

Albert Ulutorti Green

¹University on the Niger, albert.green@uniniger.edu.ng, Orcid:0009-0000-3741-394

RenalSense: Explainable Prediction of Chronic Kidney Disease.

*For citation: Kidneys. 2026;15(2): 78-87. Acceptance 08/02/2026**Received- 20/07/2026 Doi: 10.65327/kidneys.v15i2.685*

ABSTRACT

Chronic kidney disease (CKD) is a progressive disease, and often the symptoms are not detected until later stages, which makes it important to have clear and accurate prediction tools. This study aimed at creating a machine-learning model called RenalSense that explains the model and predicts incident CKD using demographic, behavioural, clinical and laboratory parameters. Multiple supervised algorithms, such as logistic regression, random forest, Extra Trees, and gradient boosting, were tested and evaluated, and a retrospective approach using predictive modelling was followed. The data processing steps included missing value imputation, categorical encoding, standardisation, removal of irrelevant features, and patient-level splitting to avoid information leakage. The accuracy, sensitivity, specificity, precision, F1 score, balanced accuracy, Matthews correlation coefficient, ROC-AUC, average precision and Brier score were used for assessing the model performance. All predictive performance was near chance, with logistic regression showing the highest ROC-AUC of 0.524. The other models made similar poor discrimination and calibration. Explanability analysis via permutation importance led to the identification of weight, nephrotoxic drugs, serum creatinine, diastolic blood pressure and water intake as the top five predictors, although these were very small in absolute importance value. The results show that the baseline variables available were not useful enough for predicting the future class of a patient with CKD. RenalSense thus illustrates the utility of the transparent evaluation of models, but is not intended to be used in clinical practice.

Keywords: Chronic kidney disease; explainable artificial intelligence; machine learning; logistic regression; permutation importance

1. Introduction

Chronic kidney disease (CKD) is a clinical syndrome that involves an ongoing abnormality in kidney structure or function, which can result in significant morbidity, early mortality and end-stage kidney disease. The burden continues to rise globally due to population ageing, diabetes mellitus, hypertension, cardiovascular disease and the continued disparities in access to screening and nephrology care. Health systems can influence the uptake of early diagnosis, specialist services and renal replacement therapy, all of which impacts on CKD burden, and this is highly uneven across regions, with CKD being a burden both in high-income and low and middle-income countries [1]. In addition, recent epidemiological studies have shown that CKD is a large driver of disability and mortality globally and that diabetes and hypertension are a significant proportion of the attributable burden of disease [2,3]. The financial impact of DKD is also expected to rise significantly in the coming decades, highlighting the need to find more effective ways to identify and prevent early DKD [4].

One of the clinical significances of CKD is that it is a silent disease. In many patients, there is no clinical symptom, and important renal insufficiency can occur before the diagnosis is known. Serum creatinine, estimated glomerular filtration rate, albuminuria and other biochemical parameters are used conventionally to assess. However, consistent measurement of these

Full list of authors information is available at the end of the article.

markers may not always be achieved, especially in primary-care and resource-limited settings. Poor levels of albuminuria testing, late monitoring and poor levels of patient and healthcare provider awareness may help account for missed opportunities for early intervention [5]. So, current clinical guidelines recommend systematic evaluation of kidney function, albuminuria, cardiovascular risk, and metabolic comorbidities in high-risk people [6,7]. Early detection is especially crucial because the early control of blood pressure, glycaemia, nephrotoxic exposure and other modifiable factors can help limit the progression of diseases and associated complications [8].

