INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

A MACHINE LEARNING METHODOLOGY FOR DIAGNNOSING CHRONIC KIDNEY DISEASE

Y.B. Gopi Chand, K.Chandana,U.Devi bhavani,K.Anil Chowdhary,M.Devagiri Raju,
Assistant professor, Priyadarshini Institute of Technology & Science, AP, India.
Under Graduate,Priyadarshini Institute of Technology & Science, AP, India.

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 a meliorate the progression of this disease. Machine learning models can effectively aid clinic iansachieve this goal due to their fast and accurate recognition performance. In this study, we propose a machine learning methodology for diagnosing CKD. The CKD dataset 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 ill 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 illing out the incomplete data set, six machine learning algorithms (logistic regression, random forest, support vector machine,k- nearestneighbor, naive Bayesclassi ierandfeedforwardneuralnetwor k)wereusedtoestablish 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 regressionand random forest by using perceptron, which could achieve an average accuracy of 99.83% after ten times of simulation.

INTRODUCTION

Chronic kidney disease (CKD) is a signi icant public health problem worldwide, especially for low and mediumincome countries. Chronic kidney disease (CKD) means that the kidney does not work as expected and cannot correctly ilter blood. About 10% of the population worldwide suffers from (CKD), and millions die each year because they cannot get affordable treatment, with the number increasing in the elderly. According to the Global Burden Disease 2010 study conducted by the International Society of Nephrology, chronic kidney disease (CKD) has been raised as an important cause of mortality worldwide with the

ISSN: 2057-5688

number of deaths increasing by 82.3% in the last two decades [1, 2]. Also, the number of patients reaching end-tage renal disease (ESRD) is increasing, which requires kidney transplantation or dialysis to save patients' lives [1, 3, 4]. CKD, in its early stages, has no symptoms; testing may be the only way to ind out if the patient has kidney disease. Early detectionof CKD in its initial stages can help the patient get effective treatment and then prohibit the progression to ESRD [1]. It is argued that every year, a person that has one of the CKD risk factors, such as a family history of kidney failure, hypertension, or diabetes, get checked. The sooner they know about having this disease, the soonerthey can get treatment. To raise awareness and to encourage those who are most susceptible to the disease to perform the tests periodically, we hope that the disease can be detected with the least possible tests and at low cost. So, the objective of this research is to provide an effective model to predict the CKD by least number of predictors.

Chronic Kidney Disease (CKD) is considered as an important threat for the society with respect to the health in the present era. Chronic kidney disease can be detected with regular laboratory tests, and some treatments are present which can prevent development, slow disease progression, reduce complications of decreased Glomerular Filtration Rate(GFR) and risk of cardiovascular disease, and improve survival and quality of life. CKD can be caused due to lack of water consumption, smoking, improper diet, loss of sleep and many other factors. This disease affected 753 million people globally in 2016 in which 417 million are females and 336 millionare males. Majority of the time the disease is detected in its inal stage and which sometimes leads to kidney failure.

Chronic kidney disease (CKD) has received much attention due to its high mortality rate. Chronic diseases have become a concern threatening developing countries, according to the World Health Organization (WHO) [1]. CKD is a kidney disorder treatable in its early stages, but it causes kidney failure in its late stages. In 2016, chronic kidney disease caused the death of 753 million people worldwide, where thenumber of males died was 336 million, while the number of females died was 417 million [2]. It is called “chronic” disease

ISSN: 2057-5688

because the kidney disease begins gradually and lasts for a long time, which affects the functioning of the urinary system. %e accumulation of waste products in the blood leads to the emergence of other health problems, which are associated with several symptoms such as high and low blood pressure, diabetes, nerve damage, and bone problems, which lead to cardiovascular disease. Risk factors for CKD patients include diabetes, blood pressure, and cardiovascular disease (CVD) [3]. CKD patients suffer from side effects,especially in the late stages, which damage the nervous and immune system. In developing countries, patients may reach the late stages, so they must undergo dialysis or kidney transplantation. Medical experts determine kidney disease through glomerular iltration rate (GFR), which describes kidney function. GFR is based on information such as age, blood test, gender, and other factors suffered by the patient [4]. Regarding the GFR value, doctors can classify CKD into ive stages. Table 1 shows the different stages of kidney disease development with GFR levels.

