

Optimizing Early Diagnosis of Chronic Kidney Disease: A Machine Learning-Based Predictive Model

Ezeji Nwamaka Georgenia, Kwubeghari Anthony

Department of Computer Engineering Enugu state University of Science and Technology (ESUT), Agbani Enugu.

DOI: <https://doi.org/10.51583/IJLTEMAS.2025.1406000121>

Received: 27 June 2025; Accepted: 03 July 2025; Published: 25 July 2025

Abstract: Chronic Kidney Disease (CKD) is a progressive condition that often remains undetected until its later stages, leading to severe health complications and increased mortality. Therefore, this study presents the implementation of a machine learning-based system for early CKD prediction using the logistic regression algorithm.

The study addresses the problem of delayed CKD diagnosis due to vague early symptoms and limited screening efficiency in traditional clinical workflows. The development of the system utilizes a clinical dataset from the UCI Machine Learning Repository made up of 400 patient records with 24 features, including demographic, clinical and laboratory parameters. Data preprocessing techniques were employed including label encoding, imputation of missing values and feature scaling for improving the quality of the data. Then, feature selection was conducted using the SelectKBest method with mutual information to identify the most relevant predictors. The logistic regression model was trained and evaluated using an 80:20 data split and the result of the implementation presents that the model achieved an accuracy of 97.5%, precision of 96.7%, recall of 98.2%, F1-score of 97.4% and a ROC-AUC score of 0.99. This work underscores the effectiveness of logistic regression in medical diagnostics and highlights the value of machine learning in facilitating early detection and timely treatment of CKD.

Keywords: Chronic Kidney Disease (CKD); Logistic Regression; Machine Learning; Medical Diagnosis; Feature Selection

I. Introduction

Chronic Kidney Disease (CKD), a major and growing worldwide public health problem, is characterised by a gradual deterioration in kidney function over time (Podkowiriska and Formanowicz, 2020). To generate urine and maintain a constant body temperature and pressure, the kidneys must filter out extra water and waste from the circulation (Wadei and Textor, 2012). The phrase "chronic kidney disease" refers to the progressive deterioration to the kidneys, which is often not recognised until it is too late (Romagnani et al., 2017). Because the body may adapt to diminished kidney function in the early stages, when symptoms are usually absent, this silent development of CKD highlights the need of early identification. The diagnosis is frequently found by accident in a blood or urine test that would have been normal otherwise, and many individuals do not receive therapy until their illness has significantly worsened (Kalantar-Zadeh et al., 2021).

The chance of developing chronic kidney disease (CKD) is greatly increased by risk factors such as diabetes, hypertension, and cardiovascular disorders (Hussain et al., 2021; Burnier and Damianaki, 2023). With an estimated 843.6 million people afflicted in 2017, the disease's prevalence has grown internationally, in part because of the growth in these risk factors (Jager et al., 2019). Even with advancements in therapy, chronic kidney disease (CKD) continues to be a major cause of death worldwide, necessitating close observation and prompt action to slow its progression and effects (Kovesdy, 2022).

To identify and track chronic kidney disease (CKD), medical professionals rely on crucial diagnostic techniques including the Glomerular Filtration Rate (GFR) and albumin urine tests (Vassalotti et al., 2016). The GFR shows how successfully the kidneys remove waste from the body based on patient data and blood test results (Inker and Titan, 2012). While a GFR below 15 indicates renal failure and the necessity for dialysis or a kidney transplant, a GFR below 60 may indicate kidney disease (Webster et al., 2011). Because a healthy kidney does not allow albumin to enter the urine, a urine test for this protein is a useful tool for diagnosing kidney illness (Martin, 2011). These tests are clearly useful, but they are not very good at forecasting how a disease will proceed or identifying the physiological mechanisms that underlie it (Levey & Levey, 2014).

