# Chronic Kidney Diseases with Machine Learning

Kriti Sharma

Galgotias University

Greater Noida, Uttar Pradesh-201310

kriti_sharma.scsebtech@galgotiasuniversity.edu.in

Ayushi Kulshrestha

Galgotias University

Greater Noida, Uttar Pradesh-201310

ayushi_kulshrestha.scsebtech@galgotiasuniversity.edu.in

**Abstract** — End stage renal disease (ESRD) describes the most severe stage of chronic kidney disease (CKD), when patients need dialysis or renal transplant. Unhealthy lifestyle is the primary reason for Chronic Kidney Disease(CKD) , CKD being the condition that leads to reduced renal function expanding over a period of months or years. It is rapidly becoming a concern of global health crises. It processes in six stages according to severity level. Chronic Kidney diseases are usually detected at a stage where the diagnosis becomes next to impossible. So for the combat of disease early prediction is necessary to provide the patient with necessary and required treatment. This study proposes the use of machine learning techniques for CKD prediction. The main purpose of this study is to provide a sustainable method that predicts CKD at an early stage with medical accuracy b creating an ensemble machine learning model including algorithms such as KNN, Decision tree, Random forest,XGBoost, Stochastic Gradient Boosting, Gradient Boosting and Extra trees. The random forest was found to be showing the highest accuracy of 0.9966 and the overall model gives an accuracy of 98 percent which is 1 percent higher than the previous studies conducted in the same domain.

Keywords: End stage renal disease, Machine learning, Prediction model, GFR, Decision Trees, Random Forests, KNN, Logistic Regression.

## 1. INTRODUCTION

Chronic Kidney Disease (CKD) is a major burden in today's time. It is majorly contributing to the global health crisis. Machine learning has the best features and libraries for the prediction of things related to clinical data. In this model data pre-processing is used for managing missing values, data aggregation and feature extraction. Various algorithms have been used in the model like KNN, Decision Tree, Gradient Boosting, Stochastic Boosting, Random Forest and extra trees. Random Forest being the most accurate of all. 10% of the global population is diagnosed with major health diseases , major part of which is dealing with Kidney diseases. While eventually leads to kidney failure.According to Global Burden of Diseases Study GBDS 1990[10],CKD was ranked 29th which has jumped to 18th rank in the last decade as the recent study of GBDS that was conducted in 2019 suggests. CKD is projected to become the 5th most common cause of death by 2040.10 percent of the global population suffers from CKD. It is estimated that the number of cases of kidney failure will increase disproportionately in developing countries like China and India.On the other hand, 15 percent of the US which is about 37 million people have CKD. CKD has affected 500 million people worldwide. It is inferred that people who have CKD are more likely to get End Stage Renal Disease which requires extensive treatments like dialysis and transplantation.The UCI clinical dataset of patients has multiple features which are selected in order to obtain high accuracy than the previous models by employing ensemble learning that uses machine learning algorithms like Decision Trees, K Nearest Neighbor, Gradient Boosting,Stochastic Gradient Boosting, Extra Trees and Random Forests. The features or attributes that are used to determine the accuracy of CKD are blood pressure, sugar, blood, urea, serum creatinine, potassium, white blood cell count, hypertension and albumin.These values are found to have positive correlation values that are swinging between 0.2 to 0.8. It is also inferred that serum creatinine is found to be very less significant in early stages of CKD where as in later stages it is found to be one of the important determining factors of CKD.From medical perspective, hypertension creates CKD and specific gravity has 0.73 correlation. The dataset is firstly preprocessed and splitted into training and testing dataset into 80/20 ratio. The features selected are on the basis of their correlation values, if they are positive then they are fit to be used else negative values are discarded. The final features obtained from the heatmap are albumin, blood pressure,hypertension,specific gravity,etc and its values are varying between 0.2 to 0.8. The proposed model consists of an ensemble machine learning model that is a combination of machine learning models such Decision Trees, Random Forest, XGBoost, Stochastic Gradient Boosting,Gradient Boosting and Extra Trees. The output of all the algorithms of their respective predictions are shown altogether in the graph so that for each stage,comparisons can be drawn between algorithms simultaneously and high accuracy results can be concluded.The system is examined and evaluated through multiclass statistical analysis, and the empirical results of KNN, decision tree, gradient boosting, stochastic boosting, random forest and extra tree algorithms found significant values of 95.8%, 97.5%, 98.3%, 99.6%, 99.6%, 99.6% with respect to accuracy metrics.The random forest algorithm outperformed all other algorithms, achieving an accuracy and precision of 99.6% for all measures.

