

Chronic Kidney Disease Prediction using Machine Learning Techniques

Nirmala M B¹, Priyamvada D K², Prerana R Shetty³ and Pallavi Singh S⁴

¹⁻⁴Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumakuru, Karnataka, India

¹mbnirmala@sit.ac.in, ²priyamvadakdongre@gmail.com, ³preranareshetty@gmail.com, ⁴pallavisingh408@gmail.com

Abstract—Predictive analytics for healthcare using machine learning is a challenging task. Development of a fast and accurate predictive analytics tool can lead to a lot of advancement in healthcare. Early detection of any disease leads to early medication. Our project is determined to solve one such real world problem i.e., early detection of chronic kidney disease, with a good accuracy and at a faster rate compared to the existing manual process. The aim of our project is to see how Machine learning algorithms such as the Support vector machine (SVM), Logistic Regression and K-Nearest Neighbor (KNN) classifiers help us to detect CKD, compare their accuracies and to use the best model among them to build a web application.

Index Terms— Chronic Kidney Disease, Support Vector Machine, Data Mining, K-Nearest Neighbor, Label Encoding, Logistic Regression, Correlation.

I. INTRODUCTION

The kidneys are an important part of the human body, they perform excretion and osmoregulation. Chronic kidney disease (CKD) or renal failure is a serious kidney disease that leads to loss of kidney functions over several years. It leads to complete kidney failure. It is therefore crucial to detect CKD on its onset, but it is not that easy as the symptoms grow gradually over the period and these symptoms are not explicit to kidney disease. Also, few patients show no symptoms. Hence, machine learning can help us to predict if the patient is having CKD or not. Data mining has a very important role in extracting hidden insights from large clinical datasets that doctors collect regularly which has the details of diagnosis of patient's symptoms, which helps them to decide the treatment to be given.

Using machine learning to predict disease in healthcare is quite challenging but would be of great help to doctors. Our project aims to use Machine learning algorithms for early CKD prediction and compare them based on accuracy, precision, f-score and recall. In the year 2017, the count of people who died due to CKD was 1.2 million worldwide and the number of cases reported was 697.5 million. More number of CKD cases were reported in countries which are in lower levels of the Socio Demographic Index (SDI). In India, Chronic Kidney Disease is a serious health issue. About 1,75,000 cases of stage 5 CKD are reported every year. It is important to accurately predict renal failure as it is referred to as a dangerous disease. This work provides an app-based solution to detect CKD using a machine learning algorithm. Information given by this model also helps in monitoring people having the risk of getting CKD, early identification of CKD and early treatments. The advantage of using this system is that the process of predicting can be done in a very short time. It will help doctors to begin treatment early and also helps in diagnosing more patients in less time.

II. LITERATURE SURVEY

The research paper [1] uses a set of real-world data available in the UCI Machine Learning repository having records of 400 people with 25 attributes. In this paper they have selected 14 important attributes from the dataset collected and predicted the accuracy using Support Vector machine and Decision tree algorithms. Decision Tree algorithm is a supervised algorithm in which data is split according to some parameter upto the last attribute. A tree can be defined by decision nodes which are intermediate nodes and leaves which are the last nodes in the decision tree. The leaves represent final decisions and the decision nodes are the part where the data is divided. From the analysis of the results, it is clear that the decision tree provides 91.75% accuracy and SVM gives an accuracy of 96.75%. Advantage of using this system is that prediction takes very less time. This system will aid doctors to treat CKD patients and also aid in diagnosing more patients in a short duration.

The aim of the proposed model [2] is to build a web-based application using naive bayes machine learning algorithm to predict CKD or not. This application can be accessed by Physicians, Receptionist, Management and even patients. This model uses the CKD dataset available in UCI repository. They have tested using other algorithms such as KNN (K Nearest Neighbour Algorithm), SVM (Support Vector Machines), Decision Trees and ANN (Artificial Neural Network) but Naïve Bayes provided amazing results with superior accuracy in comparison to other algorithms. Accuracy obtained by using naive bayes is about 94.6%.