According to Asif et al. (2023), machine learning has shown great promise in enhancing the diagnosis of a few illnesses. Specifically, the use of ML approaches has improved CKD analysis and diagnosis (Ho et al., 2024). Complex datasets may be combed through by ML models to find patterns and insights that would be hard to find using more traditional research methods (Greener et al., 2022; Ghazal et al., 2021). Nevertheless, a lot of machine learning models are infamously hard to comprehend, which is why some people refer to them as "black boxes." According to Hassija et al. (2024), a black box model produces results without disclosing or elucidating the internal decision-making procedures. These models are excellent at categorisation and prediction, but they are not transparent or interpretable, which can make it difficult to make therapeutic decisions and leave patients and physicians with unresolved issues regarding diagnosis and therapy (Rudin, 2019; Arif et al., 2024). Hence, this study presents the application of logistic regression algorithm for the early prediction of CKD using UCI ML dataset which had been pre-processed through encoding, imputation, scaling and feature selection.

II. Research Methodology

This section will give a brief overview of the steps involved in utilising a ML model to predict CKD. The procedure may be divided into six main steps, as shown in Figure 1, with data collection serving as the first step and data preparation as the second. Prior to training the machine learning model suggested for the system and testing it using various metrics, the following stage is feature selection, which entails identifying the qualities that will be most helpful for our model. Lastly, we use a performance assessment to verify the model's reliability and robustness.

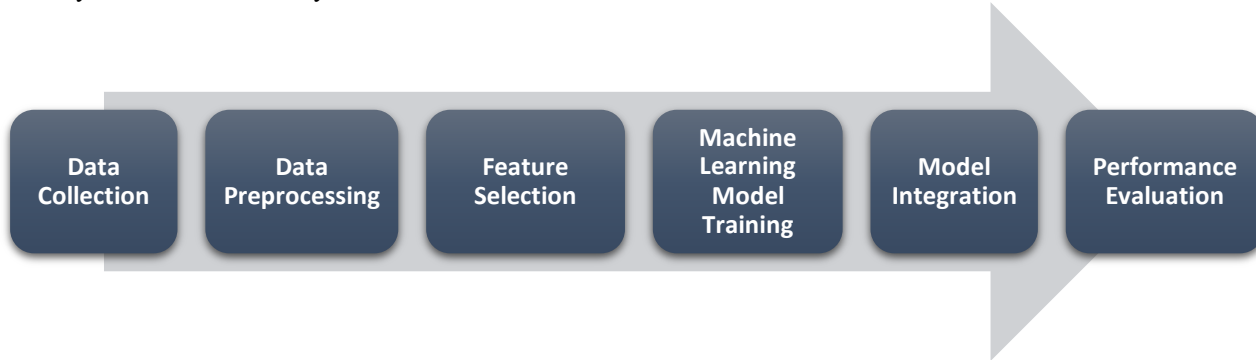


Figure 1: The Proposed Methodology Workflow

Data Collection

The UCI ML Repository provided the dataset utilised in this study (Rubini et al., 2015). It includes a variety of data formats with real-valued characteristics, mostly for classification tasks. The dataset has 25 characteristics, 400 cases, including clinical laboratory and demographic information. Each sample contains 24 predictive variables, which are distributed as follows: Thirteen nominal or category variables and eleven numerical variables. CKD and non-CKD are the two binary target classes.

Data Pre-processing

Data pre-processing is the process of cleaning and preparing raw data so that it can be efficiently and accurately used by a machine learning model. Numerous problems with medical datasets can compromise modeling's precision and effectiveness (Bellazzi and Zupan, 2008). Preprocessing is therefore required to enhance the quality of the data (García et al., 2016).

Figure 2 illustrates the three primary stages that comprise the preprocessing stage. These include data scaling, data encoding, and data imputation for missing values.