## 2. LITERATURE REVIEW

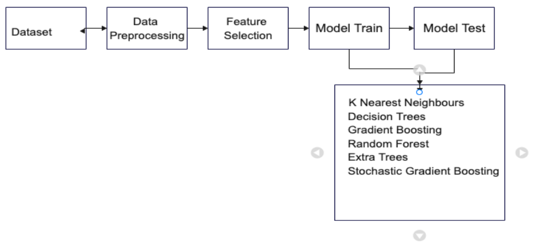
Chronic kidney disease has become a major burden worldwide and it is becoming a leading contributor to the global health crisis.People with CKD are highly prone to getting ESRD(End Stage Renal Disease) which requires extensive treatment like dialysis and transplantation.It also brings about highly adverse effects on one's psychological health. It also adds to a lot of financial load and therefore, its early prediction is highly necessary.. In Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods[1] ,a comparative analysis is carried out between three machine learning algorithms such as Decision Trees(DT), Logistic Regression(LR) and K Nearest Neighbor(KNN). This study proposed a system that included preprocessing of data,feature selection and then finally applying the ML algorithm.The data preprocessing removed all the noise and outliers so that it does not cause the model to deviate from the proper training set The data is cleaned,checked for null values and prepared for model construction.For proper training and testing, the data is split into 80/ 20 ratio of training to testing.The next step involved feature selection that is done using a heat map. The absolute values of correlation between features and class labels like blood pressure, albumin,sugar,blood,urea,serum creatinine,potassium,white blood cell count and hypertension all have positive links.All the correlations with positive values are considered The values varied between 0.2 to 0.8. From a medical perspective,hypertension causes CKD and specific gravity has a 0.73 connection to CKD. Albumin is one such protein that is assessed using the urine protein test.If high levels of albumin are present in the urine,it indicates that the filtration units in the kidneys known as nephrons are damaged.However, in order to establish the diagnosis, numerous tests need to be carried out during a course of many weeks. Creatinine is another such chemical compound that is obtained as a by-product of muscle breakdown of chemical creatine.If it is present in high quantities, it is inferred that the intake of protein diet is high,the person has diabetic issues, dehydration etc.Creatinine levels in women should be between 0.6 and 1.1 mg/dL, while those in males should be between 0.7 and 1.3 mg/dL.LR was found to be the most accurate out of the three ML Techniques in predicting CKD with an accuracy of 97%. DT gave the second best result with an accuracy of 96.25% followed by KNN with an accuracy of 71.25%. The novelty of this paper is that using the above mentioned ML techniques, it produced a prediction model with accuracy of 97%.In Chronic Kidney Diseases Prediction Using Machine Learning [12],the proposed system consists of the publicly available dataset using UCI repository which contains 400 samples of two different classes i.e. CKD and Non-CKD.there 25 attributes in the dataset out of which 11 are numeric, 13 are nominal and one is class attribute. It is inferred that CKD is caused due to diabetes and high blood pressure.For Classification, Support Vector Machine algorithm has been used to predict the disease and its performance. The proposed model also uses libraries from scikit-learn. NumPy is leveraged to perform the mathematical computations. The dataset is divided into training and testing sets. The model uses a wrapper method for feature selection using Ant Colony Optimization[4]. ACO is a meta-heuristic optimization algorithm.SVM classifies the output into two classes i.e. CKD and Non-CKD. The overall objective of this study was to use less number of determining attributes to predict the patients having CKD and achieved an accuracy of 96 percentage.In Chronic Kidney Disease Diagnosis Using Decision Tree Algorithms[5], a comparative study has been carried out between various classification algorithms.The dataset used is from University of California Irvine(UCI).It comprises of patient data which is then preprocessed and four attributes are selected such as age,race,sex, and serum creatinine that are used as in input to calculate the Glomerular Filtration Rate.(GFR)It uses CKD-EPI equation since it is more reliable to calculate the value of GFR.

