# Analysis of Multiple Machine Learning Models used toPredict Chronic Kidney Disease in Comparison

**Mona[1], Abhinav Dahiya[2], Kamaldeep Joshi[3], Rajkumar Yadav[4], and Rainu Nandal[5]**

***Department of Computer Science and Engineering, University Institute of Engineering & Technology, Maharshi DayanandUniversity, Rohtak – Haryana, 124001, INDIA***

***Email Id:vijetamalik299@gmail.com, abhinav.rs.uiet@mdurohtak.ac.in,Kamalmintwal@gmail.com, Rajyadav76@rediff.com,Rainu_nandal@yahoo.com***

**DOI Number: 10.14704/nq.2022.20.13.NQ88432** **NeuroQuantology2022;20(13): 3515-3522**

***Abstract*** - Massive potential is shown by machine learning for facilitating kidney disease decision-making. Nephrology is anticipated to see significant advancements as a result of developments in data processing and preservation. Artificial intelligence (AI) is utilized in healthcare and research for various tasks, such as disease diagnosis, managing chronic conditions, providing medical services, and finding new drugs. However, there are still many difficulties that ML and its uses in nephrology must overcome. In this regard, we work to boost the potential of ML in renal disease, which will assist patients in better comprehending the complexity of the condition and improve the ability to anticipate, diagnose, detect, and provide high-quality care for kidney disease.

***Index Terms* - *ML, Nephrology, Kidney diseases***

## I. INTRODUCTION

Machine learning (ML) is an application AI in which the system looks for observations or data or instructions, searches for patterns in data, and predicts future events based on the samples provided. ML is used across industries for various reasons: a large amount of data is being captured and made available digitally; the processing of vast amounts of data has grown to be cost-effective due to the improved computing power now offered at affordable prices; and a variety of open-source frameworks, toolkits and libraries are available that can be used to construct and implement ML applications.

Globally, kidney disease (KD) is a significant public health concern. Renal disorders are unquestionably widespread, despite estimates of their prevalence varying significantly between and within jurisdictions. Patients run the danger of suffering catastrophic repercussions, including cardiovascular disease or end-stage renal disease (ESRD), which burdens world health. Even while prevention and treatment have made great strides in recent years, more work has to be done to turn the tide. Algorithmic approaches, which are the foundation of machine learning, enable a machine to handle complicated problems without programming. The widespread application of ML in the medical sector promotes medical innovation, lowers healthcare costs, and raises the standard of care.ML helps computers enjoy the same ability to learn, identify, and judge as human beings.

Fundamentally AI devices can be categorized into two types:

1. ML category: It analyses the structured data. Examples: genetic data and imaging data. As in healthcare, ML tries to gather patient data as an individual and then find out the reasons for that disease.
2. NLP category: Information is gathered from unstructured data as from observing medical data or health journals, to boost structured health check data. The NLP processes content or data in machine-readable structured records and then is considered by machine learning (ML) procedures.



Specifically, in healthcare, ML has led to exciting new developments that improve health outcomes. ML can improve the sensitivity of recognition, add more importance to treatment decisions, and helps in personalizing treatment so that each patient gets the treatment that is best for them. The possibilities are infinite.

In recent years, the use of computer-assisted diagnosis approaches based on artificial intelligence-based solutions has increased [3]. This is because machine learning algorithms can readily find and classify relevant trends and patterns in the data [4]. Various methods, including Decision Trees (DT), K -Nearest Neighbor Algorithm (KNN), and Random Forest (RF), are used to classify the severity of the condition. The complexity of nephrotic disease necessitates a cautious approach to treatment. Data mining and classification are essential in kidney disease prediction and analysis.

This research suggests an Ensemble Solution based on Machine Learning to forecast the possibility of chronic kidneydisease. We started by implementing several Machine Learning algorithms, including Decision Tree, KNN, Random Forest, XG Boost, Adaptive Boosting, and Cat Boost. In addition, we conducted a comprehensive Exploratory Data Analysis, performed suitable scaling and data preprocessing, and gained insight into the association between various factors.

