Chronic Kidney Disease Prediction

Rahul Khadse[1][8291244084] and Dipeshkumar Rathod[2][8291614084]

Prakash Rathod[3][8291648794], Aryaman Rathor[4][8291784084]

[1] Ramrao Adik Institute Of Technology
[2]Nerul, Navi Mumbai, Maharashtra, India

Abstract. Chronic kidney disease (CKD) is a growing medical problem that impairs renal function and ultimately harms the kidneys. CKD is a highly common condition nowadays, and two life-threatening conditions that can result from it are cardiovascular infection and end-stage renal disease. These might be avoided by identifying conditions early and treating those who are in danger. The responsibility of anticipating medical issues is exceedingly challenging. One of the most fatal diseases in the medical world is specifically CKD. The prediction of risk factors is a crucial step in the initial stage before it is too late to identify CKD forecast and eliminate risks. With persistent kidney disease with a consistent growth rate, sickness has grown to be a significant problem. Because a person may only survive without their kidneys for an average of 18 days, dialysis and kidney transplants are in great demand. Effective techniques for CKD early prediction are crucial. Machine learning techniques are useful for predicting CKD. In order to predict CKD status using clinical data, this work suggests a workflow that includes data prepossessing, a method for handling missing values, collaborative filtering, and attribute selection. The study highlights the significance of incorporating domain knowledge when using machine learning for CKD status prediction as well as the practical aspects of data collection.

Keywords: Kidney, Disease, Chronic, Machine learning, Supervised, Prediction, Accuracy, Classification, Ensemble.

1 Introduction

Massive amounts of data are being generated by the healthcare sector, and these data need to be mined in order to uncover hidden information that may be used for successful prediction, diagnosis, and decision-making. Kidney disease has recently been a major issue. In India, it is one of the main causes of death. The progressive decrease of kidney function is a defining feature of chronic kidney disease (CKD). Wastes and extra fluid are removed from the circulation by the kidneys and expelled in the urine. Wastes can build up in the blood and create problems like high blood pressure, anemia, weakening of the bones, poor nutritional health, and nerve damage if the disease worsens. Additionally, kidney dysfunction raises the possibility of developing heart and blood vessel problems.According to the study, adverse consequences can be averted and prevented by early detection.

Growing out of the study of pattern recognition and computational learning theory in artificial intelligence, machine learning is a topic that focuses on the analysis and prediction of massive amounts of data with many different variables. Machine learning approaches have demonstrated efficacy in the prediction and diagnosis of many serious diseases, according to medical science. In this approach, each instance in any dataset is represented by a specific set of features. Additionally, the ability of experts and professionals to uncover hidden patterns in data is restricted. Therefore, the alternative is to analyze the raw data using computational approaches to uncover intriguing information for the decision-maker.Patients' knowledge of CKD is progressively rising but is still low. In India, chronic kidney disease is the eighth most common cause of death, according to the Global Burden of Disease (GBD) 2015 report. India is anticipated to have 57.2 million cases of diabetes by 2025, and there will also be twice as many patients there with high blood pressure, making it the country with the largest concentration of CKD cases worldwide. Therefore, primary care physicians bear the bulk of the responsibility for CKD management (PCPs). Therefore, it's crucial to have a reliable, practical, and automated CKD detection approach.This study's primary goal is to predict renal illness by examining data from those indices, using three machine learning classification algorithms to do so, and selecting the approach with the highest accuracy rate.

2 Literature survey

Existing literature was examined in order to gain the necessary knowledge about various ideas connected to the current application. The following is a list of some of the significant findings that were drawn from those.naive Bayes classifier, feed-forward neural network, support vector machine, K-nearest neighbor, logistic regression, random forest, and Random forest provide superior accuracy in these classifiers, at 99.75 percent.[1]

On the basis of a dataset of 400 occurrences, Vasquez-Morales et al. created a neural network model with a 91\% accuracy for risk prediction of the development of chronic kidney disease. Three models were used by Chen et al. on the UCI dataset. They employed these classifiers to discover the patient's risk calculation using KNN, SVM, and soft independent modeling of class analogy (SIMCA). The SVM and KNN models both achieved the highest accuracy of 95.7\%, while the SVM model is best able to withstand noise disturbance. Due to the invasiveness and expense of CKD, many patients have reached their last stages without receiving therapy. Therefore, it is still crucial to find this disease early.[2]

