Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 3): pp. 4273-4280
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.03.441

An Optimized Machine Learning Approach For Chronic Kidney Disease Prediction

Poonam Verma Department of Computer Application Graphic Era Hill University, Dehradun, Uttarakhand India, 248002 pverma@gehu.ac.in

Kumud Pant Department of Biotechnology, Graphic Era Deemed to be University, Dehradun, Uttarakhand India, 248002 kumud.pant@geu.ac.in

Abstract

As a result of the rising incidence of the condition, chronic kidney disease (CKD) has emerged as one of the pressing issues facing modern medicine. The capability of a machine vision system to diagnose chronic renal disease based on a small number of patient characteristics is the focus of this investigation. A large number of statistical tests, such as ANOVA tests, Pearson's correlation test results, and Cramer's V tests, have indeed been carried out in order to get rid of duplicate features. All of the different machine learning algorithms—including regression models, random forest, support vector machine, and gradient boosting—were trained and evaluated through a 10-fold cross-validation method. In order to attain an efficiency of 99.53, we make use of the F1 metric that is associated with the gradient augmenting classifier. In addition, it was discovered that haemoglobin has a significant role in the differentiation of CKD in the randomized forest as well as the xgboost models. Our findings represent the most impressive, despite the fact that our study had fewer distinguishing features than those of earlier research. Because of this, chronic kidney disease can be diagnosed with just three easy tests that cost a total of $26.65 just.

I. Introduction

The prevalence of chronic renal disease (CRD), in particular in middle- and low-income nations, is one of the most pressing issues affecting public health on a global scale. Chronic renal illness leads to a reduction in the kidney's capacity to properly filter blood, resulting in reduced kidney function (CKD). Around the world, 10% of the populace has chronic kidney disease, and every year millions around the world, mostly elderly people, pass away as a result of a lack of treatment options. In accordance with the Global Status report 2010 study that was conducted by the Global Alliance of Urology, renal failure (CKD) has surfaced as the major cause of death across the globe, with both the number of people killed and the prevalence of the condition increasing by 82.3% and over course of the preceding two decades [1, 2]. People who are diagnosed with edge kidney impairment (ESRD), a condition that is more prevalent [1, 3, 4], need to undergo dialysis or a lung transplant in order to have any chance of survival. In the initial stages of chronic kidney disease (CKD), there are no indications; testing is the only thing that may be required to evaluate the child's kidney function. Patients who are diagnosed with CKD at an early stage have a better chance of receiving the appropriate treatment and avoiding the development of ESRD [1]. Everyone who has one of the risk factors

for chronic kidney disease (CKD), including such diabetes, high blood pressure, or a history of kidney failure in their family, should get screened at least once a year. If treatment is started at a younger age, children will more quickly learn whether or not they have this illness. We believe that the disease may be diagnosed with the fewest number of steps and for the lowest possible cost. This will allow us to spread knowledge about the condition and persuade individuals who are most likely to be affected by it to get tested on a regular basis. The purpose of this research is to build a model that can accurately predict chronic kidney disease (CKD) with as few symptoms as possible. Figure 1 shows the proposed system architecture.

