Prediction of chronic kidney disease from 25 clinical features using machine learning

Yifan Xing

School of Public Health, Imperial College London, London, United Kingdom

yifan.xing22@imperial.ac.uk

Abstract. A growing number of people worldwide suffer from chronic kidney disease, with many individuals in developing countries lacking the necessary resources for treatment. Medical records often contain valuable information that can be utilized to predict the development of CKD, with machine learning algorithms proving particularly effective. In this study, the author analyzes a dataset of 250 participants with CKD and 150 participants without from 2015, utilizing various machine learning classifiers to determine the most significant characteristics and predict CKD development. The analysis reveals that serum creatinine, specific gravity, red blood cell count, and potassium are the four most relevant risk factors for CKD prediction. Based on these four factors, the author builds machine-learning models that can accurately predict CKD development from medical records. The results show that a combination of all the features in the original dataset achieves a similar level of accuracy as the four-feature models. This research has significant implications for clinical practice, providing doctors with a new tool to predict CKD in patients. By focusing on the most relevant features, such as serum creatinine, red blood cell count, specific gravity and potassium, physicians can make more informed decisions when treating patients with CKD.

Keywords: chronic kidney disease, feature ranking, machine learning.

1. Introduction

A major public health problem, chronic kidney disease (CKD), leads to remarkable adverse health outcomes, as well as renal failure and death [1]. According to a serious medical report, it is estimated that 324 million people suffer from chronic kidney disease worldwide, especially in developing countries [2, 3]. However, in developing countries, a lack of infrastructure and high diagnostic costs have prevented people from benefiting from early-stage CKD screenings [3].

Therefore, machine learning (ML)-based methods have been proposed to detect chronic kidney disease [4]. Different ML methods have been proposed to predict CKD utilizing the CKD dataset at the University of California, Irvine (UCI) machine learning repository [5]. Qin et al. used six classifiers to train and test the preprocessed data [5]. To address misclassification, an integrated classifier was developed to improve the accuracy to 99.83%. Meanwhile, one ML method proposed by Ebiaredoh-Mienye et al., using a modified network of sparse autoencoder (SAE) combined with a SoftMax layer, can achieve the best performance of 98% when predicting CKD [6]. Chittora et al. found that linear support vector machine (LSVM) with an N2 norm can achieve a higher accuracy of 98.46% using selected feature sets compared to complete feature sets [7]. Furthermore, another CKD prediction approach developed by Silveira et al., using decision trees with synthetic minority oversampling technique (SMOTE), has achieved an accuracy of 98.99% [8]. After evaluating 12 classifiers, Sahil Sharma explored that decision trees performed the best in predicting CKD, achieving an accuracy of 98.6%, a precision of 1, a sensitivity of 0.9720 and a specificity of 1 [9]. From these research works, various attributes can be utilized to identify patients who have a risk of developing CKD, enabling clinicians to intervene early and efficiently.

For machine learning problems to be solved, algorithms have to utilize the discriminative capabilities of features in order to classify samples. Most ML applications, particularly medical diagnosis, don't give equal weight to all input features [4]. As a result of feature selection, the redundant attributes from training data are removed, thereby providing the training algorithm with a more accurate representation of the input data. However, only a few studies have attempted to identify the most critical features that can be used to improve CKD prediction using ML [10]. Meanwhile, one of the most challenging things is to solve the imbalanced class problem when applying ML methods to clinical prediction [11].

This study proposed to fill the gaps by using several ML techniques first to predict CKD, then to rank the most relevant clinical features, and finally to aggregate both prediction and feature ranking on top features.

2. Methods

The author first described and preprocessed all the data, and then listed all methods used for binary classification, followed by ML methods for feature ranking, and finally utilized both classification and feature ranking on top clinical features. All methods were implemented using the open-source R programming language.

2.1. Survival prediction classifiers

This study employed seven different machine learning approaches to predict CKD, including Decision Tree, Random Forest, Light Gradient Boosting Machine, eXtreme Gradient Boosting, Support Vector Classifier, AdaBoost Classifier and Logistic Regression.

Common confusion matrix rates were used to measure the prediction results. For example, Matthew's correlation coefficient (MCC), receiver operating characteristic (ROC) area under the curve and F1 score. Based on the dataset imbalance, the score of MCC could reach high only if the predictor was able to perform well on both positive and negative data. Accordingly, the author did the result ranking based on Matthew's correlation coefficient rather than the other confusion matrix metrics.