Additionally, SVM machine learning classifier method testing results with an accuracy of 93\% were provided by Amirgaliyev. The use of machine learning classifier algorithms for the early diagnosis of CKD in diabetic patients was proposed by Padmanaban and Parthiban. They used Naive Bayes and Decision trees to analyze the dataset after collecting the data from a diabetic research center in Chennai. They used the Weka tool to determine the accuracy and came to the conclusion that the Naive Bayes classifier had the highest accuracy, at 91\%. De Almeida et al. 's work utilized Support Vector Machines (SVM) with linear, polynomial, sigmoid, and RBF functions in addition to Decision Trees, Random Forests, and SVM. They made use of the MIMIC-II database for their investigation.They came to the conclusion that decision trees and random forests produced the greatest results, with respective prediction accuracy rates of 80\% and 87 \%.[3]

In order to determine which machine learning classifier algorithm would be most effective given the dataset, Gunarathne et al. developed a model of multiple classifier algorithms. They made use of a UCI-provided dataset with 400 instances and 14 attributes. They came to the conclusion that the Multiclass Decision Forest algorithm, with an accuracy of 92.1\%, was best matched for the CKD dataset. The SVM technique was utilized by Polat et al. to predict CKD.[4]

They worked on a crucial component in order to get the right outcome. They combined a two-approach Wrapper and filter with an SVM algorithm to identify the appropriate feature. There was a best first search engine for the Wrapper subset evaluator and a greedy stepwise search engine for the classifier subset evaluator in the Wrapper.There were two search engines in the filter: the best first search engine for the filtered subset evaluator and the greedy stepwise search engine for the correlation feature section subset feature. Comparing the outcomes of each technique, it was discovered that SVM provided the maximum accuracy with a filtered subset evaluator or 98.5 percent.[5]

Sujata Drall, Gurdeep Singh Drall, Sugandha Singh, Bharat Drall, and others worked with the 400 cases, 25 attribute CKD dataset provided by UCI. Data were first preprocessed, and missing data was located, and filled in with 0, after which a transformation was done to the dataset. After preprocessing, the authors utilized an algorithm for significant characteristics and identified the top five features before using Naive Bayes and K-Nearest Neighbor as their classification algorithms. KNN obtained the highest degree of accuracy. 400 occurrences and 25 attributes from the CKD dataset were used by Almasoud and Ward. Haemoglobin, albumin, and specific gravity were discovered to be feature attributes in the CKD dataset after they applied the filter feature selection method to attributes. Following the selection of the features, they trained the dataset and performed 10-fold cross-validation. The algorithm that produced the highest accuracy, 92.1\%, was gradient boosting.[6]

On the same UCI dataset, Shankar et al. carried out three processes: I data preparation & feature selection; determining the correctness of the algorithms; and suggesting a diet. Two methods were used in the feature selection technique: the Wrapper method and the LASSO method. Four classification techniques were used after the feature selection method: Logistic Regression, Random Forest Tree K-Nearest Neighbors, Neural Network, and Wide and Deep Learning. The blood potassium level was used to advise a diet. Depending on its value, the blood potassium level was split into three categories: the safe zone, the caution zone, and the danger zone.[7]

The kidney function test (KFT) dataset was gathered by Vijayarani and Dhayanand from medical labs, research facilities, and hospitals. The dataset included 584 instances, 6 attributes, and the support vector machine (SVM) and artificial neural network classifier algorithms. It was discovered that ANN had the highest accuracy, coming in at 87.7\%.[8]

With the use of 9 machine learning algorithms, including XGBoost, logistic regression, lasso regression, support vector machine, random forest, ridge regression, neural network, Elastic Net, and K-nearest neighbor, Xiao et al. utilized the data from 551 patients. The linear model had the maximum accuracy, according to their evaluation of accuracy, ROC curve, precision, and recall. On the CKD Dataset, Reshma et al .'s feature selection technique was applied. The ACO approach was used to choose the features. The feature selection meta-heuristic algorithm is called ACO. It is a Wrapper method type. There were a total of 24 attributes in their dataset. Twelve features were utilized to build the model after the feature selection technique was used. The model was created using the Support Vector Machine Classifiers technique.[9]