Early diagnosis and treatment of chronic kidney disease will prevent its progression to kidney failure. %e best way to treat chronic kidney disease is to diagnose it in the early stages, but discovering it in its late stages will lead to kidney failure, which requires continuous dialysis or kidney transplantation to maintain a normal life. In the medical diagnosis of chronic kidney disease, two medical tests are used to detect CKD, which are by a blood test to check the glomerular iltrate or by a urine test to check albumin. Due to the increasing number of chronic kidney patients, the scarcity of specialist physicians, and the high costs of diagnosis and treatment, especially in developing countries, there is a need for computer-assisted diagnostics to help physicians and radiologists in supporting their diagnostic decisions. Arti icial intelligence techniques have played a role in the health sector and medical image processing, where machine learning and deep learning techniques have been applied in the processes of disease prediction and disease diagnosis in the early stages. Arti icial intelligence (ANN) approaches have played a basic role in the early diagnosis of CKD.

ISSN: 2057-5688

Machine learning algorithms are used for the early diagnosis of CKD. %e ANN and SVM algorithms are among the most widely used technologies. %ese technologies have great advantages in diagnosing several ields, including medical diagnosis. %e ANN algorithm works like human neurons,which can learn how to operate once properly trained, and its ability to generalize and solve future problems (test data) [5]. However, SVM algorithm depends on experience and examples to assign labels to the class. SVM algorithm basically separates the data by a line that achieves the maximum distance between the class data.

RELATED WORK

Base Paper

In recent years, few studies have been done on the classification or diagnosis of chronic kidney disease. In 2013, T. Di Noia et al. [5], presented a software tool that used the artificial neural network ANN to classify patient status, which is likely to lead to endstage renal disease (ESRD). The classifiers were trained using the data collected at the University of Bari over a 38-year period, and the evaluation was done based on precision, recall, and F-measure. The presented software tool has been madeavailable as both an Android mobile application and online web application. Using data from Electronic Health Records (EHR) in 2014, H. S. Chase et al. [6] identified two groups of patients in stage 3: 117 progressor patients

Where GFR is a glomerular filtration rate that commonly used to detect CKD. Based on initial lab data recorded, the authors used Naïve Bayes and Logistic Regression classifiers to develop a predictive model for progression from stage 3 to stage 4. They compared the metabolic complications between the two groups and found that phosphate values were significantly higher, but bicarbonate, hemoglobin, calcium, and albumin values were significantly lower in progressors compared to nonprogressors, even if initial eGFR alues were similar. Finally, they found that the probability of progression in patients classified as progressors was 81% (73% − 86%) and nonprogressors was 17% (13% −

ISSN: 2057-5688

23%).Later in 2016, K. A. Padmanaban and G.Parthiban [7] aimed in their work to detect chronic kidney disease for diabetic patients using machine learning methods. In their research, they used 600 clinical records collected from a leading Chennai-based diabetes research center. The authors have tested the dataset using the decision tree and Naïve Bayes methods for classification using the WEKA tool. They concluded that the decision tree algorithm outweighs the Naïve Bayes with an accuracy of 91%. A. Salekin and J. Stankovic [8] evaluated threeclassifiers: random forest, K-nearest neighbors, and neural network to detect the CKD.They used a dataset with 400 patients form UCI with 24 attributes. By using the wrapper method, a feature reduction analysis has been performed to find the

ISSN: 2057-5688

attributes that detect this disease with high accuracy. By considering: albumin, specific gravity, diabetes mellitus, hemoglobin, and hypertension as features, they canpredict the CKD with .98 F1 and 0.11 RMSE.