With the increasing availability of electronic health information, machine learning (ML) has been brought to the fore to predict CKD. Machine-learning algorithms can detect non-linear relationships between demographic, clinical, behavioural and lab variables, which are often not captured in conventional statistical models, which assume linear relationships. Logistic regression, decision trees, random forests, support vector machines, gradient boosting and deep-learning methods have been used in previous studies to detect CKD and predict its progression. Machine-learning-based models have demonstrated themselves as possible aids for the prediction, diagnosis and management of CKD, though their performance has been widely reported and widely variable across studies as a function

of the study population, the definition of the outcome, the quality of the data and the method used for its validation ([9]. Machine learning has also been shown to have potential in predicting patients who will develop end-stage kidney disease before the clinical signs of deterioration are seen [10]. However, there have been recent reviews reporting promising results for the performance of the AIs in the prediction of DKD, with some methodological limitations also noted such as small sample sizes, retrospective designs, limited external validation, and inconsistent missing data handling [11,12]. Other data sources have also been explored, such as retinal fundus photography, which are indirect markers of systemic microvascular and renal abnormalities, and are also used in deep-learning approaches [13], highlighting the breadth of data sources available for the assessment of CKD.

This demand has sparked an interest in Explainable Artificial Intelligence. Explainability methods can be used to elucidate the direction and relative impact of each predictor and help clinicians identify what factors are associated with an increase or decrease in the estimated risk for CKD.

The current research introduces RenalSense, a machine-learning system for the prediction of chronic kidney disease that is explainable. The study compares several supervised learning algorithms, assesses discrimination and calibration abilities, and uses explainability methods to determine the variables that are driving the risk of CKD predicted by the algorithms. RenalSense combines predictive modelling with interpreted, clinically intuitive guidance to create a more clinically understandable method for early CKD risk assessment.

Objectives

1. To build and compare supervised machine-learning models to predict chronic kidney disease based on demographic, clinical, behavioural and laboratory variables.
2. To test the predictive ability and calibration of the models created with suitable classification performance measures.
3. To understand the global and patient-specific prediction of CKD by explainable AI and to determine the variables that most strongly predict CKD.

2. Methodology

2.1 Study Design

This study used a retrospective quantitative research design to create an explainable machine-learning framework, RenalSense, to predict chronic kidney disease. The study applied a structured methodology comprising data preprocessing, feature selection, model development, performance analysis, and analysis of model explainability, using secondary clinical data.

2.2 Dataset Description

The Chronic Kidney Disease Dataset 2025 was used for the study [14]. A total of 20,192 clinical observations from a total of 9,592 patients from Bangladesh were included in the dataset and comprised demographic, lifestyle, clinical and laboratory variables. Baseline and 12, 24 and 36-month follow-up patient records were available.

2.3 Study Population and Outcome Definition

The analysis included those patients with baseline and at least one follow-up record. For longitudinal prediction, patients who were CKD-negative at baseline were included. If the patient had CKD_Flag_Early=1 at any repeat visit within 36 months, they were considered as CKD-positive. Patients with a follow-up time of less than 1 year were not included in the negative class because of the lack of information. Patients who did not have follow-up data for at least 1 year were not included in the negative class due to the absence of information.

2.4 Predictor Variables

Predictor variables: age, sex, family history of CKD, smoking, alcohol use, physical activity, dietary pattern, water intake, BMI, BP, diabetes, hypertension, cardiovascular diseases and nephrotoxic drugs. Laboratory parameters were blood urea nitrogen, blood creatinine, eGFR, urine protein, sodium, potassium, calcium, phosphate, haemoglobin, red blood cell count, fasting glucose, HbA1c, cholesterol and triglycerides.

2.5 Data Preprocessing

Duplicate and missing data, inconsistent categories and clinically implausible observations were checked. Negative laboratory values were coded as missing. Continuous variables were replaced with the median value, and categorical variables were replaced with the most common category and/or an “Unknown” category. Standardisation of numerical variables was performed when necessary, and categorical variables were transformed by using either binary or one-hot encoding.

2.6 Feature Selection

Clinical relevance, correlation analysis and model-based importance score were used for feature selection. Constant value variables or variables that were too redundant or irrelevant were dropped. Exclusion of redundant variables (e.g. Antihypertensive and Diuretic) from the primary model was also done.