GFR = $175 * SCr^{-1.154} * age^{-0.0203} * 0.742$(if female)

This equation can be used to calculate GFR in all stages of CKD.It is known to be appropriate only when the GFR> 60 which is in the later stages of CKD.The model incorporates predictive algorithms such as LASSO,Logistic Regression,Elastic Net, KNN, Random Forests and Extra trees. The results showed that Probabilistic Neural Models showed the highest accuracy of 96.7 percentage followed by Radial Basis Function and Logistic Regression with accuracy of 85.3 and 82 percentage respectively.

In Machine Learning Prediction Models for Chronic Kidney Disease Using National Health Insurance Claim Data in Taiwan,the model predicts a patient's risk of developing chronic kidney diseases after a period. The ML based models could be efficiently initiating public health initiatives such as early detection of CKD. The recursive feature elimination approach is used to identify which feature is most appropriate for prediction.The most important CKD features are red blood cells count, cell volume, specific gravity, hypertension and hemoglobin.The random forest classifier outperformed all the other classifiers like KNN, Decision Tree etc. The limitation of the proposed model was that it had been tested on small data sets only. The clinical data is collected from pathologist's and kaggle datasets.

## 3. PROPOSED MODEL



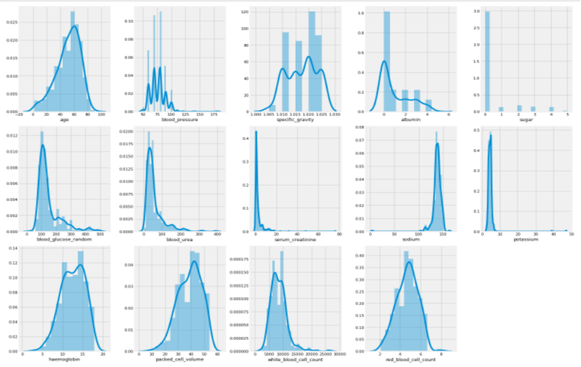
The preprocessing stage included estimating missing values and eliminating noise, such as outliers, normalization, and checking of unbalanced data. Some measurements may be missed when patients are undergoing tests, thereby causing missing values. The dataset has missing values. The simplest method to handle missing values is to ignore the record, but it is inappropriate with a small dataset. We can use algorithms to compute missing values instead of removing records. The missing values for numericals features can be computed through one of the statistical measures, such as mean, median, and standard deviation. However, the missing values of nominal features can be computed using the model method. In which the missing value is replaced, the value is replaced by the most common value of the feature. In this study, the missing numerical features were replaced by the mean method, and a mode method was applied to the place missing nominal features. The statistical analysis of the dataset, such as mean and standard deviation; max and min were introduced for the numerical features in the dataset.

Dataset

Kaggle University of California Irvine dataset was used to gather CKD data. There are 400 patient records in the data set, and some values are missing. It comprises various clinical qualities that emerge in the prognosis of chronic kidney disease, with the class attributes serving as the result of the patient's level of chronic renal failure being predicted.

Data processing

The estimation of missing values, as well as the normalization and validation of unbalanced data, were all part of the preprocessing stages. When assessing a patient, some measurements could be missing or incomplete.



Handling Missing Values

There are many completed cases in the data collection , with the remainder missing. Ignoring records is the simplest technique to deal with missing values; however this is the only practical way for small datasets. For large datasets like this the missing values are handled with the help of mean and mode method and random sampling method. The mode method was used to replace the missing values of nominal features.