The remaining sections are organized as follows: The first section includes comprehensive information about the proposed system, followed by an in-depth data analysis. Then, we analyze the EDA, followed by the outcome and conclusion, and prospects.

## II. Proposed system

This part consists of block diagrams, flow diagrams, evaluation matrices, the dataset used and the algorithms applied. First the data preprocessing is done followed by feature selection and at the end decision tree, KNN, gradient boost, ADA boost and cat boost algorithms are applied.



### *A. Dataset*

In this study, we use the "chronic kidney disease prediction" dataset provided by Kaggle [5]. Table 1 displays detailed information on the dataset's features used to differentiate between CKD and non-CKD patients. The dataset has no missing values among the 400 (250 for CKD patients and 150 for non-CKD disease) patient records. The process begins with data preprocessing, followed by comprehensive exploratory data analysis, feature selection, classification, performance evaluation, and future prospects.

Table 1. Attributes used in the prediction of ckd patients

| Attributes | Description with values | Type of test |
|---|---|---|
| age | Age (Numeric) in years | - |
| Bp | Blood pressure (Numeric)in mm/hg | - |
| Sg | Specific Gravity (Numeric)- 1.005,1.010,1.015,1.020,1.025 | Urine |
| Al | Albumin (Numeric)- 0,1,2,3,4,5 | Urine |
| su | Sugar (Numeric) -0,1,2,3,4,5 | Urine |
| rbc | Red Blood cells (Nominal) with Normal, abnormal values | Urine |
| pc | Pus Cell (Nominal) with Normal, abnormal | Urine |
| pcc | Pus Cell Clumps (Nominal)- present, not present | Urine |
| ba | Bacteria (Nominal)- Present, not present | Urine |
| bgr | Blood Glucose Random (Numeric) in mgs/dl | Blood |
| bu | Blood urea (Numeric) in mgs/dl | Blood |
| sc | Serum Creatinine in mgs/dl | Blood |
| sod | Sodium (Numeric) in mEq/l | Blood |
| pot | Potassium (Numeric) in mEq/l | Blood |
| hemo | Hemoglobin (Numeric) in gms | Blood |
| pcv | Packed Cell Volume (Numeric) | Blood |
| wbcc | White Blood Cells Count (Numeric) having cells/cumm | Blood |
| rbcc | Red Blood Cells (Numeric) Count having millions/cmm | Blood |
| htn | Hypertension (Numeric)- yes, no | - |
| dm | Diabetes Miletus (Numeric) - yes, no | - |
| cat | Coronary Artery Disease (Nominal) - yes, no | - |
| appet | Appetite (Nominal)- good, poor | - |
| pe | Pedal Edema (Nominal) - yes, no | - |
| ane | Anemia (Nominal) - yes, no | - |
| class | Class (Nominal)-ckd, notckd | - |

### B. *Data preprocessing:*

Data preprocessing is compulsory to avoid noise from the data so that it cannot hinder the model building for training purposes. It involves the handling of missing values. Data gathered always needs to be prepared and cleaned before model building and then null values get filled.

There are various methods to handle missing values. However, there are no missing values in this dataset. To increase the efficiency and accuracy of model data is divided into training and testing sets. Here, used a 70 to 30 ratio of training to testing sets.

### C. *Feature selection:*

Features selected for prediction are on the basis of heatmap. The positive correlation between the class and attributes show that they play an important role for prediction. and negative correlated factors doesn`t contribute much. Here, sodium, sugar, specific gravity, hemoglobin, WBC and RBC count are positively corelated.



### D. *Algorithms:*

KNN, Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier, Gradient Boosting Classifier, Stochastic Gradient Boosting (SGB), Xg Boost, Cat Boost Classifier Are the algorithms here applied for the prediction purpose.