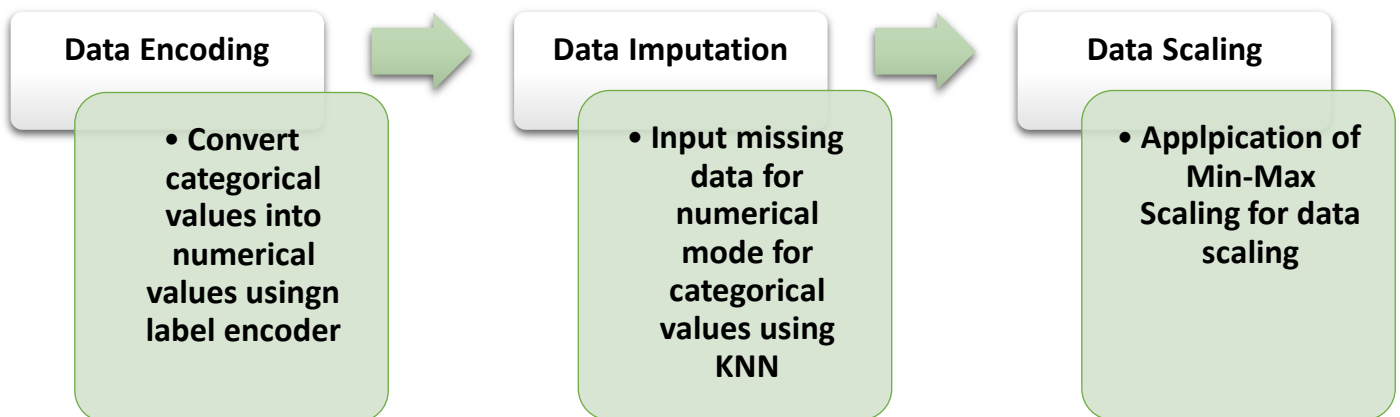


Figure 2: Data Pre-processing phase and steps

Data Encoding

We tackle this by using a label encoder because our dataset includes both numerical and category variables. To incorporate categorical data into the model, the label encoder converts them into numerical values because machine learning models only understands numbers not text. In label encoding, “yes” is encoded as “1” while “No” is encoded as “0”.

Data Imputation

Data imputation is the process of replacing missing data with substituted values so that the dataset can be used to train a machine learning model. Dealing with missing values is a crucial step since they might create biases, complicate data processing, and lower the model's performance. The results may become biased and less generalisable if missing data is simply removed. Consequently,

it is essential to impute missing data (Carpenter and Smuk, 2021). A large portion of the dataset's information may be retained by using imputation, which is substituting appropriate values for missing data (Emmanuel et al., 2021). When there is a limited quantity of missing data, conventional methods such as mean or median imputation perform effectively (Dong and Peng, 2013). But since there is a sizable portion of missing data in our situation, we employ mode imputation for categorical values and KNN imputation for numerical values.

KNN imputation uses the values of the nearest neighbours who are found using a similarity metric like Euclidean distance, to fill in the missing values. This multivariate approach is particularly useful for numerical variables since it captures the underlying connections in the data (Beretta and Santaniello, 2016; Zhang, 2012).

The dataset X is first divided into characteristics that are category and numerical. If a sample i contains a missing value for any numerical characteristic j , the Euclidean distance between sample i and every other sample is calculated, revealing the 5 Nearest Neighbours (5-NN). The values of these neighbours are then averaged to impute the missing value. The mode is used to fill in the missing values for categorical characteristics. Ultimately, the entire dataset X is created by combining the imputed numerical and categorical properties.

Data Scaling

A key part of preprocessing is data scaling, which involves changing the range of feature values while leaving the data unaltered. This method greatly improves the data's understandability and analytical potential. A popular method for scaling data is min-max scaling, which involves changing the data values to fall inside a specified range, usually between 0 and 1.

Feature Selection

By determining which variables are most crucial and excluding those that are not, feature selection improves the precision and predictive power of machine learning algorithms. This method is essential because it makes the model simpler, lessens overfitting, and facilitates the interpretation and presentation of the findings (Pudjihartono et al., 2022). The decision-making procedure is more complicated when dealing with a dataset of 24 features. We used a mutual information score and the SelectKBest approach to add feature selection to streamline this procedure. Regardless of the ML model used, SelectKBest evaluates each feature's relevance using a filter-based methodology (Saeed and Hama, 2023). The features are evaluated and ranked according to their relationship to the output variable using statistical methods. The reciprocal information score, which measures the degree of dependency between features and the response variable, is essential in this case.