The solution to CKD is explored through three different classification algorithms: random forest, KNN and neural networks in [3]. They have used feature selection methods like: the wrapper method and LASSO regularization to avoid overfitting and to select important attributes. Through this evaluation they have found that by using the reduced set of 12 important features, random forest detects CKD with greater accuracy of 99.8% among all three algorithms.

The authors of paper [4] have compared the performance of SVM with the KNN algorithm using the measures like accuracy, F-measure, precision for prediction of CKD. The dataset was taken from the UCI repository. In the proposed method, output obtained is binary classification predicting CKD or NOT CKD. From these experimental results, observation shows that KNN classifier is better compared to SVM. Support Vector Machines (SVM) is an effective, advanced algorithm based on non-linear and linear regression. The K-Nearest Neighbour algorithm (K-NN) is a machine learning algorithm that is non-parametric mainly used in regression and classification problems.

The main purpose of their research in [5] was to screen people into two categories: CKD and NOT CKD on the basis of the level of potassium in an individual's blood using data mining as well as machine learning techniques. The proposed method uses various ML algorithms like KNN and Logistic regression for predicting CKD as well as determining the method's accuracy and performance. KNN is an algorithm that stores the data of all cases and then it predicts for a new unknown case based on the similarity scale. The algorithm focuses on two distance calculating mechanisms: 1)Euclidean 2)Manhattan distance to find the minimum distance between any two points. Logistic regression is a method that is used for classification of data sets and as output, it results in two possible values i.e., {0,1} or {True or False} as its output. This research paper used a set of real-world data available in the UCI Machine Learning repository. A record of 400 people, 25 qualities.

The study in paper [6] proposes a machine learning mechanism to diagnose CKD. The used set of data is collected from the UCI ML repository. They used a Six set of Machine Learning algorithms: KNN, Random Forest, Feed Forward Neural Network, Logistic regression, Naive Bayes and SVM. The KNN algorithm was used for data pre-processing work of filling in the unavailable or missing. Tree based models generate many decision trees. SVM splits different types of samples by establishing a decision surface. KNN receives the closest among the given training samples to the unknown new sample which is the test sample, it then calculates the Euclidean or Manhattan distance between training and testing sample points. Later it decides the category of test sample by majority voting. Neural networks can analyze non-equitable relationships in data sets due to their complex structure. Naive Bayes determines the Conditional probability of a sample with CKD and NOT CKD samples for each different measurement period. Among these ML models, the Random-forest has achieved excellent performance by 99.75% accuracy.

In [7] a set of models based on Artificial Intelligence (AI) are developed and evaluated. By including a few variables like: medications, age, comorbidities and gender. These models were intended to predict if a patient is at risk of getting kidney disease in a very near time of about 6 to 12 months. Seeing what are the most informative and important attributes for the prediction, they looked at the LightGBM model. Among all available attributes, diabetes, age, viral load, and the use of sulfonamides and angiotensin, were most prominent. Among all the models, convolutional neural networks (CNN) was found to be the best performer, with AUROC metrics of 0.957 and 0.954 for 6 and 12 months, respectively.

In paper [8], they have developed and compared nine different ML models: neural network, RF, ridge regression, XGBoost, lasso regression SVM, LR, Elastic Net, K-Nearest Neighbor, to predict CKD. LR has accuracy of 82%, Elastic Net with 82%, Lasso with an accuracy of 81%, Ridge with 81%, SVM with 81% accuracy, RF with 80%, Knn 74%, nn 80% and XGBoost with 83% accuracy.

The paper [9] aims to improve the accuracy of CKD diagnoses using the Random subspace and Bagging ensemble approach on 3 base algorithms: KNN, Decision Trees and Naive Bayes. The proposed model to detect chronic disease contains various small modules including: data processing, combined model prediction, modeling and testing, performance evaluation. The database is compiled from the UCI repository. The database target consists of 2 classes that are CKD and NOT CKD. 70% of databases are considered as a training set. Three different classifiers such as: Decision Trees, Naive Bayes and KNN were used for training the model and their respective accuracy obtained were recorded. Naive Bayes is a Bayesian Classifier derived from the concept of Bayes theorem and is used to solve classification problems.