## III. EXPLORATORY DATA ANALYSIS

We conducted a comprehensive exploratory data analysis on the attributes mentioned above and obtained meaningful insights regarding the impact of these variables on kidney disease. Figure 2 shows a heatmap for the correlation between features.

*A. AGE*

There exist some intriguing analyses. Generally, it is believed that the incidence of kidney malfunctioning increases with age. But, according to the datasets, as the age increases, the number of individuals at risk for kidney damage reduces. The correlation between age and target attribute is negative.

*B. BLOOD Pressure*

Predicting whether a patient will have a kidney related disease based on resting blood pressure data is difficult. The correlation between bp and the target attribute is -0.29, which is negative and very little. Hence this attribute is dropped.

*C. Albumin*

According to Figure 2, The correlation between the variables is -0.63239, which is extremely low. Hence it is dropped.

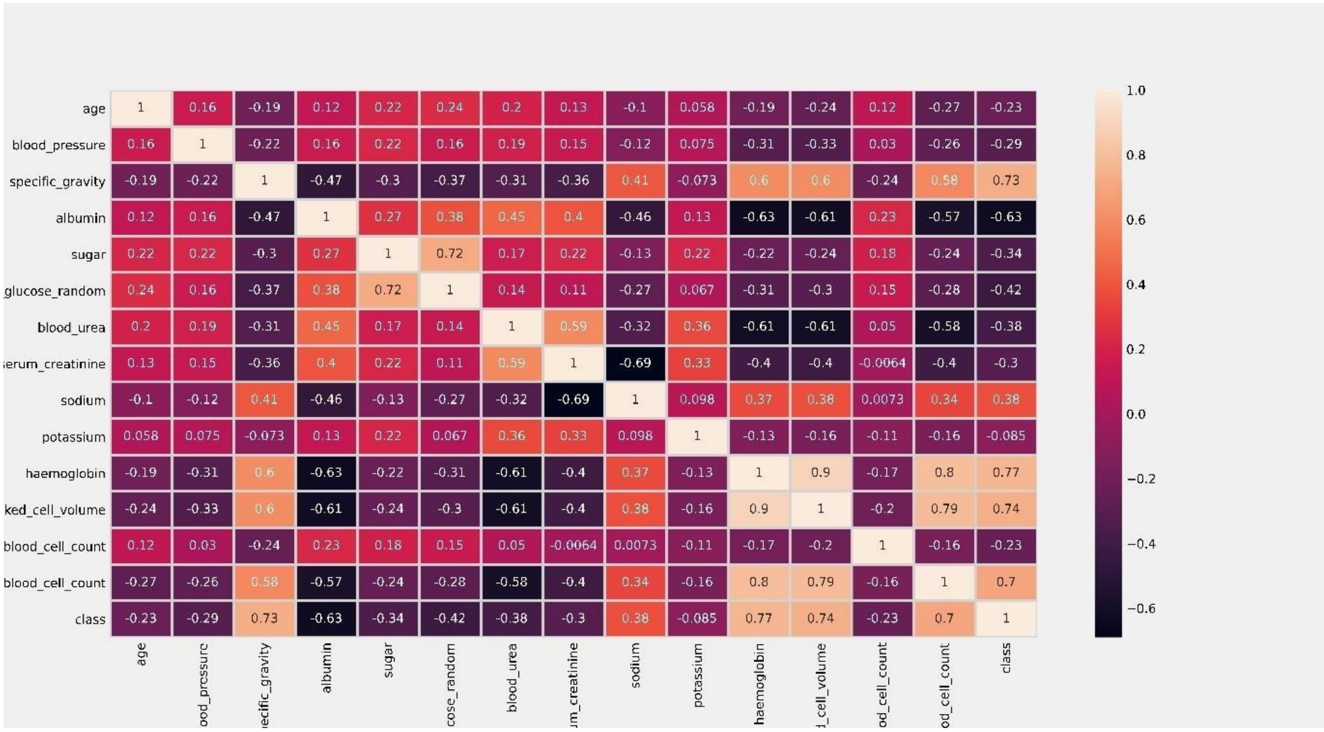




Figure 1. Heatmap

*D. Sugar*

The patient's likelihood of suffering a chronic kidney disease is proportional to the patient's sodium rate. There is a negative and moderate correlation of -0.34 between the variables.

