International Research Journal of Engineering and Technology (IRJET) -ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Analysis and Prediction of Chronic Kidney Disease

Sharanya Shankar[1], Shubhra Verma[2], Sriya Elavarthy[3], Trishala Kiran[4], Poonam Ghuli[5]

[1-4]Dept. Of Computer Science and Engineering, RV College of Engineering, Bengaluru, Karnataka, India
[5]Associate Professor, Dept. Of Computer Science and Engineering, RV College of Engineering, Bengaluru, Karnataka, India

---

Abstract - *Chronic Kidney Disease is a common term for multiple heterogeneous diseases in the kidneys. It is also known as Chronic Renal Disease. The disease affects 5 to 10 per cent of the population worldwide. Chronic Kidney Disease is a health problem around the globe. Most cases of Chronic Kidney Disease go undiagnosed or are later diagnosed in underdeveloped and developing nations; this is one of the primary reasons why a higher percentage of such cases come from developing and underdeveloped nations as opposed to developed nations where most people go through regular check-ups and diagnosis. Machine-based learning systems may be used to diagnose Chronic Kidney Disease in a timely and precise manner which will help doctors check their diagnostic results in a fairly short time, thereby allowing a doctor to attend and treat more patients in less time compared to the scenario under which he / she has to go through the diagnostic process entirely manually.*

***Key Words*: Machine-based learning, Chronic Kidney Disease (CKD), diagnosis, manual, blood potassium level, diet plan.**

1. INTRODUCTION

Chronic Kidney Disease is an exceedingly widespread public health issue. It is especially bad in low-to-middle-income countries where millions are dying because of a shortage of adequate care. CKD is very serious if it is not treated in time properly, which may be lethal. It happens because the kidneys are weakened and can't effectively absorb blood. It's a long-term disease and affects not only one kidney but both simultaneously. Chronic Kidney Disease is often called as Chronic Kidney failure, it impacts 10 percent of the population worldwide according to latest medical estimates. Loss of kidney function is a slow cycle that may take 3 months or more depending on the patients' condition and also whether the physicians have a strong method to detect symptoms which are being exhibited by the patient who is more likely to have kidney failure in advance. The healthcare sector generates vast volumes of details that needs to be analysed to identify secret knowledge for successful analysis, treatment and decision-making. It often involves a strong risk of mortality within a limited period, an individual must be treated and properly healed. Diabetes and high blood pressure are the two prominent causes of Chronic Kidney Disease. The disease may not have a clear root cause, although the degradation is usually permanent and can contribute to severe health complications. CKD testing typically starts with clinical records, diagnostic examinations, imaging scans and eventually, biopsy. Even so, this growth in the amount of information inevitably includes data that can be repossessed when necessary. Today, health administrations are able to produce and store huge bulks of data with the help of various emerging technological tools. Using data mining and various labelling methods in medical systems, partnerships and structures can be established that help in modelling and decision-making mechanisms for research and action taking, action taking may provide the DOs and DONTs of everyday life practices with instances that indicate an individual's meal or diet. For a patient with CKD, adopting a suitable diet plan may help in reducing the disease development. Therefore, an appropriate diet which is proportional to the patients' health status is important like how by determining a factor such as the approximate value of Glomerular Filtration Rate (eGFR) gives us the stage of the disease. There are five different stages: Stage 0, Stage 1, Stage 2, Stage 3 and Stage 4. Patients are found to be in the healthy spectrum till Stage 2, where the kidneys are able to deal with renal functions without accumulating potassium or excess urea in the blood from excretory materials. Subsequently, patients in Stage 0, Stage 1 and Stage 2, do not need any critical improvements to their diet schedule. Nevertheless, it is important for patients in Stage 3 and Stage 4 to preserve the balance of salts, electrolytes and liquids within the patients' body. Furthermore, to maintain a balance of salts, electrolytes and fluids in patients who are in Stage 3 and Stage 4 is difficult. The patients' diet charts do not solely depend on the stage of the disease but rather on various factors such as the amount of sodium, potassium, urea etc., which is unique and not the same for everybody.

In the past few decades, several leading researchers have effectively implemented machine learning and data mining to develop computer-aided diagnostic (CAD) systems to detect complicated health conditions with reasonable precision and efficiency.

Many scholars have been drawn to the automatic detection of different diseases. This paper outlines the analysis of various approaches and precision of prior research papers including this report. Many academic projects have been done on this disease so far.

Here, we have implemented five Machine Learning Techniques in this project, which includes Logistic Regression, Random Forest Tree, K-Nearest Neighbor, Neural Network and a recently designed hybrid algorithm by Google Inc. engineers called Wide & Deep Learning for CKD patient classification. Ultimately, we conducted a systematic study of the outcomes of all three classifiers to figure out the correct classifier for CKD diagnosis. We used Consistency, Recall, Precision and F1 Score for evaluation.