In [10] an attempt has been made to develop a model to predict End-stage renal disease (ESRD). It is the foremost serious and last-stage of CKD (fifth stage). At this stage, kidney functionality reduces to 10%-15% compared to their original functionality. Around 550,000 patient reports were analysed. They have implemented 4-Fold Cross-validation and have used a filter method for selecting the features. The most important features are: chronic conditions, medication features, medical costs, diagnosis and procedure features, demographics, and episode counts. Performance of the model is evaluated by splitting data to training and testing with a ratio 80:20. Summarizing the result, the c-statistics for the model obtained is 0.93 having sensitivity value of 0.715, specificity value of 0.958.

In paper [11], CKD is predicted by the Apriori association technique on algorithms like OneR, naive Bayes, ZeroR, J48 and KNN. 10-fold cross validation is applied. ZeroR depends on the target and ignores all predictors. It is the classification technique and predicts the majority class. OneR (One Rule) creates one rule for every predictor and chooses the rule where the total error is minimum. Decision tree J48 is the implementation of ID3 (Iterative Dichotomiser 3). The results of KNN, J48, OneR, and Naive Bayes were improved by Apriori; however, it did not make any difference to ZeroR. 99% accuracy has been recorded for KNN with Apriori.

In paper [12], Ant Colony Optimization (ACO) and the Support Vector Machine (SVM) classifier machine learning techniques are used. It uses a minimum number of features and evaluates if CKD is present or not present in a patient. The Dataset used is from the UCI repository, which has 400 samples of two different classes. In the dataset, to choose some of the important features among all the available features, ACO technique is used. To predict the disease and its performance Support Vector Machine (SVM) is used. When the training set of data is given, the model is built that categorizes any unknown point into any of these classes. Performance is described using the confusion matrix and 96 percent accuracy is obtained.

The study in paper [13] used Random Forest, XGBoost, and ResNet machine learning models for predicting the serum creatinine value which in turn helps to predict CKD. ResNet is used for regression functions. XGBoost is a set of gradient boosting algorithms. It is used for good performance and speed. Random forest is a collection of learning approaches that builds decision trees and as a result gives the average in regression of each tree. Results have been averaged from eight predictors for a good accuracy of creatinine predictor. This is known as ensemble learning. While classifying the samples for CKD, the area under the ROC curve of 0.76 is obtained.

In study in paper [14], they have created a Prediction model using the Python programming language. They have applied two classification models Naive Bayes and K-Nearest Neighbor (KNN) to the data set collected from UCI Repository. It is a dataset of 400 people with 25 attributes. K-Nearest Neighbour is a supervised machine learning algorithm which classifies new data on the basis of similarity. The Naive Bayes is an algorithm-based Bayes theorem. It is used for solving classification problems which have a high-dimensional training dataset. It is a supervised learning algorithm based on the. The results show that the Naive Bayes Algorithm has an accuracy of 96.25%, whereas KNN gives an accuracy of 100%.

In paper [15], they have used different algorithms such as K-Nearest Neighbours, Ensemble Voting Classifier, Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, Gradient Boosting, SGD (Stochastic Gradient Descent) Classifier. Random Forest (RF) and SVM are supervised machine learning algorithms which are mainly used for classification. The accuracy of the algorithms are as follows: KNN 69%, SVM 75%, SGD Classifier 75%, Random Forest 74%, Decision Tree 72%, Ensemble Voting Classifier 74%, Logistic Regression 75%, Gradient Boosting 74%.

III. IMPLEMENTATION

A. Dataset And Attributes

The dataset we used in this project is CKD dataset from UCI repository. It contains information of 400 samples each having 26 attributes in it. These attributes involve patient details such as age, albumin, serum creatinine etc. which are of data types int, float, object along with some missing values.

B. Correlation Matrix

Correlation matrix helps us understand how the attributes are related to each other and also to the target variable. It shows dependency or statistical relationship between variables.