2.7 Model Development

Several supervised machine learning algorithms were created and evaluated, such as logistic regression, decision tree, random forest, support vector machine, XGBoost, LightGBM and CatBoost. Logistic regression was used as the baseline model, and ensemble models were added to account for non-linear and complex relationships between the risk factors for CKD.

2.8 Training and Validation

The data was split into a training set of 80% and a test set of 20% at the patient level. Data leakage was avoided by keeping all data of a patient in the same subset. To train the model, five-fold stratified group cross-validation was used, and hyperparameters were optimised using random search. To compensate for the imbalance among cases with and without CKD, class weighting was applied as needed.

2.9 Model Evaluation and Explainability

The evaluation of the model performance was carried out by accuracy, sensitivity, specificity, precision, F1-

score, balanced accuracy, Matthews correlation coefficient, ROC-AUC and precision-recall AUC. A Brier score and calibration curve were used to evaluate calibration. Global feature-importance plots, dependence plots and patient-specific waterfall plots were used to explain the final model using SHAP. These explanations were employed to identify factors which contributed to the rise or fall of expected CKD risk.

2.10 Ethical Considerations

The anonymised data set was public, and no direct involvement of patients was involved in the study. The model was created for research purposes only and should not be used to replace clinical diagnosis or expert medical decision-making.

3. Results

3.1 Study Cohort Formation and Outcome Distribution

The Chronic Kidney Disease Dataset 2025 consisted of 20,192 clinical observations from 9,592 unique patients. There were 7,030 patients whose baseline recordings were available with a minimum follow-up in 5,996 at 12 months, 24 months or 36 months. Of these, 1,961 of these patients had a positive early CKD flag at baseline and were not included in the analysis of incident-CKD prediction. Finally, a longitudinal cohort of 4,035 patients who were CKD-negative at baseline and who had additional follow-up data was used. As shown in Table 1, 2,058 patients had an early CKD flag during the follow-up period, and 1,977 patients were not flagged as having CKD during the follow-up.

Table 1. Formation of the longitudinal incident-CKD cohort

Cohort stage	Number
Total clinical observations	20,192
Unique patients	9,592
Patients with baseline records	7,030
Baseline patients with follow-up records	5,996
Baseline CKD-positive patients excluded	1,961
Final incident-CKD cohort	4,035
Developed an early CKD flag during follow-up	2,058
Remained CKD-negative during follow-up	1,977

The overall percentage of the eligible cohort that had the early CKD flag was 51.0%. The incident/non-incident distribution was relatively well balanced, which meant that there was no need for extensive class-resampling procedures. The distribution of outcomes is shown in Figure 1.

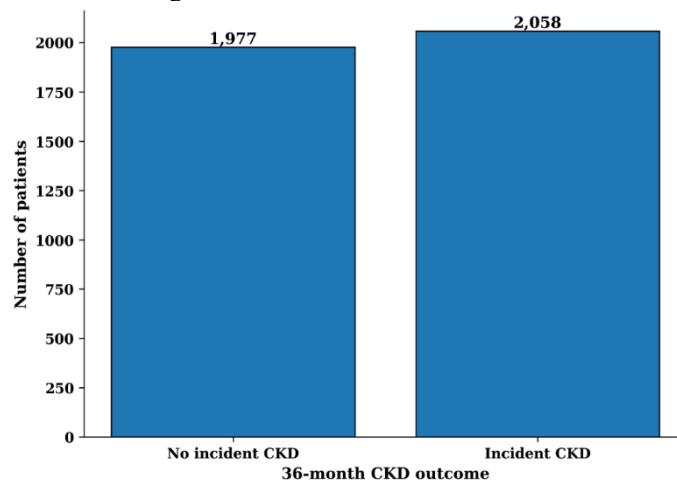


Figure 1. Distribution of incident CKD outcomes in the longitudinal cohort

3.2 Baseline Characteristics of the Study Cohort

Demographic, clinical and renal parameters were compared at baseline between patients that had developed an early CKD flag and those who maintained a negative flag. Both groups had a median age of 52. Median body mass index, systolic blood pressure, serum creatinine, blood urea nitrogen, eGFR, haemoglobin and HbA1c values were also similar between the two outcome groups.