Based on an outdated dataset of CKD, Deepika et al. developed a project for the prediction of chronic kidney disease. 24 attributes and 1 target variable were present in the dataset. They used the KNN and Naive Bayes supervised machine learning algorithms to develop the model. KNN and Naive Bayes both obtained accuracy levels of 91\% and 93\%, respectively.[9]

Ma et al .'s deep learning system was suggested for early Chronic Kidney Disease prediction. Heterogeneous Modified Artificial Neural Network Algorithm was used to create the deep neural network. The model was created using ultrasound pictures. Three different classifiers—the Support Vector Machine, the artificial neural network, and the multilayer perceptron—were used to compare the results.UI The machine learning model for early diabetic illness prediction was proposed by Haq et al.. They came to the conclusion that machine learning can be very important in the medical field.Amin et al .'s machine learning model was proposed for early Parkinson's disease prediction. SVM classifier was employed in the model's construction. Additionally, the Relief and ACO feature selection algorithms were used to extract the crucial features.The main goal of this study is to determine whether or not someone has chronic kidney disease. Seven different machine learning classifiers were used on the dataset for this perception.Both the entire features and the chosen features were active for each algorithm. All of the outcomes were recorded and oversampling was done using SMOTE. One deep neural network technique was used to compare the outcomes of every machine learning model. Two hidden layers of a deep learning neural network were employed. In order to perform computations, IBM SPSS Modeler was used. The contribution shows that, when applying deep neural networks to a dataset, the accuracy estimate is 94.6\%.[10]

3 Proposed Methodology System

3.1 Dataset:

Chronic Kidney Disease (CKD) is a medical condition in which the kidneys gradually lose function over time, leading to a buildup of toxins and waste products in the body. CKD is a serious health condition that can lead to complications such as high blood pressure, anemia, and bone disease.The dataset you have contains 26 attributes, some of which are categorical and some of which are numerical. It includes attributes such as age, gender, hemoglobin levels, and other health-related factors that may be relevant to predicting the presence of CKD.

Age: This characteristic indicates a person's age in years. It has a numerical value and functions as a predictor.

Bp: This feature, which stands for blood pressure, is both a predictive variable and a numeric number.

Sg: This property, which stands for specific gravity, has a numerical value. It serves as a predictor.

Al: Albumin is referred to as Al. It has a numerical value. It serves as a predictor.

Su: This characteristic denotes sugar. It has a numerical value. It serves as a predictor.

RBC: Red blood cells are referred to as RBCs. It has no intrinsic worth. It serves as a predictor.

Pc stands for pus cell in this attribute. It has no intrinsic worth. It serves as a predictor.

Pcc: This characteristic denotes pus cell clusters. It has no intrinsic worth. It serves as a predictor.

B: This qualifier denotes bacteria. It has no intrinsic worth. It serves as a predictor.

Bgr: The term "Bgr" stands for blood glucose random. It has a numerical value. It serves as a predictor.

Bu: This quality refers to blood urea. It has a numerical value. It serves as a predictor.

sc: Serum creatinine is the meaning of the suffix "Sc." It has a numerical value. It serves as a predictor.

Sod: This term stands for salt. It has a numerical value. It serves as a predictor.

Pot: This property refers to potassium. It has a numerical value. It serves as a predictor.

Hemo: This characteristic refers to hemoglobin. It has a numerical value.It serves as a predictor.

Pcv: The packed cell volume attribute. It has a numerical value. It serves as a predictor.

The term "Wc" stands for white blood cell count. It has a numerical value. It serves as a predictor.

Rc stands for red blood cell count in this attribute. It has a numerical value. It serves as a predictor.

HTN: Hypertension is denoted by this attribute. It has no intrinsic worth. It serves as a predictor.

Dm: Diabetes mellitus is denoted by the letter Dm. It has no intrinsic worth. It serves as a predictor.

cad: Coronary artery disease is indicated by the code CAD. It has no intrinsic worth. It serves as a predictor.