Categorical Data Encoding

Because most machine learning algorithms only accept numeric values as input, category values must be encoded into numerical values. The binary values “0” and “1” are used to represent the characteristics of categories such as “no” and “yes”.

Data Transformation

It is the process of transforming numbers on the scale so that one variable does not dominate the others. It alters the values in the data set so that they can be processed further. To improve the accuracy of machine learning models, this research requires a data normalization technique. It converts data between the -1 and +1 ranges. The converted data has a standard deviation of 1 and a mean of 0.

The formula is given below:

$$\omega = \frac{(x-\overline{x})}{\sigma}$$

ω = Standardized score

X = Observed value

$\overline{x}$ = Mean

σ = Standard Deviation

**Feature Selection**

After computing the missing values, identifying the important features having a strong and positive correlation with features of importance for disease diagnosis is required. Extracting the vector features elimi- nates useless features for prediction and those that are ir- relevant, which prevents the construction of a robust diagnostic model [25]. In this study, we used the RFE method to extract the most important features of a pre- diction. The Recursive Feature Elimination (RFE) algorithm is very popular due to its ease of use and configurations and its effectiveness in selecting features in training datasets relevant to predicting target variables and eliminating weak features. The RFE method is used to select the most sig- nificant features by finding high correlation between specific features and target (labels). RFECV lets the number of features in the dataset along with a cross-vali- dated score and visualizes the selected features is presented in Figure 3. 2.4. Classification. Data mining techniques have been used to define new and understandable patterns to construct classification templates [26]. Supervised and unsupervised learning techniques require the construction of models based on prior analysis and are used in medical and clinical diagnostics for classification and regression [27]. Four popular machine learning algorithms used are KNN, decision tree, gradient boosting and extra trees, which give the best diagnostic results. Machine learning techniques work to build predictive/classification models through two stages: the training phase, in which a model is constructed from a set of training data with the expected outputs, and the validation stage, which estimates the quality of the trained models from the validation dataset without the expected output. Algorithms are supervised algorithms that are used to solve classification and regression problems.

**Decision Tree Algorithm**

Decision tree is a type of supervised learning algorithm (having a predefined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on the most significant splitter / differentiator in input variables.

TYPES OF DECISION TREE

1. Categorical Variable Decision Tree: Decision Tree which has a categorical target variable then it is called a categorical variable decision tree.
2. Continuous Variable Decision Tree: Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree

TERMINOLOGY OF DECISION TREE:

1. Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
2. Splitting: It is a process of dividing a node into two or more sub-nodes.
3. Decision Node: When a sub-node splits into further sub-nodes, then it is called a decision node.
4. Leaf/ Terminal Node: Nodes that do not split are called Leaf or Terminal nodes.
5. Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
6. Branch / Sub-Tree: A subsection of an entire tree is called branch or sub-tree.
7. Parent and Child Node: A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.

WORKING OF DECISION TREE

Decision trees use multiple algorithms to decide to split a node into two or more sub- nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

1. Gini Index
2. Information Gain
3. Chi Square
4. Reduction of Variance

**K-Nearest Number (KNN)**

The KNN algorithm recognizes similarities between new and previous data points and categorizes fresh test points into existing related groups. The KNN method is a slow learning algorithm since it is not parametric. This means that instead of learning from Diagnostics 2022, 12, 116 8 of 22 the training data set, it should be secured. It uses K to categorize the data. The distance between the new location and the saved training point was determined using the Euclidean distance.

**Random Forest Classifier**

The random forest algorithm is based on ensemble learning, improving the model’s performance, and solving complex problems by combining several classifiers. A classifier named after the algorithm that contains multiple decision trees averaged over a database subset to improve predictions. In the forecasting process, it does not rely on a single decision tree, and the random forest algorithm creates a forecast from each decision tree that predicts the conclusion based on the majority of decision votes. The usage of several trees decreases the possibility of the model overfitting. To predict the classes in the database, the algorithm includes many decision trees, some of which can predict the proper outcome while the others cannot. As a result, there are two assumptions regarding the prediction’s accuracy. To forecast a more accurate outcome than an estimate, the algorithm must first include the actual value of the feature variable. Second, there must be an extremely low correlation between the forecasts for each tree. As a result there are two requirements for high forecast accuracy.