*E. Potassium*

If the value of this variable is between 0 and 1.5, the likelihood of suffering a kidney disease significantly. The correlation coefficient between the two variables is -0.085, negative and moderate.

*F. Sodium*

Patients with a high amount of sodium have higher risk for kidney failure. Increased bp is direct consequence for sodium sensitivity. The correlation between the two variables is 0.380937.

*G. Hemoglobin*

Patients with hemoglobin amount 11.5g/dl is sufficient for CKD patients. Above this amount of Hb can causes further consequences to CKD patient. The correlation between the two variables is 0.77798 enough strong.

*H. BLOOD_ urea*

Those with a blood urea observation value of 0 had a marginally increased risk of a kidney failure. Patients with a GFR below 60 indicated the malfunctioning of kidneys. There

*I. SERUM creatinine*

Patients with a "creatinine" value of 0.74-1.35mg/dl is is a -0.38046 correlation between the two variables. Hence it is dropped. normal for male CKD patient. should be closely monitored because the danger of having a heart attack is about twice as great as the risk of avoiding one. There is a correlation of 0.137230 between the two variables.

## IV. AREA under curve (AUC)

This shows the visualization of the sensitivity and specificity of the models we have applied. In this we applied random forest, KNN, decision tree, ada boost, xg boost and cat boost algorithms of machine learning for prediction. The more the graph towards upper left corner, the better the model.

