A Review of Chronic Kidney Disease Prediction Model Based on Machine Learning Technique

Anoop Kumar Kushwaha[1], Anurag Shrivastava[2], Vaibhav Patel[3]

[1]*MTech Scholar,* [2,3]*Assistant Professor*

CSE, Department, NIIRT Bhopal

Abstract - In India and all over the world chronic kidney disease (CKD) is an increasing and serious disease impacting public health. The symptoms of CKD are often appearing too late and many patients inevitably face pain and expensive medical treatments. The ultimate treatment is frequent dialysis or Kidney transplant. Early detection of disease through symptoms can prevent the disease progression by referral to appropriate health care services. Now days, Machine Learning (ML) techniques have been widely used in healthcare sector. ML techniques can help in identifying the potential risk by discovering knowledge from medical reports of patient. Thus helps in preventing the disease progression. In this paper review the CDK prediction models which is based on machine learning technique. The finding of the review is that several models for detecting the risk of CKD are proposed by various authors which is based on data mining (DM) technique. The use of cloud platforms for data mining tasks is new research area in the field of ML. Several literatures uses the platform confirms a considerable performance improvement in running ML tasks.

Keywords— Boosted Decision Tree, Neural Network, Chronic Kidney Disease (CKD), Data Analytics, Health Care, Microsoft Azure, Machine Learning, Prediction Model.

I. INTRODUCTION

The health care sector of the world is the most promising sector nowadays, due to pandemic. Medical Instruments, Medicines, hospital facility, and other required things are lacking everywhere. In this bad situation, it is required to enhance medical technology and its related research. Chronic kidney disease (CKD) poses a serious burden of disease worldwide with substantially increasing number of patients being diagnosed. According to the study [1] published in 2018 Global Burden of Disease study data and methodologies to describe the change in burden of CKD from 1990 to 2016 involving incidence, prevalence, death, and disability-adjusted-life-years (DALYs). Globally, the incidence of CKD increased by 89% to 21,328,972, prevalence increased by 87% to 275,929,799, death due to CKD increased by 98% to 1,186,561, and DALYs increased by 62% to 35,032,384. Figure 1.1 showing the Google search results for CKD [2]. Also, the cost related to CKD care is too high. So early detection and identification of patients with increased risk of developing CKD on the basis of symptoms can improve care by preventive measures to slow disease progression and timely initiation of nephrology care.

CKD is a known common disease, seen by the nephrologists, specialists and practitioner in other fields also.

Enormous complex data is being regularly received by healthcare division about diseases, treatment, patients, medical equipments, hospitals and claims etc. The data requires processing for extraction of knowledge. Data Mining is predominantly helpful in healthcare domain when it is difficult to deal a disease with particular treatment option. DM comprise of efficient techniques and tools to apply on healthcare data for making appropriate decisions towards taking preventive measures and predicting risks of disease.

II. LITERATURE REVIEW

The research paper [1] authors employ experiential analysis of ML techniques for classifying the kidney patient dataset as CKD or NOTCKD. Seven ML techniques together are utilized and assessed using distinctive evaluation measures such as mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), root relative squared error (RRSE), recall, precision, F-measure and accuracy. The experimental outcomes accomplished of MAE are 0.0419 for NB, 0.035 for LR, 0.265 for MLP, 0.0229 for J48, 0.015 for SVM, 0.0158 for NB Tree and 0.0025 for CHIRP. Moreover, experimental results using accuracy revealed 95.75% for NB, 96.50% for LR, 97.25% for MLP, 97.75% for J48, 98.25% for SVM, 98.75% for NB Tree, and 99.75% for CHIRP. The overall outcomes show that CHIRP performs well in terms of diminishing error rates and improving accuracy.

In paper [2] prediction is performed using Naive Bayes Classifier and K-Nearest Neigbour algorithm. The data used is collected from the UCI Repository with 400 data sets with 25 attributes. This data has been fed into Classification algorithms. The experimental results show that Naïve Bayes Algorithm gives an accuracy of 96.25%, whereas KNearest Neigbour came up with an accuracy of 100%.

In this paper [3], we conduct the statistical analysis, Machine Learning (ML) and Neural Network application on clinical data set of Uddanam CKD for prevention and early detection of CKD. As per statistical analysis we can prevent the CKD in the Uddanam area. As per ML analysis Naive Bayes model is the best where the process model is constructed within 0.06 seconds and prediction accuracy is 99.9%. In the analysis of NNs, the 9 neurons hidden layer (HL) Artificial Neural Network (ANN) is very accurate than other all models where it performs 100% of accuracy for predicting CKD and it takes the 0.02 seconds process time.

