DATA MINING FOR PREDICTING EARLY - STAGE CHRONIC KIDNEY DISEASE

Jeena Jose, S.Sheeja*

Abstract

Kidney disorders are prevalent today and are frequently referred to as silent diseases since they are hard to diagnose because the affected person does not exhibit symptoms until the disease has progressed to a more severe stage. More than 50 million people worldwide have some form of renal disease, according to studies. Utilizing big data, machine learning, and deep learning, data mining plays a crucial role in the diagnosis and prognosis of renal illness. In light of this, the study examined the advantages of using these mining algorithms, their significant contribution to the healthcare sector through their prospective prediction skills as well as some of their limits.

Keywords: Kidney Diseases, Machine Learning, Diagnosis, Data Mining, Prediction

I INTRODUCTION

The kidney is the main organ in the inhuman excretory system which takes part in the filtration of blood and urine formation. Before the blood is distributed to the remainder of the body, it has the primary responsibility of filtering blood leaving the digestive system. When you do not eat it stores glucose to keep you nourished. Numerous people experience chronic renal disease throughout the world (CKD).[1] In America, CKD affects 37 million adults, and millions more are at increased risk. Data mining algorithms contribute to the detection and prediction of CKD so that the clinical and laboratory data can do many with these algorithms for better decision taking in the case of patients moving to a quality lifestyle future. Here, in this paper, we are going through an examination of various machine-learning techniques used for CKD diagnosis [2].

II PREDICTION ANALYTICS

ML is an interdisciplinary field that uses computer algorithms to increase the accuracy of prediction of both static and dynamic data by the use of analytic/ probabilistic models using clinical data. The ability of machine learning to predict problems associated with CKD has advanced dramatically. For CKD, there are no predictive instruments that are commonly accepted so the number of patients is increasing by the lack of the same. The aging and increase in Type 2 Diabetes contribute to the elevation in the patient graph of CKD [3]. CKD is a progressive disease in some, but not likely for everyone. Data mining techniques are widely used in CKD prognosis and detection such as naive Bayes, K Nearest neighbour, decision tree, random forest, and support vector machine. The dataset was taken from different repositories by the authors like Kaggle, UCI, etc… The next step was to clean the data. Some values were missing in different attribute columns and the imputation was done by mean/median methods and sometimes with K-NN and MICE (Multivariate Imputation by Chained Equation). Different feature selection methods were implemented for the better reduction of data for reducing the burden of heterogeneous unrelated big data.

The term “instance-based learning” refers to a collection of approaches for classification and regression. Based on how closely the query resembles its closest acquaintance in the training set, it can anticipate the class label. Instance-based learning algorithms do not abstract from particular cases, in contrast to other approaches like decision trees and

[1]Department of Computer Science,
Karpagam Academy of Higher Education , Coimbatore, IndiaTamil Nadu
*Corresponding Author

neural networks another option is for them to simply keep all the information and, when a query comes in, determine the response by discussing it with the closest friend

Controllable neural network the output or the input in this case is already known. The neural network's anticipated output is contrasted with the actual output. The parameters are altered based on the fault, and the neural network is then used and then the authors used the prediction algorithms for better prediction.

III METHODOLOGY

Given its numerous benefits for patients and healthcare professionals, machine learning currently plays an important role in the industry. Clinical care recommendations pharmacological impact prediction and clinical decision assistance are some of the most typical uses of machine learning in healthcare. To stop the course of the disease to an extent, several ML algorithms were used to diagnose CKD. In this work, we analyse a few well-known algorithms and their propensity for prediction and efficiency in the detection and prediction of CKD [4].

Logistics Regression (LR)

This approach of supervised categorization is frequently employed in academic settings and other settings. It can handle a huge variety of absolute and numerical elements. It utilizes a logistic function and is a kind of linear regression. The process can provide a binary output in a little amount of time. It gives both a predictor's suitability and its connection (positive or negative). Logistic regression is not applicable when there are fewer findings and observations than characteristics. Its linear decision surface makes managing nonlinear situations challenging[5].

Naïve Bayes (NB)

This Bayes theorem-based approach assumes that all predictive factors are independent. The existence of a certain feature in this method's class is independent of the existence of any other feature. Each factor in this situation is independent[6].

Decision Tree (DT)

Here we may divide the data set according to several criteria and produce a tree structure. Every node in the tree is subjected to an attribute value test case, the choice of one of the popular supervised learning techniques is the tree approach. Both classification and regression tasks may be accomplished with it. Here a model that predicts the value of a target variable is built using straightforward decision-making principles deduced from the characteristics of the data. [7].

