Diagnostic Prediction of Chronic Kidney/Renal Disease UsingXgboost Machine Learning Algorithm

ROHINI JADHAV[1], MADHAVI MANE[2], PRASANNA GAROLE[3], PRAFULL TRIPATHI[4], NISHANT KUMAR[5]

[1, 2] *Assistant Professor, Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University) Collegeof Engineering, Pune, Maharashtra, India*

[3, 4, 5] *Student, Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University) College ofEngineering, Pune, Maharashtra, India*

Abstract— chronic kidney disease (CKD) or Chronic Renal Disease (CRD) is a general term for multiple heterogeneous diseases in the kidneys that occurs when kidneys are damaged and cannot filter blood properly. Over time this could result in the development of severe cardiovascular disease, End-Stage Renal Disease (ESRD), and may even lead to death. The disease affects 10 per cent of the population worldwide and millions of lives are lost every year with conditions worse particularly in the developing nations due to lack of early diagnosis and affordable treatment. Since CKD depends on various factors like sugar, specific gravity, anemia etc., Machine-learning based systems can be developed based on these factors so that early prediction and diagnostic results can be achieved in a fairly short time, thereby allowing doctors to treat patients more effectively. This paper proposes utilization of various techniques such as data pre- processing, data visualization, feature-selection algorithms (SelectKBest, chi2) and prediction algorithm (XGBoost-eXtreme Gradient Boosting) to determine CKD. Eventually RandomisedSearchCV is used for tuning the hyperparameters.

Indexed Terms-- chi2, Chronic Kidney Disease, End-Stage Renal Disease, SelectKBest, RandomisedSearchCV, , XGBoost

I. INTRODUCTION

Kidneys are two bean-shaped organs present on either side of spine at the lowest level of the rib cage. Kidneys are importantas they perform the following functions: 1) Removes waste products and fluids that are in excess amount, 2) Reabsorption of nutrients from blood, 3) Maintain pH, 4) Regulate blood-pressure, 5) Regulate osmolarity and water balance, 6) Regulating bone-mineralization.

Each kidney is comprising of about a million filtering units called nephrons. Each nephron includes a filter, called the glomerulus that filters your blood, and a tubule that returnsneeded substances to your blood and removes wastes.

The term Chronic is defined as something that is ongoing, persistent, and long-term. CKD implies that kidneys are ailing or harmed or are maturing and probably won't work as expectedas they used to do. CKD is the 10th leading cause of deaths worldwide according to Global Burden of Disease study-2017 [1] & every day about 360 people begin dialysis treatment for renal failure [2].

Generally, CKD refers to 5 major stages of kidney damage, ranging from Stage-1(mild) to Stage-5 (complete failure/ESRD). This depends on how well the kidneys are functioning. The table given below describes in detail about thestages of CKD [3].

TABLE I

Sr.No.	*Description*	*GFR (Glomerular Filtration Rate)*	*% of Kidney Function*
Stage 1	*Kidney damage with normal kidney function*	*90 or higher*	*90-100%*
Stage 2	*Kidney damage*	*89 to 60*	*89-60%*

	withmild loss of kidneyfunction		
Stage 3 a	*Mild to moderateloss of kidney function*	*59 to 45*	*59-45%*
Stage 3 b	*Moderate to severeloss of kidneyfunction*	*44 to 30*	*44-30%*
Stage 4	*Severe loss of kidneyfunction*	*29 to 15*	*29-15%*
Stage 5	*Kidney failure*	*Less than 15*	*Less than15%*

Stages of CKD

II. REVIEW OF LITERATURE

Machine Learning is applied in the Medical and Healthcare domain to diagnose for disorders such as CKD, Alzheimer's, Diabetes etc.

The pervasive growth rate of Covid-19 [4] over different continents has been studied using Boosting algorithms.Decision tree algorithm [5] applied on a South Korean dataset gave an outstanding 99.85% accuracy.

Shahbaz M. et al. [6] utilized six machine learning techniques such as GLM-Generalized Linear Model and Decision Trees etc. with an accuracy of 88%. On OASIS dataset Stacking, Bagging and Boosting were applied for Alzheimer Disease Prediction and 90% accuracy was obtained by Random Forest i.e., Bagging technique.