1.1 Literature Survey

In recent times, data mining technologies have made a major contribution to discovering or identifying different diseases in the healthcare sector and they have many advantages such as fraud detection in health insurance, the provision of medical services to patients at affordable rates, the discovery of better treatment approaches, the creation of innovative healthcare programs, successful hospital infrastructure, hospital infection control and improved patient care and also maintain better customer relation. The diagnosis of diseases is also one of the key fields of scientific science especially in times like these. Several diseases have emerged from the present lifestyle of individuals, working climate and diet, one of which happens to be Chronic Kidney Disease.

These are some of the first experimental research papers published which uses various machine-learning methods to treat CKD. They used the dataset available from the UCI Machine Learning Repository for multiple machine learning algorithms.

A novel approach was developed for the diagnosis of CKD using machine learning methods in the research conducted by Asif Salekin and John Stankovic. They validated their work on a dataset of 400 patient records which includes 250 early stage patients with 24 attributes observed by CKD. They used K-Nearest Neighbours, Random Forest and Neural Networks as classifiers to find a solution that fits.

M. P. N. M. Wickramasinghe.et al introduces an approach for the regulation of the condition by an effective diet chart. Classifiers are built using different algorithms in this study, such as the Multi-Class Decision Jungle, Multi-Class Decision Forest, Multi-Class Neural Network and Multi-Class Logistic Regression. Based on the patients' blood potassium levels, an appropriate potassium range is chosen. The classification algorithms propose an apt diet chart according to the potassium range of the patient.

M.K Jain and M.A Ameta in 2017 published a study of data mining techniques for CKD identification and CKD remedies such as dialysis. Evidently, the work reveals that classification is the main technique in data mining which is highly effective in CKD prediction. The analysis also revealed that various methods of choosing the attributes would further boost the results of the classification.

Ms. Astha Ameta et al primarily concentrated on data mining tools and forms of forecasting Chronic Kidney Diseases. They therefore made it apparent that data mining was a more powerful method for predicting Chronic Kidney Disease.

Xun. L. et al. used two data-mining classifiers: Artificial Neural Network (ANN) and Naive Bayes to predict CKD. ANN provided more reliable findings in their study as opposed to Naive Bayes.

S. Dilli Arasu and Dr. R. Thirumalaiselvi performed a study to fix the missing values in the CKD dataset. Lack of values in a dataset would reduce the estimated accuracy of the result. They implemented a recalculation operation on the various CKD stages to remove the missing values and the uncertain values were filled by using the new recalculated values.

1.2 Description of Dataset

The dataset used is from the UCI Repository entitled Chronic Kidney Disease. There are 400 instances in this dataset with 25 attributes among which 14 attributes were Categorical and 11 were Numeric. The output variable contains only two values, “ckd” for positive detection of Chronic Kidney Disease and “notckd” for negative detection of Chronic Kidney Disease.

2. METHODOLOGY

The proposed work was performed in four parts namely: (i) Data Pre-Processing that includes imputation of missing The proposed work was performed in four parts namely: (i) Data Pre-Processing that includes imputation of missing values (ii) Attribute selection (iii) developing and training classification models and (iv) Diet Prediction.

2.2 Pre-Processing of Dataset

First, an unsupervised attribute filter called 'Numeric Cleaner' was used to label the values missing under all the attributes. If the number of values missing was much smaller than the total instances, then instances with the missing values were deleted using the rule of 'Remove Values' under the unattended instance filter. When the number of missing values in an attribute is substantially high then the attribute will be deleted. If the number of values missing are large, then ‘Replace Missing Values' rule is used. Here, the values missing in the dataset are replaced with mean values of attributes.

2.3 Feature Selection Task

Methods of Feature or Attribute selection are used to identify and eliminate unnecessary and redundant attributes from the dataset which do not provide any

contribution in predicting models' accuracy or may actually decrease the models' accuracy. There are two general approaches to feature selection task:

i. Wrapper Method

This paper performs the wrapper method to define the most accurate subset of the 24 attributes that can produce high precision detection of CKDs. In the wrapper method, the selection of the subset attribute is carried out using the induction algorithm which behaves like a black box. The approach to the wrapper is that it conducts a search of possible parameters and because of its robustness use of 'best first search' as tool in this study. The idea is to pick the most promising collection. When the target is reached, the best first search terminates. As it is an optimization problem, the search can be terminated at any point and return the best solution found so far.

ii. Embedded Method

This project uses the embedded method to define and rate CKD detection attributes which gives high predictability and also remove unneeded and obsolete attributes. Embedded methods identify which features contribute to the accuracy during model development. Here, we have used the Least Absolute Shrinkage and Selection Operator (LASSO) which is a modification of the method least square regression.

