Prediction of Chronic Kidney DiseaseUsing Machine Learning

Mr. Narayan H.M[1], Bhumi Sri P[2], Durga Shree P[3], Asif Ahamad V[4], Muni Giri Babu E[5]
[1,2,3,4,5]*Computer Science andEngineering, M.S. Engineering CollegeBangalore, Karnataka*

***Abstract*—Chronic Kidney Disease is one of the most critical illness nowadays and proper diagnosis is required as soon as possible. Machine learning technique has become reliable for medical treatment. With the help of a machine learning classifier algorithms, the doctor can detect the disease on time. For this perspective, Chronic Kidney Disease prediction has been discussed in this article. Chronic Kidney Disease dataset has been taken from the Kaggle. Six classifier algorithms have been applied in this research such as SVM, XG Boost, Naïve bayes, logistic regression, random forest, decision tree. The important feature selection technique was also applied to the dataset.**

For each classifier, the results have been computed based on (i) full features, (ii) correlation-based feature selection, (iii) Wrapper method feature selection , (iv) Least absolute shrinkage and selection operator regression, (v) synthetic minority over-sampling technique with full features From the results, it is marked that XG Boost is giving the highest accuracy of 98.86% in synthetic minority over- sampling technique with full features. Along with accuracy, precision, recall, F-measure, area under the curve and GINI coefficient have been computed and compared results of various algorithms have been shown in the graph.

I. INTRODUCTION

Chronic kidney disease (CKD) is the serious medical condition where the kidneys are damaged and blood cannot be filtered. In the end-stage of the disease the renal disease (CKD), the renal function is severely damaged. The starting date of kidney failure may not be known, it may not recognize as an illness of the patient because it cannot show any symptoms initially. And this chronic kidney disease is also called chronic renal failure, which has become quite a serious problem in the world where the kidneys are damaged and it has become the cause of improper function of kidney organ. To overcome this issue, this project aims to kidney disease diagnosis using machine learning approaches. This is done by comparing the accuracies of different algorithms and uses the algorithm with high accuracy for prediction. Our goal is to enhance the performance of the model by removing unnecessary and insignificant attributes from the dataset and only collecting those that are most informative and useful for the classification task. Thus, the main focus of the system is to make use data analytics to predict the presence of the disease.

II. EXISTING SYSTEM

The existing system of diagnosis is based on the examination of urine with the help of serum creatinine level.Many medial methods are used for this purpose such as screening, ultrasound method. In screening, the patients with hypertension, history of cardiovascular disease, diseasein the past, and the patients who have relatives who ha kidney disease are screened. This technique includes thecalculation of the estimated GFR from the serum creatinine level, and measurement of urine albumin-to-creatinine ratio (ACR) in a first morning urine specimen. This method is oldand sow. We want to detect the kidney disease as soon as possible.

A.Disadvantages of Existing System:

There used many deep learning and machine learning algorithms to predict the CKD but that are not so accurate. The methods used previously that are old and slow.

III. PROPOSED SYSTEM

Our Aim is to predict the chronic kidney disease using machine learning algorithm. Chronic Kidney Disease (CKD)means your kidneys are damaged and can't filter blood the Way they should. The disease is called "chronic" becausethe damage to your kidneys happens slowly over a long period of time. This damage can cause wastes to build up in your body. CKD can also cause other health problems. 10% of the population

worldwide is affected by chronic kidney disease (CKD), and millions die each year because the doctors are unable diagnose the disease. The system is automation for predicting the CKD. We proposed XG Boost, Logistic Regression and Random Forest machine learning technique for kidney disease prediction of significant features.

A.Advantages:

The existing system predict CKD with six machine learning algorithms and we consider the most accurate algorithm for prediction. The proposes system will extract important features from dataset which will take less time topredict the disease.

IV. SYSTEM ARCHITECTURE

The system "design" is defined as the process of applying various requirements and permits it physical realization. Various design features are followed to develop the system. The aim is to effectively predict the chronic kidney disease using the machine learning algorithm. We proposed Decision Tree, Naïve Bayes, Support Vector Machine, Gradient Boosting, Logistic Regression and Random Forest machine learning technique for kidney disease prediction of significant features. ML process starts from a pre-processingdata phase followed by feature selection based on data cleaning, classification of modelling, performance evaluation and the results with improved accuracy. The algorithm with highest accuracy is implemented for the prediction of the chronic kidney disease.