2.2. Feature ranking

Regarding ML feature ranking, the author focused exclusively on Random Forests since it performed the best on the entire dataset. Random Forest provides one feature ranking technique: mean accuracy reduction. It generates several random Decision Trees during training, each of which contains a subset of both data examples and clinical features. Random Forests decides its final outcome based on a majority vote of the entire categorical outcomes of all the decision trees. When one risk factor is removed, the feature ranking is based on how much prediction accuracy decreases. By comparing this accuracy with the accuracy obtained by using all features, the method determines the importance of certain features: the greater the decrease in accuracy, the more important the feature is.

2.3. Aggregating feature ranking and prediction on top features

Monte Carlo cross-validation was utilized for training and testing the whole dataset. 100 execution has been set. For each execution, the training set contained 70% of the data, and the testing set included 30%. Rather than using the same balance ratio throughout the dataset, the author used a more realistic prediction.

The methods of machine learning used for ranking and classifying top features were the same as those used for CKD prediction on the entire dataset. As part of our effort to show the generalizability of this approach, a computational solution based on all seven methods was demonstrated to be valid not only with machine learning classifiers but also with different groups of data.

3. Results and discussion

This section first describes the complete data (Table 1, 2 and 3), and then interprets the results acquired for CKD prediction on the entire dataset (Table 4), feature ranking (Figure 1), and CKD prediction using only the four most critical features from the complete dataset (Table 5).

3.1. Results

3.1.1. Description of dataset. The dataset in Table 1 includes the medical records of 400 participants aged 2 to 90 years old collected by Apollo Hospitals in Tamil Nadu, India, in 2015. The participants comprised 250 CKD patients and 150 of those who did not suffer from CKD, which made an imbalanced dataset. The dataset contains one outcome variable and 24 predictors (11 numerical and 13 nominal).

Table 1. Quantitative description of the categorical characteristics.

Category feature		Full sample		Ckd patients		Notckd patients	
		#	%	#	%	#	%
sg	1.005	6.00	1.78	6.00	3.17	0.00	0.00
	1.01	73.00	21.60	73.00	38.62	0.00	0.00
	1.015	71.00	21.01	71.00	37.57	0.00	0.00
	1.02	104.00	30.77	28.00	14.81	76.00	51.01
	1.025	84.00	24.85	11.00	5.82	73.00	48.99
al	0.00	196.00	57.99	47.00	24.87	149.00	100.00
	1.00	39.00	11.54	39.00	20.63	0.00	0.00
	2.00	40.00	11.83	40.00	21.16	0.00	0.00
	3.00	38.00	11.24	38.00	20.11	0.00	0.00
	4.00	24.00	7.10	24.00	12.70	0.00	0.00
	5.00	1.00	0.30	1.00	0.53	0.00	0.00
su	0.00	280.00	82.84	131.00	69.31	149.00	100.00
	1.00	13.00	3.85	13.00	6.88	0.00	0.00
	2.00	16.00	4.73	16.00	8.47	0.00	0.00
	3.00	12.00	3.55	12.00	6.35	0.00	0.00
	4.00	14.00	4.14	14.00	7.41	0.00	0.00
	5.00	3.00	0.89	3.00	1.59	0.00	0.00
rbc	abnormal	73.00	21.60	73.00	38.62	0.00	0.00
	normal	265.00	78.40	116.00	61.38	149.00	100.00
pc	abnormal	76.00	22.49	76.00	40.21	0.00	0.00
	normal	262.00	77.51	113.00	59.79	149.00	100.00
pcc	notpresent	298.00	88.17	149.00	78.84	149.00	100.00
	present	40.00	11.83	40.00	21.16	0.00	0.00
ba	notpresent	318.00	94.08	169.00	89.42	149.00	100.00
	present	20.00	5.92	20.00	10.58	0.00	0.00

Table 1. (continued).

