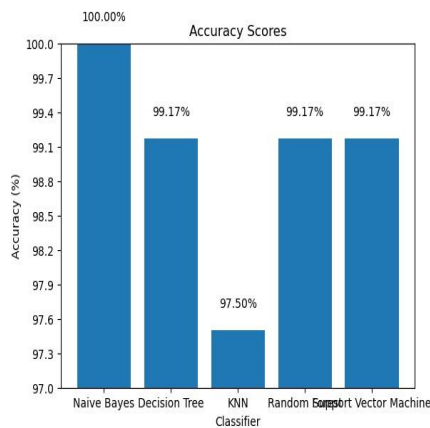
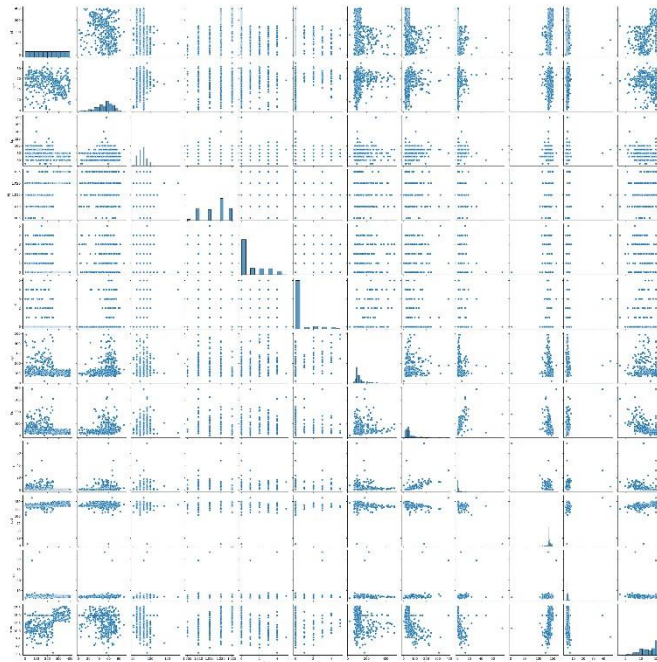
SVM aims to find an optimal hyperplane that separates the data into different classes. The stick learn package in python is used to implement SVM. The pre-processed data is divided into datasets and the training set represents 20% and 80% of the data, respectively. A support vector machine that constructs a hyperplane or set of hyperplanes in high or infinite dimensional space. Good separation is achieved by the hyperplane having the greatest distance to the nearest training point of any class, because in general, the larger the amplitude, the smaller the overall error of the classifier. In the SVM algorithm, we seek to maximize the amplitude between the data points and the hyperplane. The loss function that maximizes profits is the hinge loss. Cost is 0 if the predicted value and the actual value have the same sign. If not, then we calculate the value of the loss. We also add a regularization parameter to the cost function. The goal of the regularization parameter is to strike a balance between the maximization and the loss of profit. After adding the regularization parameter Now that we have the loss function, we take partial derivatives with respect to the weights to find the gradients. Using gradients, we can update our weights. When there is no misclassification, i.e., our model correctly predicts our data point type, we simply update the gradient from the regularization parameter. When there is a misclassification, i.e. our model makes an error in the prediction of our data point's class, we include the loss with the normalization parameter to perform the gradient update