2.3 Classification

Here, we explore 4 different types of classifiers and one new algorithm called Wide and Deep Learning.

(i) Logistic Regression

This is a linear regression model which is mostly used for binary classification. This model uses a logistic function to approximate the probabilistic distribution by calculating the relationship between the output variable and the predictor variables. It estimates the distribution between example A and boolean class label B as P(A|B).

(ii) Random Forest Tree

Random Forest is a model to the ensemble that can also be used as a type of nearest neighbour predictor. Random Forest begins with a technique known as a 'Decision Tree' which compares it to a weak learner in ensemble terms. The decision tree algorithm divides the data set repeatedly according to the given attributes that maximises the separation of data forming a tree-like structure.

(iii) K-Nearest Neighbours

KNN is a classification algorithm used for analysing examples that are not known without the need of building a model initially. This is done by looking at the closet data in the space pattern. The value of P(Y|X) is estimated by calculating the ratio class group Y with its KN neighbours of X. the benefits of this algorithm is that it is stable for large testing datasets and efficient for training data.

(iv) Neural Network

Feedforward Neural Network is an ANN algorithm that connects the neurons of different layers to each other. It consists of an input layer, an output layer and several hidden layers in between. These layers contain many neurons which form a linkage for the layers. In this model, the input signal is allowed to move only in the forward direction from the input layer passing through various hidden layers and finally reaching the output layer. As there are no loops, the predicted output cannot be fed back in model.

(v) Wide and Deep Learning

Wide and Deep Learning is a collaborative system of classification developed. This algorithm can be used for classification problems having categorical inputs and also for generic regression problems. Wide and Deep Learning algorithm is a liner model produced by the generalisation of neural network and logistic regression. These models have the added advantage of memorisation with the benefits of a feed-forward neural network. The Wide and Deep learning trains feed-forward neural networks along with a linear model which has categorical input feature transformations thereby integrating the strengths of both models and achieving both memorisation and generalisation.

2.4 Diet Plan Recommendation

The final step is to suggest a diet based on the blood potassium level which is divided into 3 groups: (i) Safe (ii) Caution and (iii) Danger.

(i) SAFE Zone: Blood Potassium Level (3.5 - 5.0)
(ii) CAUTION Zone: Blood Potassium Level (5.1 - 6.0)
(iii) DANGER Zone: Blood Potassium Level (6.1 and higher)

To get a more reliable approach, a patients' current zone is forecasted to the outcome predicted and accordingly a diet

plan is suggested. When preparing the final data set, an extra attribute called 'zone' needs to be added. The reason of including this attribute in the dataset is to help in prediction of the diet by taking consideration of other attributes values too.

3. EVALUATION

Classification in data mining includes the question of predicting which category or class a new discovery falls under. The derived model (classifier) is based on analysing a collection of training data where a class label is given for each data. The trained model (classifier) is then used to assign new, unknown data to the class mark. One of the most important terms for interpreting classification metrics is the confusion matrix.

		Predicted	
		Negative	Positive
Actual	False	True Negative (TN)	False Positive (FP)
	True	False Negative (FN)	True Positive (TP)

Fig -1: The 4 categories

Each prediction falls into one of these four categories. Let's look at what they are:

(i) True Negative (TN): Predicts data that is classified false as false.

(ii) True Positive (TP): Predicts data which is classified true as true.

(iii) False Positive (FP): Also known as "false alarm," this is a Form 1 error where the test tests a single condition and erroneously predicts a positive one.

(iv) False Negative (FN): This is a Form 2 error in which a single condition is verified, and a true instance is predicted as negative by the classifier.

3.1 Metrics

The various metrics used are:

A. Accuracy

A classifiers' accuracy is given as the percentage of total correct predictions divided by overall number of instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

B. Recall

Recall is one of the measurement criteria used frequently for an unbalanced dataset. It calculates how many of the positive actuals our model predicted to be positive (True Positive).

$$\text{Recall} = \frac{TP}{TP + FN}$$

C. Precision

Precision defines how precise or accurate our model of data mining is. Precision is often called an indicator of accuracy or consistency, or a positive predictive value.

$$\text{Precision} = \frac{TP}{TP + FP}$$

D. F1 Score

The F1 score comes into the picture when both the recall and the precision are required. It tries to combine both recall and precision. It's much better than precision, as we don't search for any real negative data with an F1 ranking. In other words, it's also a harmonic mean of precision and recall:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. RESULTS AND ANALYSIS

The outcomes of the dataset analysis to identify CKD utilising all 24 attributes with five distinct classifiers: Logistic Regression, Random Forest, K-Nearest Neighbor, Neural Networks and Wide and Deep Learning are done by state-of-the-art research which uses the root mean square error (RMSE) to predict precision of identification. This project considers Accuracy, Recall, Precision and F1 Score as output indicators to evaluate the classifiers. All tests were carried out using tenfold cross validation with 50 percent of the details as test data. Because it is not always feasible to use 24 attributes, we consider the best subset of these 24 attributes using the Wrapper Method which results in remarkable precision. Using LASSO regularisation, we list and define the attributes of CKD predictability. Finally, a diet chart is set out according to the patients' Blood Potassium range.