htn	no	221.00	65.38	72.00	38.10	149.00	100.00
	yes	117.00	34.62	117.00	61.90	0.00	0.00
dm	no	234.00	69.23	85.00	44.97	149.00	100.00
	yes	104.00	30.77	104.00	55.03	0.00	0.00
cad	no	310.00	91.72	161.00	85.19	149.00	100.00
	yes	28.00	8.28	28.00	14.81	0.00	0.00
appet	good	271.00	80.18	122.00	64.55	149.00	100.00
	poor	67.00	19.82	67.00	35.45	0.00	0.00
pe	no	280.00	82.84	131.00	69.31	149.00	100.00
	yes	58.00	17.16	58.00	30.69	0.00	0.00
ane	no	293.00	86.69	144.00	76.19	149.00	100.00
	yes	45.00	13.31	45.00	23.81	0.00	0.00

Note: #: number of participants. %: percentage of participants. Full sample: 400 patients. ckd: 250 patients.

sg: specific gravity. al: albumin. su: sugar. rbc: red blood cells. pc: pus cell. pcc: pus cell clumps. ba: bacteria. htn: hypertension. dm: diabetes mellitus. cad: coronary artery disease. appet: appetite. pe: pedal edema. ane: anemia.

The following is a brief description of some features: Specific gravity indicates the particle concentration in urine and the density of such particles relative to water. The result reveals a patient's hydration status and kidney function. Albumin is one of the most critical proteins in blood, which enters the urine when the kidneys are damaged. An increase in urine albumin levels could indicate CKD. Additionally, blood urea is a vital indicator of kidney function. The total amount of nitrogen in a patient's blood circulation was tested using blood urea nitrogen, and a high level of blood urea nitrogen indicates an abnormal kidney function. While the amount of sugar circulating in the blood was estimated using the test for random blood glucose, over 200 mg/dL helps identify diabetes. Muscles produce serum creatinine as a waste of products. High creatinine levels indicate that the kidney is not adequately filtering the blood's waste, as measured by a creatinine test.

Machine learning requires preprocessing of data. As a result, data were coded according to nominal or categorical values. In particular, the 'normal' features were scaled to 1 while the 'abnormal' features were 0. The 'present' features were coded 1, whereas the 'not present' features were 0. Also, the 'yes' were transformed into 1 while the 'no' features were into 0. Moreover, the feature with ‘good’ was scaled to 1, whereas the feature with ‘poor’ was scaled to 0.

Several missing values are also present in the dataset. A model's performance could be negatively affected if missing values are ignored or deleted. This research uses the mean imputation method to fill in the missing data. Mean imputation calculates the average of the observed values and fills in any missing values. A min–max scaling technique was employed on all other attributes except 'age' and binary features.

Based on the original data curators' representation, 400 rows (patients) and 25 columns (features) were included in a table from the dataset. The study reported quantitative characteristics of categorial and numerical features of the complete dataset in Table 1 and Table 2.

Table 2. Quantitative description of the numerical characteristics.

Numeric feature	Full sample			Ckd patients			Notckd patients		
	Mean	Median	σ	Mean	Median	σ	Mean	Median	σ
age	50.96	54.00	16.91	54.46	59.00	17.09	46.52	46.00	15.63

Table 2. (continued).

bp	75.98	80.00	13.92	79.74	80.00	16.06	71.21	70.00	8.54
bgr	149.31	121.50	79.15	182.28	157.00	92.09	107.48	107.00	18.48
bu	55.46	40.00	49.84	73.23	53.00	60.26	32.92	33.00	11.33
sc	2.65	1.20	3.77	4.05	2.20	4.57	0.87	0.90	0.26
sod	137.46	138.00	7.62	134.19	136.00	7.84	141.62	141.00	4.83
pot	4.56	4.30	3.08	4.74	4.20	4.08	4.33	4.50	0.59
hemo	12.70	13.00	2.85	10.76	11.00	2.17	15.15	15.00	1.27
pc	38.89	40.00	8.87	33.06	33.00	7.10	46.28	46.00	4.08
wc	8515.09	8050.00	3080.56	9173.54	9100.00	3641.57	7679.87	7400.00	1872.85
rc	4.58	4.60	1.02	3.95	3.90	0.83	5.38	5.30	0.61

Note: Full sample: 400 individuals. ckd: 250 individuals. notckd: 150 individuals. σ: standard deviation

age: age in years. bp: blood pressure in mm/Hg. bgr: blood glucose random in mgs/dl. bu: blood urea in mgs/dl. sc: serum creatinine in mgs/dl. sod: sodium in mEq/L. pot: potassium in mEq/L. hemo: hemoglobin in gms. wc: white blood cell count in cells/cumm. rc: red blood cell count in millions/cmm.