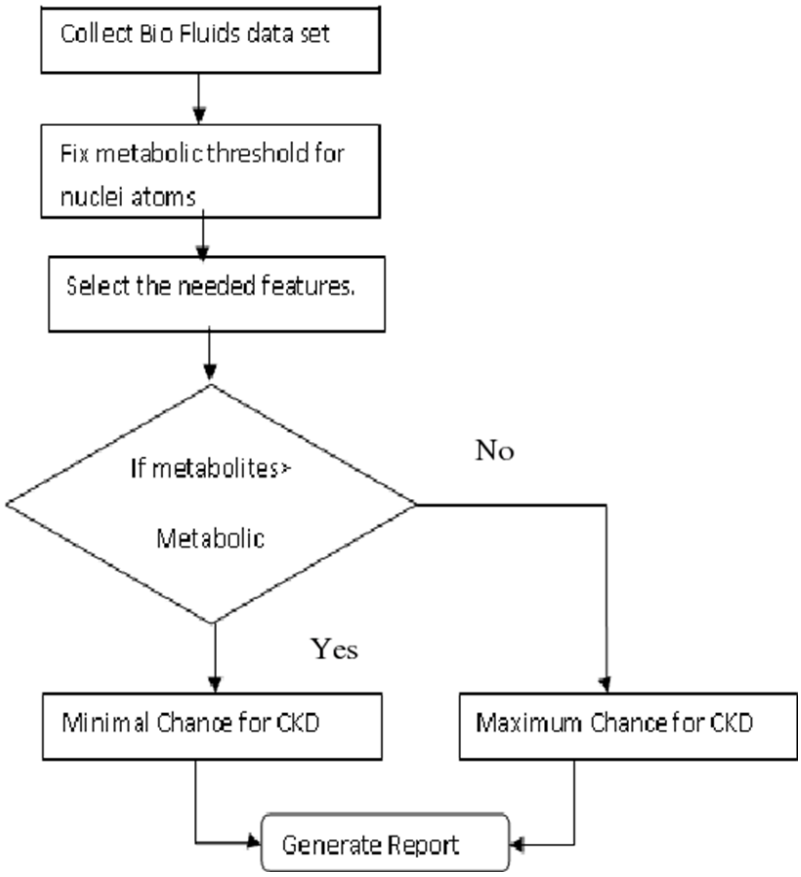

Figure 1. Architecture of the proposed system

II. Related Work

Recent years have not seen a large number of studies devoted to the categorization or diagnosis of progressive renal illness. T. Di Noia et al. [5] classified patients in 2013 in accordance with the probability that they will develop edge kidney damage using an Artificial Neural Networks (ANN). This classification was based on the chance that the individual would develop ESRD (ESRD). The classifiers were evaluated based on their accuracy, recall, and F-measure after having previously been trained using data gathered at the University of Bari over a period of 38 years. An interactive online app and a smartphone app are both potential utilizations of the software instrument in question. In 2014, H. S. Chase et al. [6] used data from Health Care Delivery (EHR) to select two groups of people who were in stage 3: the first group consisted of patients who had been diagnosed with stage 2 colon cancer. Patients who were affected comprised 117 patients who were considered progressors (their eGFR dropped by more than

0.1 m) and 364 patients who were considered quasi patients (their eGFR dropped by 1 ml/min/1.73m2). The urine output, commonly abbreviated as GFR, is a common diagnostic tool for chronic kidney disease (CKD). The researchers have constructed a forecasting model for the progression of the illness between phases three and four by merging columned Bayesian or linear discriminant classifiers with the original lab data that was collected. In spite of the fact that the preliminary eGFR measurements were similar, the researchers investigated the differences in metabolic issues here between 2 categories and discovered that the modern person had significantly lower levels of hypoxia, leukopenia, calcium, and nutrients than the non-progressors did; however, the non-progressors had significantly higher levels of phosphate. In conclusion, the researchers discovered that the chance of advancement was 81% (73%-86%) for patients who have been categorised as progressors, but only 17% (13%-23%) for patients who were considered to be non-progressors. The dataset that was used in this investigation is a relatively small sample that suffers from some degree of imbalance; this aspect of the dataset will be discussed in the section titled "Dataset." As a direct consequence, this dataset is plagued with issues such as asymmetries, informational clutter, and questions with generalisation or generalisation in general. P. Yang et al. [7] came to the conclusion in their evaluation that the predicated-on analysis methods have the benefit of resolving the issue of comparatively small data besides trying to merge and averaged nearly over many classifier to decrease the likelihood of overfitting. This conclusion was reached after the researchers conducted their review. In addition, Deng and colleagues [8] discovered that the predictive algorithm has the potential to alleviate the class imbalance and enhance predictive accuracy when calculating the locations of various molecular sites.

M. Fatima and M. Pasha discovered in a study conducted [9] that the advantages of classifiers and randomization made it possible for SVM to make more accurate predictions regarding heart illness. The information required for this research was collected in 2015 from patients who were attending Apollo Hospital in Delhi over two separate time periods. The Chronic Kidney Failure Set of data that is maintained at the California state University, Irvine [10] has the relevant information in its entirety. These 400 samples contain both null values and clutter in various amounts. There are a total of 350 patient records in this data set, along with 150 documents for individuals who do not have CKD. As a consequence of this, 62.5% of individuals in each category have CKD, compared to 37.5% of individuals who do not. These findings cover a broad age range, ranging from two years old to ninety years old. The CKD dataset has a total of 24 features, 11 of which are quantitative and 13 of which are nominal. These features are outlined in Table I below. The presence or severity of chronic kidney disease (CKD) can be determined by the 25th feature. The proposed research consists of the following steps:

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd

Table 1. Features in the data set

III. Proposed System

A. Data preprocessing:

First and foremost, prepare the data In today's modern different datasets, healthcare datasets, in especially, are susceptible to incomplete information, inaccurate data, data duplication, and data inconsistency. The outputs will reflect the quality of the incoming data, so pay attention to both. In order to have the dataset prepared for the simulation phase, the first step of every machine tool for learning is to analyse the dataset and gain a knowledge of the features it contains. "Data pre-processing" is the name given to this method of processing.