## 4. RESULT/DISCUSSION

The proposed model was created using various features given in the data. The accuracy was calculated. The confusion matrix was also utilized to evaluate performance by using True Positives(TP), True Negatives(TN), False Positives(FP), and False Negatives(FN). All the null values are filled using random sampling in the confusion matrix to evaluate the performance.Twenty-four numerical and nominal features were introduced from 400 patients with CKD. Due to the neglect of some tests for some patients, some computation methods were applied to solve this problem. To solve the missing numerical values, the mean method was used; for missing nominal values, the mode method was used. There is a positive correlation, for example, between specific gravity with red blood cell count, packed cell volume, and hemoglobin; between sugar with blood glucose random; between blood urea and serum creatinine; and between hemoglobin with red blood cell count and packed cell volume. There is also a negative correlation, for example, between albumin and blood urea with red blood cell count, packed cell volume, and hemoglobin and between serum creatinine and sodium. CKD has been accurately categorized.

Accuracy

It refers to the proportion of correct guesses to total predictions. Accuracy can be described as the ability to accurately predict the outcome of a situation.

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

Recall

The recall calculates the proportion of accurately predicted positive observations to the total number of observations in the class, as shown in the following equation.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision

As stated in the equation below, this metric represents the proportion of accurately predicted positive observations to total predictive positive observations.

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

F-Measure

Precision and Recall are weighted averaged in the F-measure. False positives and false negatives are part of the process. F-measure is a term that is defined as

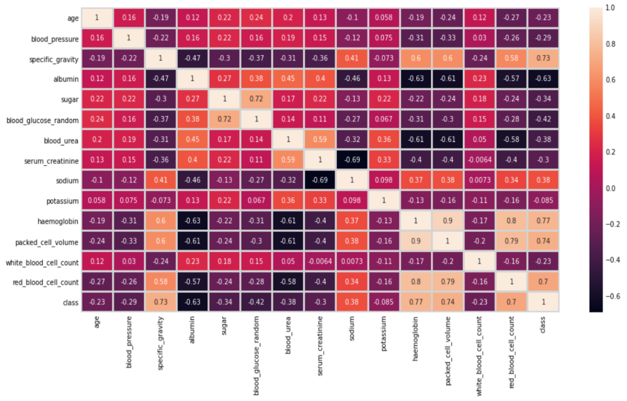
$$\text{F-Measure} = \frac{Two \times (Precision \times Recall)}{(Precision + Recall)}$$

The F-Measure values lie from 0 to 1.

## Comparative Analysis

In this section we present the proposed model. The data set is split into 70% training and 30% test data set

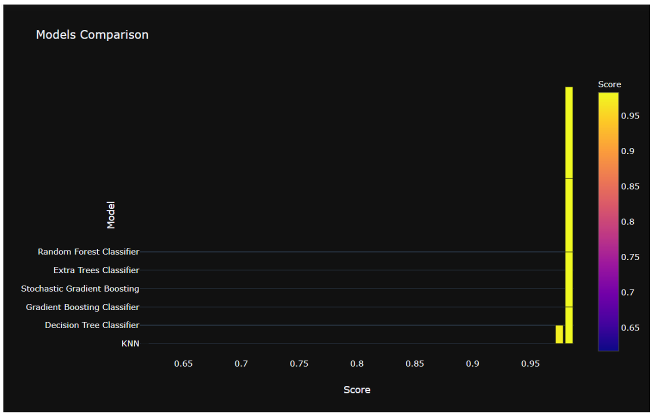
Heatmap of the whole data set is shown in the diagram.