## Naïve Bayes:

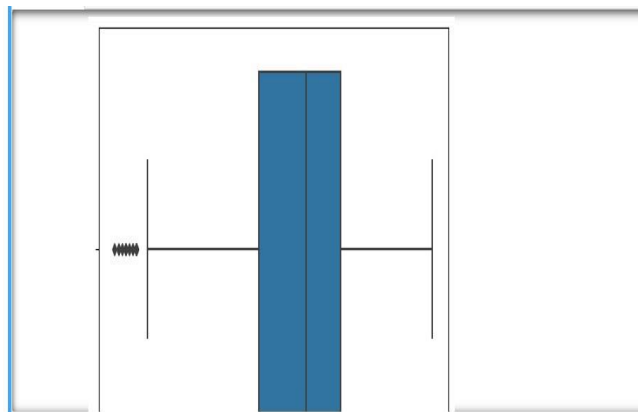
Bayesian classifier is a statistical classifier. They can predict class membership probabilities, for instance, the probability that a given sample belongs to a particular class. Bayesian classification is created on the Bayes theorem. Studies comparing the classification algorithms have found a simple Bayesian classifier known as the naive Bayesian classifier to be comparable in performance with decision tree and neural network classifiers. Bayesian classifiers have also displayed high accuracy and speed when applied to large databases. Naive Bayesian classifiers adopt that the exact attribute value on a given class is independent of the values of the other attributes. This assumption is termed class conditional independence. It is made to simplify the calculations involved, and is considered “naive”. Bayesian belief networks are graphical replicas, which unlike naive Bayesian classifiers allow the depiction of dependencies among subsets of attributes. Bayesian belief can also be utilized for classification.



**Experiment Results:**



**RESULTS AND DISCUSSION:**



In this project, we have used different classification models for classification of chronic kidney disease in patients. The four machine learning techniques that were used they are Decision Tree, KNN, Naïve Bayes and Random Forest. The accuracy of different algorithms on the dataset was evaluated. The dataset

contains 400 rows. It contains data of 250 not chronic kidney data and 150 chronic kidney disease data. The dataset spitted into training and testing data. Accuracy was evaluated based on TP, TN, FP, FN. The accuracy achieved by naïve bayes is 100%. This project could be useful in the current medical field with advancement in sciences and new emerging technologies it would be of good help.

### Conclusion and Future work:

From this we conclude that for the given dataset Random Forest and Naïve Bayes classifier that gives us the best accuracy results. But the same may not be true if a different testing set is used the accuracy will vary. In the real world, if the size of dataset increases by a large amount in the future. Then the overfitting problem of the models will be resolved.

### V.REFERENCES

1. Rahman, T. M., Siddiqua, S., Rabby, S. E., Hasan, N., & Imam, M. H. (2019, January). Early detection of kidney disease using ECG signals through machine learning based modelling. In 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) (pp. 319-323). IEEE.
2. Kriplani, H., Patel, B., & Roy, S. (2019). Prediction of chronic kidney diseases using deep artificial neural network technique. In Computer aided intervention and diagnostics in clinical and medical images (pp. 179-187). Springer, Cham.
3. Devika, R., Avilala, S. V., & Subramaniaswamy, V. (2019, March). Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest. In 2019 3rd International conference on computing methodologies and communication (ICCMC) (pp. 679-684). IEEE.
4. Rady, E. H. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked*, 15, 100178.
5. Raju, N. G., Lakshmi, K. P., Praharshitha, K. G., & Likhitha, C. (2019, May). Prediction of chronic kidney disease (CKD) using Data Science. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 642-647). IEEE.
6. Almansour, N. A., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J., ... & Olatunji, S. O. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. *Computers in biology and medicine*, 109, 101-111.
7. Elhoseny, M., Shankar, K., & Uthayakumar, J. (2019). Intelligent diagnostic prediction and classification system for chronic kidney disease. *Scientific reports*, 9(1), 1-14.
8. Almasoud, M., & Ward, T. E. (2019). Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *International Journal of Soft Computing and Its Applications*, 10(8).
9. Snegha, J., Tharani, V., Preetha, S. D., Charanya, R., & Bhavani, S. (2020, February). Chronic kidney disease prediction using data mining. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-5). IEEE.
10. Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020). Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of personalized medicine*, 10(2), 21.
11. Aldhyani, T. H., Alshebami, A. S., & Alzahrani, M. Y. (2020). Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms. *Journal of healthcare engineering*, 2020.