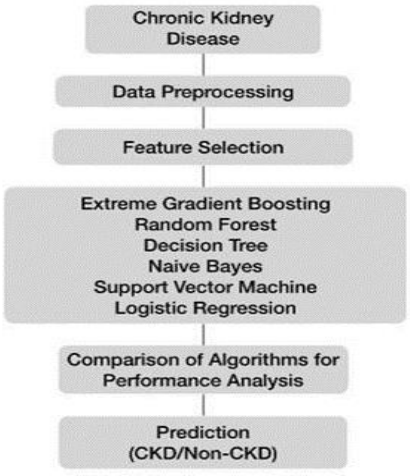

Fig 4.1 System Architecture

V. METHODOLOGY

Once the design aspect of the system is finalizes the system enters into the coding and testing phase. The coding phase brings the actual system into action by converting the designof the system into the code in a given programming language. Therefore, a good coding style has to be taken whenever changes are required it easily screwed into the system.

ALGORITHMS

i. Logistic Regression: Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

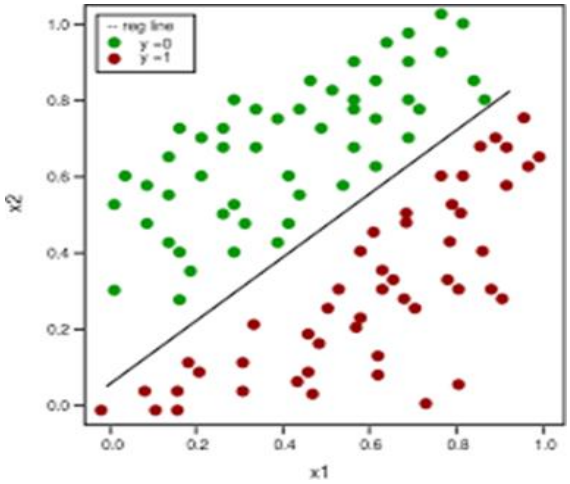

Fig 5.1 Logistic Regression

Logistic Regression is basically a supervised classification algorithm. In a classification problem, the target variable, y, can take only discrete values for a given set of features, X.

ii. Random forest: is like bootstrapping algorithm with Decision tree (CART) model. Say, we have 1000 observation in the complete population with 10variables. Random forest tries to build multiple CARTmodel with different sample and different initial variables. For instance, it will take a random sample of 100 observation and 5 randomly chosen initial variables to build a CART model. It will repeat the process (say) 10 times and then make a final prediction on each observation. Final prediction is a function of each prediction. This final prediction can simply be the mean of each prediction

iii. XGBoost: is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction.

iv. Support Vector Machine Algorithm: Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

v. Naïve Bayes Classifier Algorithm: Naïve Bayes algorithm is a supervised learning algorithm, which isbased on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learningmodels that can make quick predictions. It is a probabilistic classifier, which means it predicts on thebasis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles

VI. IMPLEMENTATION

Implementation is the stage of the project where the theoretical design is turned out into a working system. Thus, it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning,investigation of the existing system and its constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

Modules:

1. Data Pre-Processing
2. Algorithm Implementation
3. Feature Selection
4. Prediction

Module Descriptions:

1. Data pre-Processing:

Our Kidney Disease project dataset are collected from kaggle.com. Chronic kidney disease data is pre-processed after collection of various records. The dataset contains a more number of patient records, where some records are with some missing values. Those missing records have been removed from the dataset and the remaining patient records are used in pre-processing. After that we remove some columns based on feature selection.

2. Algorithm Implementation: The Classification Algorithmsto produce the best results. We are using XG Boost, Logistic regression and Random Forest Algorithm to predict the kidney disease using ML. On an analysis conducted within various algorithms, the Random Forest was found to provide highest efficiency. Then, the classifiers are applied to each clustered dataset in order to estimate its performance. The best performing models are identified from the above results based on their low rate oferror.

•Logistic Regression
• Random Forest Classifier
• XG Boost

3. Feature Selection:
   We have used feature selection to make the model more effective and taking only the most importantdata.
4. Prediction:
   Several standard performance metrics such as accuracy, precision and error in classification have been considered for the computation of performance efficacy of this model. Preprocessed data are trained and input given by the user goes to the trained dataset.