The proposed model is compared with other algorithms, including KNN, Decision Tree, Gradient Boosting, Stochastic Gradient Boosting, Random Forest and Extra Trees. No parameter adjustments were made for the algorithms to show improved performance. The default values were used . All models are evaluated using f1 score. The table shows the result of the proposed models when tested on CKD data set.

| **Method** | **Recall** | **Precision** | **F-Measure** | **Accuracy** |
|---|---|---|---|---|
| KNN | 0.97 | 0.96 | 0.97 | 0.9583 |
| Decision Tree | 0.99 | 0.97 | 0.98 | 0.975 |
| Gradient Boosting | 1.00 | 0.97 | 0.98 | 0.9833 |
| Stochastic Gradient Boosting | 0.99 | 0.97 | 0.99 | 0.9964 |
| Extra Trees | 0.97 | 1.00 | 0.98 | 0.9963 |
| Random Forest | 1.00 | 0.98 | 0.99 | 0.9966 |

The figure shows the accuracy graph comparing the performance of the classification algorithms to the proposed approach for CKD prediction.



The accuracy of KNN, Decision Tree, Gradient Boosting, Stochastic Gradient Boosting, Random Forest and Extra Trees is 0.95, 0.97, 0.98, 0.994, 0.99, 0.996, respectively. The proposed model was found to be the most accurate, with a 99.6% accuracy rate. The results of all six classifiers are shown in the table.

The F-1 score, recall, and precision of KNN were all 0.97, 0.97, 0.96 respectively. The F-1 score, recall, and precision of Decision Tree were all 0.98, 0.99, 0.97 respectively. The F-1 score, recall, and precision of Gradient Boosting were all 0.98, 1.00, 0.97 respectively. The F-1 score, recall, and precision of Stochastic Gradient Boosting were all 0.99, 0.99, 0.97 respectively. The F-1 score, recall, and precision of Extra trees were all 0.98, 1.00, 0.97 respectively. The F-1 score, recall, and precision of Random forest were all 0.99, 1.00, 0.98 respectively.

## 5. CONCLUSION

A deep learning model for early diagnosis of chronic Kidney disease is presented in this paper. This study provided insight into the diagnosis of CKD patients to tackle their condition and receive treatment in the early stages of the disease. The dataset was collected from 400 patients containing 24 features. The dataset was divided into 70% training and 30% testing and validation. The dataset was processed to remove outliers and replace missing numerical and nominal values using mean and mode sta- tistical measures, respectively. The most essential CKD features are packed red blood cell count, albumin, cell volume, serum creatinine, specific gravity, hemoglobin, and hypertension. Different metrics, including classification accuracy, recall,precision and f-measure, are used for the estimation of comparative analysis. The RFE algorithm was applied to select the most strongly representative features of CKD. Selected features were fed into classification algorithms: KNN, decision tree, gradient boosting, stochastic gradient boosting, random forest and extra trees. The parameters of all classifiers were tuned to perform the best classification, so all algorithms reached promising results. The random forest  algorithm outperformed all other algorithms, achieving an accuracy and precision of 99.6% for all measures. The system was examined and evaluated through multiclass statistical analysis, and the empirical results of KNN, decision tree, gradient boosting, stochastic boosting, random forest and extra tree algorithms found significant values of 95.8%, 97.5%, 98.3%, 99.6%, 99.6%, 99.6% with respect to accuracy metrics.

