International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

A Machine Learning Methodology for Diagnosing Chronic Kidney Disease

Prof. Rashmi Patil[1], Bhagyashri Deshmukh[2], Dhanshri Lonkar[3], Mahima Kumari[4]

[1] Assistant Professor, [2,3,4] Students

Department of E&TC, Sinhgad Academy of Engineering, Savitribai Phule Pune University, Pune, Maharashtra

Article Info
Volume 9, Issue 2
Page Number : 295-301

Publication Issue
March-April-2022

Article History
Accepted : 01 April 2022
Published : 15 April 2022

ABSTRACT

Chronic kidney disease (CKD) is a global health problem with high morbidity and mortality rate, and it induces other diseases. Since there are no obvious symptoms during the early stages of CKD, patients often fail to notice the disease. Early detection of CKD enables patients to receive timely treatment to ameliorate the progression of this disease. Machine learning models can effectively aid clinicians achieve this goal due to their fast and accurate recognition performance. In this study, we propose a machine learning methodology for diagnosing CKD. The CKD data set was obtained from the University of California Irvine (UCI) machine learning repository, which has a large number of missing values. KNN imputation was used to fill in the missing values, which selects several complete samples with the most similar measurements to process the missing data for each incomplete sample. Missing values are usually seen in real-life medical situations because patients may miss some measurements for various reasons. After effectively filling out the incomplete data set, six machine learning algorithms (logistic regression, random forest, support vector machine, k-nearest neighbor, naive Bayes classifier and feed forward neural network) were used to establish models. Among these machine learning models, random forest achieved the best performance with 99.75% diagnosis accuracy. By analyzing the misjudgments generated by the established models, we proposed an integrated model that combines logistic regression and random forest by using perceptron, which could achieve an average accuracy of 99.83% after ten times of simulation. Hence, we speculated that this methodology could be applicable to more complicated clinical data for disease diagnosis.

Keywords: Chronic kidney disease (CKD), KNN, University of California Irvine (UCI), disease diagnosis

I. INTRODUCTION

Chronic kidney disease (CKD) is a global public health problem affecting approximately 10% of the world's population. The percentage of prevalence of CKD in China is 10.8% , and the range of prevalence is 10%-15% in the United States. According to another study, this percentage has reached 14.7% in the Mexican adult general population. This disease is characterised by a slow deterioration in renal function, which eventually causes a complete loss of renal function. CKD does not show obvious symptoms in its early stages. Therefore, the disease may not be detected until the kidney loses about 25% of its function. In addition, CKD has high morbidity and mortality, with a global impact on the human body. It can induce the occurrence of cardiovascular disease. CKD is a progressive and irreversible pathologic syndrome. Hence, the prediction and diagnosis of CKD in its early stages is quite essential, it may be able to enable patients to receive timely treatment to ameliorate the progression of the disease.

Machine learning refers to a computer program, which calculates and deduces the information related to the task and obtains the characteristics of the corresponding pattern. This technology can achieve accurate and economical diagnoses of diseases; hence, it might be a promising method for diagnosing CKD. It has become a new kind of medical tool with the development of information technology and has a broad application prospect because of the rapid development of electronic health record . In the medical field, machine learning has already been used to detect human body status , analyze the relevant factors of the disease and diagnose various diseases. For example, the models built by machine learning algorithms were used to diagnose heart disease , diabetes and retinopathy, acute kidney injury , cancer and other diseases. In these models, algorithms based on regression, tree, probability, decision surface and neural network were often effective. In the field of CKD diagnosis, Hodneland et al. utilized image registration to detect renal morphologic changes. Vasquez-Morales et al. established a classifier based on neural network using large-scale CKD data, and the accuracy of the model on their test data was 95%.In addition, most of the previous studies utilized the CKD data set that was obtained from the UCI machine learning repository. Chen et al. used k-nearest neighbor (KNN), support vector machine (SVM) and soft independent modelling of class analogy to diagnose CKD, KNN and SVM achieved the highest accuracy of 99.7% . In addition, they used fuzzy rule-building expert system, fuzzy optimal associative memory and partial least squares discriminant analysis to diagnose CKD, and the range of accuracy in those models was 95.5%-99.6% . Their studies have achieved good results in the diagnosis of CKD. In the above models, the mean imputation is used to fill in the missing values and it depends on the diagnostic categories of the samples. As a result, their method could not be used when the diagnostic results of the samples are unknown. In reality, patients might miss some measurements for various reasons before diagnosing. In addition, for missing values in categorical variables, data obtained using mean imputation might have a large deviation from the actual values. For example, for variables with only two categories, we set the categories to 0 and 1, but the mean of the variables might be between 0 and 1. Polat et al. developed an SVM based on feature selection technology, the proposed models reduced the computational cost through feature selection, and the range of accuracy in those models was from 97.75%-98.5%.