3.1.2. CKD prediction on all features of the complete dataset. In order to predict the development of CKD, the author employed several methods. 100 executions were applied for each method, and the score of the mean result was reported (Table 3). The author randomly selected 60% of the dataset to train each classifier, 20% to validate it, and 20% to test it for methods that needed hyperparameter optimization. Based on a grid search, the author selected the hyper-parameters that generated the highest MCC. For the other methods, the dataset was divided into 80% to train the model and 20% to test it.

From the original dataset, 100 random data instances were selected for training, testing, and validation (in the case of hyperparameter optimization). Models were trained on training sets (and validated on validation sets for hyper-parameter optimization). A test set was then applied to the script by the author. Each dataset split produced slightly different results because the data instances were selected differently.

Table 3. Scores of CKD prediction results on all characteristics – mean of 100 executions.

Method	MCC	Accuracy	F1 score	TP rate	TN rate	ROC AUC
RFC	0.985*	0.993*	0.994*	1.000*	0.983*	0.994*
XGB	0.971	0.985	0.987	1.000*	0.967	0.987
lightGBM	0.956	0.978	0.980	1.000*	0.951	0.981
SVC	0.955	0.978	0.981	0.987	0.966	0.979
DTC	0.928*	0.963	0.967	1.000*	0.921	0.968
ADC	0.858	0.926	0.932	0.986	0.864	0.934
LR	0.805	0.904	0.916	0.922	0.881	0.903

Among the prediction methods, Random Forests performed the best, obtaining the best MCC (0.985), the best accuracy (0.993), the best F1 score (0.994), the best TP rate (1.000), the best TN rate (0.983) and the best ROC AUC (0.994).

Since the dataset consisted of 62.5% positive elements and 37.5% negative elements, all the selected methods achieved higher prediction scores on the true positive rate than on the true negative rate (Table

4). As a result, algorithms are more able to recognize the profiles of patients without CKD during testing since they can see more positive elements during training.

3.1.3. Feature ranking results. The accuracy reduction identified serum creatinine, red blood cell count, specific gravity, and potassium as the top four risk factors of importance in the dataset based on Random Forests feature ranking (Figure 1).

After ranking the features based on their importance, the author sought to find features of minimum numbers (and which features should be contained) to be still able to predict CKD progression accurately. Medical doctors in the hospital can use this method when they can access just a few features of a patient's electronic health record.

After observing that serum creatinine, red blood cell count, specific gravity and potassium were the four most important features in the Random Forests analysis, the author decided to examine whether these four clinical features alone could be used to predict patients' survival.

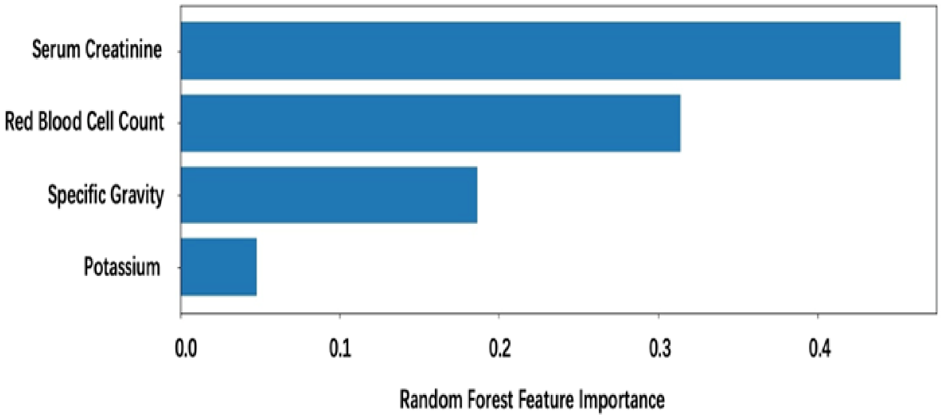

Figure 1. Feature selection using random forest.

3.1.4. Survival machine learning prediction. The goal was to determine whether ML methods are able to accurately predict CKD development by utilizing only the four most significant features. As a result, the author elaborated a second computational pipeline based on feature importance, then used different ML classifiers for prediction according to the top four characteristics (Table 4).