VII. ER/UML DIAGRAMS

1. Use Case Diagram:

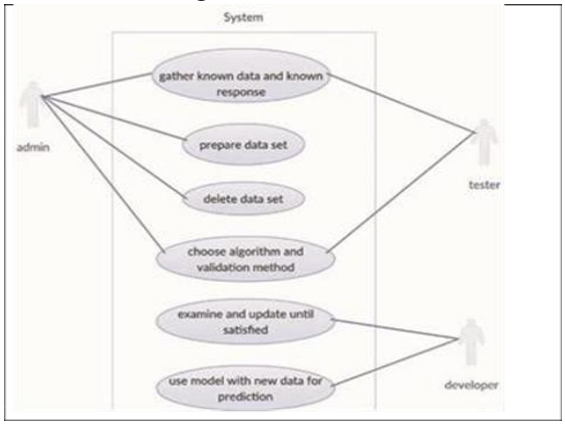

Fig.7.1 Use Case Diagram

A use case diagram in the Unified Modeling Language(UML) is a type of behavioral diagram defined by and created from a Use case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals and any dependencies between those use cases. The main purpose of a use case diagram is to show what systems functions are performed for which actor. Roles of the actors in the system can be depicted.

2. Sequence Diagram:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. .It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

Fig.7.2 Sequence Diagram

3. Deployment Diagram:

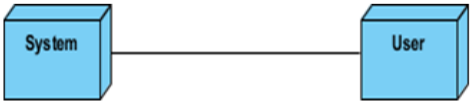

Fig.7.3. Deployment Diagram

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations(or methods), and the relationships among the classes. It explains which class contains information

4. ER Diagram:

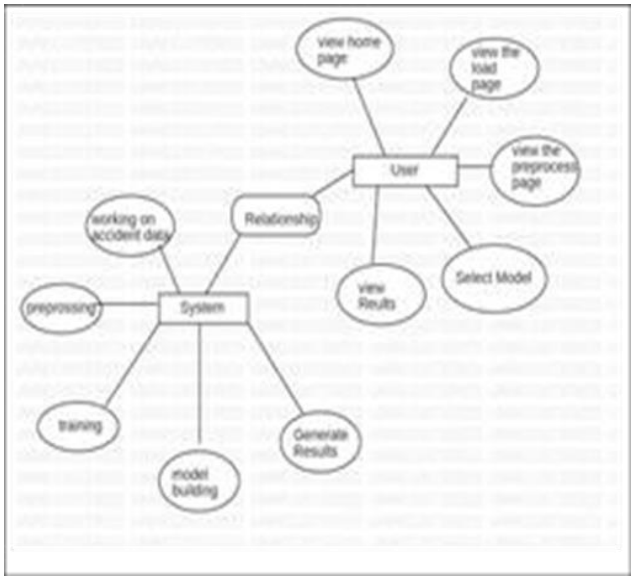

Fig.7.4 ER Diagram

A sequence diagram in Unified Modeling Language (UML) isa kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct ofa Message Sequence Chart. Sequence diagrams aresometimes called event diagrams, event scenarios, and timing diagrams.

VIII. CONCLUSION

The research offers a practical method for CKD prediction using six different ML algorithms. In addition to pre- processing and feature selection, the dataset is trained and evaluated with precise results. The most crucial features among the 25 attributes from the dataset of 400 records that are required for the prediction were provided by the select K best test and chi-square test, which are the two main tests utilized for effective feature selection. With the help of this technique, XG Boost has improved accuracy and production. The suggested approach can be effectively used to identifythe illness and fast improve one's health.

REFERENCES

[1] Ahmed, R. M., & Alshebly, O. Q. (2019), "Prediction and Factors Affecting of Chronic Kidney Disease Diagnosis using Artificial Neural Networks Model and Logistic Regression Model," IRAQIJOURNAL OF STATISTICAL SCIENCES, 16(28),140–159.

[2] Wang C, CuiCui L, Gong W, Tanqi L (2013). "Newurinarybiomarkers for diabetic kidney disease". Biomarker Res; 1(9):1-4.