Guneet Kaur [7], 2017 proposed a system for predicting the CKD using Data Mining Algorithms in Hadoop. They use two data mining classifiers like KNN and SVM. Here the predictive analysis is performed based upon the manually selected data columns. SVM classifier gives the best accuracy than KNN in this system

Abhijit Pathak, Most. Asma Gani, Abrar Hossain Tasin, SanjidaNusrat Sania, Md. Adil, Suraiya Akter proposed a system for diagnosing kidney disease using a number of machine learningalgorithms such as the Support Vector Machine (SVM) and the Bayesian Network (BN) and to select the most effective oneto assess the extent of CKD patients. The results show that SVM classification has proven its performance in predicting the best results in terms of accuracy and minimum execution time [8].

Maryam Soltanpour Gharibdousti, Kamran Azimi, Saraswathi Hathikal, Dae H Won [9] applied different machine learning classification algorithm such as Decision Tree, Linear Regressing, Super Vector Machine, Naive Bayesian and NeuralNetwork. Results show that, first except of Neural Network which is sensitive to scale of data, performance of other classifiers are almost the same for original and normalized dataset.

III. DATA DESCRIPTION AND ANALYSIS

A multivariate dataset obtained from UC Irvine Machine Learning Repository sourced from Apollo Hospitals, TN, India- containing about 400 data instances of people aged ranging from 2-90 has been used in this paper [10].

There are about 25 features in total, majority of which are clinical with some also being physiological.

We use 24 features + class = 25 (11 numeric ,14 nominal)

TABLE II

Sr.No.	*Feature*	*Abbreviation*	*Description*
1	*class*	*class*	*class - (ckd,notckd)[nominal]*
2	*age*	*age*	*age in years [numerical]*
3	*bp*	*blood pressure*	*bp in mm/Hg [numerical]*
4	*sg*	*specific gravity*	*Sg-(1.005,1.010,1.015,1.020,1.025)*

			[nominal]
5	*al*	*albumin*	*al - (0,1,2,3,4,5) [nominal]*
6	*su*	*sugar*	*su - (0,1,2,3,4,5) [nominal]*
7	*rbc*	*red blood cells*	*rbc(normal,abnormal) [numerical]*
8	*pc*	*pus cell*	*pc - (normal,abnormal) [nominal]*
9	*pcc*	*pus cell clumps*	*pcc - (present,notpresent) [nominal]*
10	*ba*	*bacteria*	*ba - (present,notpresent) [nominal]*
11	*bgr*	*blood glucose random*	*bgr in mgs/dl [numerical]*
12	*bu*	*blood urea*	*bu in mgs/dl [numerical]*
13	*sc*	*serum creatinine*	*sc in mgs/dl [numerical]*
14	*sod*	*sodium*	*sod in mEq/L [numerical]*
15	*pot*	*potassium*	*pot in mEq/L [numerical]*
16	*hemo*	*haemoglobin*	*hemo in gms [numerical]*
17	*pcv*	*packed cell volume*	*pcv in % (L/L) [numerical]*
18	*wc*	*white blood cell count*	*wc in cells/cumm [numerical]*
19	*rc*	*red blood cell count*	*rc in millions/cmm [numerical]*
20	*htn*	*hypertension*	*htn - (yes,no) [nominal]*
21	*dm*	*diabetes mellitus*	*dm - (yes,no) [nominal]*
22	*cad*	*coronary arterydisease*	*cad - (yes,no) [nominal]*
23	*appet*	*appetite*	*appet - (good,poor) [nominal]*
24	*pe*	*pedal edema*	*pe - (yes,no) [nominal]*
25	*ane*	*anemia*	*ane - (yes,no) [nominal]*

Features: Abbreviations & Description

IV. PROPOSED SYSTEM

ML models can be constructed from different features/variableslike potassium, anemia, pedal edema etc. which are proportionate in the development of Renal disorders. Out of thenumerous features, the ones that contribute the most in determination are known as best features. Identifying them is carried out utilizing SelectKBest, chi2 after performing label encoding for all data that is qualitative in nature. Based on the feature scores, we select the best 8 columns only. Function file is constructed independently which is therafter invoked in main. Forecast of model is finished utilizing eXtreme Gradient Boosting, and to hypertune the params randomized search CV is employed.

Implementation:

Firstly, a .CSV file containing dataset for training and testing isattached. The dataset is split into two halves in the ratio of 80:20, where 80% is used to train the model and the remaining 20% will be used to test/validate the result/prediction.

Since the data is raw in nature and not in the appropriate form to use as it is, we need to transform it into suitable form. Then,data pre-processing is done which involves renaming the columns according to text description file, for ex: al to albuminand changing the datatypes of required columns for computational purpose like, object to float64.

