Kidney failure prediction at an early stage using Machine Learning: A Comparative Study

Vijay Kumar¹, Amrita ticiku², Rachna Narula³

1. Assistant professor, Bharati Vidyapeeth College of Engineering, New Delhi, CSE Department
2. Assistant professor, Bharati Vidyapeeth College of Engineering, New Delhi, CSE Department
3. Assistant professor, Bharati Vidyapeeth College of Engineering, New Delhi, CSE Department

Abstract:
Chronic kidney disease (CKD) is a medical complication of a person due to which the kidney can't filter the blood due to which the body fills with extra water and waste products. It can lead to stroke, heart attack, heart failure, swelling of the feet and kidney failure, which can lead to death. The global health problem is growing rapidly as more and more people are being diagnosed with CKD. With advancing technology, as well as ongoing medical research, machine learning is being used in the healthcare sector to diagnose many diseases early. ML algorithms and decoding methods have been very useful in extracting, analyzing data and making predictions when a person is positive or negative about a disease based on the given data sets. ML algorithms and in-depth reading have been proven to be very true in detecting CKD early. Machine learning algorithms, Cat boost classifier, Support Vector Machine (SVM), Decision Tree (DT), Random Forest, KNN, ANN were studied and applied in this work to perform comparative analysis to shape a ML model which will accurately predict if a person is positive to CKD or not. This paper uses pre-data processing, including background and above-mentioned machine learning algorithms to build the most accurate model to accurately detect this disease CKD and perform a comparative research of various ML algorithms for prognosis of CKD.

Keywords: Machine Learning, Data Mining, ANN, KNN, Decision Trees, Logistic Regression, Support Vector Machine, Data Preprocessing, Feature extraction.

I. Introduction

When the blood filtering capability of the kidney is hindered, the disease is called Chronic Kidney disease. Chronic Diseases refers to diseases persisting for a long time. When the kidney fails to strain the waste products and blood, fluid starts building up in the body due to which excess fluids get deposited in the body and several consequences are seen, like swelling in feet, irritation, and high blood pressure due to which there is an increased risk of heart related diseases (Heart Attack, Heart Failure, Difficulty in breathing). These fluid buildup can cause deposition of some salts in the body which can cause cystic ovaries in women. Some complications of CKD are:

- High BP
- Low blood count
- Weaker bones and muscles
- Nerve damage
- Poor nutritional health
- Fluid retention
- Uremia

CKD is evaluated by measured glomerular filtration rate (GFR) and creatinine level which is calculated by blood test KFT. All such parameters like creatinine level, GFR value are calculated using KFT (kidney function test). In severe cases, CKD may lead to kidney transplants or permanent dialysis. This paper is made to detect CKD at an early stage, as in later stages, the chances of survival and recovery are very low. We have used various ML algorithms like Cat boost Classifier, KNN, ANN, Decision Tree, SVM, and Random Forest and also indicate a more accurate prediction model by performing a comparative analysis on the basis of efficiency of predicted outcomes by different techniques.

I. Literature Survey:
Md. Rashed-Al-Mahfuz [2] introduces a reliable method of classifying CKD and selecting annotations with simplified enhancement and cost effectiveness. In this study the classifiers were constructed using different algorithms such as RandomForest, Gradient-boosting, and extreme gradient-boosting, logistic-regression, and support-vector machines. The outcome of this study showed that the key features recognized by SHAPE were congruous with present-day medical diagnosis. It was described that the RandomForest method provided maximum accuracy with pathologically divided sets of attributes. Therefore the proposed RF separator and reduced diagnostic criteria can be used to reduce money and time.

Anusorn Charleonnan [3] presented a paper in which Machine learning strategies were used to predict CKD using clinical data. These speculative models are based on a database of CKD and the efficiency of these models is put together to select the best CKD forecast.

Gazi Mohammed Ifraz [3] presents a paper in which they have done a comparative analysis of different Machine Learning algorithms and When contradictory verification measures are used in the diagnosis of chronic kidney disease, the LR method goes beyond other
D.M. Perera[4] shows an approach to control the CKD using a proper diet plan by using the classification techniques on the test result obtained from the patient's medical records. In this research classifiers are constructed using different algorithms like Multiclass-Decision-Jungle, Multiclass-Decision-Forest, and Multiclass-Neural-Network and Multiclass-Logistic-Regression[3].

