Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Prediction of Chronic Kidney Disease Based on Comparison of Machine Learning Models

Pai Li*

School of Applied Science and Civil Engineering, Beijing Institute of Technology Zhuhai, Zhuhai, Guangdong, China

*Corresponding author: lpai040910@gmail.com

Abstract. Chronic kidney disease (CKD) has become an important issue affecting global public health security due to its complex and variable pathogenic factors. Machine learning models with good predictive performance for CKD can break the limitations of traditional diagnostic methods and effectively control the burden of CKD on patients and the harm to human health security. In this paper, eXtreme Gradient Boost (XGBoost), logistic regression, and Support Vector Machine (SVM) were adopted for the prediction training of CKD dataset. This research comprehensively evaluated the model performance through accuracy rate, precision, recall rate, F1 value, AUC value, ROC curve, and scatter plot based on the T-SNE algorithm. Finally, the research concluded that XGBoost had the best performance. Subsequently, the research statistically analyzed features that were more important for the research through the plot_importance function and plotted a horizontal bar chart of the top 10 features in terms of importance. This research can help improve the efficiency of relevant practitioners and researchers in diagnosing CKD, contribute to reducing the burden on patients and enhancing the control of the incidence of CKD, and make contributions to public health issues.

Keywords: Chronic kidney disease; machine learning; disease prediction model.

1. Introduction

Over the past 10 years, chronic Kidney disease (CKD) has become a serious problem that endangers worldwide public health [1]. Its incidence rate has been increasing year by year, having a significant impact on human health. According to research, the prevalence rate of CKD in China is approximately 10% [2]. Due to its long incubation period, complex and uncertain pathogenic factors, long treatment cycle and difficulty in completely cured, CKD imposes a great burden on patients [3]. Therefore, timely detection in the early stage of CKD and early intervention in medical treatment can effectively reduce the harm of CKD. However, due to the lack of good predictive diagnostic methods, most patients have already entered stage 5 of CKD when they were diagnosed with CKD [4]. The kidney function of these patients is usually badly impaired, leading to a significant impact on the quality of their lives [5]. Therefore, a good CKD prediction based on machine learning can help medical workers diagnose diseases as early as possible, making timely interventions, thereby improving the medical quality and reducing the burden on patients [6].

Unlike the limitations of traditional research methods [7], machine learning algorithms have better prediction and stability through processing and analysis based on big data [8]. This research was conducted based on the CKD patient case dataset shared by Rabie EI Kharoua. First, the research conducted data cleaning and preprocessing on the dataset to improve the data quality for subsequent research. Then, the research carried out training respectively based on eXtreme Gradient Boost (XGBoost), logistic regression and Support Vector Machine (SVM). The research compared the performance of the three models for this dataset through accuracy rate, precision, recall rate, F1 value, AUC value, receiver operating characteristic curve (ROC) and the scatter plot based on the T-SNE algorithm.

Based on the research, medical workers and related researchers can screen out more suitable machine learning prediction models in the prediction tasks of CKD, understand the features that are more

important for the prediction. This research is helpful to improve the medical diagnosis efficiency of CKD, assist the early prevention of CKD, and contribute to the field of public health.

2. Introduction to the dataset and research method

2.1. Introduction to the Dataset

This research based on the dataset which was shared by Rabie EI Kharoua. This dataset contains detailed information on 1659 patients diagnosed with CKD. This dataset contains 54 features, representing various indicators of each patient, it contains a total of 89586 pieces of data, providing rich and multi-dimensional information for a comprehensive exploration of the related factors of CKD. Its contents mainly include: the basic information of patients such as age, gender, ethnicity, etc., all of which may have an impact on the risk of CKD; The patients' lifestyle-related factors, such as smoking, alcohol consumption, diet quality, etc., these information are important for kidney health; The disease-related medical history records of patients, such as family history of kidney diseases, family history of hypertension, and family history of diabetes, these characteristics related to genetic factors are helpful for identifying the genetic susceptibility of CKD; The physiological indicators of patients, such as systolicBP and diastolicBP, fasting blood sugar, GFR and other indicators, can directly or indirectly reflect the renal function and the metabolic state of the body. In addition, the dataset also includes features such as heavy metals exposure, water quality, and medical checkup frequency. These features are helpful for comprehensively reflecting the potential impact of external environmental factors and individual health management behaviors on CKD. To sum up, this dataset, with its rich and comprehensive feature information, provides a solid data foundation for the CKD prediction model.