In the study carried out by W. Gunarathne, K. Perera, and K. Kahandawaarachchi [9], Microsoft Azore has been used to predict the patient status of CKD. By considering

14 attributes out of 25, they compared four different algorithms, which were Multiclass Decision Forest, Multiclass Decision Jungle, Multiclass Decision Regression, and Multiclass Neural Network. After comparison, they found that Multiclass Decision Forest performed the best with 99.1% accuracy. H. Polat, H. D. Mehr, and A. Cetin [10] in their research used SVM algorithm along with two feature selection methods: filter and wrapper to reduce the dimensionality of the CKD dataset with two different evaluations for each method. For the wrapper approach, the ClassifierSubsetEval with the Greedy Stepwise search engine and WrapperSubsetEval with the Best First search engine were used. For the Filter approach, CfsSubsetEval with the Greedy Stepwise search engine and FilterSubsetEval with the Best First search engine were used. However, the best accuracy was 98.5% with 13 features using FilterSubsetEval with the Best First searchengine using the SVM algorithm without mentioning which features were used.

P. Yildirim [11] studied the effect of sampling algorithms in predicting chronic kidney disease. The experiment was done by comparing the effect of the three sampling algorithms: Resample, SMOTE, and Spread Sup Sample on the prediction by multilayer perceptron classification algorithm.

A. J. Aljaaf et al. [12] examined in their study the ability of four machine learning (ML) models for early prediction of CKD, which were: support vector machine (SVM), classification and regression tree (CART), logistic regression (LR), and multilayer perceptron neural network (MLP). By using the CKD dataset from UCI and seven features out of 24, they compared the performance of these ML models. The results showed that the MLP model had the highest

ISSN: 2057-5688

AUC and sensitivity. It was also noticeable that logistic regression almost had the same performance as MLP butwith the advantage of the simplicity of the LR algorithm. Therefore, in our study, we can use the LR algorithm as a start or a benchmark and then use more complex algorithms. Lastly in 2019, J. Xiao et al. [13] in their study established and compared nine ML models, including LR, Elastic Net, ridge regression lasso regression SVM,

RF, XGBoost, knearest neighbor and neural network to predict the progression of CKD. They used available clinical features from 551 CKD follow-up patients. They conclude that linear models have the overall predictive power with an average AUC above 0.87 and precision above 0.8 and 0.8, respectively

Pujari et al. [7] presented a system for detecting the stages of CKD through ultrasonography (USG) images. %e algorithm works to identify fibrotic cases during different periods. Ahmed et al. [8] proposed a fuzzy expert system to determine whether the urinary system is good or bad. Khamparia et al. [9] studied a stacked autoencoder model to extract the characteristics of CKD and used Softmax to classify the final class. Kim et al. [10] proposed a genetic algorithm (GA) based on neural networks in which the weight vectors were optimized by GA to train NN. %e system surpasses traditional neural networks for CKD diagnosis. Vasquez-Morales et al. [11] presented a model based on neural networks to predict whether a person is at risk of developing CKD. Almansour et al. [12] diagnosed a CKD dataset using ANN and SVM algorithms. ANN and SVM reached an accuracy of 99.75% and 97.75%, respectively. Rady and Anwar [13] applied probabilistic neural networks (PNN), multilayer perceptron (MLP), SVM, and radial basis function (RBF) algorithms to diagnose CKD dataset. %e

PNN algorithm outperformed the MLP, SVM, and RBF algorithms. Kunwar et al. [14] applied two algorithms—naive Bayes and artificial neural networks (ANN)— to diagnose a UCI dataset for CKD. Naive Bayes algorithm outperformed ANN. %e accuracy of the naive Bayes algorithm was 100%, while the ANN accuracy was 72.73%. Wibawa et al. [15] applied correlation-based

ISSN: 2057-5688

feature selection (CFS) for feature selection, and AdaBoost for ensemble learning was applied to improve CKD diagnosis. %e KNN, naive Bayes, and SVM algorithms were applied for CKD atasetdiagnosis

ISSN: 2057-5688

class is 62.5% with CKD and 37.5% without CKD. The ages of these observations are varied from 2 to 90 years old.