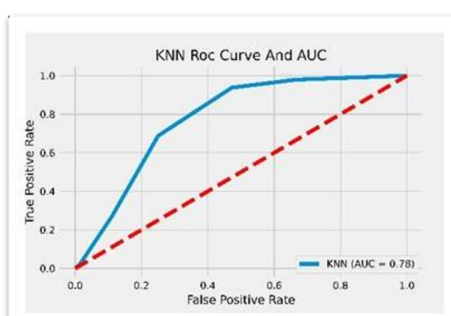


Figure 2. AUC of KNN

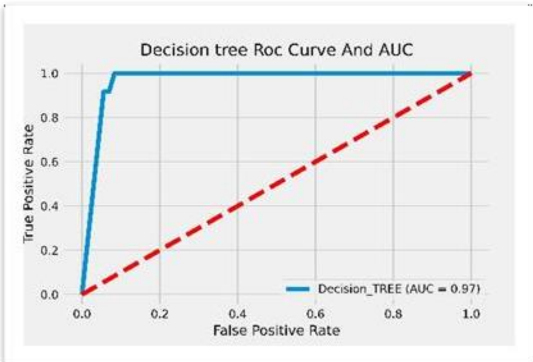


Figure3. AUC of Decision Tree

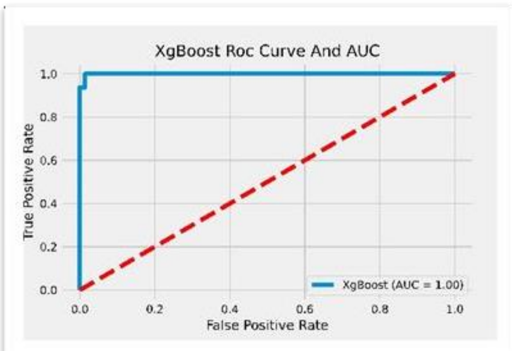


Figure4. AUC of XgBoost

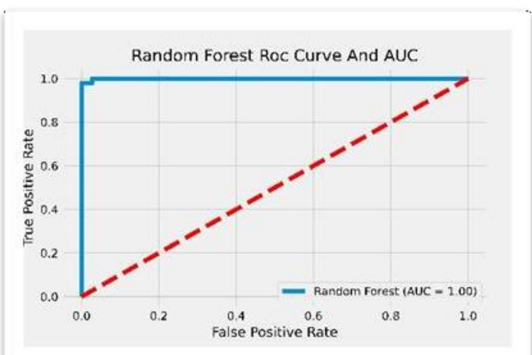


Figure 5. AUC of Random Forest

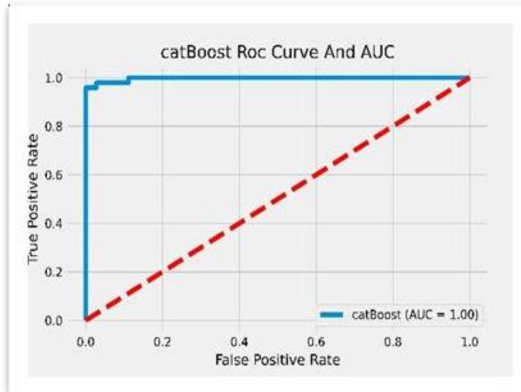


Figure6. AUC of Cat Boost

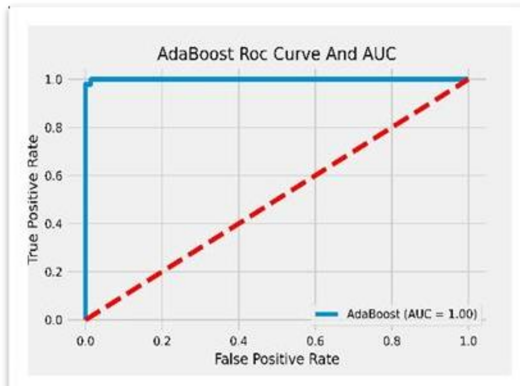


Figure7. AUC of AdaBoost

## V. RESULTS

We evaluated a range of techniques, from simple classifiers such as Logistic Regression to complicated boosting and bagging algorithms such as XGBoost, AdaBoost, and CatBoost. Table 2 includes all the metrics on the basis of which machine learning algorithms performance is calculated on the test data.

Table 2 metrics for evaluating machine learning algorithms prediction

| Metrics | Description | Equation |
|---|---|---|
| Accuracy | Proportion of correct predictions made by the model or the number of correct forecasts out of total predictions. | (TP+TN)/(TP+TN+FP+FN) |
| Sensitivity | It is the proportion of true positives that were correctly classified across all predictions | TP/(TP+FN) |
| Specificity | Fraction of total number of accurate negative predictions to the total number of negatives in the dataset. Also called True negative rate. | TN/(TN+FP) |
| Positive predictive value | Ratio of total number of correct positive predictions to the total number of positives predictions. Also called Precision. | TP/(TP+FP) |
| Negative predictive value | Ratio of total number of correct negative predictions to the total number of negative predictions. | TN/(TN+FN) |
| F1 score | Harmonic mean of precision and recall. Used when there is problem in large class positive/negative imbalances | 2.{Precision X recall/Precision +recall} |
| AUROC | The ROC curve for an algorithm is created with true positive rate against the false positive rate at various probability thresholds. | |

We evaluated a range of techniques, from simple classifiers such as Decision Tree to complicated boosting and bagging algorithms such as XG Boost, AdaBoost, and Cat Boost. Table 3 includes all the classifiers we employed and their accuracy, recall, precision, and F1 score on test data.

The Random Forest model gives an accuracy of 97.5 % and an AUC score of 0.92.