From the figure 1 we can clearly see that hemo and pcv are highly correlated (correlation value - 0.9) so we can eliminate one of them (we have eliminated pcv) and also, we can see that pot has low correlation with the target variable (correlation value - 0.085). Hence, we have dropped pcv and pod attributes.

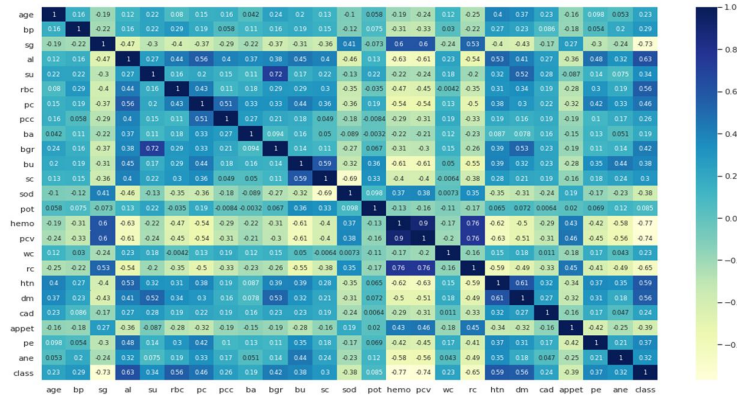


Figure 1. Correlation Matrix

C. Data Pre-Processing

- **Label encoding:** It is used to convert categorical data such as “Yes”, “no”; “good”, “poor”; “present”, “not present” into numerical values. This is done to convert our dataset into machine readable form.
- **Filling the missing dataset:** A missing dataset may lead to loss of important information and may lead to some invalid conclusions. Hence, we used Median-imputation to replace the missing values in our dataset.
- **Feature Scaling:** The dataset contains values of various ranges, scaling from very high to low. Hence there is a chance that attributes with a large scale may get more weightage than a more important attribute with a small scale. Therefore, it is necessary to standardize the features present in the dataset. We have used the StandardScaler() function imported from sklearn library to standardize the values.
- **Training and Testing Dataset:** The entire dataset is split into training and testing dataset. 70% of data is taken for training purposes and the remaining 30% for testing.

IV. ML ALGORITHMS AND RESULTS

A. Logistic Regression

It is a supervised classification algorithm. It is a very simple and most commonly used classification algorithm. Initially we imported a Logistic regression class from the sklearn library. Then created a classifier object lg. To estimate the model's efficiency, we used a test data set and obtained an accuracy of **99.16%**.

The AUC value varies from 0.5 to 1 where 0.5 indicates that our classifier is bad and 1 denotes that our classifier is able to accurately distinguish between the two classes. Higher the AUC score, better our model is at distinguishing between patients with CKD and not CKD. From the Figure 2, the roc_auc_score obtained for Logistic Regression is **0.988**. We created a confusion matrix as shown in Figure 3 which indicates the number of correct and incorrect predictions made by the logistic regression classifier. Here 42 True positive and 77 True Negative values are obtained i.e., 42+77=119 correct predictions were made whereas 1 False Negative and 0 False Positive i.e., 1 incorrect prediction was obtained. This model gave an accuracy of **99.16%**.

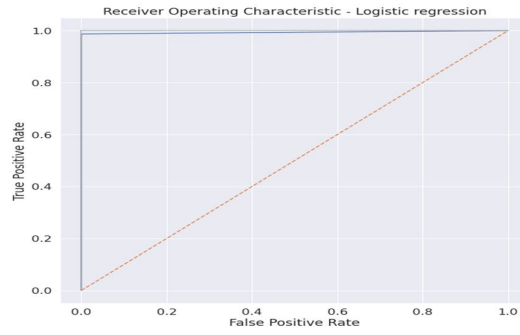


Figure 2.Receiving Operating Characteristic (ROC) curve



Figure 3. Confusion Matrix for Logistic Regression

TABLE I. THE CLASSIFICATION REPORT OF LOGISTIC REGRESSION MODEL

	Precision	Recall	F-Score	Support
0.0	0.96	1.00	0.98	27
1.0	1.00	0.98	0.99	53
Accuracy			0.99	80
Macro avg	0.98	0.99	0.99	80
Weighted avg	0.99	0.99	0.99	80