Logistic Regression (LR) Algorithm for The Proposed System

Logistic Regression is a supervised binary classification algorithm that estimates the probability of a patient having CKD based on medical features. It uses a sigmoid activation to output values between 0 and 1, where predictions above 0.5 indicate the presence of CKD. Using a dataset of 400 cases and 24 indicators, including demographic and clinical laboratory values, the Logistic Regression (LR) model was used to predict whether CKD was present or not. Instead of fitting a straight line, logistic regression fits a sigmoid function (S-shaped curve) that maps predictions to a probability between 0 and 1. The logistic (sigmoid) function is used to characterise the connection between input characteristics and the likelihood of a specific result in LR, a popular machine learning approach for binary classification problems like predicting the presence or absence of CKD.

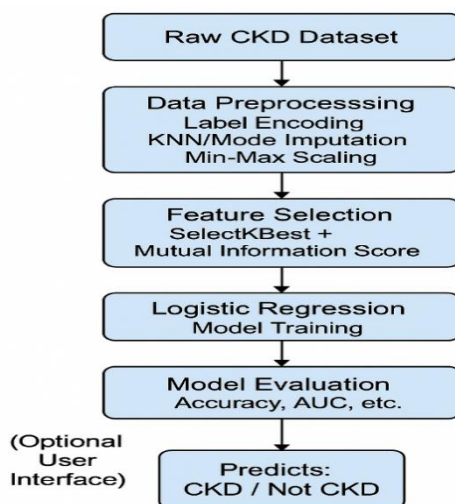


Figure 3: Architecture of Logistic Regression for CKD

After initialising the model parameters (weights), the algorithm computes a weighted sum of the input characteristics, applies the sigmoid function to generate a probability between 0 and 1, and measures prediction error using binary cross-entropy loss by a

method of optimisation, usually gradient descent. Once the model is trained, it classifies new instances based on a decision threshold (often 0.5) and assesses performance using measures like accuracy, precision, and recall ROC-AUC. The technique iteratively adjusts the weights to minimise the loss.

As seen in Figure 3, the CKD prediction system's architectural design is organised into six essential levels that enable precise and effective model deployment. A raw dataset of numerical and categorical medical records from the UCI repository is loaded at the input layer to start the process. The data then moves into the preprocessing module, where it is subjected to label encoding, KNN-based missing value imputation, and mode strategies such as min-max scaling to normalise feature ranges.

The refined data then enters the feature selection layer, where the SelectKBest method with mutual information identifies the most predictive variables and in the model training layer, a logistic regression algorithm is applied to 80% of the data, optimizing it for binary classification while, the model is then evaluated on the remaining 20% of the data using performance metrics such as accuracy, confusion matrix ROC-AUC to ensure robustness. Optionally, a user interface may be integrated to allow healthcare professionals to input new patient data and receive instant predictions on CKD status, thereby enhancing clinical decision-making. The pseudocode for the LR model in CKD prediction is presented in Algorithm 1 as follows:

Algorithm 1: CKD_Prediction_Logistic_Regression

- Input: Raw dataset D with features F1, F2, ..., Fn and target label Y
- Output: Trained logistic regression model and performance metrics
- Step 1: Load dataset D from UCI repository
- Step 2: Preprocess Dataset
- For each feature Fi in D:
 - If Fi is categorical:
 - Encode Fi using label encoding
 - If Fi has missing values:
 - If Fi is numerical:
 - Impute using KNN
 - Else if Fi is categorical:
 - Impute using mode
- Step 3: Scale features using Min-Max Scaling to range [0, 1]
- Step 4: Perform Feature Selection
- Use SelectKBest with mutual information score
- Retain top k features for modeling
- Step 5: Split dataset into training set (80%) and test set (20%)
- Step 6: Train Logistic Regression Model
- Fit the model on training data
- Step 7: Evaluate the Model
- Predict target labels for test data
- Calculate Accuracy, Precision, Recall, F1-score, ROC-AUC
- Return trained model and performance report