Table3. classifiers with results

| Classifier | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| KNN | 76.70 % | 78.50 % | 74.45 % | 76.30 % |
| Decision Tree | 97.14 % | 94.74 % | 81.82 % | 87.80 % |
| Random Forest | 97.50% | 90.48 % | 86.36 % | 88.37 % |
| XGBoost | 98.30% | 97.40% | 99.00% | 95.32% |
| AdaBoost | 97.50 % | 91.30 % | 95.45 % | 93.33 % |
| CatBoost | 92.80 % | 97.24 % | 95.91 % | 99.02 % |

Table 4 compares the models with the prior one models found in the referenced papers. The table indicates that the boosting algorithms performed well on same dataset as used before. Decision tree with accuracy of 97.14% performed well as compared to both of the referenced papers. KNN models have least accuracy among all with value of 76.7%.

Table4.Models comparison

| Model (this paper) | Accuracy (%) | Model (reference paper) | Accuracy (%) |
|---|---|---|---|
| KNN | 76.7 | KNN [Ref. 15] | 71.25 |
| Decision Tree | 97.14 | Decision Tree [Ref. 15] | 96.25 |
| XG boost | 98.3 | Logistic Regression [Ref. 15] | 97 |
| Random forest | 97.5 | Decision tree [Ref. 14] | 73.2 |
| ADA boost | 97.5 | CNN [Ref. 14] | 89 |
| Cat boost | 92.8 | Random forest [Ref. 14] | 72.5 |

## VI. Challenges and Future Prospects:

Surprisingly, the remaining instances of the use of ML in healthcare are still in the early stages and pertain to kidney illness. Both machine learning and nephrology applications face difficulties.

Challenges in ML:

It is still important and necessary to improve ML algorithms in order to face the challenges that are faced during clinical practices. As ML models' logic is like black box, still difficult to elaborate to the doctors and limitations in development of ML algorithms. Moreover, the ethics of ML must be taken care off. It is obligatory to develop clear guidelines as soon as possible to standardize ML clinical application. Also, there is issue related to the probability of training dataset, as ML performs well with huge training set. So, there must be a balance between regulation of dataset and privacy so that accuracy of ML models can be improved.

Challenges for nephrology:

First and foremost, challenge is the data collection and processing as in data collection it is hard to collect enough electronic medical record information. Different medical institutions have different regulations on data due to which there can be lack of good quality data and ultimately it makes difficult to extract useful information. Mostly there is lack of external verification that makes it difficult to prove the accuracy and practicability of the model. To better analyze the results data must be integrated and authentic. Data prepared and preprocessed before modelling as medical data are large, confusing and complex.

It is quite challenging to create and attain high quality data as inaccurate data can limit the accuracy of the algorithms for prediction and a lot of efforts invested in improving the algorithm will go in vain. Also, cost of implementation and modelling will be wasted. Therefore, investment of funds for data collection and modelling is also challenging.

Although AI is not known by most of the nephrologists.so, as a future scope we can collaborate with them and AI can build a big database for kidney related diseases which will help in building a high efficiency model that can be used on a large scale for kidney disease diagnosis and treatment.



## VI. Conclusion

ML is amongst the most trending and one of the biggest technology that changed human and machine interaction. It has many applications in health sector, agriculture sector, e-commerce, smart living, infrastructure, etc. in almost every aspect of life. In this paper, the importance and use of ML in the healthcare sector is studied to find out the better tools and techniques. The accuracy and generalization of algorithms have been discussed in this survey paper. Different methods like SVM, NB, DT, and KNN are widely used for the diagnosis of various diseases. Also, other methods are used for disease detection, risk identification, physician review, and ICU data.

Prediction of kidney disease in nephrology at early stages may help in:

1. Reducing diagnosing costs.
2. It May increase patient satisfaction.
3.Can reduce the burden on hospitals as of patient readmissions.
4.It May help insurance firms to find genuine nephrotic patients.

REFERENCES

[1] P. Chittora, "Prediction of Chronic Kidney Disease - A Machine Learning Perspective", IEEE Access, vol. 9, pp. 17312-17334, 2021.MayoClinic, "heart attack," June162020.Accessed on: October 01, 2022. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106.

