Chronic Kidney Disease (CKD) Detection

Md. Ashab Mohiuddin, Md. Younus Hossain Ahsan, Md. Abid Rahman, and Fahim Ahmed
Department of Computer Science And Enginnering
Ahsanullah University Of Science And Technology
{190104128, 190104131, 190104141, 190104149}@aust.edu

***Abstract*—One of the most leading causes of death, with a vast number of patients throughout the world, Chronic Kidney Disease, in short CKD, is very hard to identify at an early stage as it has no symptoms. But if found at an early stage, that patient can be diagnosed and can be cured. So early detection of this disease has a greater impact and can save many lives from death. Various researches have been carried out using machine learning techniques on the detection of CKD at the premature stage. Their focus was not mainly on the specific stages of prediction. So, in our study, both binary and multi classification for stage prediction have been carried out. The prediction models used include Random Forest (RF), Support Vector Machine (SVM) and Decision Tree (DT). At last, we implement a confusion matrix to measure the accuracy, precision, recall and f1-score and compare all the models. By this study, we want to achieve the best possible outcome to help and contribute in identifying Chronic Kidney Disease at an early stage.**

***Index Terms*—Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Confusion matrix.**

I. Introduction

Chronic kidney disease, also known as chronic renal disease or CKD, is a condition characterized by a gradual loss of kidney function over time [1]. The main risk factors for developing kidney disease are diabetes, high blood pressure, heart disease, and a family history of kidney failure. Diabetes and high blood pressure are the most common causes of kidney disease [2]. The kidneys' main job is to filter extra water and waste out of your blood to make urine. To keep your body working properly, the kidneys balance the salts and minerals—such as calcium, phosphorus, sodium, and potassium—that circulate in the blood. Your kidneys also make hormones that help control blood pressure, make red blood cells, and keep your bones strong. Kidney disease often gets worse over time and may lead to kidney failure. If your kidneys fail, you will need dialysis or a kidney transplant to maintain your health [3].

According to the report for World Kidney Day 2019, at least 2.4 million people die every year due to kidney-related diseases. Currently, it is the 6th fastest-growing cause of death worldwide. CKD is becoming a challenging public health problem with increasing prevalence worldwide. Its burden is even higher in low-income countries where detection, prevention, and treatment remain low [4].

The National Kidney Foundation classifies stages of CKD into five based on abnormal kidney function and reduced glomerular filtration rate (GFR), which measures a level of kidney function. The mildest stage (stages 1 and 2) is known with only a few symptoms, and stage 5 is considered end-stage or kidney failure. The cost of renal replacement therapy (RRT) for total kidney failure is very expensive. The treatment is also not available in most developing countries, like Ethiopia. As a result, the management of kidney failure and its complications is very difficult in developing countries due to a shortage of facilities and physicians and the high cost of treatment [5][6].

Predictive analysis using machine learning techniques can be helpful for early detection of CKD and efficient and timely interventions [7]. In this study, Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) have been used to detect CKD. Most of the previous research focused on two classes, which makes treatment recommendations difficult because the type of treatment to be given is based on the severity of CKD.

II. Related Works

Different machine-learning techniques have been used for effective classification of chronic kidney disease from patients' data.

Charleonnan et al. [8] did a comparison of the predictive models such as K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and decision tree (DT) on the Indian Chronic Kidney Disease (CKD) dataset in order to select the best classifier for predicting chronic kidney disease. They have identified that SVM has the highest classification accuracy of 98.3% and the highest sensitivity of 0.99.

Salekin and Stankovic [9] evaluated classifiers such as K-NN, RF, and ANN on a dataset of 400. Wrapper feature selection was implemented, and five features were selected for model construction in the study. The highest classification accuracy is 98% by RF and a RMSE of 0.11.

S. Tekale et al. [10] worked on "Prediction of Chronic Kidney Disease Using Machine Learning Algorithms" with a dataset consisting of 400 instances and 14 features. They have used decision trees and support vector machines. The dataset has been preprocessed, and the number of features has been reduced from 25 to 14. SVM is stated as a better model with an accuracy of 96.75%.