Next data cleaning is done. This involves removing unnecessary parameters, differentiating & separating attribute columns of various datatypes and removing NAN (Not a Number) values (imputation). Now, this dataset can be used forour prediction model.

Data visualization is done with the help of matplotlib-plotly interface and seaborn graphical libraries of python. Histograms, Count plots, Kernel Density Estimate plots etc. help us study the skewness, data-correlation, aberration and distribution.

We utilize sci-kit learn library as it provides tools for feature selection, classification-regression and model selection. Categorical variables are converted into

numerical form so thatit becomes machine readable. This is known as Label Encoding.

Feature Selection:

1. SelectKBest:

It's a filter based univariate selection technique selecting best features on the basis of rank order. The correlation of target variable with features help determine statistical scores which further determine the ranks. Variables that are related strongly with output variable can be determined and features with highest K-score are retained.

Parameters: score_func-This function takes two arrays X, Y where X alludes to the indicators and Y alludes to the objective variable. It returns a couple of clusters (scores, pvalues) for each highlight. The primary K highlights of X on the basis of highest score are retained [11].

1. chi2:

chi-squared statistics test is used with SelectKBest and basically works as a scoring function. If after computingthe score we find its value is less, it means that feature is independent of class label and alternatively high value indicates that the feature is not independent of class label. By doing this we can find the best fit features withhighest scores.
This function can be called as:

- select = SelectKBest(score_func=chi2, k=3)

We'll fit and transform method on training X and Y data.

- Z = select.fit_transform(X,Y)

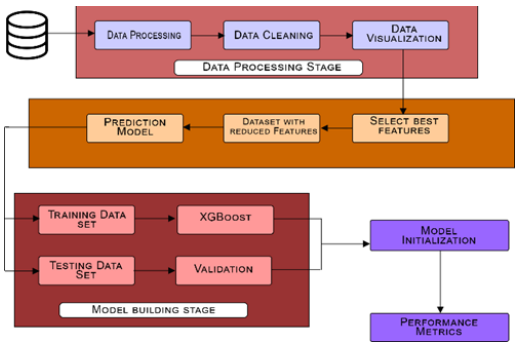

Fig. 1. CKD Prediction System Architecture

XGBoost:

eXtreme Gradient Boosting is a supervised machine learning algorithm based on Decision Trees (CART-Classification and Regression Trees) as Base Learners.

Boosting is an ensemble learning technique where the models are built sequentially, where the previous model's output serves as an input to the next. In each iteration, the error of previous model is corrected by the next one. The correctly predicted results are given a lower weight, and the ones predicted wrong are assigned a higher weight. Then final result is produced usinga weighted average.

In Gradient Boosting the errors are minimized using a gradientdescent algorithm. Effectively it's an iterative optimization algorithm which is applied to minimize a loss function. This loss function is a measure of difference between our prediction fromthe actual result for a given data point. Also, weights are adjusted based on a gradient (slope/ 2^{nd} differential of univariate function) reflected by the direction in the loss function where the loss reduced the quickest.

Thus, working of XGBoost can be concluded as firstly 'k' bestfeatures are selected, then trees that are small with fewer splitsare built. Subsequently other trees are built with considering errors present in previous updated in residual errors [12].

We use the xGBClassifier model for binary classification (CKD-yes/no) with objective loss function "binary:logistic".

Before the learning process begins parameters are set that are external to the model, these can be fine-tuned and can directly affect the performance of the model. Such parameters are known as Hyperparameters. For Tree-based learners' common parameters are learning_rate, max_depth, gamma, colsample_bytree, min_child_weight. We use Scikit-learn's hyperparameter optimizer function-RandomisedSearchCV for Hyperparameter tuning. It uses a large range of hyperparameters values, and randomly iterates a number of times (can be specidied) over combinations of those values.

Testing and Performance Metrics:

20% of the dataset is used for testing and validation purpose. Here we utilize a Confusion matrix [13] to map the correctly andincorrectly predicted values as well as other performance metrics such as Precision, Recall, Accuracy & F-measure.

Type	*Actual Positive*	*Actual Negative*
Predicted Positive	*TP (11)*	*FP (10)*
Predicted Negative	*FN (01)*	*TN (00)*

Fig. 2. Confusion Matrix: Binary Classification

True Negative (TN) 00, False Negative (FN) is represented by01, False Positive (FP) is represented by 10 and 11 represents True Positive (TP).