## 6. REFERENCES

1. Gazi Mohammad Ifraz,Muhammad Hasnath Rashid, Tahia Tazin, Sami Bourouis and Mohammad Monirujjaman Khan, “Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Method”,Computational and Mathematical Methods in Medicine, vol. 2021, Article ID 6141470, 10 pages, 2021. https://doi.org/10.1155/2021/6141470
2. World Health Organization, Preventing Chronic Disease: A Vital Investment, WHO, Geneva, Switzerland, 2005.
3. B. Bikbov, N. Perico, and G. Remuzzi, “Disparities in chronic kidney disease prevalence among males and females in 195 countries: analysis of the global burden of disease 2016 study,” Nephron, vol. 139, no. 4, pp. 313–318, 2018.
4. https://towardsdatascience.com/the-inspiration-of-an-ant-colony-optimization-f377568ea03f
5. Ilyas H, Ali S, Ponum M, et al. Chronic kidney disease diagnosis using decision tree algorithms. BMC Nephrol. 2021;22(1):273. Published 2021 Aug 9. doi:10.1186/s12882-021-02474-z
6. Z. Chen, X. Zhang, and Z. Zhang, “Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models,” International Urology and Nephrology, vol. 48, no. 12, pp. 2069–2075, 2016.
7. Glomerular Filtration Rate (GFR), National Kidney Foun- dation, New York, NY, USA, 2020, https://www.kidney.org/ atoz/content/gfr.
8. T. H. Al Dhyani, A. S. Alshebani, and M. Y. Alzahrani, “Soft computing model to predict chronic diseases,” Information Science and Engineering, vol. 36, no. 2, pp. 365–376, 2020.
9. T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, “Support vector machine classification and validation of cancer tissue samples using microarray expression data,” Bioinformatics, vol. 16, no. 10, pp. 906–914, 2000.

10. https://www.healthdata.org/gbd/2019
11. R. M. Pujari and V. D. Hajare, “Analysis of ultrasound images for identification of Chronic Kidney Disease stages,” in Proceedings of the 2014 First International Conference on Networks & Soft Computing (ICNSC2014), pp. 380–383, IEEE, Guntur, India, August 2014.
12. S. Ahmed, M. T. Kabir, N. T. Mahmood, and R. M. Rahman, “Diagnosis of kidney disease using fuzzy expert system,” in Proceedings of the 8th International Conference on Software, Knowledge, Information Management and Applications
13. Journal of Healthcare Engineering
14. (SKIMA 2014), pp. 1–8, IEEE, Dhaka, Bangladesh, December 2014.
15. Reshma S , Salma Shaji , S R Ajina , Vishnu PriyaS R, Janisha A, 2020, Chronic Kidney “DiseasePrediction using Machine Learning,INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY” (IJERT) Volume 09, Issue 07 (July 2020)
16. Machine Learning Prediction Models for Chronic Kidney Disease Using National Health Insurance Claim Data in Taiwan(May 2021) by Surya Krishnamurthy ,Kapelesh KS ,Erik Dovgan ,Mitja Luštrek ,Barbara Gradišek Piletič ,Kathiravan Srinivasan ,Yu-Chuan (Jack) Li ,Anton Gradišek and Shabbir Syed-Abdul,
17. G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, and J. A. Recio-Garcia, “Explainable pre- diction of chronic renal disease in the Colombian population using neural networks and case-based reasoning,” IEEE Ac- cess, vol. 7, pp. 152900–152910, 2019.
18. N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Atheeb, R. E. Juri, and J. Alhiyafi, “Neural network and support vector machine for the prediction of chronic kidney disease: a comparative study,” Computers in Biology and Medicine, vol. 109, pp. 101–111, 2019.
19. E. H. A. Rady and A. S. Anwar, “Prediction of kidney disease stages using data mining algorithms,” Informatics in Medicine Unlocked, vol. 15, Article ID 100178, 2019.
20. V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, “Chronic kidney disease analysis using data mining classification techniques,” in Proceedings of the 2016 6th International Conference-Cloud System and Big Data Engineering (Con- fluence), pp. 300–305, IEEE, Noida, India, January 2016.
21. M. S. Wibawa, I. M. D. Maysanjaya, and I. M. A. W. Putra, “Boosted classifier and features selection for enhancing chronic kidney disease diagnose,” in Proceedings of the 2017 5th international conference on cyber and IT service man- agement (CITSM), pp. 1–6, IEEE, Denpasar, Indonesia, August 2017.
22. E. Avci, S. Karakus, O. Ozmen, and D. Avci, “Performance comparison of some classifiers on chronic kidney disease data,” in Proceedings of the 2018 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1–4, IEEE, Antalya, Turkey, March 2018.
23. R. K. Chiu, R. Y. Chen, S. A. Wang, Y. C. Chang, and L. C. Chen, “Intelligent systems developed for the early de- tection of chronic kidney disease,” Advances in Artificial Neural Systems, vol. 2013, 2013.
24. A. K. Shrivas, S. K. Sahu, and H. S. Hota, “Classification of chronic kidney disease with proposed union based feature selection technique,” SSRN Electronic Journal, vol. 26, 2018.
25. M. Elhoseny, K. Shankar, and J. Uthayakumar, “Intelligent diagnostic prediction and classification system for chronic kidney disease,” Scientific Reports, vol. 9, no. 1, pp. 1–14, 2019.
26. A. Abdelaziz, M. Elhoseny, A. S. Salama, and A. M. Riad, “A machine learning model for improving healthcare services on cloud computing environment,” Measurement, vol. 119, pp. 117–128, 2018.
27. C. Z. Xiong, M. Su, Z. Jiang, and W. Jiang, “Prediction of hemodialysis timing based on LVW feature selection and ensemble learning,” Journal of Medical Systems, vol. 43, no. 1, pp. 1–8, 2019.
28. S. Ravizza, T. Huschto, A. Adamov et al., “Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data,” Nature Medicine, vol. 25, no. 1, pp. 57–59, 2019.
29. S. B. V. Sara and K. Kalaiselvi, “Ensemble swarm behavior based feature selection and support vector machine classifier for chronic kidney disease prediction,” International Journal of Engineering & Technology, vol. 7, no. 2, p. 190, 2018.
30. D. Dua and C. Graff, UCI Machine Learning Repository, University of California, School Of Information and Computer Science, Irvine, CA, USA, 2019, http://archive.ics.uci.edu/ml.
31. L. N. Sanchez-Pinto, L. R. Venable, J. Fahrenbach, and M. M. Churpek, “Comparison of variable selection methods for clinical predictive modeling,” International Journal of Medical Informatics, vol. 116, pp. 10–17, 2018.
32. T. H. Al Dhyani, A. S. Alshebani, and M. Y. Alzahrani, “Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms,” Healthcare Engineering, vol. 16, Article ID 4984967, 2020.
33. J. Joshi, R. Doshi, and J. Patel, “Diagnosis and prognosis of breast cancer using classification rules,” International Journal of Engineering Research and General Science, vol. 2, no. 6, pp. 315–323, 2014.