Akash Maurya[5] presents Machine Learning techniques and suggests a proper diet plan for CKD patients using classification algorithmson medical test information. Dietary recommendations for patients will be given according to the potassium level calculated using potassium levels in the blood to delay the progression of CKD.

H. Zhang et al[6] investigatesthe effectiveness of Artificial Neural Network (ANN) models while using the survival forecast for Chronic Kidney Disease (CKD) patients.

T. G. Kaur et al[7] predicted chronic kidney disease applying several data mining techniques in the Hadoop environment. ML models, KNN and SVM were used in the study. Creatinine-level, and GFR value in the blood is a significant factor used for the prognosis for the disease. This model will help to reduce costs and increase the accuracy of results.

Jiong Ming Qin[8] proposed CKD diagnostic method based on data attribution and sampled diagnostics. After a random drop in non-data values set by KNN compliance, the integrated model can achieve satisfactory accuracy.

Qin, J.[9] developed a classification model for anticipating transitional intervals of Kidney disease stages 3 to 5 and also used Decision Tree, K-nearest neighbor, Naïve Bayes, and Artificial neural networks to obtain information and create a classification model with a selected set of features.

S. Vijayarani et al[10] predicted kidney diseases by using Support Vector Machine and Artificial Neural Network. The analysis incorporates the behavior of the two techniques above for accuracy and performance time.

S Krishnamurthy[12] In this paper, he established and assessed a series of models based on Artificial Intelligence taking into account small variables such as gender, age, comorbidities, and medications. These models anticipate patients' risk of developing CKD within six or twelve months. Among the several techniques tested. Highlights contain diabetes, age, gout, and the use of sulfonamides and angiotensin, all of which make sense according to CKD.

Qiong Bai[13] 5 ML techniques, incorporate Logistic Regression, Naïve Bayes, Random Forest, Decision Tree and K-nearest neighbor have been taught and examined using five-fold verification. The strength of each model is comparable to that of the Kidney Failure Risk Equation (KFRE). The three ML models, which include retreat, Naive Bayes and the Random Forest, display similarity and susceptibility compared to KFRE. KFRE has high accuracy and precision. This analysis demonstrated the power of ML in evaluating CKD predictions based on readily available findings. Three ML algorithms with optimal performance and sensitive points propose potential use in patient evaluation.

Marwa Almasoud[14] This paper aims to analyze the ability of machine learning methods for predicting CKD using a small set of features. Various mathematical analysis were performed on it to remove obsolete features such as ANOVA tests, Pearson correlations, and Cramer's V tests. The more accurate results come from the Gradient Boosting classifier.

Arsh K. Jain[15] In this paper performed a structured evaluation of the impact of neutral-pH, low-GDP, collate to standard-PD solutions. In this analysis, neutral pH, low GDP solutions caused a large amount of urine over 6 months and better RRF retention during all treatment periods.

Giorgina Barbara Piccoli[16] This study aimed to identify factors associated with the risk of adverse effects associated with pregnancy in females with CKD, with special significance of stage 1 CKD. We aim to work on whether the adverse pregnancy results in females with stage 1 CKD were due to high blood pressure, proteinuria, systemic disease, or other clinically unrelated CKD-related factors.

Kunihiro Matshushita[17] In this, Two important steps for CKD, eGFR and albuminuria, enhance the prognosis of CVD risk in addition to the usual risk factors, especially CVD death diseases. For the same purpose, the purpose of this study was to predict various stages of CKD using machine-readable algorithms found in the medical records of the affected people. Specifically, this study used Random Forest and J48 algorithms to obtain a stable and effective model for determining the various stages of CKD with complete medical accuracy. A relative study of the outcome uncovered that J48 predicted CKD in all categories better than the random forest. Manal.

AAbdel-Fattah[19] this paper proposes a combination of machine learning strategies that include feature selection and how to classify a learning machine based on largedatasets used to diagnose CKD. Feature selection methods, i.e., Relief-F and Chi-squared, were used to select key features. Six distinctive methods were applied in this study: Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), vector support machine (SVM), and Gradient-Boosted Trees as integrated learning
algorithms. Or in each algorithm, the opposite confirmation results and test results are computer-based based on full features. Relief-F and selected chi-squared features. The outcome shows that SVM, DT, and GBT Classifiers with selected features obtained excellent performance and greater accuracy.