II. LITERATURE SURVEY

- **Convolutional Neural Network for Paraphrase Identification.**

In [1] the new deep learning architecture Bi-CNN-MI paraphrase identification (PI). The PI compares two sentences on multiple levels of granularity. In this BI-

CNN means two CNN and MI is Multigranular interaction. They determine whether paraphrase roughly have the same meaning. They are closely related to NN for sentence representation and text matching. They are mainly based on Convolutional sentence model. The parameters of the entire model are optimized for PI. Use of language modeling task is to address the lack of training data. Results on the MSRP corpus surpass that of previous NN competitors. The Bi-CNN-MI can be used for sentence matching, question answering in future. The new deep learning architecture Bi-CNN-MI Paraphrase Identification (PI). The PI contemplates two sentences on various levels of granularity. They choose if rephrase by and large has a similar importance. The parameters of the considerable number of models are updated for PI. Usage of vernacular showing task is to address the nonattendance of planning data.

- **Machine Learning Techniques for Data Mining: A Survey**

In [2] they have compared machine learning algorithms like Decision Tree, Bayes algorithm, Support Vector Machine and Nearest Neighbor. These algorithms are used for classification mainly. They are used for predicting group membership for data instances. They provide a relative analysis of various algorithms. In data mining they extract the hidden predictive data from the large database. They have analyzed machine learning calculations like Decision Tree, Bayes algorithm, Support Vector Machine (SVM) and Nearest Neighbor. These figuring are used all together generally. They are used for anticipating group enlistment for data illustrations. They give a relative examination of various calculations. In information mining they remove the covered insightful data from the sweeping database.

- **Mining electronic health records: towards better research applications and clinical care.**

In [3] the clinical data demonstrate the categories and treatment of patients that represent the under used data sources which are much greater in research potential than the currently which is realized. The potential of EHR (Electronic Health Record) is for establishing the new patients by revealing the unknown disease correlation. In EHR and mining of it a broad range of ethical, legal and technical reasons may hinder the systematic deposition. The potential for the medical research and clinical health care by using EHR data and the challenges which can be overcome before this becomes a reality. The capacity of Electronic Health Record (EHR) is for setting up the new patients by revealing the dark sickness connection. In EHR and its mining a sweeping extent of good, honest to goodness and particular reasons may keep the systematic declaration. The tele-health administrations are being used which are known as the tele-health cautioning organizations. They are generally used as a piece of metropolitan urban communities.

- **Optimal Big Data Sharing Approach for Tele-Health in Cloud Computing.**

In [4] the tele-health services are been used which are known as the telephone health advisory services. They are mostly used in metropolitan cities. Due to tele-health services the patients can get a help easily. Rapid increase in tele-health system has received various techniques like cloud computing and big data. They have proposed a dynamic programming to produce optimal solutions so that data sharing mechanisms can be handled. In this it considers the transmission probabilities, the timing constraints, and also the maximizing network capacities. Due to tele-health organizations the patients can get help effortlessly. A quick incremental in the tele-health structure has become diverse strategies like distributed computing and enormous information. They have a dynamic programming to make perfect game plans with the objective that data sharing frameworks can be dealt with. In this it contemplates the transmission probabilities, the arranging objectives, and moreover increasing as far as possible.

- **Recurrent Convolutional Neural Networks for Text Classification.**

In [5] for a content conclusion examination with jointed Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) engineering, taking the upsides of both like course grained neighborhood highlights features which are made by CNN and long-separate conditions learned by methods for the RNN. The provincial perpetual infection has been engaged.

- **Combination of Convolutional and Recurrent Neural Network for Sentimental Analysis of Short Texts**

In[6] consideration has been paid on both organized and unstructured information. It utilizes a maximum pooling layer that consequently judges, which words assume an essential part in content arrangement to catch the key segments in writings. The data contains in characteristics with missing data regards are basic in improving decision- making system of an affiliation. The learning procedure on every event is fundamental as it may contain a couple of remarkable learning. There are distinctive procedures to manage missing data in choice tree learning. The proposed credit figuring depends on the genetic count that uses space regards for that property as pool of courses of action. Survival of the fittest is the start of hereditary calculation. The health work is gathering precision of an event with credited a motivation on the decision tree.