Appet: This characteristic refers to appetite. It has no intrinsic worth. It serves as a predictor.

Pe: This characteristic refers to pedal edema. It has no intrinsic worth. It serves as a predictor.

ane: Anemia is indicated with the suffix "ane." It has no intrinsic worth. It serves as a predictor.

Class: The response attribute is what it is. Here, a person with chronic kidney disease or one without it is described. And this variable's type is nominal.

3.2 Pre-Processing:

A total of 24 attributes are used in this dataset, 24 of which are predictive variables and 1 is a response variable. Some of the 24 predicted qualities are numerical, while others are nominal. As a result, use a mapping function to translate nominal attributes into numerical attributes. The nominal value in this dataset includes the following attributes: rbc, applet, pc, pcc, ba, dm, htn, cad, pe, and ane. We converted this attribute to a numeric value using a mapping function. Our data set is now filled with numerical values. Additionally, dividing the dataset into 20% for testing and 80% for training.

3.3 Methodology:

Chronic Kidney Disease (CKD) is a major public health problem worldwide. Early diagnosis of CKD is crucial for the management and prevention of complications. Machine learning (ML) algorithms have proven to be useful in predicting CKD. In this report, we will discuss four ML algorithms - Decision Tree, Random Forest, Grid Search, and xgBoost - and their application in predicting CKD.

Decision Tree: A Decision Tree is a simple yet powerful algorithm that is widely used in classification and regression tasks. It works by recursively partitioning the feature space into regions that correspond to different classes or target values. Each split in the tree is chosen to maximize the information gain, which measures the reduction in entropy (or impurity) of the target variable. Decision Trees are easy to interpret and can handle both categorical and continuous variables. However, they are prone to overfitting and can be unstable.

Random Forest: Random Forest is an ensemble method that combines multiple Decision Trees to improve performance and reduce overfitting. It works by generating several random subsets of the training data and constructing a Decision Tree on each subset. The final prediction is obtained by averaging the predictions of all the trees. Random Forest is a robust and scalable algorithm that can handle high-dimensional data and non-linear relationships. However, it can be computationally expensive and less interpretable than a single Decision Tree.

Grid Search: Grid Search is a hyperparameter optimization technique that is used to find the best set of hyperparameters for a given ML algorithm. Hyperparameters are parameters that are not learned from data but are set by the user before training the model. Grid Search works by specifying a grid of hyperparameter values and training the model on all possible combinations of values in the grid. The performance of each model is evaluated using a validation set, and the hyperparameters that yield the best performance are chosen. Grid Search is a simple and effective way to fine-tune ML models, but it can be computationally expensive for large grids.

xgBoost:xgBoost is a gradient-boosting algorithm that is designed to improve the accuracy and speed of Decision Trees. It works by iteratively adding new Decision Trees to correct the errors of the previous trees. The predictions of all the trees are combined using a weighted sum, where the weights are determined by the gradients of the loss function. xgBoost is a state-of-the-art algorithm that has won many Kaggle competitions and is widely used in the industry. It can handle missing data, categorical variables, and large datasets. However, it can be sensitive to hyperparameters and require more expertise in tuning.