Support Vector Machine (SVM)

Static support vector machine is another popular modern machine learning approach (SVM). Support-Vector machine in machine learning algorithms that examine data used for regression and classification analyses. SVM may effectively do nonlinear classification in addition to linear classification by simple translation of their inputs into high-dimensional feature spaces. This technique is known as the kernel trick. In essence, it establishes boundaries between the classes. The margins are designed to have the shortest possible distance between them and the classes, which minimizes the classification error. There will be more support vectors in huge data sets. [8].

Random Forest (RF)

It is a procedure that several decision trees select the majority of them as the greatest value to get the best results in a random selection of features; it seeks the most desirable feature [9]. A classifier called random forest uses many decision trees on different subsets of a given data set and the average of the outputs to increase the data sets predicted accuracy.

K- Nearest Neighbour (KNN)

The algorithm places the new instance in the category that matches the available categories the most by assuming a resemblance between new data and existing data. We can state the items that are comparable to one another are nearby. It determines the distinction between two records. [10].

The advantages and disadvantages of the aforementioned six algorithms were identified and their common operating principles were studied.

Algorithm	Advantages	Disadvantages
Logistic Regression	Easier to put into action Simple to train May encompass several types For small data sets fast segmentation of unknown records. Provides both association and direction (positive or negative)	When there are fewer observations than features, it cannot be used Creates linear limits. Be restricted to discrete function prediction. Problems with nonlinearity couldn't be resolved
Naïve Bayes	Saves time and is quick to use. Addressing challenges of multiple class prediction. Very little training data Easily attainable.	Independent predictors are assumed. Here, it is assumed that each property is independent of the others The zero-probability issue
Decision Tree	Normalization is not necessary It is possible to implement without scaling the data No need to incompetent incomplete data handles category and numerical variables more quickly	Extra RAM is needed for the calculation The tree structure can be drastically altered by a minor variation in the data The intricacy of space and time is quite tremendous.
SVM	More accuracy in prediction Fewer parameters are required Fast target estimate	When compared to other models will take longer to train on huge data sets Expensive
Random Forest	Less likely to overfit It will take both continuous and categorical data values It automates the detection of data missing values Data normalization is not necessary	Calculations are more difficult More time for calculation The class is determined using a vast number of decision trees
K-Nearest Neighbour	If samples are large classifications are more beneficial. May be used with any distributions data forms	Testing is pricey The tasks will be done during the testing period without any training.

Table 1:
Benefits and Drawbacks of
Different Machine Learning Algorithms

IV LITERATURE REVIEW

Classification algorithms are important in the hypothesis testing of computer-generated medical health diagnoses Here we have examined a few CKD papers from google scholar and observed how different algorithms perform when it comes to identifying HCC and other related liver conditions.

Another study included 14 distinct characteristics and machine learning techniques including the Support vector machine and decision tree. The decision tree and SVM both achieved an accuracy of 91.75% and 96.75% respectively. S numerous techniques, including clustering and classification algorithms, have been studied by Dilli Arasu and D and R. Thirumalaiselvi [11]. They discussed the algorithm that was applied in a separate republication in this study. They looked for a more accurate classification system that might be used to determine the stages of chronic renal disease. Chronic renal disease has been predicted by Drs. S Vijayarani and S Dayand made use of the support vector machine and Naïve Bayes techniques. The Support Vector Machine's performance to predict CKD was 3.22, whereas the Naïve Bayes classifier's execution performance was 1.29. using the perception, Artificial Neural Network, and C4.5 algorithms, Tabassum S, Mamatha Bai BG, and Jarna Majumdar created a system for predicting chronic renal disease. EM achieved 70% accuracy, and C4.5 achieved 96.75% accuracy [12].[13] Pushpa M Patil used data mining methods to forecast chronic kidney disease. This study examined the algorithms employed in a separate study and sought to identify the

superior one. The confusion matrix, random forest, nave Bayes, SVM, K- nearest neighbours, and Radial Basis Functions are a few classifiers that show the best accuracy.

For the identification of chronic renal illness, Ruey Key constructed three distinct neural network models in 2015: the Backpropagation Neural Network (BPN), the Generalized Feed-Forward Neural Network (GRNN), and the Modular Neural Network (MNN). By incorporating GA evolutionary algorithms into each of the models corresponding brain components, he further applied these models in his study, all three models used in this trail have higher accuracy or greater than 85%. The backpropagation network has the best accuracy as per observing (BPN) compared to the other two models. Manish Kumar used six distinct data mining approaches, including Random Forest Classifiers, Sequential Minimal Optimization, Naïve Bayes Radial Basis Function, Recursive Neural Classifier (MLPC), and simple logistic regression (SLG) to predict chronic kidney disease [14]. He utilized 400 records in total for the training and prediction of the algorithm. He has discovered the random forest technique has the highest accuracy among these methods.