2.2. Model Introduction

This research built a model based on Python and divided the dataset into 70% of the training set and 30% of the test set for training and testing the model performance. The research selected 3 representative models: XGBoost, logistic regression and SVM. These 3 models have their own unique algorithmic characteristics. In this paper, the performance of 3 models was evaluated by the standards of accuracy rate, precision, recall, and F1 value. The most suitable model for the prediction of this dataset was obtained through comparison, thereby further improving the accuracy and reliability of the prediction. The following is the introduction to each model used in the research respectively.

XGBoost, also known as the Extreme Gradient Boosting algorithm, is based on the idea of the gradient boosting algorithm. It gradually improves the model's performance by continuously adding weak learners such as decision trees. During the iterative process, a new weak learner is trained each time to correct the model's previous prediction deviations, thereby continuously approaching the true values of the model's predicted values. Compared with the traditional gradient boosting algorithm, both the training speed and accuracy have been greatly improved, and it has outstanding performance in handling classification and regression problems [9].

Logistic regression, also known as logistic regression analysis, is a generalized linear regression analysis model. Based on the principle of linear regression, it transforms the output of linear regression into a probability value between 0 and 1 through logarithmic transformation, thereby achieving the prediction of binary classification problems.

SVM is a well-performing binary classifier, and its core lies in the distinction between two types of problems. The essential idea is to find an optimal hyperplane in the feature space that can separate data points of different categories to the greatest extent, thereby achieving the purpose of classification [10]. The optimal hyperplane helps classification to have better generalization ability and effectively improves the classification.

3. Research's result and analysis

3.1. Data preprocessing

3.1.1. Dataset Import and preliminary preprocessing

In this research, the read_csv function in the pandas is used to read CKD dataset mentioned above for subsequent preprocessing. The content of preprocessing includes viewing the basic information of the dataset such as column names, data types, etc., so as to have a better understanding of the overall dataset. Checking the missing values and duplicate rows in the dataset to ensure the quality of subsequent research, dividing the dataset into feature matrices and label vectors, and encoding the labels to ensure they can be correctly processed by the model.

3.1.2. Dataset Partitioning and Feature Engineering

In this research, the dataset was divided into 70% of the training set and 30% of the test set, and random seeds were set to ensure the reproduction of the results. In feature processing, for classification features, OneHotEncoder was adopted for unique encoding to convert them into numerical form. For numerical features, StandardScaler was used for standard language processing, each feature is scaled with a mean of 0 and standard deviation of 1, thereby preventing certain features from affecting the model training results due to excessive range. The ColumnTransformer was used to uniformly manage and transform different types of features to ensure data consistency.

3.2. Model Training and Evaluation

As mentioned above, the 3 different models, XGBoost, logistic regression and SVM, were respectively selected for training, and then the performance of each model in predicting CKD was compared. The research conducted hyperparameter tuning for each model through GridsearchCV and cross-validation, thereby finding the optimal combination of model parameters. During the model training process, the research conducts performance training on the model through the training set, and then uses the test set to evaluate the model's performance. The evaluation parameters, as mentioned above, were accuracy rate, precision, recall, F1 value and AUC value. In addition, this research helps to evaluate and compare the performance of different models by drawing the ROC curve.

3.3. Data Visualization Based on T-SNE

The CKD dataset contains 54 features and belongs to high-dimensional data, it's difficult to intuitively understand the relationships between each data point and the distribution of the whole data. Therefore, in this research, the T-SNE algorithm was used to perform dimensionality reduction processing on the dataset, mapping the high-order data on a two-dimensional plane, and drawing scatter plots using different colors based on different label values. To help researchers better display the distribution of the data, visual images can provide guidance for subsequent research and improvement.