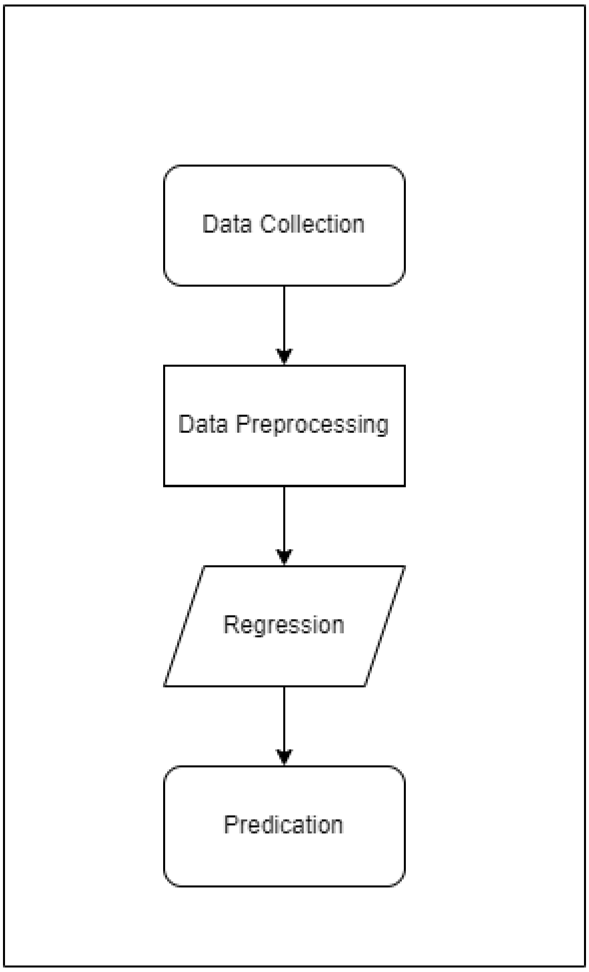

Fig. 1. Flow of prediction

4 Architecture

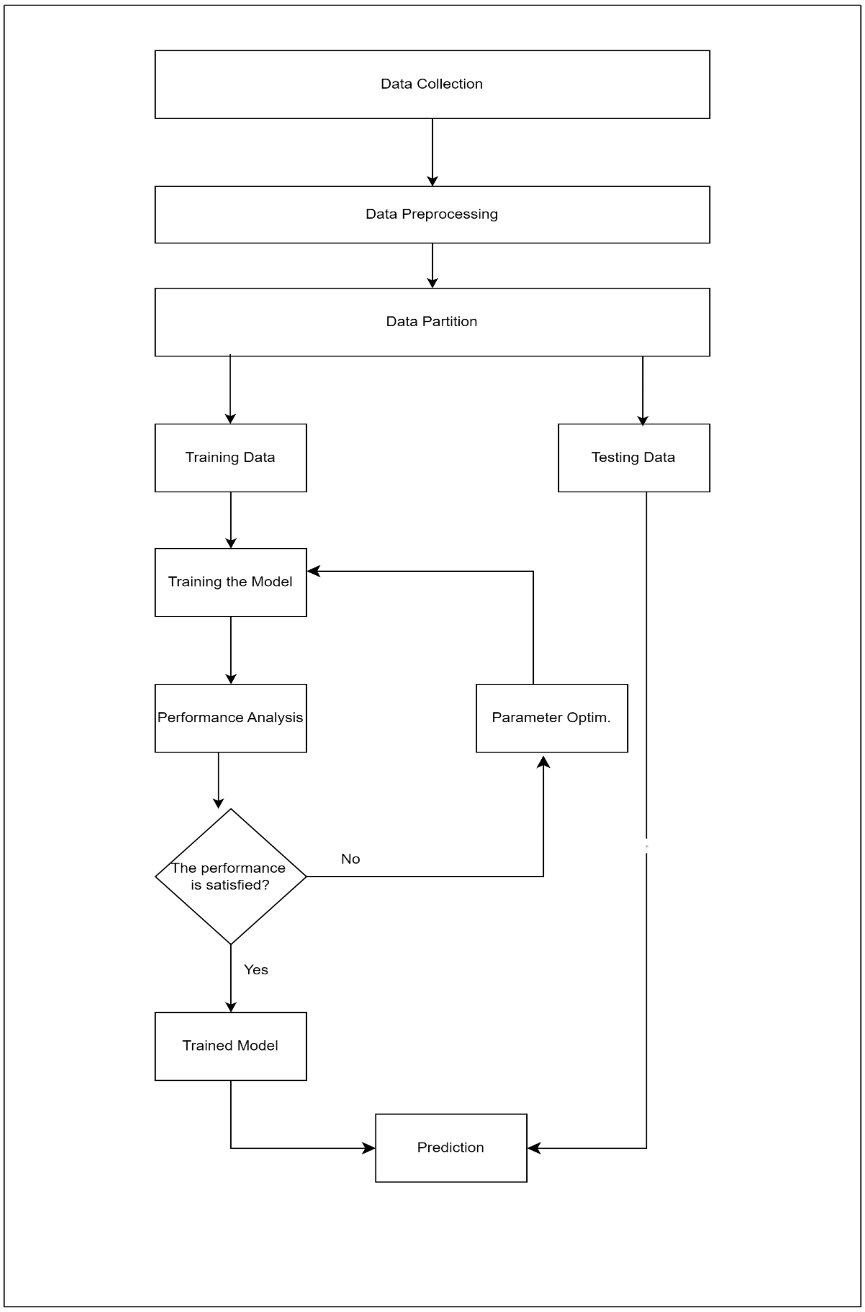

Fig. 2. Architecture of CKD

An abstract representation of a system's structure and behavior is called a system architecture. A system is formally represented by it. A system architecture can refer to either a model used to explain the system or a process used to create the system, depending on the context. Building a suitable system architecture aids in project analysis, particularly at the beginning.

- Requirement analysis: This stage is involved gathering the system's needs. The creation of documents and requirement analysis steps in this procedure. \
- System Design: The system specifications are converted into a software representation while keeping the needs in mind. The designer places emphasis on things like algorithms, data structure, software architecture, etc. at this phase.
- Implementation: The actual coding or programming of the software is done during the implementation phase. The libraries, executables, user guides, and supplementary software documentation are frequently the outcome of this phase.
- Testing: All programs (models) are integrated and tested at this phase to make sure the entire system complies with the software specifications. The testing focuses on validation and verification.

5 Result Analysis

The user or the patient provides all the values of the attributes as input to the system model, The machine model takes the input and analysis it on the five types of algorithms we have used i.e The decision tree algorithm, Random Forest algorithm, Grid Search algorithm, and xgBoost. These algorithms perform the analysis of these attributes and provide the user or the patient with the output. The output contains the data on the condition of their kidney and gives the user the prediction of the chronic kidney disease in "YES" or "NO" terms. If the patient's kidney is affected by the disease then the system will further predict the stage up to Which the disease is increased. The machine model system bifurcates the disease mainly into the first stage, second stage, and third stage also call it as the last stage.

The following Figures 6.1, figure 6.2, and figure 6.3 shows the analysis of the attributes provided by the user as input and it is converted into the graphical model to understand the actual meaning of it. Figure 6.1 gives the numerical feature distribution of the data values, figure 6.2 shows the categorical feature distribution of the attributes provided as the input, and Figure 6.3 shows the heat map among all 24 attributes.