Data Preprocessing

Today's real-world datasets are susceptible to missing, noisy, redundant, and inconsistent data, especially clinical datasets. Working with low-quality data leads to low-quality results. Therefore, the first step in every machine learning application is to explore the dataset and understand its characteristics in order to make it ready for the modeling stage. This process is commonly known as data preprocessing.

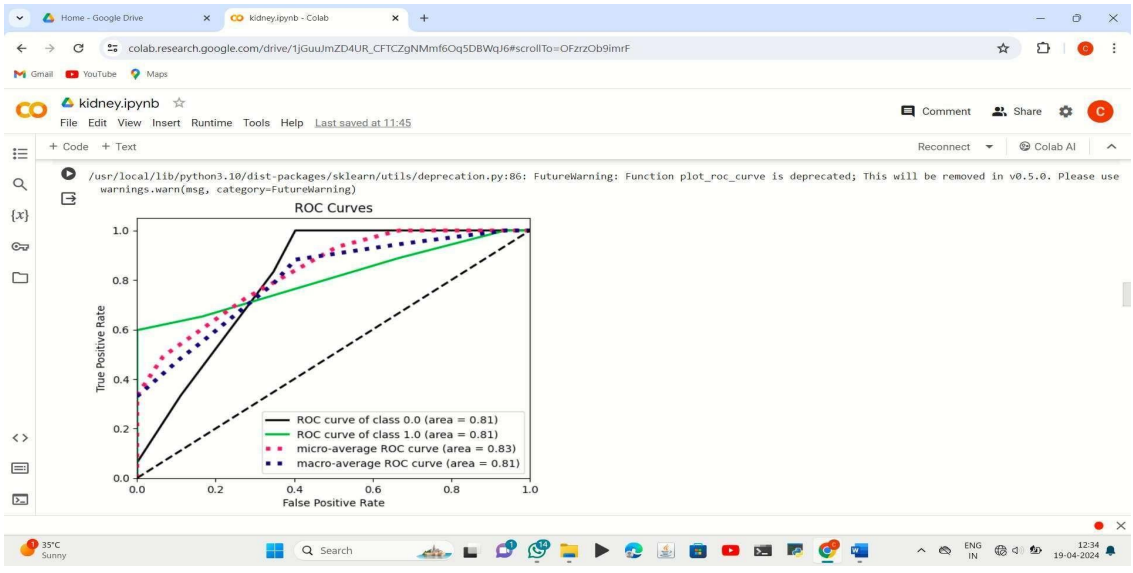
Missing Values

In real-world datasets, missing data is a very common issue, especially in the medical area. Usually, every patient record and every attribute contains some missing values Feature Selection

The process of selecting the most discriminating features in a given dataset is known as feature selection. This process is enhancing the model's performance, reducing overfitting, and reducing the cost of building a model. Filter feature selection methods