These CKD measures for predicting CVD, especially if the data are already limited to other clinical indicators. When the risk profile is to be continuously improved other biological symptoms may be helpful, and indicating the pathophysiological process of CVD such as coronary artery calcium and high-sensitivity troponin T heart appear specifically promising. Future clinical recommendations may need to be updated on how to include CKD values and other biological indicators into CVD predictors, relying on the results of an interesting CVD, target population, and the possibility of such measures and biomarkers.

Hamida Ilyas[18] prior identification and treatment of CKD is highly advisable as it can lead to the avoidance of undesirable side effects. A machine learning method is widely recommended for prior identification of symptoms and for the prognosis of several And heart failure. It seems valid to consider.

II. Methodology

1. Decision Tree

It is a form of supervised learning. By using this we can solve both retrieve and edit problems. This separator acts as a tree data structure, where internal nodes display attributes or datasets, branches showing decision rules and leaf nodes indicating the result. It includes two types of nodes, name Decision and Leaf. The decision tree is based on the CART (Classification and Regression Tree) algorithm. It simply asks the question and is based on the result and divides the tree into a few small trees.

Confusion Matrix:

<table>
<thead>
<tr>
<th></th>
<th>Predicted No</th>
<th>Predicted Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual No</td>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>Actual Yes</td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>

2. K-nearest neighbors (KNN):

Used for stage prediction problems. It is a type of supervised learning strategy. It does not think about the data provided, it is also known as a non-parameter learning algorithm. It is also known the learning algorithm is lazy because it does not learn anything during training from the training set, instead performs a database function as required. It keeps data in training and separates data continuously when it receives new data.
3. **Random Forest:**

The informal forest forms many decision-making trees and based on their results further advances the diversification and reversal of processes. A few decision trees are made using random subset sets of training databases. Algorithm run time is faster and allows for missing data. Random Forest makes a random algorithm. Decision stage is the mode of classes produced by cutting trees.

<table>
<thead>
<tr>
<th></th>
<th>Predicted No</th>
<th>Predicted Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual No</td>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>Actual Yes</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>

4. **Support Vector Machine:**

A model for partitioning and reversing a Vector-support machine (SVM) that can be used to solve specific and indirect problems. This data is divided using hyperplanes. In this case, each item of data is sorted as a point in the n-d space, the value of each item becomes the sum of the object.

<table>
<thead>
<tr>
<th></th>
<th>Predicted No</th>
<th>Predicted Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual No</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>Actual Yes</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

5. **Logistic Regression:**

Indentation is a mathematical method of predicting a binary outcome, such as yes or no, based on a predetermined data set. The systematic regression model predicts the variance of the dependent data by analyzing the relationship between one or more independent variables present.

6. **Neural Networks:**

It is a series of algorithms that attempt to find basic relationships in a set of data through a process that mimics how the human brain works. In this sense, neural networks refer to neuron systems, which can be an organic or synthetic environment.

7. **Artificial Neural Network:**

ANN models are extremely flexible for human emotional systems. ANN incorporates computational units similar to those of neurons of the biological nervous system known as synthetic neurons.

ANN is capable of learning and modeling indirect and complex relationships as most relationships between input and output are not interconnected. After training, ANN is able to determine non-binding relationships from intangible data, and as a result is generally made.

<table>
<thead>
<tr>
<th></th>
<th>Predicted No</th>
<th>Predicted Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual No</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>Actual Yes</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>
III. DATASET:

We did some research on a set of chronic kidney disease data downloaded from kaggel. This database contains 25 attributes (including 10 categories, 14 numbers and a class attribute). Specific 'category', possibly 'ckd' or 'not ckd' - ckd = chronic kidney disease. There are 400 lines. Data needs to be cleaned up: because it has NaNs and numerical features need to be forced to float. Basically, we were instructed to delete ALL LINES with Nans, without limitation - that is, any line with one NN, is deleted.