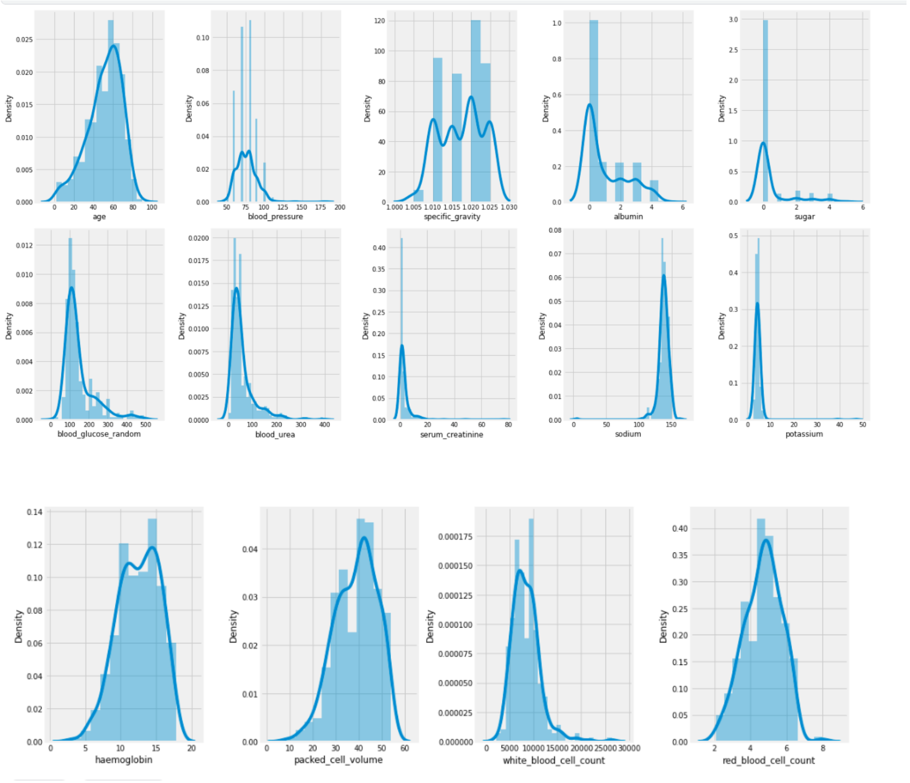

Fig. 3. Data Correlation

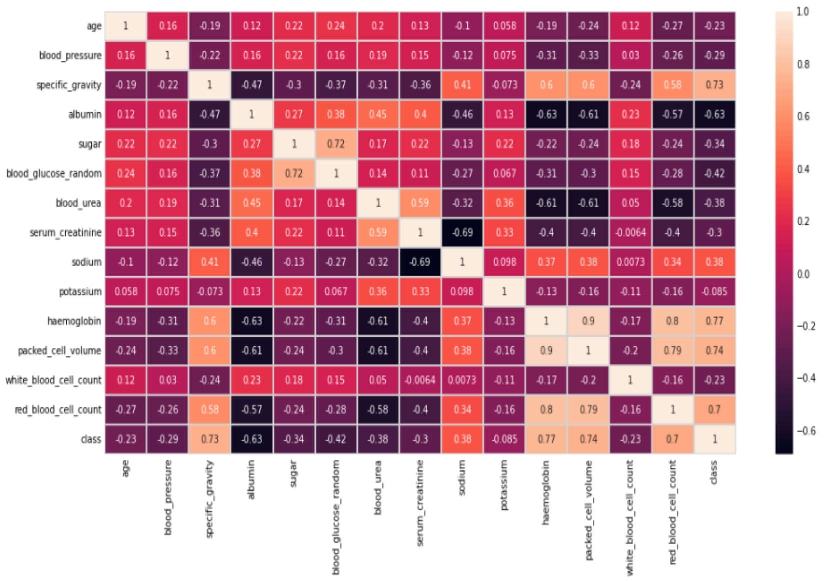

Fig. 4. Correlation Matrix

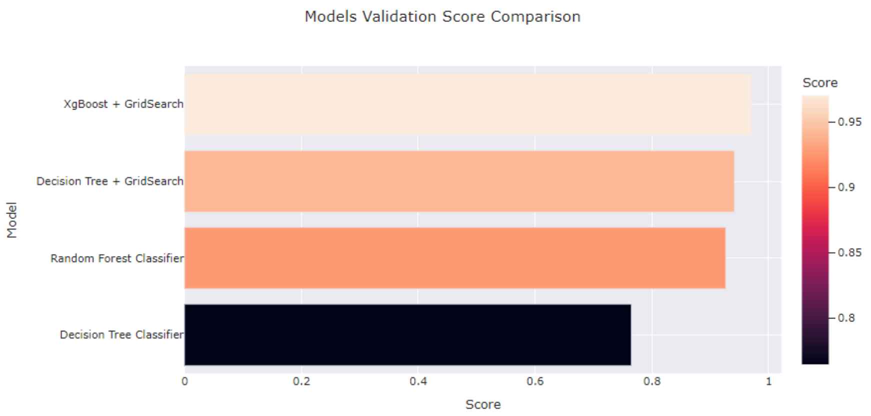

Fig. 5. Result of Predication

6 Conclusion

Progressive loss of kidney function over time is a feature of chronic renal disease. Since the majority of victims show no symptoms, it is a quiet illness. The medical community faces a serious challenge in the early diagnosis and treatment of CKD, and they turn to machine learning theory to develop an effective solution. In the current work, a supervised learning methodology is given that focuses primarily on probabilistic, tree-based, and ensemble learning-based models to build effective models for predicting the risk of developing CKD. Additionally, GridSearch + XgBoost: Achieving an accuracy of 97\% is an excellent performance, indicating that the combination of GridSearch and XgBoost model has yielded highly accurate predictions on your data. This suggests that the XgBoost model with optimized hyperparameters from GridSearch is a strong performer for your data. Random Forest: With an accuracy of 92\%, the Random Forest model has also performed well and is a reliable choice for your data analysis. It may be slightly less accurate compared to the GridSearch + XgBoost model, but still provides a respectable level of accuracy. Decision Tree: The Decision Tree model achieving an accuracy of 76\% indicates that it may not be the best-performing model compared to the other two models. It may be less accurate and may require further optimization or consideration of other models.GridSearch + Decision Tree: The Decision Tree model with hyperparameter optimization using GridSearch achieving an accuracy of 94\% is a notable improvement compared to the basic Decision Tree model, but it still falls slightly short of the performance of the GridSearch + XgBoost and Random Forest models. Based on these results, the conclusion can be drawn that the GridSearch + XgBoost and Random Forest models are the top performers in terms of accuracy for your data. These models can be considered for further analysis or deployment in your web application.

In our upcoming work, we intend to focus our research on Deep Learning techniques using CNN and LSTM and examine the potential performance improvement that these models may offer. We set out to go in two different paths to fully utilize the potential of these models. Before supplying the little dataset to the ML models, the former will use a data augmentation technique, such as an SVR-based additive input doubling technique, to improve it. In the latter, we'll start off by experimenting with a sizable, non-synthetic dataset.