34. E. M. Senan and M. E. Jadhav, “Analysis of dermoscopy images by using ABCD rule for early detection of skin cancer,” Global Transitions Proceedings, vol. 2, no. 1, 2021.
35. S. Hore, S. Chatterjee, R. K. Shaw, N. Dey, and J. Virmani, “Detection of chronic kidney disease: a NN-GA-based approach,” in Proceedings of the Nature Inspired Computing, pp. 109–115, Springer, Singapore, December 2018.
36. A. Ogunleye and Q. G. Wang, “XGBoost model for chronic kidney disease diagnosis,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 6, pp. 2131–2140, 2019.
37. B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, “An empirical evaluation of machine learning techniques for chronic kidney disease prophecy,” IEEE Access, vol. 8, pp. 55012–55022, 2020.
38. P. Chittora, S. Chaurasia, P. Chakrabarti et al., “Prediction of chronic kidney disease-a machine learning perspective,” *IEEE Access*, vol. 9, pp. 17312–17334, 2021.
39. O. A. Jongbo, A. O. Adetunmbi, R. B. Ogunrinde, and B. Badeji-Ajisafe, “Development of an ensemble approach to chronic kidney disease diagnosis,” Scientific African, vol. 8, Article ID e00456, 2020.
40. K. Harimoorthy and M. Thangavelu, “Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system,” Journal of Ambient Intelligence and Humanized Computing, vol. 1, 2020.