Xiao et al. [11] proposed prediction of chronic kidney disease progression using logistic regression, Elastic Net, lasso regression, ridge regression, support vector machine, random forest, XGBoost, neural network, and k-nearest neighbor and compared the models based on their performance. They have used 551 patients' history data with proteinuria and 18 features and classified the outcome as mild, moderate, or severe. They

have concluded that logistic regression performed better with an AUC of 0.873 and sensitivity and specificity of 0.83 and 0.82, respectively.

III. Dataset

A. Data Description

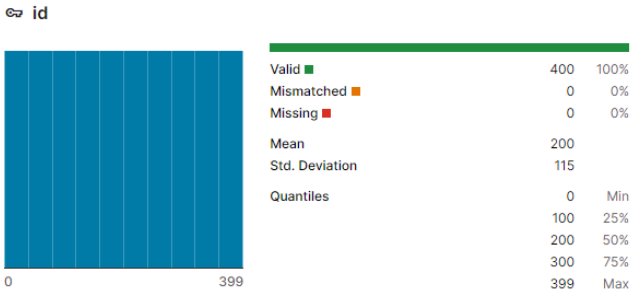
We have a dataset with 1200+ instances and 10 attributes that are labeled as id, age, bp (blood pressure), sg (specific gravity), al (AL amyloidosis), su (stroke unit), rbc (red blood cell), pc (post-cigarette), pcc (prothrombin complex concentrate), and ba (bronchial asthma). Patient age, blood pressure, and sugar level are vital attributes for detecting whether CKD is in the body or not. The main test for chronic kidney disease is a blood test. The test measures the levels of a waste product called creatinine in our blood. Using blood test results, age, size, gender, and ethnic group, calculate how many milliliters of waste kidneys should be able to filter in a minute.

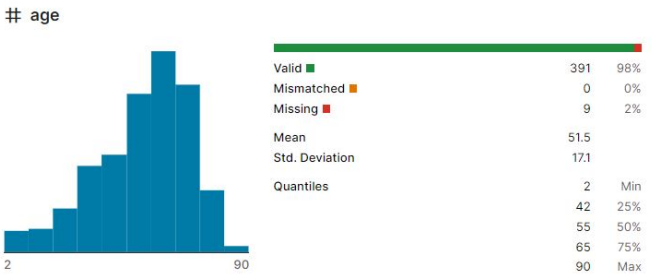
B. Data Collection

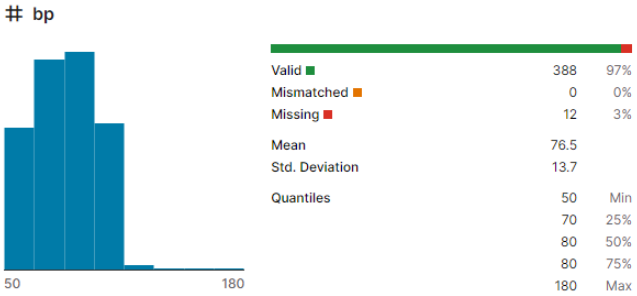
Our source for collecting our dataset is vastly dependent on the website Kaggle, some GitHub repositories, and some other hospital resources.

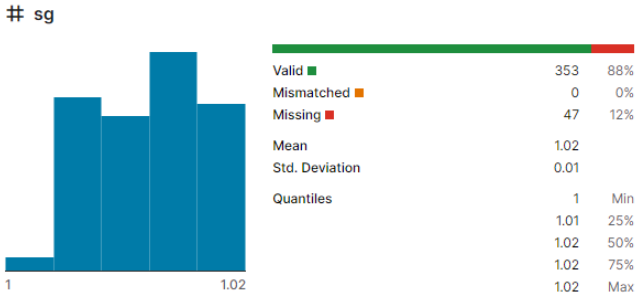
C. Data Analysis

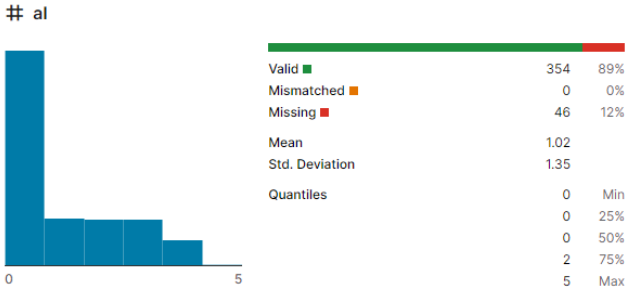
A graphic representation of all the attributes of the CKD dataset is given below:

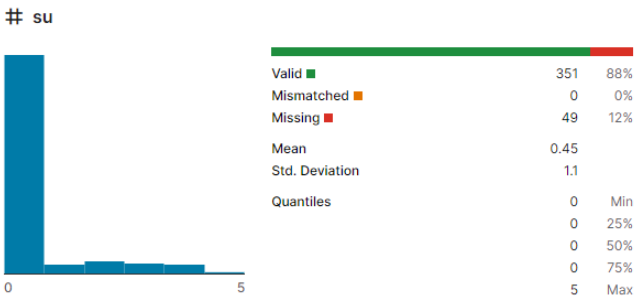


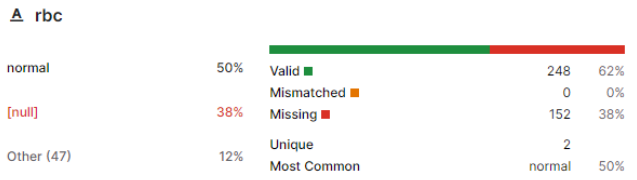


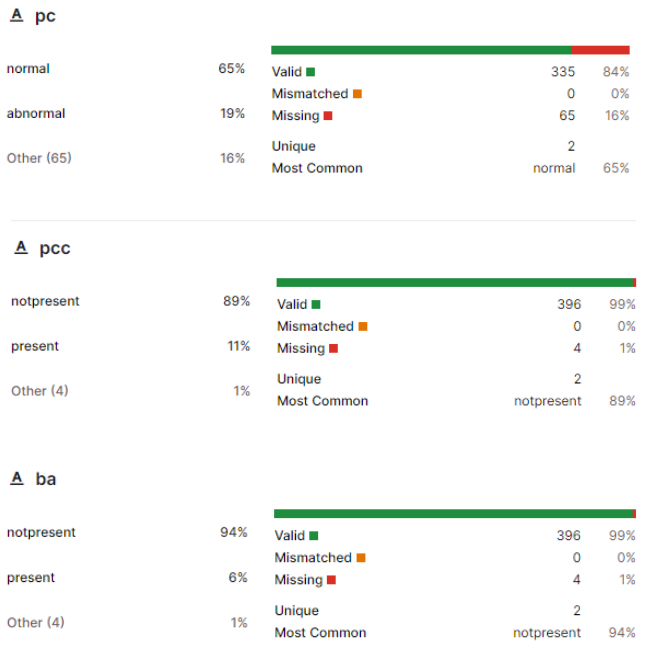












Figure(s): Attributes of the CKD dataset

IV. Methodology

A. Feature extraction

Feature extraction performs several tasks related to data preprocessing, feature selection, and visualization using the Pandas, scikit-learn, Seaborn, and Matplotlib libraries. 27 columns amidst 26 are features, and 1 column named 'classification' is the target.

The columns (rbc, pc, pcc, ba, htn, dm, cad, appet, pe, ane, classification) have categorical values that need to be converted. Before using them for the chi-square feature extraction method. Chi Square provides the attributes, which provide more information related to classification. A graph is given below of chi-square values:

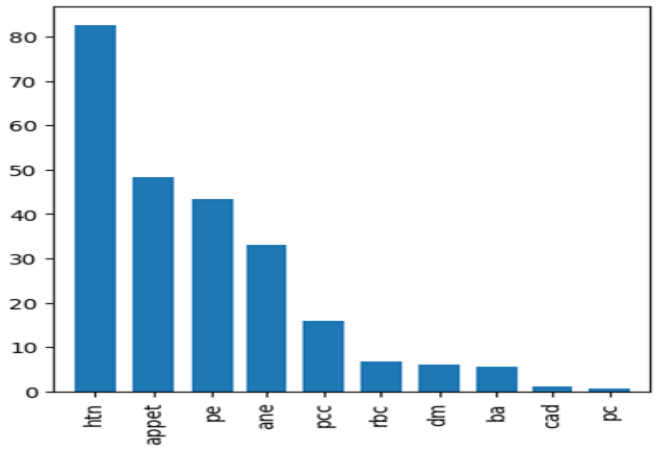


Figure: Graphical representation of chi-square values

From the above graph, the columns htn, appet, pe, ane, and pcc give much information to classify the target. So, the first few, like 3 or 4 columns, could be enough for a model to reach a verdict with high probability.