Table 1. Performance comparison of the models

Model	Accuracy/%	Precision/%	Recall/%	F1 Score	AUC
XGBoost	94.58	95.45	98.93	97.16	0.84
Logistic Regression	93.98	94.32	99.57	96.88	0.83
SVM	93.78	93.78	100.00	96.79	0.80

As can be seen from Table 1, in the prediction performance on the CKD dataset, the accuracy rate of XGBoost reached 94.58%, leading among the three models. The accuracy rate refers to the proportion of samples correctly predicted by the model among the total samples. A higher accuracy rate means that the model can better predict whether a patient is a potential CKD patient. By comparing the precision of each model, the precision of XGBoost is 95.45%, which is also higher than that of logistic

regression and SVM. Precision reflects the proportion of actually diseased samples among the predicted diseased samples of the model. High precision means the possibility of misdiagnosis of the model is smaller, that is, determining non-CKD patients as patients. This means that the XGBoost model can more precisely determine diseased samples. Recall rate refers to the proportion of samples that are actually diseased and correctly predicted by the model. A higher recall rate means that the model correctly identifies a large number of real diseased samples. As can be seen from Table 1, all 3 models achieved extremely high recall rates. F1 value considers both precision and recall rate. Among them, the F1 value of XGBoost reached 97.16, indicating that it performed best in balancing precision and recall. The AUC value is measured by the model's ability to distinguish between positive and negative samples. Its value range is from 0 to 1. The closer it is to 1, the stronger the model's discrimination performance. Among them, XGBoost led the other models with an AUC value of 0.84. The ROC curves of 3 models were also plotted in this experiment, as shown in Figure 1.

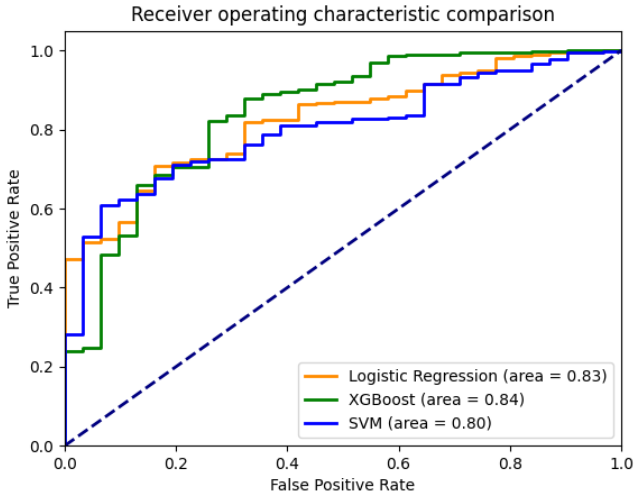

Fig. 1 ROC curves of 3 models (Picture credit: Original)

As can be seen from Figure 1, the prediction performance of XGBoost model is better than other models.

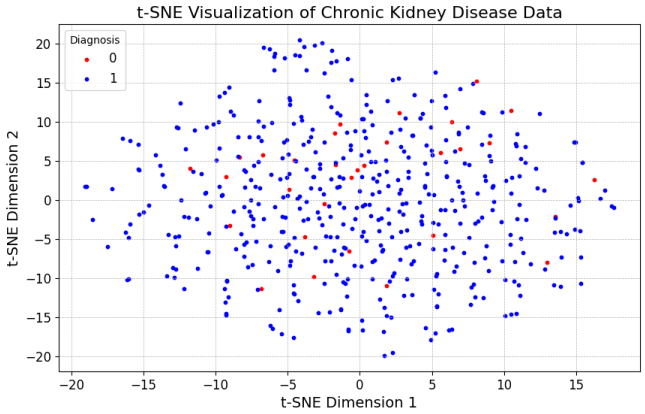

Fig. 2 Data visualization based on T-SNE algorithm (Picture credit: Original)

Figure 2 shows the T-SNE visualization of the CKD prediction, its horizontal and vertical coordinates represent the 2 dimensions after the T-SEN algorithm reduces the high-dimensional data in the CKD dataset to a two-dimensional space. Among them. 0 represents patients without CKD, and 1 represents patients diagnosed with CKD, which is helpful for observing the distribution of different types of data in the two-dimensional space. By observing Figure 4, it's found that the distribution of confirmed CKD data points predicted by the model is relatively reasonable and has a high matching degree with the real label, further confirming that XGBoost has good performance for this model.