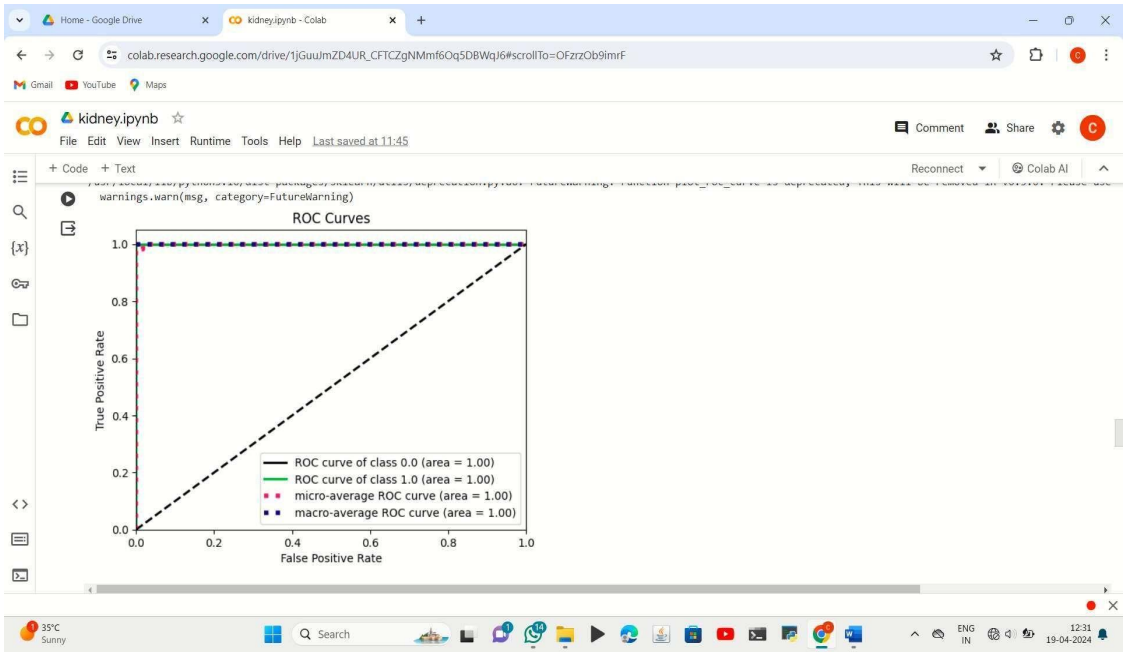
[25] selects features that have a stronger relationship with the outcome variable independent to the learning model. Therefore, use a measure or test independent to the learning algorithm to assess a subset of features.

RESULTS

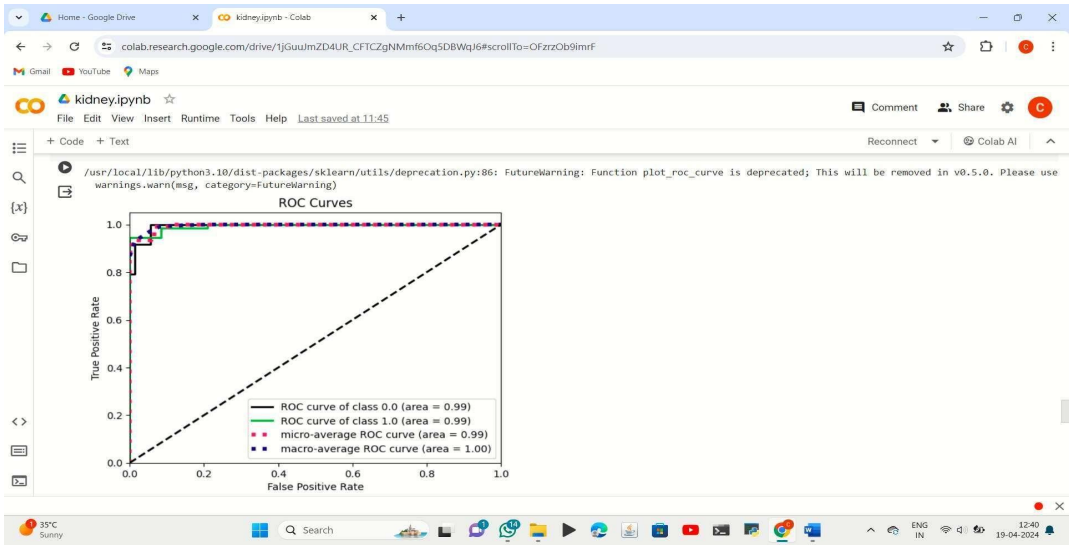


KNN Algorithm accuracy graph . we got 73% accuracy in this algorithm

ISSN: 2057-5688



RANDOM FOREST Algorithm Accuracy graph. we got 98% in this algorithm



NAÏVE BAYES Algorithm Accuracy graph. We got 94% in this algorithm

CONCLUSION

This work examines the ability to detect CKD using machine learning

ISSN: 2057-5688

algorithms while considering the least number of tests or features. We approach this aim by applying four machine learning classifiers: logistic regression, SVM, random forest, and gradient boosting on a small dataset of 400 records. In order to reduce the number of features and remove redundancy, the association between variables have been studied. A filter feature selection method has been applied to the remaining attributes and found that there are haemoglobin, albumin, and specific gravity have the most impact to predict the CKD.