B. Support Vector Machine

The main goal of SVM is to create the best possible boundary that can separate n-dimensional space into different classes. So that whenever we encounter a new datapoint, we can precisely classify it. First, we imported the svc class from the sklearn.svm library. We used kernel="linear" for linearly separable data. The accuracy of the model was estimated by predicting the values of the test data set, accuracy came out to be **97.5%**. Confusion matrix obtained is as shown in Figure 4 which indicates the number of correct and incorrect predictions made by the logistic regression classifier. Here 42 True positive and 78 True Negative values are obtained i.e., 42+78=120 correct predictions were made where as 0 False Negative and 0 False Positive i.e., no incorrect prediction was obtained. The total accuracy of this model is obtained as **97.5%**.

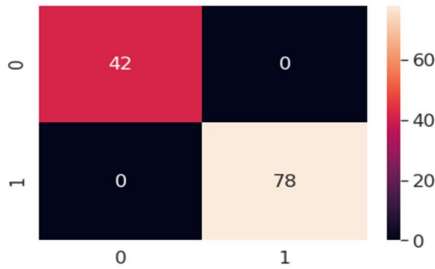


Figure 4. Confusion matrix for Support Vector Machine

TABLE II. THE CLASSIFICATION REPORT OF SUPPORT VECTOR MACHINE MODEL

	Precision	Recall	F-Score	Support
0.0	1.00	1.00	1.00	42
1.0	1.00	1.00	1.00	78
Accuracy			1.00	120
Macro avg	1.00	1.00	1.00	120
Weighted avg	1.00	1.00	1.00	120

C. K Nearest Neighbors

KNN is a supervised machine learning algorithm. When an unknown data point is encountered, it calculates the distance of that point to already classified points to find out k nearest neighbors and hence classifies the new data point. We used 3 nearest neighbors, leaf size=30 for this model. The model's accuracy tested using test dataset is about **97.5%**. Confusion matrix obtained for KNN is as shown in Figure 5 which indicates the number of correct and incorrect predictions made by the logistic regression classifier. Here 42 True positive and 76 True Negative values are obtained i.e., 42+76=118 correct predictions were made whereas 2 False Negative and 0 False

Positive i.e., 2 incorrect prediction was obtained.

Precision, Recall, and F-score is calculated in Table 2, Table 3 and Table 4 using the formulas:

$$\text{Precision} = (\text{True Positive}) / (\text{True Positive} + \text{False Positive})$$

$$\text{Recall} = (\text{True Positive}) / (\text{True Positive} + \text{False Negative})$$

$$\text{F-Score} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

In Table I, Table II and Table III, Presence of CKD is denoted as 1.0 and Absence of CKD is denoted as 0.0.

Table IV gives the accuracy obtained

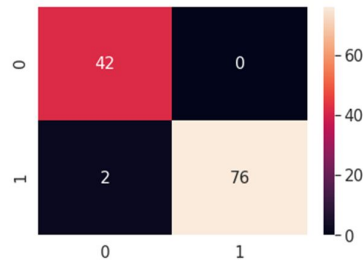


Figure 5. Confusion matrix for KNN

TABLE III. THE CLASSIFICATION REPORT OF KNN MODEL

	Precision	Recall	F-Score	Support
0.0	0.95	1.00	0.98	42
1.0	1.00	0.97	0.99	78
Accuracy			0.98	120
Macro avg	0.98	0.99	0.98	120
Weighted avg	0.98	0.98	0.98	120

TABLE IV. ACCURACIES OF ALL USED MODELS

Sl. no	Algorithm	Accuracy
1	Logistic Regression	99.1
2	Support Vector Machine	97.5
3	K Nearest Neighbour's	97.5

V. CONCLUSION

This paper gives the classification, whether patients have CKD or not using various machine learning algorithms by analysing different attributes of CKD patients. We have used Logistic regression, SVM, KNN algorithms and found that the classification model built using Logistic regression gave the best results with an accuracy of 99.1%. The main advantage of this system is that CKD in a patient can be detected early with a good accuracy rate. Hence, it aids doctors in diagnosing more patients in a short duration.