2) Outliers: Critical values that are situated at a location that is remote from the concept's centre tendency are referred to as outliers. The origin of data noise can be traced back to inaccurate outliers brought by by errors during data entry [11]. When it comes to dealing with anomalies, health records cannot be treated in the same way as other data since the outliers in question can be real (valid) or significant. Each and every one of the outliers that are found in the CKD dataset is investigated in depth in order to determine whether or not they are credible. Severe pieces of information in this study that fall outside of the allowed limits in terms of healthcare have indeed been regarded as incomplete information and later adjusted, as will be detailed in the chapter on missing information. These latest results will be discussed in more detail in the section on data. In order to identify abnormalities in the CKD sample, graphs have been utilised. As can be seen in Figure 1, there were a number of blood glucose random readings that above 500 mg/dl that were classified as outliers. In 2008, a participant's blood sugar levels could not have been higher than unions and professional mg/dl and yet have had a chance of survival, as stated in [12]. Because of this, given that these outliers are correct, we shouldn't make any changes to them. However, in the case of both potassium and sodium, 3 data sets at the extremes are not sufficient. The highest level of potassium that has ever been discovered is 7.6 mEq/L [13]. This demonstrates that the values of 38 and 46 for the blood concentrations in Fig. 2 are illogical and most likely the result of an error in the data collection process. As is the case with potassium, calcium only has single data point that falls into the extreme category; this point is 4.5, as seen in Figure 2. If the child's sodium content is less below 135 mEq/L [14], then the patient is diagnosed with hyponatremia. So because average range of sodium intake is between between 135 and 145 mEq/L, a finding of 4.5 is therefore inappropriate or impossible to implement because of this.

3) Missing Values: In different datasets, discrepancies are a highly widespread problem, particularly in the healthcare industry. This problem is especially prevalent in the United States. In most cases, some values are absent from each attribute as well as the patient record . In contrast to this, the dataset for chronic kidney disease contains 96% of its features with missing or partial data. In 60.75 % (243) of the occurrences, there is most one binary value, and 10% of all ideas are not present. The amount of missing information varies greatly depending on the category, ranging anywhere from 0.3% all the way up to 38%. When it came time to simulate the CKD dataset, the researchers from [14] relied on single-imputation methods such as average or median. Nevertheless, the results of little's test with a p-value indicate that the

missing information inside the CKD dataset are not totally absent due to a random occurrence .

4) Data Reduction: The technique of cutting down on the number of features in an analysis while still producing reliable findings is known as "data reduction." Research has been done on the elimination of duplicate information, focusing on the selection of features, their relationships, and their correlations.

a) Feature Associations: Finding correlations among datasets has been accomplished through the use of tests such as Regression analysis, Cramer's V, and ANOVA. It is abundantly obvious that mean corpuscular capacity and hemoglobin, as well as haematocrit and red blood cells count, are substantially connected. The correlation coefficients for these two sets of variables are 0.89 and 0.79, correspondingly. In addition, the p-value link among anaemia and PCV was found to be confirmed by the Anova analysis. Because they have a bigger impact on the class feature than its related aspects, we have come to the conclusion that hemoglobin and creatinine levels should be kept, whereas the rest of the attributes should be removed because they are unnecessary.

b) Feature Selection

The term "feature selection" is used to describe the process of choosing the characteristics that have greatest computation impact on our output or making changes. In this research, Ant colonies optimization (ACO) was utilised to identify the most significant aspects of the information. It's a method for efficiently finding ways to navigate graphs for solving computational issues. Multi-agent systems, "Artificial Ants," are modelled after the cooperative strategies of real ants. Pheromones play a crucial role in the communication of ants in the wild. In the past, numerous optimization issues with a graph have been solved with the help of Synthetic Ants and search engine methods. Pheromone concentrations are assessed after each cycle as opposed to being accumulated. The proposed method decreases the number of attributes in the chosen subgroups by selecting the most suitable ants. Classification techniques are required to evaluate the efficiency of the wrappers evaluation function's constituent parts. Ant colony optimisation (ACO) requires recasting the optimization model as one of locating the shortest route on a weighted graph. At the outset of each cycle, the ants randomly produce a solution, which is the recommended sequence in which to traverse the map's edges. In the second part, we examine the various courses taken by the ants and make comparisons. Finally, the concentration of pheromones at each edge is modified. Each ant needs a way to traverse the graph, thus it's up to us to figure out a solution. An ant will take its current pheromones concentration and the distance of each nearby edge into account before deciding which one to explore next.