REFERENCE

[1] J. Radhakrishnan et al, "Taming the chronic kidney disease epidemic: a globalview of surveillance efforts," Kidney Int., vol. 86, (2), pp. 246-250, 2014.

[2] R. Lozano et al, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010," The Lancet, vol. 380, (9859), pp. 2095- 2128, 2012.

[3] R. Ruiz-Arenas et al, "A Summary of Worldwide National Activities in Chronic Kidney Disease (CKD) Testing," Epic, vol. 28, (4), pp. 302,2017.

[4] Q. Zhang and D. Rothenbacher, "Prevalence of chronic kidney disease in population-based studies: systematic review," BMC Public Health, vol. 8, (1), pp2008. [5] T. Di Noia et al, "An end stage kidney disease predictor based on an artificial neural networks ensemble," Expert Syst. Appl., vol. 40, (11), pp. 4438-4445, 2013. [6] H. S. Chase et al, "Presence of early CKD-related metabolic complicationspredict progression of stage 3 CKD: a case-controlled study," BMC Nephrology, vol.15, (1), pp. 187, 2014.

[7] K. A. Padmanaban and G. Parthiban, "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease," Indian Journal of Science and Technology, vol. 9, (29), 2016.

[8] A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in Healthcare Informatics (ICHI), 2016 IEEE International Conference On, 2016.

[9] W. Gunarathne, K. Perera and K. Kahandawaarachchi, "Performance

ISSN: 2057-5688

evaluation on machine learning classification techniques for disease classification and forecasti through data analytics for chronic kidney disease (CKD)," in Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference On, 2017.

ISSN: 2057-5688

[10] H. Polat, H. D. Mehr and A. Cetin, "Diagnosis of chronic kidney disease basedon support vector machine by feature selection methods," J. Med. Syst., vol. 41, (4), pp. 55, 2017.

[11] P. Yildirim, "Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction," in Computer Software andApplications Conference (COMPSAC), 2017 IEEE 41st Annual, 2017.

[12] A. J. Aljaaf et al, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in 2018 IEEE Congress on EvolutionaryComputation (CEC), 2018.

13] J. Xiao et al, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of Translational Medicine,vol. 17, (1), pp. 119, 2019.

[14] P. Yang et al, "A review of ensemble methods in bioinformatics," Current Bioinformatics, vol. 5, (4), pp. 296-308, 2010.

[15] L. Deng et al, "Prediction of protein-protein interaction sites using an ensemble method," BMC Bioinformatics, vol. 10, (1), pp. 426, 2009.

[16] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9, (01), pp. 1, 2017.

[17] S. Karamizadeh et al, "Advantage and drawback of support vector machine functionality," in 2014 International Conference on Computer, Communications, andControl Technology (I4CT), 2014.

[18] L. Rubini. (2015). Chronic_Kidney_Disease DataSet, UCI Machine Learning
Repository.Available:https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Diseas
e. [19] J. D. Kelleher, B. Mac Namee and A. D'arcy, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press, 2015.

[20] Michael and B. (2015). Highest blood sugar

ISSN: 2057-5688

level. Available:

http://www.guinnessworldrecords.com/world-records/highest-bloodsugar-level/.

[21] Prasadu Peddi (2018), “A STUDY FOR BIG DATA USING DISSEMINATED FUZZY DECISION TREES”, ISSN: 2366- 1313, Vol 3, issue 2, pp:46-57.