Based on the above research and comparison of the performance of the 3 models, it can be found that XGBoost performs the best and is the most suitable for the prediction task of the CKD dataset.

3.4. Sorting of Important Features

This research used the plot_importance function in XGBoost model. The plot_importance function can measure the importance of features through the number of times a special node is split in the weight statistical model. The research listed the top 10 important characteristics, as shown in Figure3, which can assist researchers in making better decisions for patient treatment and thereby improve the quality of medical care.

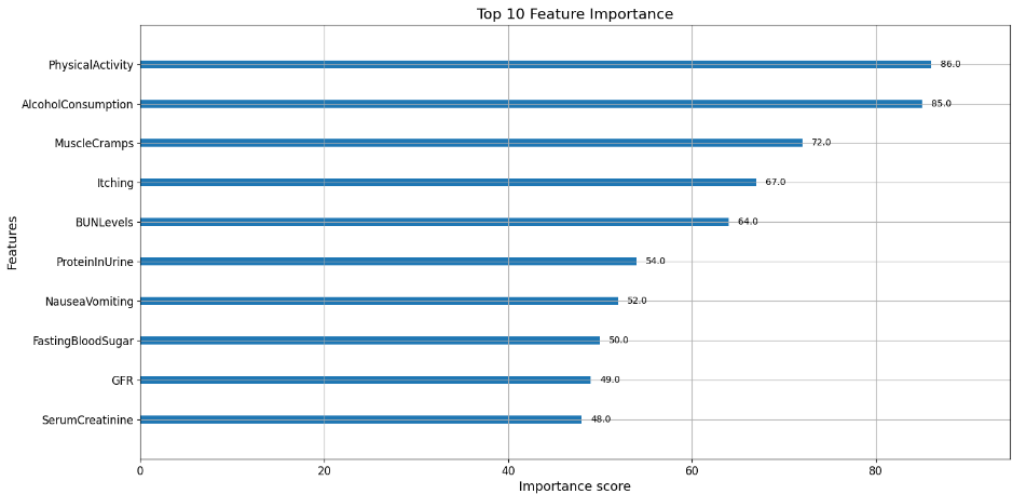

Fig. 3 Ranking of top 10 important features (Picture credit: Original)

PhysicalActivity, AlcoholConsumption, MuscleCramps, Itchin, BUNLevels, ProteinInUrine, NauseaVomiting, FastingBloodSugar, GFR, SerumCreatinine is more important than other features. Therefore, medical personnel can give priority to these features when diagnosing whether a patient has CKD.

4. Conclusion

Due to the long latency and complex pathogenic factors of CKD, most patients are often close to or have already entered stage 5 of CKD, which is the terminal stage of CKD, when they were diagnosed with CKD. At this stage, the kidney function has been severely damaged. Usually, patients with end-stage CKD suffer a lot of pain and it is difficult to be completely cured. Therefore, early determination of whether a patient is at risk of CKD, early screening and diagnosis, and early subsequent treatment based on the results are of great significance for controlling the condition of CKD and can make contributions to global public health.