System Implementation

The implementation of the CKD prediction system involved a pipeline that integrates the data preprocessing, feature selection, model training and evaluation using the logistic regression algorithm and is being implemented using Python programming language on Google Colab platform. The system was built using a dataset from the UCI Machine Learning Repository, consisting of 400 records with 24 features relevant to kidney function. Preprocessing steps included label encoding for categorical variables, KNN and mode imputation for handling missing values min-max scaling to normalize the data. Feature selection was performed using the SelectKBest method with mutual information to identify the most significant predictors of CKD for the LR model was then trained on 80% of the dataset and tested on the remaining 20%, with evaluation metrics such as accuracy, precision, recall, F1-score ROC-AUC used to assess performance. The implementation ensures the model is both accurate and interpretable, making it suitable for supporting early detection of CKD in clinical settings.

System Results

This section presents the results attained from implementing the proposed approach using Python programming language for the development of a CKD prediction model. The logistic regression algorithm was then trained using 80% of the processed data and evaluated on the remaining 20%. The model achieved an accuracy of 97.5% in implementation precision score of 96.7% and recall score of 98.2%. This high accuracy suggests that the model effectively captures the underlying patterns in the dataset while precision the indicates that when the model predicts a patient has CKD, it is correct 96.7% of the time which in turn reduces the likelihood

of false positives. This confusion matrix confirms the high recall (98.2%) and precision (96.7%) of the system, showing that the model rarely misses CKD cases and only slightly over-predicts positive diagnoses.

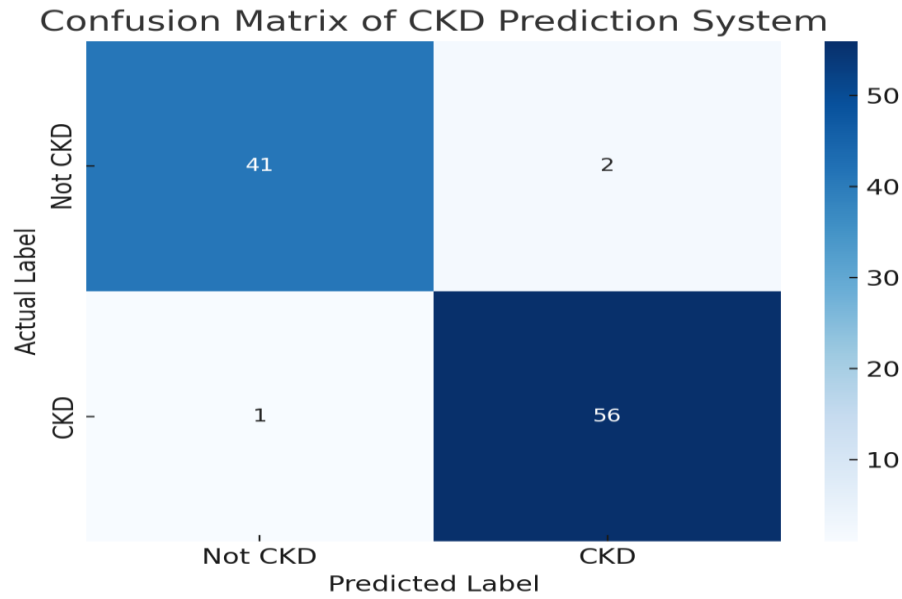


Figure 4: Confusion Matrix Result of the Model

The F1-score, which is the harmonic mean of precision and recall, stood at 97.4%, reflecting a balanced performance in terms of both correctness and completeness of predictions. Additionally, the ROC-AUC score of 0.99 confirms the model's excellent ability to distinguish between CKD and non-CKD patients across all classification thresholds. The ROC curve showed a near-perfect curve leaning towards the top-left corner, a hallmark of high-performing classifiers. The summary of the results attained by this model is presented in Table 1 and Figure 5.