IV. Results and Discussion

Each classifier's input has been evaluated using a wide variety of criteria, and the findings have been validated using 10-fold cross-validation to prevent overfitting. Adjustments to the model's properties were also made with the help of multilayer cross-validation. The tests are executed in Python 3.3 via the Jupyter Notebook web-based software. Sciket-learn is a popular free Python deep learning package that has been used to create numerous other libraries. Accuracy

(as measured by the F1-measure), sensitivity, specificity, and the area beneath the curve were used as markers of efficacy in this investigation (AUC). Different models produce different sets of results when their parameters are given different values. The most accurate LR result achieved using layered classification technique has a 98.9 percent success rate when utilising the F1 measurement, C=1000, and punishment settings. Multiple "C," "gamma," and "kernel" combinations have been tested with the SVM model. We found that C=1, gamma=3, kernel="RBF" (rbf function), and an F1 score of 97.9 percent gave the best results for SVM.

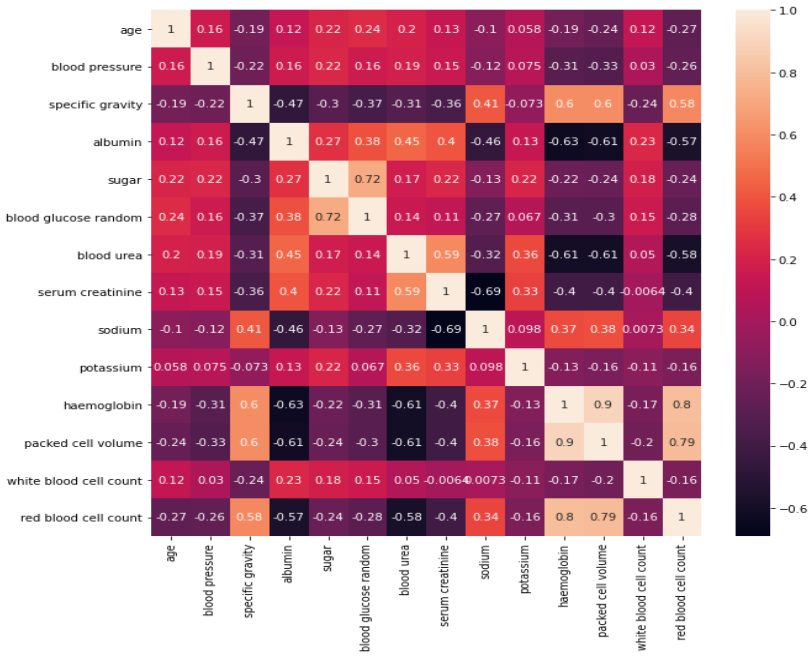

Figure 2. Confusion matrix

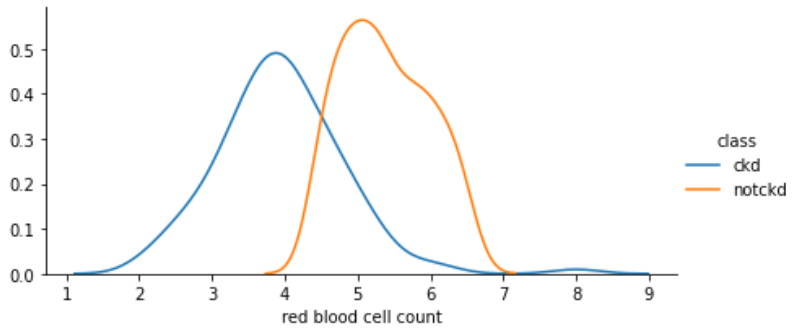

Figure 3. Detected Values vs RBC count

We also look at the significance of the qualities in each method, since the RF and Bg produced the best outcomes. Figure 3 shows that between RF and GB, haemoglobin is the highest and albumin is the lowest rated protein. There is a narrow range between 0.29 and 0.44 between how important various elements are, as determined by RF. Nevertheless, the relevance of haemoglobin (0.77), compared to other variables in GB, varies greatly. Based on our findings, we conclude that haem has proven useful in the diagnosis of CKD. However, the dataset this research relied on has a few caveats. One potential issue is that the small size of the dataset

(about 400 cases). Second, evaluating the consequences of the datasets requires obtaining complementary information with the same attributes. Figure 4 shows the training and testing accuracy of three different methods.