[2] Z. Wang, J. Won Chung, X. Jiang, Y. Cui, M. Wang, and A. Zheng, "Machine Learning-Based Prediction System for Chronic Kidney Disease Using Associative Classification Technique", International Journal of Engineering & Technology, vol. 7, no. 436, p. 1161, 2018.

[3] Freeprojectsforall.com. 2022. [online] Available at:<http://www.freeprojectsforall.com/wp-content/uploads/2017/12/Disease-Prediction by-Machine-Learning-over-Big-Data-from-Healthcare-Communities.pdf> [Accessed 2september 2022].

[4] Li, Q., Fan, Q., Han, Q., Geng, W., Zhao, H., Ding, X., Yan, J. and Zhu, H., 2020. Machine learning in nephrology: scratching the surface. Chinese Medical Journal, (6), pp.687-698.

[5] H. Information, K. Disease, C. (CKD), C. (CKD) and N. Health, "Chronic Kidney Disease (CKD) | NIDDK", National Institute of Diabetes and Digestive and Kidney Diseases, 2021. [Online]. Available: https://www.niddk.nih.gov/healthinformation/kidney-disease/chronic-kidney-disease-ckd. [Accessed: 18- sept- 2022].

[6] S. Tekale, P. Shingavi, and S. Wandhekar, "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm", IJARCCE, vol. 7, no. 10, pp. 92-96, 2018. [3] S. Krishnamurthy et al., "Machine Learning Prediction Models for Chronic Kidney Disease Using National Health Insurance Claim Data in Taiwan", Healthcare, vol. 9, no. 5, p. 546, 2021.

[7] J. Xiao,Ruifeng ding, Haochen Guan, Xinhui Feng,Tao Sun, Sibo Zhu and Zhibin Ye, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression", Journal of Translational Medicine, vol. 17, no. 1, 2019.

[8] H. Chang, C. Wu, S. Lee, Y. Chen, W. Su, and S. Su, "A predictive model for progression of CKD", Medicine, vol. 98, no. 26, p. e16186, 2019.

[9] Mayo Clinic, "chronic kidney disease", June 16 2020. [Online]. Available: https://www.mayoclinic.org/diseases- conditions/heart-attack/symptoms-causes/syc-20373106 [Accessed on: July 16 ,2022].

[10]Yu, S. M. W., & Bonventre, J. V. (2018). Acute Kidney Injury and Progression of Diabetic Kidney Disease. Advances in Chronic Kidney Disease, 25(2), 166–180. https://doi.org/10.1053/j.ackd.2017.12.005.

[11]Z. Wang, J. Won Chung, X. Jiang, Y. Cui, M. Wang, and A. Zheng, "Machine Learning-Based Prediction System for Chronic Kidney Disease Using Associative Classification Technique", International Journal of Engineering & Technology, vol. 7, no. 436, p. 1161, 2018.

[12]M. Almasoud and T. E, "Detection of Chronic Kidney Disease Using Machine Learning Algorithms with Least Number of Predictors", International Journal of Advanced Computer Science and Applications, vol. 10, no. 8, 2019.

[13]P. Chittora, "Prediction of Chronic Kidney Disease - A Machine Learning Perspective", IEEE Access, vol. 9, pp. 17312-17334, 2021.

[14]Ifraz, G.M. et al. (2021) “Comparative analysis for prediction of kidney disease using intelligent machine learning methods,” Computational and Mathematical Methods in Medicine, 2021, pp. 1–10. Available at: https://doi.org/10.1155/2021/6141470.

[15]S. Krishnamurthy et al., "Machine Learning Prediction Models for Chronic Kidney Disease Using National Health Insurance Claim Data in Taiwan", Healthcare, vol. 9, no. 5, p. 546, 2021.