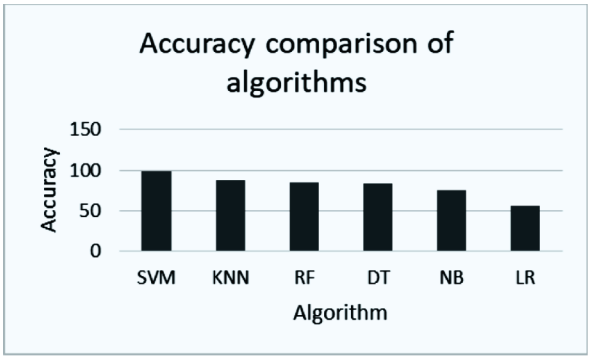

Fig.1 Accuracy of Different Algorithms

V CONCLUSION

We discussed the likelihood of developing CKD and the necessity to identify the most effective algorithmic technique for kidney disease early detection in the survey. Different prediction techniques were examined and used in the study. Six distinct machine learning algorithms working principles have been examined here, and it was discovered that the SVM algorithm performs more accurately than the others and is more frequently employed in prediction analysis.

REFERENCES

[1] https://www.kidney.org/atoz/content/about-chronic-kidney-disease.

[2] Arasu, S. Dilli, and R. Thirumalaiselvi. "Review of chronic kidney disease based on data mining techniques." International Journal of Applied Engineering Research 12, no. 23 (2017): 13498-13505.

[3] Hippisley-Cox, J., and Coupland, C., 2010, “Predicting the Risk of Chronic Kidney Disease in Men and Women in England and Wales: Prospective Derivation and External Validation of the QKidney® Scores,” Hippisley-Cox and Coupland BMC Family Practice, 11-49.

[4] Shouval, R., Bondi, O., Mishan, H., Shimoni, A., Unger, R., & Nagler, A.: Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in SCT. Bone marrow transplantation, 49(3), 332-337 (2014).

[5] Logistic Regression, Retrieve from: https://www.saedsayad.com/logistic_regression.htm last accessed 2021/06/08.

[6] Webb, Geoffrey I., Eamonn Keogh, and Risto Miikkulainen. "Naïve Bayes." Encyclopedia of machine learning 15 (2010): 713-714.

[7] Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for

machine learning." Journal of Applied Science and Technology Trends 2, no. 01 (2021): 20-28.

[8] Mohammadi, Mokhtar, Tarik A. Rashid, Sarkhel H. Taher Karim, Adil Hussain Mohammed Aldalwie, Quan Thanh Tho, Moazam Bidaki, Amir Masoud Rahmani, and Mehdi Hosseinzadeh. "A comprehensive survey and taxonomy of the SVM-based intrusion detection systems." Journal of Network and Computer Applications 178 (2021): 102983.

[9] Speiser, Jaime Lynn, Michael E. Miller, Janet Tooze, and Edward Ip. "A comparison of random forest variable selection methods for classification prediction modeling." Expert systems with applications 134 (2019): 93-101.

[10] Xing, Wenchao, and Yilin Bei. "Medical health big data classification based on KNN classification algorithm." IEEE Access 8 (2019): 28808-28819.

[11] S.Dilli Arasu, Dr. R.Thirumalaiselvi,"Review of Chronic Kidney Disease based on Data Mining Techniques", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13498-13505 ©Research India Publications.

[12] Tabassum S, [2] Mamatha Bai BG, [3] Jharna Majumdar," Analysis and Prediction of Chronic Kidney Disease using Data Mining Techniques," International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 4, Issue 9, September 2017.

[13] Pushpa M. Patil. (2016). Review on Prediction of Chronic Kidney Disease Using Data Mining Techniques. International Journal of Computer Science and Mobile Computing, 5(5), 135-141.

[14] Jeffersonrson, Klinsega, Manish Kumar, Lord win Jeyakumar and Raghav Yadav. "A Combined Machine-Learning Approach for Accurate Screening and Early Detection of Chronic Kidney Disease." In International Conference on Machine Intelligence and Signal Processing, pp. 271-283. Springer, Singapore, 2019.