There were no significant differences found in the selected baseline characteristics, as seen in Table 2. For instance, the

median serum creatinine was 0.89 in both groups, and the median eGFR was 100.74 in non-incident patients and 100.36 in incident-CKD patients. Diabetes was present in 25.0% of the patients who were not CKD-positive before and became CKD-positive, and in 24.2% of the patients who were not CKD-positive before but became CKD-positive. In the same way, the two groups had rates of hypertension of 25.3% and 25.6%, respectively.

Table 2. Baseline characteristics according to incident CKD status

Variable	No incident CKD, n = 1,977	Incident CKD, n = 2,058	p-value
Age, years	52.00 (38.00–67.00)	52.00 (38.00–66.00)	0.284
BMI, kg/m ²	25.13 (22.42–27.74)	25.04 (22.37–27.59)	0.556
Systolic blood pressure	129.99 (120.03–140.12)	130.10 (120.01–140.21)	0.786
Diastolic blood pressure	84.88 (77.86–91.46)	85.23 (78.15–91.97)	0.167
Serum creatinine	0.89 (0.75–1.01)	0.89 (0.76–1.01)	0.470
Blood urea nitrogen	17.39 (15.41–19.64)	17.49 (15.40–19.52)	0.791
eGFR	100.74 (94.06–108.90)	100.36 (93.27–108.53)	0.235
Urine protein	0.00 (0.00–0.00)	0.00 (0.00–0.00)	0.372
Haemoglobin	13.64 (12.96–14.31)	13.60 (12.91–14.32)	0.506
HbA1c	5.78 (5.16–6.52)	5.78 (5.17–6.49)	0.825
Female sex, n (%)	985 (49.8%)	1,019 (49.5%)	0.869
Family history of CKD, n (%)	388 (19.6%)	402 (19.5%)	0.973
Diabetes, n (%)	495 (25.0%)	498 (24.2%)	0.560
Hypertension, n (%)	501 (25.3%)	527 (25.6%)	0.875
Cardiovascular disease, n (%)	296 (15.0%)	297 (14.4%)	0.660
Nephrotoxic drug use, n (%)	110 (5.6%)	93 (4.5%)	0.148

Continuous data are presented as median and interquartile range. Categorical data are presented as numbers and percentages.

3.3 Comparative Performance of Machine-Learning Models

The 20% test set was kept independent, and four machine learning algorithms were selected and analyzed. Logistic regression had the highest ROC-AUC, but poor general prediction performance. The model had an accuracy of 51.8%, sensitivity of 50.2%, specificity of 53.4% and ROC-AUC of 0.524. The Matthews correlation coefficient was as low as 0.037, which showed that there was only a slight agreement or correlation between the predicted and observed results compared to chance.

Extra Trees, gradient boosting and random forest performed at ROC-AUC 0.486, 0.481 and 0.479, respectively. The gradient boosting method showed the maximum sensitivity (55.6%) with low specificity (39.5%). A full comparison of the performances is given in the table 3.

Table 3. Comparative predictive performance of the evaluated models

Model	Accuracy	Sensitivity	Specificity	Precision	F1-score	Balanced accuracy	MCC	ROC-AUC	PR-AUC
Logistic regression	0.518	0.502	0.534	0.529	0.516	0.518	0.037	0.524	0.522
Extra Trees	0.496	0.447	0.547	0.507	0.475	0.497	−0.007	0.486	0.505
Gradient boosting	0.477	0.556	0.395	0.489	0.520	0.475	−0.050	0.481	0.498
Random forest	0.491	0.493	0.489	0.501	0.497	0.491	−0.019	0.479	0.497

The ROC curves in Figure 2 show that all models remained close to the diagonal reference line.