In this research, machine learning training was conducted on CKD patient dataset based on 3 models: XGBoost, logistic regression, and SVM respectively. The models were evaluated through accuracy, precision, recall, F1 value and AUC value. XGBoost performed well in the evaluation of various indicators. Among them, the accuracy was 94.58%， the precision was 95.45%, the recall rate was 98.93%, the F1 value was 97.16, and the AUC value was 0.84. Meanwhile, combined with the visualization results of the T-SNE algorithm, the predicted values of the XGBoost model were relatively reasonably distributed. It had a high matching degree with the real labels. Combined with the above content, this research concludes that XGBoost performs best in predicting whether patients are likely to have CKD. The plot_importance function in the XGBoost model was used to rank the importance of features in the dataset, listing the top 10 important features in terms of their influence on predicting CKD.

There are still deficiencies in this research. Firstly, due to the complex pathogenic factors of CKD, this dataset may still have limitations. The dataset adopted in this research has 54 features, but there may be potential influencing factors for predicting CKD that have been ignored. Secondly, the model used in the research performed well for this dataset, but its performance prediction for datasets still needs to be tested. Finally, the models tested in the research are limited, and it is possible that some models with better performance in predicting CKD have been ignored.

Considering the deficiencies in the research, the following directions will be continuously explored in future studies. Firstly, select multiple datasets with larger samples, greater representativeness, and more comprehensive disease cases for testing, so as to ensure that good predictive performance can be achieved for CKD disease prediction in different populations. Secondly, more models such as neural networks, random forests, and comparative experiments are conducted to find the model that is more suitable for the prediction task. Finally, attempting to integrate the prediction results of different models, combining the advantages of different models, thereby, further enhancing the predictive ability of the models for different samples, such as different races of people and different regions of people.

In conclusion, this research has helped to screen out the XGBoost model that is more suitable for CKD prediction among common types of models, which is beneficial for medical personnel to better diagnose patients, such as predicting whether a patient has CKD through the model. Meanwhile, researchers can also focus on relevant patients based on characteristics such as PhysicalActivity and AlcoholConsumption, which is helpful to improve the medical quality and the cure rate of CKD.

References

[1] Yang C., Gao B. X., Zhao X. J., et al. “Executive Summary for China Kidney Disease Network (CK-NET) 2016 Annual Data Report.” Kidney International, Wiley, 2020.

[2] Gao X., Mei C. L. “Interpretation of ‘Guidelines for Early Screening, Diagnosis and Prevention of Chronic Kidney Disease (2022 Edition).’” Chinese Journal of Practical Internal Medicine, 2022.

[3] Huang X., He S., Xi H. H., et al. “Multi-Disease Risk Prediction Based on XGBoost and Keras Frameworks.” Intelligent Computing and Applications, 2020.

[4] Belokurova E. V., Alekseeva T. V., Mishina E. N., et al. “Prospects for Applying Ultrasonic Processing of Semi-Finished Products in Baking Technology.” IOP Conference Series: Earth and Environmental Science, vol. 640, no. 2, IOP Publishing, 2021, Article 022061.

[5] Cheng Y., Chen Y. J. “Application of Machine-Learning-Based Ultrasound Imaging and SWE Prediction Models in Early Chronic Kidney Disease.” Imaging Science and Photochemistry, 2024.

[6] Wang R. Q. “Study and Application of Machine-Learning–Based Auxiliary Diagnostic Algorithms for Chronic Kidney Disease.” PhD diss., Zhengzhou University, 2021.

[7] Jiang Y. P., Yu C., Lin Y. R., et al. “Early Screening Methods for Chronic Kidney Disease Based on Ensemble Learning Algorithms.” Journal of Southwest University (Natural Science Edition), 2020.

[8] Hasan Z. K. M., Hasan Z. M. “Performance Evaluation of Ensemble-Based Machine Learning Techniques for Prediction of Chronic Kidney Disease.” In: Shetty N., Patnaik L., Nagaraj H., et al., eds. Emerging Research in Computing, Information, Communication and Applications. Singapore: Springer, 2019, pp. 415–426.

[9] Chen T., Guestrin C. “XGBoost: A Scalable Tree Boosting System.” CoRR, vol. abs/1603.02754, 2016.

[10] Long Y. H. “Key Technologies for Speaker Verification Based on SVM.” PhD diss., University of Science and Technology of China, 2011.