For continuous values, they have to be extracted using the co-relation matrix method. A graph is given below:

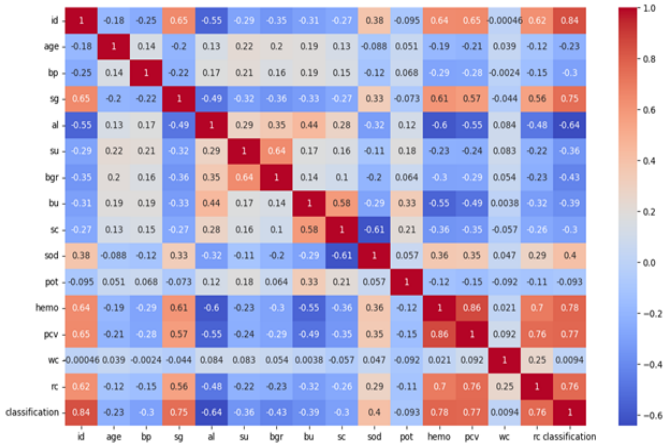


Figure: Graphical representation of co-relation matrix

B. Machine Learning Models

We use different machine learning models such as Logistic Regression, KNN, SVM, Decision Tree, etc.

- **Logistic Regression (LR):** Logistic regression predicts binary outcomes by calculating the probability using feature weights and the sigmoid function. Training adjusts these weights to minimize the difference between predicted probabilities and actual outcomes.
- **K-Nearest Neighbors (KNN):** K-Nearest Neighbors (KNN) predicts by selecting (k) closest training examples to a new input and deciding based on the majority class or average value. Used for classification and regression, it doesn't need explicit training but requires tuning (k) and distance metrics.
- **Support Vector Machine (SVM):** SVM optimally separates classes by maximizing margin with support vectors. It handles nonlinear data via kernels and accommodates misclassifications with a soft margin. Efficient and versatile, SVM suits classification and regression tasks.
- **Decision Tree:** Decision Trees divide data using features, constructing a tree for outcome prediction at leaf nodes. They iteratively select the best feature to split the data, creating a structured decision path. This process continues until a D-F1-Score stopping point is reached, providing a clear and interpretable decision-making model.
- **Random Forest:** Random Forest combines multiple decision trees using bootstrapped samples and random feature selection to enhance predictions. It reduces overfitting and provides more accurate results by aggregating the

predictions of individual trees. This ensemble technique is versatile and effective for various machine-learning tasks.

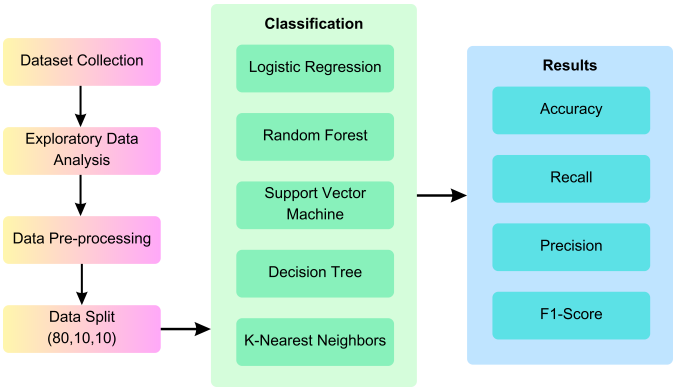


Figure: Methodology

V. Result Analysis

A. Accuracy:

Accuracy is the proportion of correctly predicted instances, but it's not the main metric for our imbalanced dataset. Additionally, TP, TN, FP, and FN represent the counts of true positives, true negatives, false positives, and false negatives, respectively. The mathematical representation of accuracy is presented below:

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

B. Recall:

Recall calculates the proportion of true positive predictions among all actual positive instances in the dataset. This metric is essential for understanding the performance of a classification model, especially when dealing with imbalanced datasets. The mathematical representation of recall is shown below:

Recall = TP / (TP + FN)

C. Precision:

Precision measures the proportion of true positive predictions among all positive predictions made by the model. This metric is also essential for understanding the performance of a classification model, especially when dealing with imbalanced datasets. The mathematical representation of precision is shown below:

Precision = TP / (TP + FP)

D. F1-Score :

The F1 score is a single evaluation metric that combines precision and recall to provide a balanced measure of a model's performance in binary classification tasks. It is particularly valuable when dealing with imbalanced datasets. F1 score for a particular class follows below:

F1 score = (2 x Recall x Precision) / (Recall + Precision)

Table for Result

Machine learning Models	Accuracy
Logistic Regression	75.23%
Random Forest	78.55%
SVM	79.51%
Decision Tree	70.35%
KNN	73.87%

Classification Report For Logistic Regression

	Precision	Recall	F1-Score
0	0.73	0.92	0.81
1	0.95	0.04	0.02

Classification Report For Support Vector Machine

	Precision	Recall	F1-Score
0	0.66	0.91	0.72
1	0.69	0.29	0.40

Classification Report For Decision Tree

	Precision	Recall	F1-Score
0	0.67	0.66	0.66
1	0.49	0.50	0.50

Classification Report For KNN

	Precision	Recall	F1-Score
0	0.68	0.74	0.71
1	0.53	0.46	0.49

VI. Conclusion And Future Work

Early prediction is crucial for both experts and patients to prevent and slow down the progression of chronic kidney disease to kidney failure. In this study, three machine-learning models, like Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, and KNN, were used to build the proposed models, with Support Vector Machine getting the highest accuracy of 79.51% and Random Forest getting closer to this with an accuracy of 78.55%. In the future, we will like to add more data to better classify and detect chronic kidney disease.

References

1 "Chronic Kidney Disease (CKD) - NIDDK." - National Institute of Diabetes and Digestive and Kidney Diseases, — niddk.nih.gov/health-information," 2023, [Accessed 22 Aug. 2023].
2 "Chronic Kidney Disease (CKD) - NIDDK." - National Institute of Diabetes and Digestive and Kidney Diseases, — niddk.nih.gov/health-information," 2023, [Accessed 22 Aug. 2023].
3 "Chronic Kidney Disease (CKD) - NIDDK." - National Institute of Diabetes and Digestive and Kidney Diseases, — niddk.nih.gov/health-information," 2023, [Accessed 22 Aug. 2023].
4 "George C, Mogueo A, Okpechi I, Echouffo-Tcheugui JB, Kengne AP. Chronic kidney disease in low-income to middle-income countries: The case f increased screening. BMJ Glob Heal. 2017;2(2):1–10."
5 "Stanifer JW, et al. The epidemiology of chronic kidney disease in sub-Saharan Africa: A systematic review and meta-analysis. Lancet Glob Heal. 2014;2(3):e174–81."
6 "AbdElhafeez S, Bolignano D, D'Arrigo G, Dounousi E, Tripepi G, Zoccali C. Prevalence and burden of chronic kidney disease among the general population and high-risk groups in Africa: A systematic review. Bmj Open. 2018;8:1."
7 "Agrawal A, Agrawal H, Mittal S, Sharma M. Disease Prediction using Machine Learning. SSRN Electron J. 2018;5:6937–8."

8 “Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S, Ninchawee N. Predictive analytics for chronic kidney disease using machine learning techniques. Manag Innov Technol Int conf MITicon. 2016;80–83:2017.”
9 “Salekin A, Stankovic J. Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes. In: Proc. - 2016 IEEE Int. Conf. Healthc. Informatics, Ichi 2016, pp. 262–270, 2016.”
10 “Tekale S, Shingavi P, Wandhekar S, Chatorikar A. Prediction of chronic kidney disease using machine learning algorithm. Disease. 2018;7(10):92–6.”
11 “Xiao J, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. J Transl Med. 2019;17(1):1–13.”