This paper review paper [4] the machine learning (ML), deep learning (DL) and other intelligent cloud based diagnosis models for the three diseases. A detailed introduction to the cloud based healthcare system is also provided. Besides, the reviews of the techniques are made with respect to aim, underlying technique, diagnosed disease and experimental results. At the end of the survey, a detailed comparative study has also been made to examine the possible future work for the readers.

The authors [12] synthesized systematic reviews of risk prediction models for CKD and externally validated few models for a 5-year scope of disease onset. Authors worked on ~234 k patients' data of UK. Seven relevant CKD risk prediction models were identified. All models distinguished well between patients developing CKD or not, with Receiver Operating Characteristic curve (ROC) around 0.90. But, it is concluded that most of the models were poorly calibrated and substantially over-predicting the risk.

The authors [13] predicted CKD using two classification techniques: Naive Bayes and Artificial Neural Network (ANN). The experiment is conducted using Rapidminer tool over dataset containing 400 instances with 25 attributes including class. The dataset from UCI repository [4] is used. The results [27] revealed that Naive Bayes produced more accurate results than ANN.

In study [14] CKD is diagnosed with Adaboost Ensemble Learning (EL) method. For diagnosis Decision tree based classifiers is used. The classifier performance is evaluated using several metrics including area under curve (AUC).The main observation of paper [5] is that Adaboost EL method provides better performance than individual classification. The dataset from UCI repository [4] is used.

III. DATASET DESCRIPTION

In various research papers authors used the dataset for experiment purpose. CKD dataset from UCI ML repository [8] is used. The dataset includes 400 instances with 24 attributes and a class attribute. The CKD dataset used in this study is taken from the UCI Machine Learning Repository [8]. The data was donated by Soundarapandian et al. and collected for nearly 2-month period. The dataset comprise of 400 samples represented by 11 numeric and 10 nominal attributes and a class descriptor which is also nominal. Out of 400 samples, 250 samples belong to the CKD group, and the other 150 samples belong to the non-CKD group. Details are more discussed in [12].

IV. NEED ANALYSIS OF PREDICTION MODEL

Prevention is better than cure, diagnosis the disease has done before damage increased the chances of survival among people. Through detection a doctor can check and lessen the effect of progression of disease. The academicians utilize potential of Machine Learning (ML) for getting insights by analyzing medical datasets. Data scientists are in huge demand in healthcare researches. ML has empowered wellness among people by analyzing the data collected through various sources including reviews by practitioners. Prediction data analytics based on ML can enhance health prospects and discover the reasons of diseases, by analyzing patient's reviews, symptoms, history and other related information in real-time.

In spite of various innovations and inventions in health care sector worldwide, efforts to improve early detection of CKD often remains less futile. Several academicians and researchers in health care division worldwide are working on the detection of CKD problem and trying to develop efficient models to predict and classify the CKD patient, so that the necessary care and preventive measures can be provided to patient. The challenge of detecting CKD risk is handled by data mining techniques that are actually based on ML algorithms. Several models have been developed to predict CKD onset, but most have not been validated outside the setting in which they were developed [5, 6]. We can say that, most of the models only classify the disease as 'CKD' and 'NOT-CKD'. So, there is always a need for Prediction tools for early medical diagnosis and detection.

To address the onset of diseases, researchers are applying ML to principles of medical sciences. For this purpose they are using data and information explosion. Comprehensive disease classification means fitting the people in suitable classes of disease or not-disease or positive or negative. The prediction models are necessary for easy diagnostics and finding symptoms. These models can be also used as recommender systems to improve lives. In past decade, several techniques are applied which are reviewed in this paper.

The application of prediction models to extract large medical datasets grows stronger with the extraordinary raise in the patient logs. So, need of cloud platforms for such tasks are established in next section. Such ML based Prediction models are generally designed to gain insights from large unstructured and structured datasets also.

V. FINDING OF THE REVIEW

The finding of the review process is that despite various innovations and inventions in health care sector worldwide, efforts to improve early detection of CKD often remains less ineffective. Several academicians and researchers in health care division worldwide are working on the detection of CKD problem and trying to develop efficient models to predict and classify the CKD patient. The challenge of detecting CKD risk is handled by data mining techniques that are actually based on ML algorithms. Several models have been developed to predict CKD onset, but most of the models only classify the disease as 'CKD' and 'NOT-CKD'. So, there is always a need for prediction tools for early medical diagnosis.