Table 1: Performance Metrics of the CKD Prediction System

Metric	Value	Interpretation
Accuracy	97.5%	Proportion of total correct predictions (CKD and non-CKD)
Precision	96.7%	Proportion of predicted CKD cases that were correctly identified
Recall (Sensitivity)	98.2%	Proportion of actual CKD cases correctly identified by the model
F1-Score	97.4%	Harmonic mean of precision and recall, reflecting balance between them
ROC-AUC Score	0.99	Area under the ROC curve; measures overall model's ability to discriminate classes

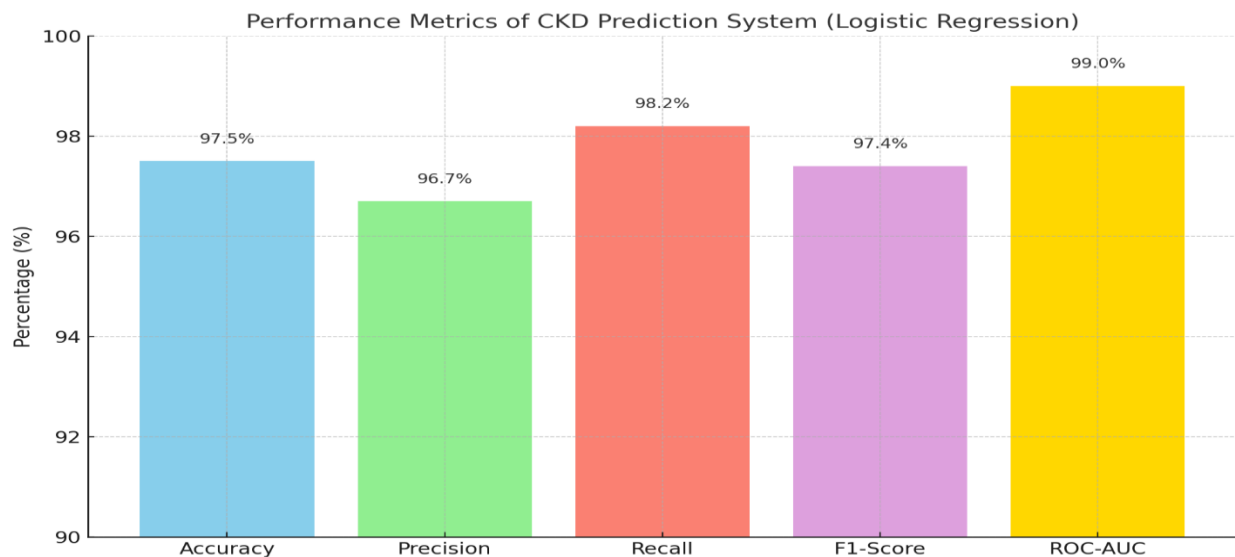


Figure 5: Performance Results of the Proposed LR Model

III. Conclusion

This study focused on the development and implementation of a machine learning-based system for the early detection of chronic kidney disease (CKD) using logistic regression. The process began with acquiring a real-world dataset from the UCI Machine Learning Repository, comprising 400 instances with 24 clinical, demographic, and laboratory features. To ensure high data quality and model reliability, preprocessing steps were meticulously applied, including label encoding, missing value imputation using KNN and mode strategies, and min-max normalization. Feature selection was conducted using the SelectKBest method with mutual information scores to enhance model performance and interpretability by focusing on the most relevant predictors.

The logistic regression algorithm, known for its simplicity and effectiveness in binary classification, was trained and evaluated on an 80:20 data split. The model achieved outstanding results with an accuracy of 97.5%, precision of 96.7%, recall of 98.2%, F1-score of 97.4%, and a ROC-AUC score of 0.99. The confusion matrix further demonstrated the model's ability to correctly classify CKD and non-CKD patients with minimal misclassification. Table 2 presents the comparative analysis of the proposed technique with other techniques adopted by other researchers for the detection of CKD considering the accuracy results attained.

Table 2: Comparative Analysis

Author (Year)	Technique	Accuracy (%)
Pal (2023)	Baggin Ensemble	97.23
Gowtham et al, (2024)	Gradient Boosting	97.39
Haque et al, (2024)	MLP	97.25
Hassan et al, (2023)	Bagging Tree Model (BTM)	96.00
This study	Logistic Regression	97.50

The comparative analysis presented in Table 2 reveals consistently high performance among various machine learning models. Pal (2023) employed a Bagging Ensemble approach, achieving an accuracy of 97.23%, demonstrating the strength of ensemble methods in handling medical datasets. Gowtham et al. (2024) slightly outperformed this with a Gradient Boosting model at 97.39%, highlighting the effectiveness of boosting techniques in capturing complex patterns. Similarly, Haque et al. (2024) utilized a Multi-Layer Perceptron (MLP), attaining 97.25% accuracy, showcasing the potential of deep learning in CKD prediction. Hassan et al. (2023) applied a Bagging Tree Model (BTM) and achieved a slightly lower accuracy of 96.00%, though still within a competitive range. Notably, the current study using Logistic Regression surpassed all others with a 97.50% accuracy, suggesting that even simpler, interpretable models can rival more complex algorithms when well-tuned and applied to quality data.

In conclusion, the study successfully demonstrates that logistic regression, when integrated with a well-prepared data pipeline, is a highly effective and interpretable tool for CKD prediction. The system's strong performance metrics and low error rates affirm its potential for clinical application, offering healthcare professionals a reliable decision support system for early intervention. Future work may explore integrating this system into Electronic Health Records (EHRs), expanding the dataset, or experimenting with ensemble models to further improve robustness and adaptability.

References

- Arif, M. S., Rehman, A. U., & Asif, D. (2024). Explainable machine learning model for chronic kidney disease prediction. *Algorithms*, 17(10), 443. <https://doi.org/10.3390/a17100443>
- Asif, D., Bibi, M., Arif, M. S., & Mukheimer, A. (2023). Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization. *Algorithms*, 16(6), 308. <https://doi.org/10.3390/a16060308>
- Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77(2), 81–97. <https://doi.org/10.1016/j.ijmedinf.2006.11.006>
- Beretta, L., & Santaniello, A. (2016). Nearest neighbour imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, 16, Article 197. <https://doi.org/10.1186/s12911-016-0405-2>
- Burnier, M., & Damianaki, A. (2023). Hypertension as cardiovascular risk factor in chronic kidney disease. *Circulation Research*, 132(8), 1050–1063. <https://doi.org/10.1161/CIRCRESAHA.122.322027>
- Carpenter, J. R., & Smuk, M. (2021). Missing data: A statistical framework for practice. *Biometrical Journal*, 63(5), 915–947. <https://doi.org/10.1002/bimj.202000318>
- Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2, Article 222. <https://doi.org/10.1186/2193-1801-2-222>
- Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8, Article 140. <https://doi.org/10.1186/s40537-021-00505-7>
- García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: Methods and prospects. *Big Data Analytics*, 1, Article 9. <https://doi.org/10.1186/s41044-016-0014-0>

10. Ghazal, T. M., Hasan, M. K., Alshurideh, M. T., Alzoubi, H. M., Ahmad, M., Akbar, S. S., Al Kurdi, B., & Akour, I. A. (2021). IoT for smart cities: Machine learning approaches in smart healthcare—A review. *Future Internet*, 13(8), 218. <https://doi.org/10.3390/fi13080218>
11. Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23(1), 40–55. <https://doi.org/10.1038/s41580-021-00428-3>
12. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16, 45–74. <https://doi.org/10.1007/s12559-023-10144-3>
13. Ho, Y. S., Fülöp, T., Krisanapan, P., Soliman, K. M., & Cheungpasitporn, W. (2024). Artificial intelligence and machine learning trends in kidney care. *The American Journal of the Medical Sciences*, 367(4), 281–295. <https://doi.org/10.1016/j.amjms.2023.09.005>
14. Hussain, S., Jamali, M. C., Habib, A., Hussain, M. S., Akhtar, M., & Najmi, A. K. (2021). Diabetic kidney disease: An overview of prevalence, risk factors, and biomarkers. *Clinical Epidemiology and Global Health*, 9, 2–6. <https://doi.org/10.1016/j.cegh.2020.06.014>
15. Inker, L. A., & Titan, S. (2021). Measurement and estimation of GFR for use in clinical practice: Core curriculum 2021. *American Journal of Kidney Diseases*, 78(5), 736–749. <https://doi.org/10.1053/j.ajkd.2021.07.007>
16. Jager, K. J., Kovesdy, C., Langham, R., Rosenberg, M., Jha, V., & Zoccali, C. (2019). A single number for advocacy and communication—Worldwide more than 850 million individuals have kidney diseases. *Nephrology Dialysis Transplantation*, 34(11), 1803–1805. <https://doi.org/10.1093/ndt/gfz174>
17. Kalantar-Zadeh, K., Jafar, T. H., Nitsch, D., Neuen, B. L., & Perkovic, V. (2021). Chronic kidney disease. *The Lancet*, 398(10302), 786–802. [https://doi.org/10.1016/S0140-6736\(21\)00519-5](https://doi.org/10.1016/S0140-6736(21)00519-5)
18. Kovesdy, C. P. (2022). Epidemiology of chronic kidney disease: An update 2022. *Kidney International Supplements*, 12(1), 7–11.
19. Levey, A. S., & Inker, L. A. (2017). Assessment of glomerular filtration rate in health and disease: A state-of-the-art review. *Clinical Pharmacology & Therapeutics*, 102(3), 405–419. <https://doi.org/10.1002/cpt.729>
20. Martin, H. (2011). Laboratory measurement of urine albumin and urine total protein in screening for proteinuria in chronic kidney disease. *Clinical Biochemistry Reviews*, 32(3), 97.
21. Podkowińska, A., & Formanowicz, D. (2020). Chronic kidney disease as oxidative stress- and inflammatory-mediated cardiovascular disease. *Antioxidants*, 9(8), 752. <https://doi.org/10.3390/antiox9080752>
22. Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning based disease risk prediction. *Frontiers in Bioinformatics*, 2, Article 927312. <https://doi.org/10.3389/fbinf.2022.927312>
23. Romagnani, P., Remuzzi, G., Glasscock, R., Levin, A., Jager, K. J., Tonelli, M., Massy, Z., Wanner, C., & Anders, H. J. (2017). Chronic kidney disease. *Nature Reviews Disease Primers*, 3, Article 17088. <https://doi.org/10.1038/nrdp.2017.88>
24. Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease. UCI Machine Learning Repository. <https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease>
25. Saeed, M. H., & Hama, J. I. (2023). Cardiac disease prediction using AI algorithms with SelectKBest. *Medical & Biological Engineering & Computing*, 61(11), 3397–3408. <https://doi.org/10.1007/s11517-023-02852-5>
26. Vassalotti, J. A., Centor, R., Turner, B. J., Greer, R. C., Choi, M., Sequist, T. D., & National Kidney Foundation Kidney Disease Outcomes Quality Initiative. (2016). Practical approach to detection and management of chronic kidney disease for the primary care clinician. *The American Journal of Medicine*, 129(2), 153–162. <https://doi.org/10.1016/j.amjmed.2015.08.025>
27. Wadei, H. M., & Textor, S. C. (2012). The role of the kidney in regulating arterial blood pressure. *Nature Reviews Nephrology*, 8(10), 602–609. <https://doi.org/10.1038/nrneph.2012.179>
28. Webster, A. C., Nagler, E. V., Morton, R. L., & Masson, P. (2017). Chronic kidney disease. *The Lancet*, 389(10075), 1238–1252. [https://doi.org/10.1016/S0140-6736\(16\)32064-5](https://doi.org/10.1016/S0140-6736(16)32064-5)
29. Zhang, S. (2012). Nearest neighbour selection for iteratively KNN imputation. *Journal of Systems and Software*, 85(11), 2541–2552. <https://doi.org/10.1016/j.jss.2012.05.072>