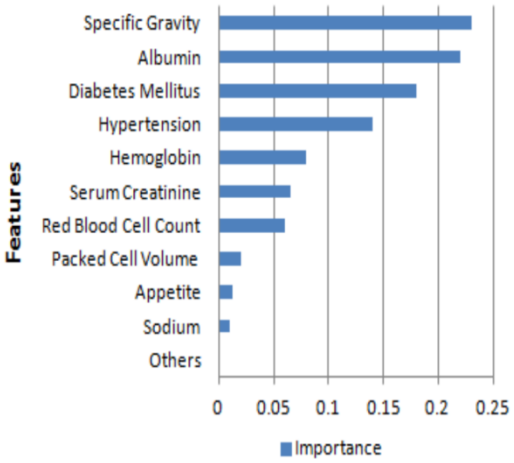

Chart-1: The 4 categories

A confusion matrix is a table that is sometimes used to define the performance of a classification models' (or "classifier") on a collection of test data for which the true values are defined. We evaluated the effects of Logistical Regression, Feed-Forward Neural Network and the Wide and Deep model for both instances. We used the aforementioned 10 hidden neuron combinations for both the Feed-Forward Neural network and for the deep portion of the Wide and Deep model as well. It is seen that from Table I, the Feed-Forward Neural network produces 0.98 as F1 Score, 0.92 as the Precision value, 0.98 as the Recall value and 0.99 as the Accuracy value. We may also conclude that the Feed-Forward model once again dominates the other models by analysing all of the above findings, we may make a quantitative analysis which shows that the Feed-Forward Neural network performs the best among balanced data cases. The confusion matrix shown below shows that there are totally 40 patients who are to undergo diagnosis and out of the 40 patients, 29 suffer from CKD and the other 11 don't.

Table -1: Result of all the algorithms used

Algorithms	Accuracy	Precision	F1 Score	Recall
Wide & Deep Learning	0.98	0.94	0.97	0.98
Neural Networks	0.99	0.92	0.98	0.98
Random Forest	0.99	0.99	0.99	0.99
KNN	0.97	0.96	0.95	0.94
Logistic Regression	0.96	0.87	0.95	0.95

$$\begin{bmatrix} 29 & 0 \\ 0 & 11 \end{bmatrix}$$

Fig –1.1: Confusion Matrix

5. CONCLUSION

By utilising several machine-learning algorithms, we applied a novel method towards detecting CKD. We did an evaluation on a group of 400 patients, 250 of which had early stage CKD. This collection of data has certain noisy values which are missing. Thus, classification algorithms are provided with the ability to manage missing and noisy values. To find a suitable response for this problem, we tested other classification algorithms, such as Logistic Regression, Random Forest and Neural Networks. We used two methods to conduct feature reduction i.e Wrapper Method and LASSO Regularisation, they helped in removing overfitting as well as to identify the most significant predictive attributes for CKD. Notably, this projects' results bring new variables that classifiers will use to classify CKD more accurately than state-of-the-art formulas.

REFERENCES

1. M. P. N. M. Wickramasinghe, D. M. Perera and K. A. D. C. P. Kahandawaarachchi, "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms," 2017 IEEE Life Sciences Conference (LSC), Sydney, NSW, 2017, pp. 300-303.

2. H. A. Wibawa, I. Malik and N. Bahtiar, "Evaluation of Kernel-Based Extreme Learning Machine Performance for Prediction of Chronic Kidney Disease," 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 2018, pp. 1-4

3. U. N. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5

4. H. Zhang, C. Hung, W. C. Chu, P. Chiu and C. Y. Tang, "Chronic Kidney Disease Survival Prediction with Artificial Neural Networks," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2018, pp. 1351-1356

5. J. Aljaaf et al., "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," 2018 IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, 2018, pp. 1-9.

6. Arif-Ul-Islam and S. H. Ripon, "Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-6.

7. G. Kaur and A. Sharma, "Predict chronic kidney disease using data mining algorithms in hadoop," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 973-979

8. N. Tazin, S. A. Sabab and M. T. Chowdhury, "Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique," 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), Dhaka, 2016, pp. 1-6.