A Journey of Machine Learning Algorithms for the prediction of Long-Term Kidney Disease

Jayashree M, Associate Professor, EPCET, Bangalore, India, jayashree.raju2006@gmail.com

Abstract: Long-term kidney disease (CKD) is becoming major health problem all over the world. Long-term kidney disease(CKD) is the one in which the function of the kidney is slowly getting decreased, and goes unnoticed till the end stage.CKD has becoming more common in children presenting unique features. These features not only influence the child's health but have long term impact on the life of the adult that they will become. With this respect the focus is on the unique issues of the pediatric kidney disease. If the factors and symptoms related to the PCKD are seen during the childhood then proper treatment is required so that CKD progression can be prevented. The data mining techniques can be used to observe the historical data and to discover patterns in order to find out future decisions. the efficient machine learning algorithms can be used to process this data , enabling the use of computationally intensive methods for data analysis. Various data mining classification approaches and machine learning algorithms are applied for prediction of Long-term diseases.

Keywords: Long-term Kidney disease, Random forest, Ada Boost, Naïve Bayes, K-nearest neighbour, Gradient boost

I. Introduction

Long-term Kidney disease (CKD) is a condition in which the kidneys are permanently damaged. It affects people of all age groups. The challenging task is to diagnose CKD at the early stage ,as in it's earliest stage symptoms are not visible. When symptoms appear, it is mandatory to find the stage of the kidney disease. There are two different kinds of Kidney failure exist.

Short-term renal disease. Short-term kidney disease starts suddenly. In some cases, after some duration kidneys start to work normally. i.e function of the kidney can be reversible.

Long-term kidney disease. This type gets worse slowly over at least 3 months and It can lead to permanent kidney failure.

The causes of CKD in different age groups

In infants and children : , congenital abnormalities, and hereditary diseases, Birth defects, polycystic kidney disease, are the most common causes of CKD. Frequent urinary tract infections (UTIs) in children should be promptly treated and further evaluated, as urinary tract abnormalities could potentially lead to CKD.

In teenagers: In children over the age of 12, glomerulonephritis (inflammation of the kidneys) is the most frequent cause of kidney failure. Other conditions that may damage the kidneys, like nephrotic syndrome, or diseases that affect many organs, like lupus, are also common causes.

In adults*:* CKD is mainly caused by diabetes and high blood pressure. In contrast to adults, high blood pressure does not usually cause kidney failure in children, but often is caused by the renal failure. Hence it is important to note that many of the risk factors for CKD, such as obesity leading to type 2 diabetes, start in childhood and may contribute to progressive kidney disease in adulthood

Short-term kidney disease[7] may be caused by:

- Reduced blood flow to the kidneys for a period of time.
- A blockage in the urinary tract Taking medicines that may cause kidney problems Any condition that may slow or block oxygen and blood to the kidneys, such as cardiac arrest[7].
- Hemolytic uremic syndrome. This is usually caused by an E. coli infection. Kidney failure develops because small structures and vessels in the kidney are blocked.
- Glomerulonephritis. This is a type of kidney disease that happens in parts of the kidneys called glomeruli. The glomeruli become inflamed and harm how the kidney filters urine.

Long-term Renal disease may be caused by:

- A long-term blockage in the urinary tract
- **Alport syndrome.** This is an hereditary disorder. It causes hearing problem, kidney damage and eye defects.
- **Nephrotic syndrome.** This is a condition that causes protein in the urine, low protein in the blood, high cholesterol levels, and tissue swelling.
- **Polycystic kidney disease.** This is a genetic disorder. It causes many cysts filled with fluid to grow in the kidneys.

DOI : 10.18231/2454-9150.2019.0594

- **Cystinosis.** This is an inherited disorder. The amino acid cystine collects in cells in the kidney called lysosomes.
- **Other Long-term conditions.** Conditions such as diabetes or high blood pressure can lead to kidney problems. If these aren't treated, less oxygen and blood can get to the kidneys.
- **Untreated Short-term kidney disease.** Short-term kidney disease may turn into Long-term kidney disease if not treated.