<table>
<thead>
<tr>
<th>age</th>
<th>bp</th>
<th>sg</th>
<th>al</th>
<th>rbc</th>
<th>pc</th>
<th>pco2</th>
<th>ha</th>
<th>hgr</th>
<th>ptv</th>
<th>xpt</th>
<th>doc</th>
<th>dm</th>
<th>cad</th>
<th>apper</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>405</td>
<td>80.0</td>
<td>120</td>
<td>10.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>good</td>
</tr>
<tr>
<td>1</td>
<td>70</td>
<td>50.0</td>
<td>120</td>
<td>40.0</td>
<td>3.0</td>
<td>normal</td>
<td>normal</td>
<td>30.0</td>
<td>6000</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>good</td>
</tr>
<tr>
<td>2</td>
<td>120</td>
<td>80.0</td>
<td>100</td>
<td>20.0</td>
<td>3.0</td>
<td>normal</td>
<td>normal</td>
<td>450.0</td>
<td>31.0</td>
<td>7500</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>poor</td>
</tr>
<tr>
<td>3</td>
<td>403</td>
<td>70.0</td>
<td>150</td>
<td>40.0</td>
<td>3.0</td>
<td>normal</td>
<td>abnormal</td>
<td>170.0</td>
<td>30.0</td>
<td>6000</td>
<td>3.0</td>
<td>0.0</td>
<td>0.0</td>
<td>poor</td>
</tr>
<tr>
<td>4</td>
<td>610</td>
<td>80.0</td>
<td>100</td>
<td>20.0</td>
<td>3.0</td>
<td>normal</td>
<td>normal</td>
<td>180.0</td>
<td>35.0</td>
<td>7000</td>
<td>4.0</td>
<td>0.0</td>
<td>0.0</td>
<td>good</td>
</tr>
</tbody>
</table>

Chronic Kidney Disease Dataset Information:

We use the following presentations to collect data:
- years - years
- bp - blood pressure
- sg - certain gravitational force
- albumin
- su - sugar
- rbc - red blood cells
- pc - pus cell
- pco2 - pus cell clumps
- bacterla
- bgr - random blood sugar
- blood urea
- sc - serum creatinine
- salt - sodium
- pot - potassium
- hemoglobin - hemoglobin
- pcv - full cell volume
- wc - the number of white blood cells
- rc - the number of red blood cells
- htn - high blood pressure
- dm - diabetes
- Cad - a disease of the coronary artery
- hunger - appetite
- pedal edema
- anemia
- class - class

IV. Conclusion:

This paper presents a comparative analysis of predictive algorithms to determine the most effective predictive algorithm to detect chronic kidney failure in advance. The database shows input parameters collected from symptomatic patients and models are trained and validated the input parameters provided. Decision Tree, Random Forests, KNN, ANN, Cat-boost classifiers, SV machine learning models are designed to test CKD. Model performance is assessed based on predictive accuracy by building up a confusion matrix for individual models. Research results have shown that the Cat boost Classifier model and the Decision tree classifier predict better CKD compared to artificial neural networks and Vector Support machines. Comparisons can also be made based on the time of the study, the selection of the feature set as an improvement of this study.
V. Results and Analysis:

Comparative analysis of various attributes of dataset:
Heat map:
A Heat map is a way of showing a certain type of matrix structure. To use the heat map data must be in matrix form. By matrix we meant that the index name and column name must be identical in some way for the data we fill with in the cells to be relevant.

Violin plot of red blood cell:
The violin plot is a combination of box structure and kernel density plot, which reflects the height of the data. It is used to visualize the distribution of numerical data. Unlike the box structure that can only display summary statistics, the violin sections show summary statistics and the density of each variation.

Scatter plot:
Scatter structure is a type of structure or diagram using Cartesian links to display values usually in two dynamic data sets. When points are coded, one additional variation can be displayed.
KDE plot:
The kernel density estimate (KDE) structure is a way of detecting the distribution of visuals in a database, analogous in a histogram. KDE stands for data using a continuous curve in one or more directions.

Bar graph:

Preprocessing:
When you create a machine learning project, it is not always a problem for us to meet clean and formatted data. And while any data processing is done, it is compulsory to clean it and format it in a format. So in this case, we are using a data processing function.
Feature extraction:
Feature extract refers to the process of converting raw data into numerical features that can be processed while retaining information in the original data set. It produces better results than using direct machine learning over raw data.

Confusion Matrices of used algorithms:

```
Confusion matrix of CatBoostClassifier
[[25  3]
 [ 0 12]]
Accuracy score is 0.925

Confusion matrix of Support Vector Machine
[[28  0]
 [12  0]]
Accuracy score is 0.7

Confusion matrix of Decision Tree
[[26  2]
 [ 0 12]]
Accuracy score is 0.95

Confusion matrix of ANN
[[27  1]
 [ 0  6]]
Accuracy score is 0.825

Confusion matrix of Random Forest
[[26  2]
 [ 1  1]]
Accuracy score is 0.925
```
Result:

![Image of a chart showing Sorted Accuracy of different Models]

References:


