

*Article*

**Chronic Kidney Disease Prediction Using Machine Learning Techniques (XGBoost Classifier)**

# Marina Reda Abdullah , Aly Maher Abdel Fattah, Maram Ashraf Elbanna, Mohammed Ahmed Fathi, Hamza Nashaat Abdelbaki and Omar Adly Mahmoud

# Abstract

Chronic kidney disease (CKD) is a growing health concern worldwide, affecting millions of people. Early detection of CKD can help prevent its progression and improve patient outcomes. Machine learning techniques have shown promise in predicting the onset of CKD. In this study, we aimed to improve the performance of classification algorithms in diagnosing CKD by using XGBoost feature selection. To achieve this goal, We utilized data from the UCI machine learning database which contain 400 sample and apply XGBoost classifier on it in the feature selection method. Our results showed that the XGBoost classifier achieved an accuracy of 98.5% after training on 12 features. Our findings suggest that machine learning techniques can aid in the early detection of CKD and may prove to be a valuable tool for healthcare providers in identifying patients at high risk of developing this disease.

**Keywords:** chronic kidney disease; data balancing; machine learning; supervised learning; Feature Selection;

Machine Learning; XGBoost Classifier; Missing Data

# 1 Introduction

Studies have shown that COVID-19 is more severe in Chronic Kidney Disease (CKD) patients, with a mortality rate of 44.6% among those who tested positive. Kidney function in CKD patients is measured using blood and urine tests, where a glomerular filtration rate (GFR) of 60 is considered normal function, and values between 15 and 60 indicate substandard function. Early diagnosis is crucial for reducing mortality rates, as late diagnosis can lead to renal failure, dialysis, or transplantation. However, with a shortage of specialized physicians and high diagnosis and treatment costs, machine learning algorithms, such as support vector machine (SVM), XGBoost classifier, K-nearest neighbor (KNN) and Random Forest (RF), can aid in early CKD diagnosis. However, the dataset

may contain missing values, which can impact the accuracy of machine learning models. Therefore, data reduction, via feature selection methods, is necessary to reduce dimensionality and enhance diagnostic performance. Previous studies have performed feature selection after missing data imputation, but this can result in biased prediction. Hence, in this study, we aim to perform feature selection without missing data imputation for CKD diagnosis. And to do this, we use the XGBoost classifier as a feature selection method, along with support vector machine and K-nearest neighbor classifiers for CKD prediction. The paper is divided into four sections: a review of existing studies in Section 2, dataset description and XGBoost feature selection method in Section 3, experimental design, results, and discussions in Section 4, and finally, conclusions and future works in Section 5.

# Literature Review

In the context of Chronic Kidney Disease (CKD) diagnosis, feature selection methods have been utilized in previous studies. However, CKD datasets often contain a high rate of missing data, which can complicate feature selection. One study by Elhoseny et al. [11] (2019) employed Density-based Feature Selection (DFS) and Ant Colony Optimization (D-ACO) algorithms to perform wrapper feature selection after imputing missing data. This approach led to improved CKD prediction performance using a smaller number of features. Another study by Ogunleye and Wang [12] (2019) combined recursive feature elimination, extra tree classifier, and univariate feature selection methods. They achieved 100% accuracy in CKD prediction using the XGBoost classifier, after imputing the dataset with median. Polat et al. [13] (2019) used both filter and wrapper feature selection methods and achieved 98.5% accuracy with an SVM classifier. They also imputed the missing data before performing feature selection. In a study by Almansour et al. [14] (2020), missing data was imputed with the mean of features, followed by recursive feature elimination to achieve 99.75% accuracy with an ANN. Manonmani and Sarojini Balakrishnan [15] (2020) used a teaching-learning-based algorithm for feature selection after imputing missing data with KNN. They obtained 95.25% accuracy using the CNN classifier. Tazin et al. [16] (2021) selected the top ten features based on the information gain score and achieved 99% accuracy with decision tree algorithms. However, they also performed feature selection after imputing missing data. XGBoost feature selection has been successfully applied in protein-protein interactions prediction, driving assessment, and risk prediction. However, to the best of our knowledge, XGBoost feature selection has not yet been applied in CKD diagnosis. Therefore, in this study, we aim to perform XGBoost feature selection in CKD diagnosis with classification algorithms. Our approach will consider incomplete datasets without imputing missing data, which could have important implications for CKD diagnosis in clinical practice.

1. **Materials and Methods**

## Dataset

The dataset used in this study was obtained from the UCI Machine Learning Repository, which is a reliable and commonly used source for machine learning datasets. The dataset includes 400 records and 25 features, including categorical and numerical variables, with class attributes indicating whether the patient has CKD or NOTCKD. Out of the 25 features, 14 were categorical and 11 were numerical. The dataset contained a significant number of missing values, with only 158 records having complete data. Additionally, there was a class imbalance in the dataset, with 62.5% of the observations indicating the presence of CKD and 37.5% indicating the absence of CKD. Table 1 provides descriptions and important information on the features, while Figure 1 shows a visualization of the missing data. Overall, this dataset presents challenges for machine learning model development due to its missing data and class imbalance.





**Figure 1.** Visualization of missing values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Si no. | Name of the attribute | Abbreviation of the attribute | Value type | Number of missing values |
| 1 | Age | age | in years | 9 |
| 2 | Blood pressure | bp | mm/Hg | 2 |
| 3 | Specific gravity | sg | Nominal | 47 |
| 4 | Albumin | al | Nominal (1,2,3,4,5,…) | 46 |
| 5 | Sugar | su | Nominal (1,2,3,4,5,…) | 49 |
| 6 | Red blood cells | rbc | Normal, abnormal | 152 |
| 7 | Pus cell | pc | Normal, abnormal | 65 |
| 8 | Pus cell clumps | pcc | Present, not present | 4 |
| 9 | Bacteria | ba | Present, not present | 4 |
| 10 | Blood glucose random | bgr | mgs/dl | 44 |
| 11 | Blood urea | bu | mgs/dl | 19 |
| 12 | Serum creatinine | sc | mgs/dl | 17 |
| 13 | Sodium | sod | mEq/L | 87 |
| 14 | potassium | pot | mEq/L | 88 |
| 15 | Hemoglobin | hemo | gms | 52 |
| 16 | Packed cell volume | pcv | numerical | 70 |
| 17 | White blood cell count | wc | celles/cumm | 105 |
| 18 | Red blood cell count | rc | millions/cmm | 130 |
| 19 | Hypertension | htn | yes, no | 2 |
| 20 | Diabetes mellitus | dm | yes, no | 2 |
| 21 | Coronary artery disease | cad | yes, no | 2 |
| 22 | Appetite | appet | good, poor | 1 |
| 23 | Pedal edema | pe | yes, no | 1 |
| 24 | Anemia | ane | yes, no | 1 |
| 25 | Class | class | CKD, NCKD | 0 |

**Table 1.** Attribute descriptions.

## Feature Selection with XGBoost classifier

There are three main categories of feature selection methods: filter, embedded, and wrapper methods. Filter methods evaluate each feature based on a statistical measure and select the top n features. Examples of filter methods include information gain, Pearson correlation, Fisher score, and chi-square. Wrapper methods, on the other hand, use the performance of a machine learning model to assess the feature subset. This involves iterating through different feature subsets and selecting the one that produces the best model performance. Common wrapper methods include forward feature elimination, backward feature elimination, and metaheuristic methods such as particle swarm optimization algorithm, genetic algorithm, and differential evolution algorithm. Embedded methods, on the other hand, incorporate feature selection as part of the machine learning model training process. Decision tree-based algorithms such as Random Forest, CART, C4.5, and more recently, XGBoost can be classified as embedded methods. These algorithms are capable of evaluating the importance of each feature during the model training process, making them a popular choice for feature selection in machine learning.

The XGBoost algorithm was introduced by Chen and Guestrin in 2016 and has gained widespread attention for its effectiveness in machine learning competitions. It is a type of gradient boosting decision tree ensemble machine learning algorithm that creates a new decision tree at each iteration to enhance the performance of the current tree. To evaluate the performance of the decision tree, XGBoost uses an objective function that is calculated by a loss function l that measures the accuracy of the predicted values (y^i) as compared to the actual values (yi). The penalty term for the complexity of the decision tree structure is represented by Ω.

𝑂𝑏j𝑒𝑐𝑡i𝑣𝑒 = ∑ 𝑙(𝑦^i, 𝑦i) + ∑ Ω(ƒ𝑘) (1)

The XGBoost classifier calculates the feature importance score using the gain score during the training process. The gain score for each split in the decision tree can be calculated using Equation 2, where IR and IL represent the right and left nodes, respectively. 𝑔i is the first-order gradient on the loss function, and ℎi is the second-order gradient on the loss function. The regularization parameters 𝛾 and 𝜆 are also included in the equation.



(2)

The feature importance score in XGBoost is determined by the average gain of all trees. A feature that is more significant and useful for the XGBoost classifier model will have a higher feature importance score. To select the most important features, we identified the top n percent of features with the highest feature importance score using XGBoost.

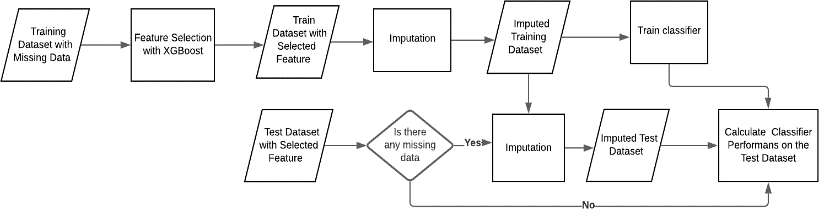
## Classifiers

We utilized three classifiers, namely Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and XGBoost, to verify the effectiveness of XGBoost feature selection on the CKD dataset. SVM is a well-established and robust classification algorithm that was developed by Cortes and Vapnik. Its primary objective is to segregate classes using a hyperplane by maximizing the margin between support vectors. If the dataset is not linearly separable, kernel functions are utilized to transform it into a linearly separable one. In contrast, KNN is a non-parametric, lazy learning-based classifier that predicts the class of new instances by utilizing the class information of the k-nearest neighbors in the training data. We have already presented a comprehensive explanation of the XGBoost classifier in the previous section.

# Experiments and Results

## Experimental design

To assess the accuracy of our model, we performed 10-fold cross-validation on the CKD dataset. We partitioned the dataset into ten subsets, out of which nine subsets were utilized for training, while the remaining subset was reserved for testing. This process was repeated ten times to ensure robustness. For each iteration, we followed the proposed CKD diagnosis process, as presented in Figure 2, and determined the average performance of the classifier. Based on the XGBoost feature importance score, we selected feature subsets with 20%, 30%, 40%, and 50% of the highest-scoring features. Missing values in the dataset were replaced with the mean imputation method, and the classifiers were trained for each feature subset.



**Figure 2.** Proposed CKD diagnosis process

## Evaluation metrics

We evaluated the performance of the proposed model using several metrics, including classification accuracy, sensitivity, specificity, and Kappa statistic. These metrics were calculated using 10-fold cross-validation and are represented by equations 3, 4, and 5, respectively.

𝑇𝑃 + 𝑇𝑁

(3)

𝐴𝑐𝑐𝑢𝑟𝑎𝑐𝑦 =

𝑇𝑃 + 𝑇𝑁 + 𝐹𝑃 + 𝐹𝑁

(4)

𝑆𝑒𝑛𝑠i𝑡i𝑣i𝑡𝑦 = 𝑇𝑃

𝑇𝑃 + 𝐹𝑁

𝑇𝑃

(5)

𝑆𝑝𝑒𝑐iƒi𝑐i𝑡𝑦 = 𝑇𝑃 + 𝐹𝑃

𝑂𝐴 − 𝐸𝐴

(6)

𝐾𝑎𝑝𝑝𝑎 𝑆𝑡𝑎𝑡i𝑠𝑡i𝑐 = 100 − 𝐸𝐴

The scores for accuracy, sensitivity, and specificity lie between 0 and 1, while the Kappa statistic ranges from -1 to 1. The TP, TN, FP, and FN values are illustrated in the confusion matrix displayed in Table 1. In equation 4, the observed agreement is represented by OA, and the expected agreement is represented by EA. The observed agreement is equal to the overall accuracy, represented as a percentage. The expected agreement is calculated as (%(TP+FP)\*%(TP+FN)) + (%(FN+TN)\*%(FP+TN)).

**Table 2.** Confusion Matrix

|  |  |  |
| --- | --- | --- |
| predection | CKD | Not CKD |
| CKD | **TP** | FP |
| Not CKD | FN | TN |

True Class

|  |  |  |
| --- | --- | --- |
| Predicted Class | True Positive (TP) | False Positive (FP)  True Negative (TN) |
| False Negative (FN) |

**Figure 3.** Confusion matrix and performance metrics formula.

## Parameter settings

In the feature selection stage, we utilized XGBoost's default parameters to compute the feature importance scores and selected the top 50% of these features. To optimize the classifiers' performance, we performed a grid search with a nested 10-fold cross-validation to tune the hyperparameters.

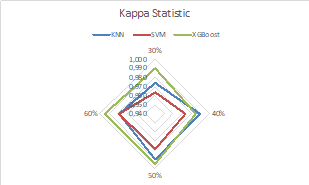
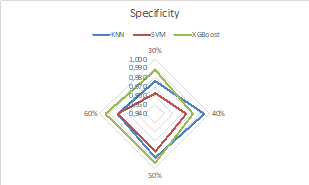
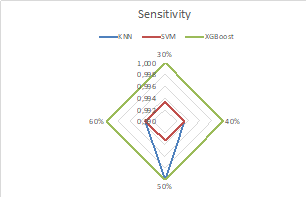
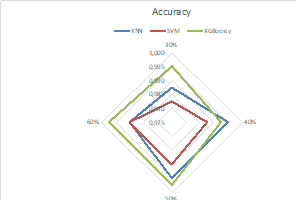
## Results

The performance of various classifiers and feature selection percentages using the XGBoost feature selection method is presented in Table 3. The results indicate that the XGBoost classifier trained on 50% and 60% of the selected features outperforms other classifiers in all performance metrics. The study's objective is to reduce the number of features, and thus the XGBoost classifier trained on the top 50% of the features is selected as the best CKD diagnosis model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Percentage of selected features | accuracy | sensitivity | Kappa statistics |
| KNN | 30% 40% 50% 60% | 0.9875 0.995 0.955 0.99 | 0.933 0.993 1.000 0.993 | 0.974 0.989 0.989 0.979 |
| SVM | 30% 40% 50% 60% | 0.9825 0.9875 0.99 0.99 | 0.962 0.975 0.981 0.981 | 0.963 0.974 0.979 0.979 |
| XGBoost | 30% 40% 50% 60% | 0.995 0.9925 0.9975 0.9975 | 0.988 0.982 0.994 0.994 | 0.989 0.984 0.995 0.995 |

**Table 3.** Experimental results of proposed model

The performance metrics of the proposed model according to the percentage of selected features are depicted in Figure 2. It is evident from the figure that the XGBoost classifier outperforms other classifiers, indicating its superiority.



**Figure 4.** Performance metrics of the proposed model according to the percentage of the selected feature

In Table 4, we have compared the results of our proposed model with the previous studies that used the same CKD dataset based on the number of selected features and accuracy. The proposed model in this study achieved higher accuracy with fewer features compared to the models presented in [16], [13], and [11]. Furthermore, our proposed model achieved the same accuracy as the model presented in [14] with the same number of selected features. We can also argue that our proposed model is competitive with the model presented in [12].

|  |  |  |
| --- | --- | --- |
| Model | Number of Feature | Accuracy |
| OgunleyeandWang[12] | 13 | 100% |
| Almansour et al. [14] | 12 | 99.75% |
| Tazin and et al. [16] | 15 | 99% |
| Polat et al. [13] | 13 | 98.5% |
| Elhoseny et al. [11] | 14 | 95% |
| Proposed Model | 12 | 99.75% |

**Table 4.** Comparison of the proposed model with the existing works

# Conclusion

This research employed XGBoost feature selection technique without performing imputation on missing data to diagnose CKD. Four feature subsets containing the top 20%, 30%, 40%, and 50% of features with the highest importance score were selected. support vector machine (SVM), XGBoost classifier, K-nearest neighbor (KNN) classifiers were trained using these feature subsets. The XGBoost classifier trained on the top 50% of features exhibited the highest accuracy of 98.5%. Our study highlights that XGBoost feature selection without missing data imputation can be a promising alternative to other feature selection methods that require imputing missing data. Further research can be carried out to investigate the effectiveness of XGBoost feature selection on larger CKD datasets and to determine if it can achieve similar performance.

# References

[1] Akgül G, Çelik A, Ergül Aydın Z, Kamışlı Öztürk Z., “Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı”. Bilişim Teknolojileri Dergisi, 13(3), 255-268, 2020.

[2] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. “Machine Learning and Data Mining Methods in Diabetes Research”. Computational and Structural Biotechnology Journal, 15,104-116, 2017.

[3] Moradi E,Pepe A, Gaser C, Huttunen H, Tohka J. “Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects”. NeuroImage, 104, 398-412, 2015.

[4] “The top 10 causes of death, World Health Organization”. https://www.who.int/news-room/fact- sheets/detail/the-top-10-causes-of-death (30.01.2020)

[5] Amirgaliyev Y, Shamiluulu S, Serek A. “Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods”. 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, 2018

[6] Tikariha P, Richhariya P. “Comparative Study of Chronic Kidney Disease Prediction Using Different Classification Techniques”. Proceedings of International Conference on Recent Advancement on Computer and Communication, Singapore, 2018.

[7] Wang W,Chakraborty G, Chakraborty B. “Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm”. Applied Sciences, 11, 202, 2021.

[8] Subasi A, Alickovic E, Kevric J. “Diagnosis of Chronic Kidney Disease by Using Random Forest”. CMBEBIH 2017, Singapore, 2017.

[9] Jain D, Singh V. “Feature selection and classification systems for chronic disease prediction: A review”. Egyptian Informatics Journal, 19(3), 179-189, 2018.

[10] Seijo-Pardo B, Alonso-Betanzos A, Bennett KP, Bolón-Canedo V, Josse J, Saeed M, Guyon I.” Biases in feature selection with missing data”. Neurocomputing, 342,97-112, 2019.

[11] Elhoseny M, Shankar K, Uthayakumar J. “Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease” Sci Rep, 9, 9583, 2019.

[12] Ogunleye A, Wang QG. “XGBoost Model for Chronic Kidney Disease Diagnosis” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol 17(6), 2131-2140, 2020.

[13] Polat H, Danaei Mehr H, Cetin A. “Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods” J Med Syst, 41,55, 2017.

[14] Almansour NA, Syed HF, Khayat NR, Altheeb RK, Juri RE, Alhiyafi J, Alrashed S, Olatunji SO. “Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study” Computers in Biology and Medicine, 109, 101-111, 2019.

[15] Manonmani M, Balakrishnan S. “Feature Selection Using Improved Teaching Learning Based Algorithm on Chronic Kidney Disease Dataset”. Procedia Computer Science, 171, 1660-1669, 2020.

[16] Tazin N, Sabab SA, Chowdhury MT. “Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique”. 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), Dhaka, 2016.

[17] Chen C, Zhang Q, Yu B, Yu Z, Lawrence PJ, Ma Q, Zhang Y. “Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier”. Computers in Biology and Medicine, 123, 103899, 2020.