REFERENCES

- [1] Siddheshwar Tekale, Pranjal Shingavi, Sukanya Wandhekar, Ankit Chatorikar. "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm". In 2018 International Journal of Advanced Research in Computer and Communication Engineering.
- [2] Amogh Babu K A, Priyanka K, Raghavendra Babu T M." Chronic Kidney Disease Prediction Based On Naïve Bayes Technique". In 2019 International Research Journal of Engineering and Technology (IRJET).
- [3] Salekin, Asif, and John Stankovic. "Detection of chronic kidney disease and selecting important predictive attributes." In 2016 IEEE International Conference on Healthcare Informatics (ICHI), pp. 262-270. IEEE, 2016.
- [4] Sinha, Parul, and Poonam Sinha. "Comparative study of chronic kidney disease prediction using KNN and SVM." International Journal of Engineering Research and Technology 4, no. 12 (2015): 608-12.
- [5] S.D.Harish, K.Vinay Kumar, K.Taraka Ram, G.Pradeepini."Chronic Kidney Disease Prediction based on Blood Potassium Levels using Machine Learning". In [2019] International Journal of Engineering Research & Technology (IJERT).
- [6] Qin, Jiongming, Lin Chen, Yuhua Liu, Chuanjun Liu, Changhao Feng, and Bin Chen. "A machine learning methodology for diagnosing chronic kidney disease." IEEE Access 8 (2019): 20991-21002.
- [7] Krishnamurthy, Surya, Kapelesh Ks, Erik Dovgan, Mitja Luštrek, Barbara Gradišek Piletič, Kathiravan Srinivasan, Yu-Chuan Jack Li, Anton Gradišek, and Shabbir Syed-Abdul. "Machine Learning Prediction Models for Chronic Kidney Disease using National Health Insurance Claim Data in Taiwan." In Healthcare, vol. 9, no. 5, p. 546. Multidisciplinary Digital Publishing Institute, 2021.
- [8] Xiao, Jing, Ruifeng Ding, Xiulin Xu, Haochen Guan, Xinhui Feng, Tao Sun, Sibozhu, and Zhibin Ye. "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression." Journal of translational medicine 17, no. 1 (2019): 1-13.

- [9] Jongbo, Olayinka Ayodele, Adebayo Olusola Adetunmbi, Roseline Bosede Ogunrinde, and Bukola Badeji-Ajisafe. "Development of an ensemble approach to chronic kidney disease diagnosis." *Scientific African* 8 (2020): e00456.
- [10] Segal, Zvi, Dan Kalifa, Kira Radinsky, Bar Ehrenberg, Guy Elad, Gal Maor, Maor Lewis, Muhammad Tibi, Liat Korn, and Gideon Koren. "Machine learning algorithm for early detection of end-stage renal disease." *BMC nephrology* 21, no. 1 (2020): 1-10.
- [11] Zixian Wang, Jae Won Chung, Xilin Jiang, Yantong Cui, Muning Wang, Anqi Zheng, "Machine Learning- Based Prediction System For Chronic Kidney Disease Using Associative Classification Technique." 2020 *International Journal of Engineering & Technology*.
- [12] Reshma S, Salma Shaji, S R Ajina, Vishnu Priya S R, Janisha A. "Chronic Kidney Disease Prediction using Machine Learning." In 2020 *International Journal of Engineering Research & Technology (IJERT)*.
- [13] Wang, Weilun, Goutam Chakraborty, and Basabi Chakraborty. "Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm." *Applied Sciences* 11, no. 1 (2021): 202.
- [14] Sujata Drall, Gurdeep Singh Drall, Sugandha Singh, Bharat Bhushan Naib. "Chronic Kidney Disease Prediction Using Machine Learning: A New Approach" 2018.
- [15] Bhattacharyya, Debnath, Bhanu Prakash Doppala, and N. Thirupathi Rao. "Prediction and Forecasting o Persistent Kidney Problems Using Machine Learning Algorithms." *Int J Cur Res Rev* Vol 12, no. 20 (2020):134.