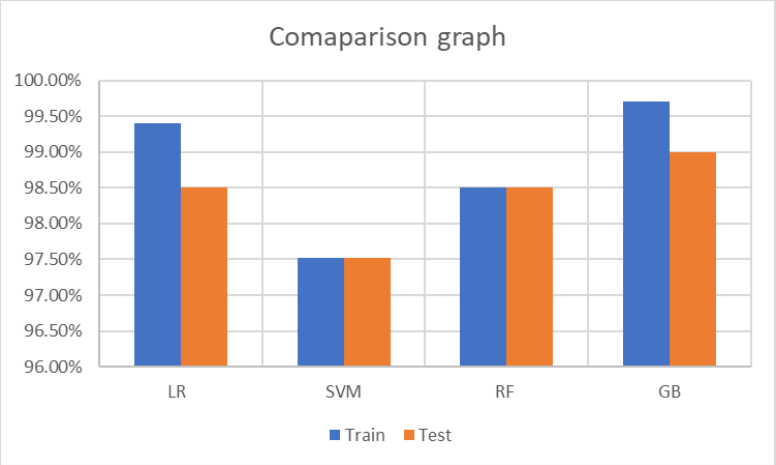

Figure 4. Comparison of three models

Conclusion

The overarching purpose of this research is to establish how well machine learning algorithms can detect CKD with the minimum number of tests and features. On a tiny dataset of 400 records, we test out four different machine learning classifier ideas: engine optimization is the process, randomized forest, regression analysis, and support vector machine (SVM). As a means of streamlining the list of traits and eliminating unnecessary ones, the interdependencies among them have been studied. After applying a filter feature selection method to the remaining variables, it was determined that haemoglobin, antibody, and relative density were the most influential in terms of being able to predict the start of chronic renal disease. Ten-fold cross-validation was used to train, test, and verify the classifiers. The gradient boosting technique improved in all three metrics we used to evaluate performance: F1-measure (99.1%), hypersensitive (98.8%), and specific (99.3%). In comparison to other experiments with fewer features and, thus, lower expenses, this one yielded the best results. As a consequence, we conclude that the presence of CKD can be determined by monitoring just three variables. Furthermore, we discovered that the RF and GB models for detecting CKD each have their own unique role for haemoglobin and albumin. To make up for the small sample size of this study, we hope to replicate the results with a more representative data set in the future. Furthermore, we aim to use relevant data to predict if a person with CKD risk variables such mellitus, hypertensive, and a personal history of kidney failure will develop CKD inside the future.

References

[1]. J. Radhakrishnan et al, "Taming the chronic kidney disease epidemic: a global view 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," Ejifcc, 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), pp. 117, 2008.
[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 complications predict progression of stage 3 CKD: a case-controlled study," BMC Nephrology, vol. 15, (1), pp. 187, 2014.
[7]. P. Yang et al, "A review of ensemble methods in bioinformatics," Current Bioinformatics, vol. 5, (4), pp. 296-308, 2010.
[8]. L. Deng et al, "Prediction of protein-protein interaction sites using an ensemble method," BMC Bioinformatics, vol. 10, (1), pp. 426, 2009.
[9]. S. Karamizadeh et al, "Advantage and drawback of support vector machine functionality," in 2014 International Conference on Computer, Communications, and Control Technology (I4CT), 2014,
[10]. L. Rubini. (2015). Chronic_Kidney_Disease DataSet, UCI Machine Learning Repository. Available: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.
[11]. 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.
[12]. Michael and B. (2015). Highest blood sugar level. Available: http://www.guinnessworldrecords.com/world-records/highest-bloodsugar-level/.
[13]. G. Gheno et al, "Variations of serum potassium level and risk of hyperkalemia in inpatients receiving low-molecular-weight heparin," Eur. J. Clin. Pharmacol., vol. 59, (5-6), pp. 373-377, 2003.
[14]. D. A. Henry, "In The Clinic: Hyponatremia," Ann. Intern. Med., vol. 163, (3), pp. ITC1-ITC19, 2015. DOI: 10.7326/AITC201508040.