[3] Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker,S., Plattner, B., Saran, R., Wang,A. Y. M., & Yang, C. W. (2013),"Chronic kidney disease: global dimension and perspectives," The Lancet, 382(9888), 260–272.

[4] Abraham, G., Varughese, S., Thandavan, T., Iyengar, A., Fernando, E., Naqvi, S. A. J., Sheriff, R., Ur-Rashid,H., Gopalakrishnan, N., &Kafle, R. K. (2015),"Chronic kidney disease hotspots in developing countries in South Asia," Clinical Kidney Journal, 9(1),135–141.

[5] Ifraz, G. M., Rashid, M. H., Tazin, T., Bourouis, S., &Khan, M. M. (2021),"Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods," Computational and Mathematical Methods in Medicine, 2021, 1–10.

[6] H. Nasri, "World kidney day 2014; chronic kidney disease andaging: a global health alert," Iranian Journal of Public Health, vol. 43, no. 1, pp. 126-127, 2014.

[7] Centers for Disease Control and Prevention. "Chronic kidney disease in the United States, 2019." Atlanta, GA:US Department of Healthand Human Services, Centers for Disease Control and Prevention3(2019).

[8] Lv, J.C. and Zhang, L.X., 2019. "Prevalence and disease burden of chronic kidney disease. Renal Fibrosis: Mechanisms and Therapies", pp.3-15.

[9] Murshid, G., Parvez, T., Fezal, N., Azaz, L. and Asif, M., 2019." Datamining techniques to predict chronic kidney disease." Int. J. Scientific Res. Comput. Sci., Eng. Inf. Technol, 5(2), pp.1220-6.

[10] Adejumo, O.A., Akinbodewa, A.A., Okaka, E.I., Alli, O.E. and Ibukun, I.F., 2016."Chronic kidney diseasein Nigeria: Late presentation is still the norm." Nigerian Medical Journal: Journal of the Nigeria MedicalAssociation, 57(3), p.185.

[11] Haratian, A. (2022, December 16), "Detection of factors affecting kidney function using machine learning methods".

[12] Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D. (2015)," Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. Big Data," 3(4), 277–287.

[13] Ilyas, H., Ali, S., Ponum, M. et al. "Chronic kidney disease diagnosis using decision tree algorithms." BMC Nephrol 22, 273(2021).https://doi.org/10.1186/s12882-021-0247

[14] Almasoud, M. and Ward, T.E., 2019. "Detection of chronic kidney disease using machine learning algorithms with least number of predictors." International Journal of Soft Computing and Its Applications, 10(8).

[15] Anantha Padmanabhan, K. R., & Parthiban, G. (2016). "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease." Indian Journal of Science and Technology, 9(29).

[16] Senan EM, Al-Adhaileh MH, Alsaade FW, Aldhyani THH, Alqarni AA, Alsharif N,Uddin MI, Alahmadi AH, Jadhav ME, Alzahrani MY. "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques". J Healthc Eng. 2021 Jun 9;2021:1004767.
doi: 10.1155/2021/1004767.PMID: 34211680; PMCID: PMC8208843.

[17] Nishanth A, Thiruvaran T." Identifying Important Attributes for Early Detection of Chronic Kidney Disease." IEEE Rev Biomed Eng. 2018;11:208-216. [18] Snegha, J., Tharani, V., Preetha, S., Charanya, R., & Bhavani, S. (2020). "Chronic Kidney Disease Prediction Using Data Mining." 2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic- ETITE).

[19] Praveen, S.P., Jyothi, V.E., Anuradha, C., Venugopal, K., Shariff, V. and Sindhura,S., 2022. 'Chronic Kidney Disease Prediction Using ML-Based Neuro-Fuzzy Model." International Journal of Image and Graphics, p.2340013.

[20] Alabi, O. (2022, December 9). "Comparative Study of Chronic Kidney Disease Predictor Performance Given Insufficient Training Dataset, Information Technology and Management Science." https://itms- journals.rtu.lv/article/view/itms-2022-0001.

[21] Tekale, S., Shingavi, P., Wandhekar, S. and Chatorikar, A., 2018. "Prediction of chronic kidney disease using machine learning algorithm". International Journal of Advanced Research in Computer and Communication Engineering, 7(10), pp.92-96.