The Kidney disease can be identified when the following symptoms are observed.

The symptoms for Short-term and Long-term kidney disease may be different. These are the most common symptoms. But symptoms may be a bit different for each child.

Symptoms of Short-term Kidney disease can includeBleeding(hemorrhage),Fever,Rash,Bloodydiarrhea,Severevomiting,Stomachpain, urine,Paleskin,Swelling of the tissues, Inflammation of the eye, Stomach mass.

Symptoms of Long-term kidney disease can include appetite, Vomiting, Bone pain, Headache, less growth, Malaise ,Lots of urine or no urine, Repeated urinary tract infections.

CKD is diagnosed many ways. diagnosis of CKD can be done Based on a child's health issues or symptoms[7], his or her pediatrician may run the following tests:

- **Analysis of urine:** A child's urine will be collected to check for protein. Protein in the urine may be a sign of kidney damage.
- **Blood tests:** Blood tests can help show many things, including kidney function level, blood chemical levels, and red blood cell levels, all of which the kidneys help to control. Sometimes, there are also specialized blood tests that may help diagnose specific kidney diseases such as lupus.
- **Ultrasound and X-rays:** Pictures of the kidneys help show any damage to the kidney and surrounding structures. They may also give hints about what caused the kidney problem.
- **Kidney biopsy:** A small piece of kidney tissue is taken out and examined under a microscope to determine the cause of and extent of damage to the kidneys.

II. LITERATURE SURVEY

Many data mining classifiers are used to predict and detect the Long-term kidney disease problem. Some of the work carried out regarding the prediction of the disease at the early stage is listed out.

Chin-Chi Ku et al.(2019) The author analyzed and Predicted eGFR through convolutional neural networks. For predicting continuous eGFR, the aggregated model was used and achieved a correlation of 0.741 and a mean absolute error (MAE) of 17.605 on the testing dataset. NN architecture is considered for kidney function estimation, based on kidney sonographic images[1]. A neural network architecture discussed in this paper included 33 residual blocks (100 convolution layers) in total.The relationship between predicted and actual eGFRs was visualized using a scatter plot. Results showed that the accurate prediction of the CKD can be done using CNN efficiently.

Jing Xiao et al.(2019) in this paper the linear models including Elastic Net, lasso regression, and the ridge regression is discussed and showed that the rate of the prediction was high in logistic regression[2], with an average AUC and a precision above 0.87 and 0.8, respectively. The sensitivity and specificity was found to be 0.83 and 0.82, respectively. The author summarized that the model with the highest sensitivity was Elastic Net (0.85), while XGBoost showed the highest specificity (0.83). The developed online tool can facilitate the prediction of proteinuria progress.

Q.Zheng et al(2019). In this paper A transfer learning method was used to extract features of kidneys from ultrasound images and classified diseased and normal kidneys, support vector machine classifiers were built on the extracted features using transfer learning imaging features[3] from a pre-trained deep learning model, conventional imaging features.[7].These classifiers were compared, and their diagnosis performance was measured with respect to accuracy, specificity, and sensitivity. The results shows that the classifiers built on the transfer learning features and conventional image features could distinguish abnormal kidney images.

K. Sindhya et al(2017). Used the Genetic Algorithms (GA) as an effective search method, which is a search heuristic that imitates the process of natural evolution. It is used to generate useful solutions to optimization and search problems[4]. The authors perform an inference with the CKD features to find the best discrimination. GA algorithm has been implemented in MATLAB and the different values are tested. Results proved that the performance of the GA has been up to the mark for the classification of the CKD

III. REVIEW OF METHODS USED

As medicine is playing a vital role in human life, knowledge extraction should be done automatically from huge medical data sets. Research on knowledge extraction and analysis of medical data is growing fast. The activities in medicine includes : screening, diagnosis, treatment, monitoring and management. As the healthcare industry is becoming more and more important on technology, machine learning methods are required to assist the physicians in identifying and curing abnormalities at early stages. Medical diagnosis is one of the important activities

DOI : 10.18231/2454-9150.2019.0594

of medicine. The accuracy of the diagnosis contributes in deciding the right treatment and subsequently in cure of diseases. The very first task in analyzing the disease is by considering the dataset.