TABLE III

True Negative:	*When the real and predicted labels of a sample are negative.*
True Positive:	*When the real and predicted labels of a sample are positive.*
False Negative:	*When the real value of a sample is positive while its predicted label is negative.*
False Positive:	*When the real value of a sample is negative while its predicted label for the sample is positive.*

Confusion Matrix Terms

The formulas and description of other performance metrics areas below:

Precision: Accuracy alludes to the proportion of rightly classified classes of CKD amongst all positive classes.

Precision=TP/(TP+FP)

Recall: Review addresses the proportion of rightly classified classes to the rightly predicted classes.

Recall = TP/(TP+FN)

Accuracy: It is the most natural presentation measure alluding to the proportion of rightly classified classes amongst all classes.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

f-B eta-Measure: The f-Beta measure is an exhibition metric used to contrast two models and differing review and accuracy by punishing the outrageous qualities.

F-measure = $((1+\beta^2)*Recall*Precision))/(\beta^2)*(Recall+Precision)$

Picking a reasonable Beta worth relies upon the application area. Whenever Beta worth is 1, equivalent weightage is givento both accuracy and review. At the point when more accentuation is given on accuracy, a worth lesser than 1 is picked and when review is given higher significance, a worth more noteworthy than 1 is picked.

V. RESULTS

The confusion matrix obtained is:

```
from sklearn.metrics import confusion_matrix,accuracy_score
confusion_matrix(ytest,ypred)
```

```
array([[51,  1],
       [ 1, 27]], dtype=int64)
```

The performance metrics are as follows:

```
from sklearn import metrics

print("Accuracy score:",metrics.accuracy_score(ytest, ypred))
print("Precision score:",metrics.precision_score(ytest, ypred))
print("Recall score:",metrics.recall_score(ytest, ypred))
print("F1 Score :",metrics.f1_score(ytest, ypred))
```

```
Accuracy score: 0.975
Precision score: 0.9642857142857143
Recall score: 0.9642857142857143
F1 Score : 0.9642857142857143
```

The system gave astonishing results with a high accuracy of 97.5%. Precision and Recall are both 0.96 as our FP and FN areequal i.e., 1. Since both values are very close to 1, this proves that our model is an excellent classifier. Also, as our Recall andPrecision are equal, the F1 score is also 0.96 which is again extremely close to 1.

VI. USE-CASE DIAGRAM

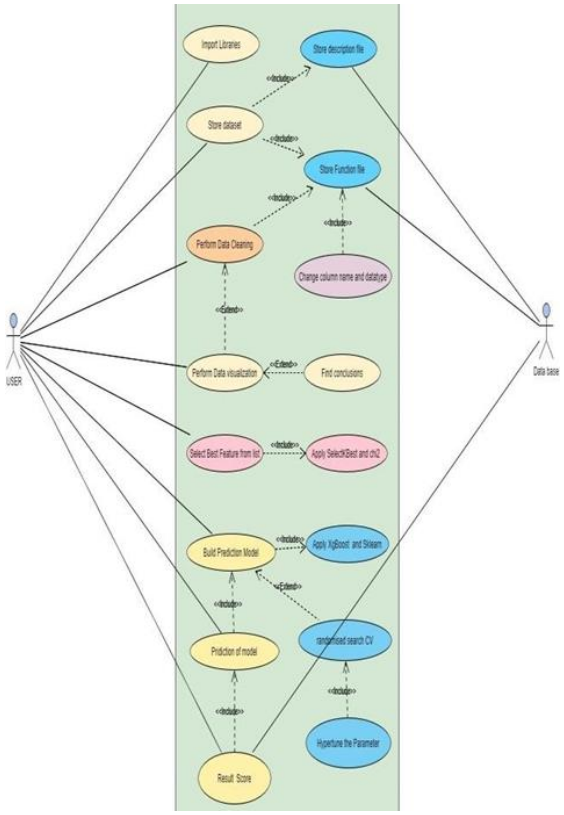

Fig. 3. USE Case UML Diagram

The diagram shown above is a USE CASE diagram that describes our project execution and interaction with the other entities. Although we can use other UML Diagrams too like Object diagram, Class diagram, Component diagram etc. but as for the requirement we have chosen USE CASE diagram.

Marked ovals, stick figures and straight lines address use cases, actors in the process, and involvement of actors in the system respectively. Thus, the diagram can be deemed as an outline ofthe connection between the aforementioned terms.