- **Multiple Imputation of Missing Data with Genetic Algorithm based Techniques.**

The health work in [7] is gathering precision of an event with credited a motivation on the decision tree. The overall chase framework used as a piece of hereditary calculation is depended upon to get ideal arrangement. Our procedure reasonably melded zone finding out about the therapeutic portrayal of both ailment and EHRs into a data driven approach. Exploratory results on a certified dataset from a mending office showed the practicality of our proposed procedure. Their approach merged endeavor relatedness, i.e., how every disease relates with others, appropriately, which incited an adjustment in the perceptive execution.

- **Learning text representation using recurrent convolutional neural network with highway letters.**

In [8] the joining of zone finding out about the helpful request of EHRs was also effective. Plus, the delayed consequences of the examinations of the malady particular judicious features not simply contained revelations relentless with existing therapeutic territory adapting, yet furthermore conveyed a couple of hypothetical proposition. Their strategy and results could be capable to update the perception of sickness specific settings and moreover to improve the insightful execution in mortality showing in intense healing center care.

- **Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care.**

In [9] they join illness particular settings into mortality displaying by detailing the mortality forecast issue as a multi-errand learning issue in which an undertaking relates to an ailment. Our technique viably coordinates restorative area information relating to the similitude among illnesses and the likeness among Electronic Health Records (EHRs) into information driven approach by joining chart Laplacians into the regularization term to encode these likenesses. The test comes about on a genuine dataset from a healing facility support the viability of the proposed strategy. The Acute Hospital Care (AUC) of a few baselines was enhanced, including calculated relapse without multi-errand learning and a few multi-undertaking learning strategies that don't consolidate the area information. Moreover, we show some fascinating outcomes relating to disease specific prescient highlights, some of which are not just steady with existing medicinal area learning, yet in addition contain suggestive theories that could be approved by facilitate examinations in the medicinal area.

- **Septic shock prediction for patients with missing data.**

In [10] Sepsis and septic shock are normal and possibly lethal conditions that regularly happen in Intensive Care Unit (ICU) patients. Early expectation of patients in danger for septic stun is hence significant to limiting the impacts of these entanglements. Potential signs for septic stun hazard traverse an extensive variety of estimations, counting physiological information assembled at various fleeting resolutions and quality articulation levels, driving to a nontrivial forecast issue. Past deals with septic stun forecast have utilized little, deliberately curated datasets or clinical estimations that may not be accessible for some ICU patients. The current accessibility of a huge, rich ICU dataset called MIMIC-II has given the chance to broader demonstrating of this issue. Be that as it may, such an extensive clinical dataset definitely contains a significant sum of missing information. We examine how extraordinary ascription choice criteria and strategies can overcome the missing information issue. Our outcomes demonstrate that attribution techniques in conjunction with prescient displaying can prompt exact septic stun expectation, regardless of whether the highlights are confined essentially to noninvasive estimations.

III.PROBLEM STATEMENT

A. Problem Statement:

In this section, we first design A Machine Learning Methodology for Diagnosing Chronic Kidney Disease.

B. Goals & Objectives:

The aim of the present study was to systematically review published economic models that simulated long-term outcomes of kidney disease to inform cost-effectiveness evaluations of CKD treatments.

IV. PROPOSED SYSTEM

A. Proposed Work:

An In the present study, a number of different ML classifiers are experimentally validated to a real data set, taken from the UCI Machine Learning Repository, and our findings are compared with the findings reported in the recent literature.

The results are quantitatively and qualitatively discussed and our findings reveal that the random forest (RF) classifier achieves the near-optimal performances on the identification of CKD subjects.

Hence, we show that ML algorithms serve important function in diagnosis of CKD, with satisfactory robustness, and our findings suggest that RF can also be utilized for the diagnosis of similar diseases.

System Architecture:

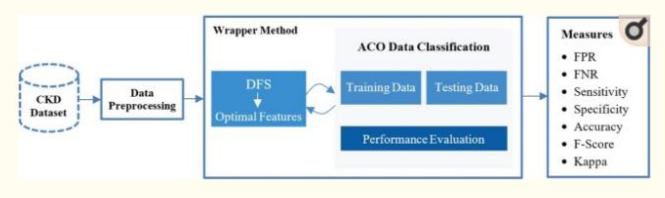

Fig.1 : System Architecture

V. METHODOLOGY

1. Machine Learning

Machine Learning is such a _eld which gives an ability to learn without being explicitly programmed. They mainly focus on the prediction. Statistics and Machine Learning are closely related fields. They can be divided into three categories: 1) Supervised Learning 2) Unsupervised Learning 3) Reinforcement Learning.