Dataset

The dataset for diagnosis of Long-term kidney disease is collected from medical reports of the patients. different attributes related to kidney disease like PID (patients ID), Age, Gender, Weight, Serum-albumin, Serum- sodium, Blood urea nitrogen, Serum creatinine, Serum uric acid, Sodium urine, Urine urea nitrogen, Urine creatinine, Urine uric acid ,eGFR are considered. Considering these above attributes the different machine learning algorithms were used for predicting the disease. and they are implemented as classification task for accurate diagnosis of CKD based on different performance evaluation measures.

Artificial Neural Network: Artificial Neural Network:

The mathematical model or computational model which is based on biological neural networks called An artificial neural network (ANN), often just called a "neural network" (NN), is one of the concept used for prediction[5]. It contains interconnected group of artificial neurons and processes information by classification. the given object is allocated to a group based on certain attributes and classification involves pattern recognition and detection which helps in diagnosis and treatment.

Back Propagation Neural Network:

It is a supervised learning algorithm for multi layered neural networks. they are feed forward networks, in which the information flow from the input layer through the hidden layer(s) and then to the output layer. The weights are assigned to each link to determine the node's strength. Calculate the sum ,when sum reaches a threshold value the node produces the output 1 or 0. The training process makes the network to learn where network acquires the knowledge same as human brain from learning experience .Here the network is trained using data for which inputs as well as outputs are known .In this the error between the desired output and the actual output is reduced.

Random Forest:

It is a supervised classification algorithm, the algorithm creates a number of tree to form a forest and makes it random. higher the number of trees in the forest it gives more accuracy .Random forest works in two steps: Creation of random forest and prediction using the classification. In the first stage it selects K features among M features. From the K features it selects the node D using the best split. process will continue for N times to build N number of trees[5]. In the second stage test features are taken and the rules of randomly created decision is used to predict the outcome. the vote for each predicted target is calculated, and highly voted target becomes the final prediction.

Decision Tree

It is a prediction method used for constructing the classification or regression model in a tree structured manner. It is a tree like graph having nodes representing the attributes and edges with answers ,leaves representing output. each node act as test case for some attribute ,It is divided into smaller subsets and develops decision nodes and leaf nodes. The two types of Decision trees used in data mining are classification tree analysis and Regression tree analysis for different predicted outcome .Decision tree are of two types : classification tree and regression tree

K-nearest neighbour Classification:

The K-Nearest Neighbour algorithm (K-NN) is a method used for classification and regression. here the input consists of the K closest training examples in the feature space. K-NN is a type of instance-based learning. In K-NN Classification, the output is a class membership. Classification is done by a majority vote of neighbours[10]. The shortest distance between any two neighbours is always a straight line and the distance is known as Euclidean distance. The constraint of the K-NN algorithm is it's sensitive to the local configuration of the data. The process of transforming the input data to a set of features is known as Feature extraction.

Bayesian Classification

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem. it works on conditional probability, the probability of an event can be calculated using its prior knowledge. the formula for calculating the conditional probability is P(H|E)=P(E|H)*P(H) /P(E). Naive Bayes is a kind of classifier which uses the Bayes Theorem and predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. This is also known as **Maximum A Posteriori (MAP)**..An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters[8] (means and variances of the variables) n necessary for classification.

Ada boost algorithm

Ada boosting algorithm is a powerful predictive learning tasks. in short it can be named as adaptive boosting. it is a classification algorithm whose aim is to convert a set of weak classifiers into strong classifiers. given a dataset containing n points. -1 denote the negative class and +1 denote the positive class. weak classifiers are fit into data set and one with lowest weighted classification error is selected .weight error is calculated[12]. if it is higher than 50% the weight is positive. the more accurate the classifier the larger the weight .after M iteration final prediction is obtained by adding up the weighted prediction of each classifier.

Gradient Boosting

It is one of the machine learning technique considered for regression and classification problems it produces the weak

DOI : 10.18231/2454-9150.2019.0594

prediction model in the form of decision trees. This can be interpreted as an optimization algorithm for cost function[11] .it fits a decision tree to pseudo residuals. Let Jm be the number of leaves. the tree divides the input space into jm disjoint regions and predicts a constant value in each region.

IV.EVALUATION METRICS

The analysis can be done by considering the following metricsand the performance of the model can be evaluated.

Classification accuracy: which is the ratio of number of correct predictions to total number of predictions.

True Positive: Predicted as YES and actual output is also YES

True Negative: Predicted as NO and actual output is also NO

False Positive: Predicted as YES ,actual output is NO

False Negative: Predicted as NO, actual output is YES.

Sensitivity: TP/(FN+TP) Proportion of positive data points considered as positive, with respect to all positive data points.

Specificity: FP/(FP+TN) Proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

V.CONCLUSION

The various machine learning algorithms were considered for analyzing and predicting the disease at the early stage. Evaluation metrics were discussed corresponding to each algorithm. The newer version of machine learning algorithm such as adaboost, gradient boosting is also discussed. the machine learning algorithms playing a vital role in the prediction of disease ,which has become a challenging task in the health care industry.

REFERENCES

[1]Chin-Chi,Kuo, Chun-Min,Chang, Kuan-Ting Liu, Wei-Kai, Lin. “Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning”. *Digital Medicine* volume **2**, Article number:29, 2019.

[2] Jing Xiao,Ruifeng Ding, Xiulin Xu, Haochen Guan, Xinhui Feng Tao Sun,Sibo Zhu, and Zhibin Ye“Comparison and development of machine learning tools in the prediction of Long-term kidney disease progression.” J Transl Med. 2019 Apr 11 17: 119. Published online 2019 Apr 11.

[3]. Q. Zheng, S.L. Furth, G.E. Tasian, Y. Fan’ “Computer-aided diagnosis of congenital abnormalities of the kidney and urinary tract in children based on ultrasound imaging data by integrating texture image features and deep transfer learning image features”. February,2109 volume 15,Issue 1,Pages 75.e1-75.e7.

[4] K. Sindhya, Dr. R. Rangaraj “Design and Development of the Novel Genetic Algorithm Framework for Long-term Kidney Disorder Classification” International Journal of Scientific Research in Computer Science, Engineering and Information Technology 2017 IJSRCSEIT ,Volume 2 , Issue 4 , ISSN : 2456-3307 170

[5] S.ramya, Dr. N.Radha “Diagnosis of Long-term Kidney Disease Using Machine Learning Algorithms”
International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016.

[6] Guneetkaur, Ajaysharma “predict Long-term Kidney disease using Data Mining algorithms in hadoop” International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-5, Issue-4, Apr.-2018.

[7] C.D.W. Kaspar R. Bholah T.E. Bunchman “A Review of Pediatric Long-term Kidney Disease”Advances in CKD 2016 41:211–217.

[8]Asif Salekin John Stankovic “Detection of Long-term Kidney Disease and Selecting Important Predictive Attributes” , IEEE International Conference on Healthcare Informatics.

[9] Parul Sinha, Dr. Poonam Sinha “Performance evaluation of Classification Techniques on Prediction of Long-term Kidney Disease”

[10] Parul Sinha and Poonam Sinha, “Comparative study of Long-term kidney disease prediction using KNN and SVM”, International Journal of Engineering Research & Technology (IJERT), December 2015 (ISSN: 2278-0181), Vol. 4, Issue 12, PP. 608-612,

[11] R. Zemel and T. Pitassi “A Gradient Based Boosting Algorithm For Regression Problems” In NIPS, pages 696–702. 2001.

[12].Robert E Schapire and yoram Singer Boostexter “A Boosting- Based System for Text Categorization” Machine Learning,ToAppear.

DOI : 10.18231/2454-9150.2019.0594