Components in USE CASE diagram:

- Actors
- Include
- Extend
- USE CASES

Implementation: In the provided use case diagram, the user actas an actor which interacts with our system (System here means the whole process including interactions and actions of actors with the system). In the diagram we can see Extend and Includevery often.

Extend use case is usually optional and can be triggered conditionally whereas the base use case is deficient without theincluded use case. The included use case is required compulsorily and not discretionary. We can see storing data description, function file are few processes that are mandatory for our data set and thus uses Include whereas finding conclusion or applying visualization is not mandatory and thus uses Extend. The whole diagram shows a high-level overview of the relation and action of actor with the system.

VII. CONCLUSION

The advancements and innovations in the field of science and technology pave the path to explore different possibilities in multidisciplinary domains particularly healthcare & wellness. While building this system we have implemented with the state- of-the-art prediction algorithm XGBoost along with feature selection algorithms- SelectKBest and chi2 to predict whethera person is suffering from CKD or not. The hyperparameters were tuned using RandomisedSearchCV. The Results gave a astonishing 97.5% accuracy along with high precision, recall and f1 score all being equal to 0.96(extremely close to 1). Thisproved that the system is an excellent classifier and fulfills our objective of successfully predicting CKD/CRD. Due to this the early identification of high-risk patient may enable timely management and lead to improved outcomes and reduce healthcare expenditure.

The scope of this system is limited to dataset from India context. Subjecting the system to datasets and records collected from all over the word will increase the learning power thus also increasing the prediction ability.

Future scope includes,

1. The model can be used with other datasets available and toanalyze its performance.
2. Other available feature selection techniques can be utilized. 3.Risk Estimation Analysis & Preventive Prescription can beadded as in-built features.

ACKNOWLEDGMENT

We would like to express gratitude to our guide Prof. Rohini Jadhav, who generously provided knowledge and expertise and without whom the project and paperwork could not have been possible.

We also took the help and guidance of Prof. Madhavi Mane, who provided us valuable feedback and constructive suggestions, conveying our gratitude to you.

REFERENCES

[1] Worldwide, local, and public weight of ongoing kidney infection, 1990-2017: a methodical investigation for the Global Burden of Disease Study2017

[2] CDC-Centers for Disease Control and Prevention,Retrieved from https://www.cdc.gov/kidneydisease/basics.html#:~:te xt=Every%2024%20hours%2C%20360%20people,out %20of%204%20 new%20cases

[3] National Kidney Foundation, New York, NY 10016, Retrieved from https://www.kidney.org/sites/default/files/01-10-7278_HBG_CKD_Stages_Flyer3.pdf

[4] A.S. Albahri, “Role of Biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (Covid-19): A Systematic Review”, Journal of Medical Systems,Vol. 44, pp. 1-11, 2020.

[5] L.J. Muhammad, “Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients’ Recovery”, SN Computer Science, Vol. 1, No. 4, pp. 1-7, 2020

[6] M. Shahbaz, S. Ali, and A. Umer, "Order of Alzheimer's Disease utilizing Machine Learning Techniques", Proceedings of International Conference on Mining and Multimedia Data, pp. 296-303, 2019.

[7] Guneet Kaur, "Foresee Chronic Kidney Disease utilizing Data Mining in Hadoop, International Conference on Inventive Processing and Informatics,2017.

[8] Z. Xu and Z. Wang, "A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selectionof Random Timberland and XGBoost Ensemble Classifier", Proceedings of Worldwide Conference onAdvanced Computational Insight, pp. 278-283, 2019.

[9] Maryam Soltanpour Gharibdousti, Kamran Azimi, Saraswathi Hathikal, Dae H Won, "Forecast of Chronic Kidney Disease Using Data MiningTechniques" Industrial and Frameworks Engineering Conference, 2017

[10] L.Jerlin Rubini (2015), UCI Machine Learning Repository: Chronic_Kidney_Disease Data Set [https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease]. Irvine, CA: University of California, School of Information and Computer Science.

[11] Sayank Paul, "Fledgling's Guide to Feature Selection in Python",Retrieved from https://www.datacamp.com/tutorial/feature-selection-python

[12] Jason Brownlee, "A Gentle Introduction to Xgboostfor applied AI", Retrieved from https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

[13] Sarang Narkhede, "Figuring out Confusion Matrix",Retrieved from https://towardsdatascience.com/understandingconfusi onmatrix-a9ad42dcfd62