2. Healthcare

Healthcare is preserving or improving the health through prevention, diagnosis and treatment of that particular disease. Healthcare contributes beyond the delivery of services to the patients. It contributes to the part of country economy. It is mainly regarded as to determine in promoting the physical and mental health around the world.

3. Naive Bayesian

The Naive Bayes is a classification technique based on the Bayes Theorem. They are easy to build and useful for large data set. It is even a highly sophisticated classification method. It is used to predict the multi-class prediction. It performs well for categorical input variables. Naive Bayes is a simple technique for constructing classifier: models that assign class label to problem instances, represented as vector of feature values, where the class labels are drawn from some finite set. The application is the real time prediction, text classification, multi-class prediction.

4. K Nearest Neighbor

The k-Nearest Neighbor stores all the cases and classifies new class based on the similarity measures. The output of k-NN is the class membership. The object is been classified on the basis of majority votes of its neighbors. The values of output are by averaging the values of its k nearest neighbor. It is a special case of variable bandwidth. They are used for both classification and regression. The neighbors are taken from the set of objects for which the class or the object property value is known. It is sensitive to the local structure of the data.

VI. CONCLUSION

The proposed CKD diagnostic methodology is feasible in terms of data imputation and samples diagnosis. After unsupervised imputation of missing values in the data set by using KNN imputation, the integrated model could achieve a satisfactory accuracy. Hence, we speculate that applying this methodology to the practical diagnosis of CKD would achieve a desirable effect. In addition, this methodology might be applicable to the clinical data of the other diseases in actual medical diagnosis. However, in the process of establishing the model, due to the limitations of the conditions, the available data samples are relatively small, including only 400 samples. Therefore, the generalization performance of the model might be limited. In addition, due to there are only two categories (ckd and notckd) of data samples in the data set, the model can not diagnose the severity of CKD. In the future, a large number of more complex and representative data will be collected to train the model to improve the generalization performance while enabling it to detect the severity of the disease. We believe that this model will be more and more perfect by the increase of size and quality of the data.

VII. ACKNOWLEDGMENT

To develop A Machine Learning Methodology for Diagnosing Chronic Kidney Disease System using symptom detection that accurately updates databases according to the syptoms of user and maintains transparency in the system and prevents forgery.

VIII. REFERENCES

[1] Z. Chen et al., “Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers,” Chemometr. Intell. Lab., vol. 153, pp. 140-145, Apr. 2016.

[2] A. Subasi, E. Alickovic, J. Kevric, “Diagnosis of chronic kidney disease by using random forest,” in Proc. Int. Conf. Medical and Biological Engineering, Mar. 2017, pp. 589-594.

[3] L. Zhang et al., “Prevalence of chronic kidney disease in china: a crosssectional survey,” Lancet, vol. 379, pp. 815-822, Aug. 2012.

[4] A. Singh et al., “Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration,” J. Biomed. Inform., vol. 53, pp. 220-228, Feb. 2015.

[5] A. M. Cueto-Manzano et al., “Prevalence of chronic kidney disease in an adult population,”

Arch. Med. Res., vol. 45, no. 6, pp. 507-513, Aug. 2014.

[6] H. Polat, H.D. Mehr, A. Cetin, “Diagnosis of chronic kidney disease based on support vector machine by feature selection methods,” J. Med. Syst., vol. 41, no. 4, Apr. 2017.

[7] C. Barbieri et al., “A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis,” Comput. Biol. Med., vol. 61, pp. 56-61, Jun. 2015.

[8] V. Papademetriou et al., “Chronic kidney disease, basal insulin glargine, and health outcomes in people with dysglycemia: The origin study,” Am. J. Med., vol. 130, no. 12, Dec. 2017.

[9] N. R. Hill et al., “Global prevalence of chronic kidney disease - A systematic review and meta-analysis,” Plos One, vol. 11, no. 7, Jul. 2016.

[10] M. M. Hossain et al., “Mechanical anisotropy assessment in kidney cortex using ARFI peak displacement: Preclinical validation and pilot in vivo clinical results in kidney allografts,” IEEE Trans. Ultrason. Ferr., vol. 66, no. 3, pp. 551-562, Mar. 2019.

Cite this article as :

Prof. Rashmi Patil, Bhagyashri Deshmukh, Dhanshri Lonkar, Mahima Kumari, "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 9 Issue 2, pp. 295-301, March-April 2022.
Journal URL : https://ijsrst.com/IJSRST229258