12. Reshma S , Salma Shaji , S R Ajina , Vishnu Priya S R, Janisha A, 2020, Chronic Kidney Disease Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 07 (July 2020)
13. I. U. Ekanayake and D. Herath, "Chronic Kidney Disease Prediction Using Machine Learning Methods," 2020 Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, 2020, pp. 260-265, doi: 10.1109/MERCon50084.2020.9185249.
14. Jongbo, O. A., Adetunmbi, A. O., Ogunrinde, R. B., & BadejiAjisafe, B. (2020). Development of an ensemble approach to chronic kidney disease diagnosis. *Scientific African*, 8, e00456.
15. Ghosh, P., Shamrat, F. J. M., Shultana, S., Afrin, S., Anjum, A. A., & Khan, A. A. (2020, November). Optimization of prediction method of chronic kidney disease using machine learning algorithm. In 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) (pp. 1-6). IEEE.
16. Gudeti, B., Mishra, S., Malik, S., Fernandez, T. F., Tyagi, A. K., & Kumari, S. (2020, November). A novel approach to predict chronic kidney disease using machine learning algorithms. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1630-1635). IEEE.
17. Islam, M. A., Akter, S., Hossen, M. S., Keya, S. A., Tisha, S. A., & Hossain, S. (2020, December). Risk factor prediction of chronic kidney disease based on machine learning algorithms. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 952-957). IEEE.
18. Emon, M. U., Islam, R., Keya, M. S., & Zannat, R. (2021, January). Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches. In 2021 6th International Conference on Inventive Computation Technologies (ICICT)(pp. 713-719). IEEE.
19. Roy, M. S., Ghosh, R., Goswami, D., & Karthik, R. (2021, May). Comparative analysis of machine learning methods to detect chronic kidney disease. In *Journal of Physics: Conference Series* (Vol. 1911, No. 1, p. 012005). IOP Publishing.
20. Senan, E. M., Al-Adhaileh, M. H., Alsaade, F. W., Aldhyani, T. H., Alqarni, A. A., Alsharif, N., ... & Alzahrani, M. Y. (2021). Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. *Journal of Healthcare Engineering*, 2021.
21. Ifraz, G. M., Rashid, M. H., Tazin, T., Bourouis, S., & Khan, M. M. (2021). Comparative analysis for prediction of kidney disease using intelligent machine learning methods. *Computational and Mathematical Methods in Medicine*, 2021.
22. Twarish Alhamazani, K., Alshudukhi, J., Aljaloud, S., & Abebaw, S. (2021). Implementation of Machine Learning Models for the Prevention of Kidney Diseases (CKD) or Their Derivatives. *Computational Intelligence and Neuroscience*, 2021.
23. Khan, P. F., Reddy, M. R., Samatha, K., Chowdary, R. A., & Rao, P. P. (2021). Predictive Analytics Of Chronic Kidney Disease By Using Machine Learning.
24. Mane, P., Thoutam, N., Tiwari, N., Mandlik, G., & Pandey, N. (2021). Chronic kidney disease prediction using machine learning. *International Journal of Advanced Science and Technology*, 30(4), 1834-1844.
25. Abdel-Fattah, M. A., Othman, N. A., & Goher, N. (2022). Predicting Chronic Kidney Disease Using Hybrid Machine Learning Based on Apache Spark. *Computational Intelligence and Neuroscience*, 2022.



26. Poonia, R. C., Gupta, M. K., Abunadi, I., Albraikan, A. A., Al- Wesabi, F. N., & Hamza, M. A. (2022, February). Intelligent diagnostic prediction and classification models for detection of kidney disease. In *Healthcare* (Vol. 10, No. 2, p. 371). MDPI.
27. Bai, Q., Su, C., Tang, W., & Li, Y. (2022). Machine learning to predict end stage kidney disease in chronic kidney disease. *Scientific reports*, 12(1), 1-8.
28. Ebiaredoh-Mienye, S. A., Swart, T. G., Esenogho, E., & Mienye, I. D. (2022). A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease. *Bioengineering*, 9(8), 350.
29. Dritsas, E., & Trigka, M. (2022). Machine learning techniques for chronic kidney disease risk prediction. *Big Data and Cognitive Computing*, 6(3), 98.
30. Debal, D. A., & Sitote, T. M. (2022). Chronic kidney disease prediction using machine learning techniques. *Journal of Big Data*, 9(1), 1-19.