Table 4. Scores of CKD prediction results on top four characteristics – mean of 100 executions.

Method	MCC	Accuracy	F1 score	TP rate	TN rate	ROC AUC
lightGBM	0.985*	0.993*	0.994*	1.000*	0.983*	0.994*
ADC	0.985	0.993	0.994	1.000*	0.983	0.994
RFC	0.971	0.985	0.987	1.000*	0.967	0.987
XGB	0.942	0.971	0.974	1.000*	0.935	0.974
DTC	0.888	0.941	0.946	1.000*	0.879	0.949
SVC	0.871	0.934	0.940	0.986	0.877	0.940
LR	0.871	0.934	0.940	0.986	0.877	0.940

Light Gradient Boosting Machine performed the best among the prediction methods with the best MCC (0.985), accuracy (0.993), F1 score (0.994), TP rate (1.000), TN rate (0.983) and ROC AUC (0.994). The results are highly similar to those of completed data analyzed by Random Forests. Since the dataset was imbalanced, the classifiers obtained better recall value of true positive rate than specificity value of true negative rate in this application.

3.2. Discussion

The results show that it might be possible to predict CKD development only from their serum creatinine, red blood cell count, specific gravity and potassium and that predictions made on these four features alone can achieve similar accuracy when predicting CKD development utilizing the entire dataset. Especially encouraging is from the hospital's perspective: even if a patient's electronic health record does not contain a number of results from laboratory tests or clinical features, doctors can still estimate patient disease status by only analyzing the ejection fraction, specific gravity, and potassium levels of the patient. However, the author acknowledges that further confirmation studies are needed before applying this machine-learning procedure to actual clinical practice.

4. Conclusion

Researchers found that machine learning methods could be used accurately and effectively to classify patients with or without chronic kidney disease according to their electronic health records. The small dataset (400 patients) of the present study is a limitation: a more extensive dataset would have permitted more reliable results to be obtained. A comprehensive assessment of the patient's physical characteristics, disease history and socioeconomic status would have helped identify additional risk factors. A validation cohort from a different geographical area would have been used if other external datasets with the same exposure factors were available. Future research plans will focus on alternative datasets of CKD and other diseases (colorectal cancer, lymphoma, breast cancer, and ovarian cancer).

References

[1] Bikbov B, et al. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet, 2020, 395: 709-733.

[2] Bhaskar N, et al. Time Series Classification-Based Correlational Neural Network With Bidirectional LSTM for Automated Detection of Kidney Disease. IEEE Sensors Journal, 2021, 21(4): 4811-4818.

[3] Ali S I, et al. Ensemble Feature Ranking for Cost-Based Non-Overlapping Groups: A Case Study of Chronic Kidney Disease Diagnosis in Developing Countries. IEEE Access, 2020.

[4] Ebiaredoh-Mienye S A, et al. A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease. Bioengineering, 2022, 9(8): 350.

[5] Qin J, et al. A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. IEEE Access, 2020, 8: 20991-21002.

[6] Ebiaredoh-Mienye S A, et al. Integrating Enhanced Sparse Autoencoder-Based Artificial Neural Network Technique and Softmax Regression for Medical Diagnosis. Electronics, 2020, 9(11): 1963.

[7] Chittora P, et al. Prediction of Chronic Kidney Disease - A Machine Learning Perspective. IEEE Access, 2021, 9: 17312-17334.

[8] Silveira, A, et al. Exploring Early Prediction of Chronic Kidney Disease Using Machine Learning Algorithms for Small and Imbalanced Datasets. Applied Sciences, 2022, 12(7): 3673.

[9] Sharma S, et al. Performance based evaluation of various machine learning classification techniques for chronic kidney disease diagnosis. arXiv preprint arXiv:1606.09581, 2016.

[10] Motwani A, et al. Novel Machine Learning Model with Wrapper-Based Dimensionality Reduction for Predicting Chronic Kidney Disease Risk. Singapore: Springer Singapore, 2021.

[11] Aruleba K, et al. Applications of Computational Methods in Biomedical Breast Cancer Imaging Diagnostics: A Review. Journal of Imaging, 2020, 6(10): 105.