Also, the healthcare data is increasing at a large rate and contain some noise in it. Hence there is a requirement of large storage, memory and processing power to process this data efficiently. The rapid advances allow us to access large storage devices at low cost, on the other hand we require RAM and Processors to manage and develop the efficient methods to analyze such data. Analyzing such critical data in reliable manner is another issue. The complexities can be managed through Cloud Computing platforms that provide computational resources on demand and on pay-as-you-go model. The CC platform also provides tools and languages for complete KDD (Knowledge Discovery in Databases) process.

VI. CONCLUSION

A generalized prediction model for CKD risk detection is needed which is based on machine learning technique and implement in cloud platform. The model is having utility in healthcare domain. The prediction model work will definitely provide significant insight into risk prediction for other diseases also. Health care data is increasing at a large huge and contain some noise in it. Hence there is a requirement of large storage, memory and processing power to process this data efficiently. Analyzing such critical data in reliable manner is another issue. The complexities can be managed through Cloud Computing platforms that provide computational resources on demand. The CC platform also provides tools and languages for complete KDD process. The model can be further tested for more parameters and extended for batch prediction by supplying huge dataset.

REFERENCES

[1] Bilal Khan 1, Rashid Naseem 2, Fazal Muhammad 3,Ghulam Abbas 4, (Senior Member, IEEE), and Sunghwan Kim 5 "An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy" Received March 2, 2020, accepted March 14, 2020, IEEE 2020

[2] 1Sujata Drall, 2Gurdeep Singh Drall, 3Sugandha Singh, 4Bharat Bhushan Naib "Chronic Kidney Disease Prediction Using Machine Learning: A New Approach" International Journal of Management, Technology And Engineering ISSN NO : 2249-7455 Volume 8, Issue V, MAY/2018

[3] K. B. Anusha, T. PanduRanga Vital, K. Sangeeta "Machine Learning Models and Neural Network Techniques for Predicting Uddanam CKD" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2, July 2019

[4] R. Raja1 and B. Ashok2 "An Effective Analysis on Cloud Computing in Healthcare Diagnosis and Prediction Systems" International Journal of Advanced Science and Technology Vol. 29, No. 03, (2020), pp. 12016 - 12029

[5] Jameson K, Jick S, Hagberg KW, Ambegaonkar B, Giles A, O'Donoghue D., "Prevalence and management of chronic kidney disease in primary care patients in the UK". Int J Clin Pract. 2014;68 (9):1110–21.

[6] www.google.co.in/search?q=Chronic+kidney+disease

[7] A. S. Levey, K. Eckardt, U. Tsukamoto, A. Levin, J. Koresh, J. Rossert, D. D. Zeeuw, T. H. Hostetter, N. Lameire and G. Eknoyan, "Definition and classification of chronic kidney disease: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO)," Kidney International, Vol. 67, pp. 2089-2100, 2005.

[8] P. Soundarapandian and L. J. Rubini, http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease, UCI Machine Learning Repository, Irvine, 2015.

[9] Collins GS, Omar O, Shanyinde M, Yu L-M. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. J Clin Epidemiol. 2013; 66(3):268–77.

[10] Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. Remuzzi G, editor. PLoS Med. 2012; 9(11):e1001344.

[11] J. K. Han, Micheline, Data mining: concepts and techniques: Morgan Kaufmann, 2001.

[12] I. H. a. F. Witten, Eibe Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann 2005.

[13] C. M. Bishop, Pattern recognition and machine learning: Springer, 2006.

[14] P. Simon, Too Big to Ignore: The Business Case for Big Data: John Wiley & Sons, 2013.

[15] Paolo Fraccaro, Sabine van der Veer, Benjamin Brown, Mattia Prosperi, Donal O'Donoghue, Gary S. Collins, Iain Buchan and Niels Peek, "An external validation of models to predict the onset of chronic kidney disease using population-based electronic health records from Salford, UK", RESEARCH ARTICLE, BMC Medicine (2016) 14:104.

[16] V. Kunwar, K. Chandel, A. S. Sabitha and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, 2016, pp. 300-305.

[17] M. D. Başar, P. Sarı, N. Kılıç and A. Akan, "Detection of chronic kidney disease by using Adaboost ensemble learning approach," 2016 24th Signal Processing and Communication Application Conference (SIU), Zonguldak, 2016, pp. 773-776.

[18] Z. Sedighi, H. Ebrahimpour-Komleh and S. J. Mousavirad, "Featue selection effects on kidney desease analysis," 2015 International Congress on Technology, Communication and Knowledge (ICTCK), Mashhad, 2015, pp. 455-459.

[19] Sumit Basu, “Empirical Results on the Generalization Capabilities and Convergence Properties of the Bayes Point Machine”, Technical Report, December, 1999