References

1. [1]Suresh, C., Pani, B. C., Swatisri, C., Priya, R., & Rohith, R. (2020). A Neural Network based Model for Predicting Chronic Kidney Diseases. 2020 Second International Conference on Inventive Research in

Computing Applications (ICIRCA). doi:10.1109/icirca48905.2020.9183318

2. [2]Ekanayake, I. U., & Herath, D. (2020). Chronic Kidney Disease Prediction Using Machine Learning Methods. 2020 Moratuwa Engineering Research Conference (MERCon). doi:10.1109/mercon50084.2020.9185249
3. [3]Islam, M. A., Akter, S., Hossen, M. S., Keya, S. A., Tisha, S. A., & Hossain, S. (2020). Risk Factor Prediction of Chronic Kidney Disease based on Machine Learning Algorithms. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). doi:10.1109/iciss49785.2020.9315878
4. [4]Maurya, A., Wable, R., Shinde, R., John, S., Jadhav, R., & Dakshayani, R. (2019). Chronic Kidney Disease Prediction and Recommendation of Suitable Diet Plan by using Machine Learning. 2019 International Conference on Nascent Technologies in Engineering (ICNTE). doi:10.1109/icnte44896.2019.8946029
5. [5]Suresh, C., Pani, B. C., Swatisri, C., Priya, R.,& Rohith, R. (2020). A Neural Network based Model for Predicting Chronic Kidney Diseases. 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). doi:10.1109/icirca48905.2020.9183318
6. [6]Suresh, C., Pani, B. C., Swatisri, CEkanayake, I. U., & Herath, D. (2020). Chronic Kidney Disease Prediction Using Machine Learning Methods. 2020 Moratuwa Engineering Research Conference (MERCon). doi:10.1109/mercon50084.2020.9185249
7. [7]Desai, M. (2019). Early Detection and Prevention of Chronic Kidney Disease. 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA). doi:10.1109/iccubea47591.2019.9128424
8. [8]Asra, T., Setiadi, A., Safudin, M., Lestari, E. W., Hardi, N., & Alamsyah, D. P. (2021). Implementation of AdaBoost Algorithm in Prediction of Chronic Kidney Disease. 2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST). https://doi.org/10.1109/iceast52143.2021.9426291
9. [9]Vashisth, S., Dhall, I., & Saraswat, S. (2020). Chronic Kidney Disease (CKD) Diagnosis using Multi-Layer Perceptron Classifier. 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence). doi:10.1109/confluence47617.2020.9058178
10. [10]Wibawa, H. A., Malik, I., & Bahtiar, N. (2018). Evaluation of Kernel-Based Extreme Learning Machine Performance for Prediction of Chronic Kidney Disease. 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS). doi:10.1109/icicos.2018.8621762