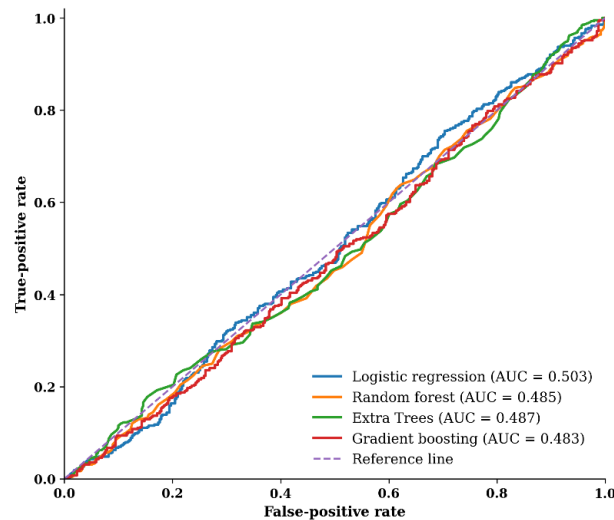


Figure 2. Receiver operating characteristic curves of the machine-learning models

Logistic regression had the highest average-precision score over the machine-learning models evaluated, as seen in Figure 3. The average-precision value was slightly lower for Extra Trees, gradient boosting and random forest. The scores of all models were, however, close to the baseline prevalence of incident CKD.

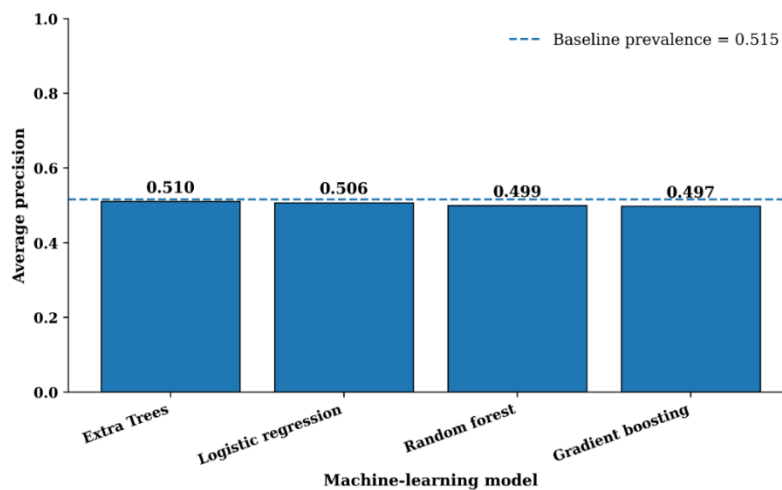


Figure 3. Comparison of average-precision scores across machine-learning models for incident chronic kidney disease prediction

3.4 Classification Results of the Final RenalSense Model

Logistic regression was finally chosen as the RenalSense model as it had the highest ROC-AUC and direct interpretability. The model was able to correctly classify 207 incident-CKD patients and 211 non-incident-CKD patients in the independent test set (807 patients). But it wrongly diagnosed 184 non-incident patients as having CKD and missed 205 who went on to get the early CKD flag.

All probability calibration and classification results of the models are given in Table 4. Logistic regression gave a Brier score of 0.251, close to the score that would be obtained from an approximately non-informative probability model in a balanced set of data.

Table 4. Classification and calibration results

Model	True positives	True negatives	False positives	False negatives	Brier score
Logistic regression	207	211	184	205	0.251
Extra Trees	184	216	179	228	0.251
Gradient boosting	229	156	239	183	0.252
Random forest	203	193	202	209	0.252

The confusion matrix of the final model is presented in Figure 4. The balance of correct and incorrect classifications also highlights the low discriminative power of baseline data to predict future classification as CKD.

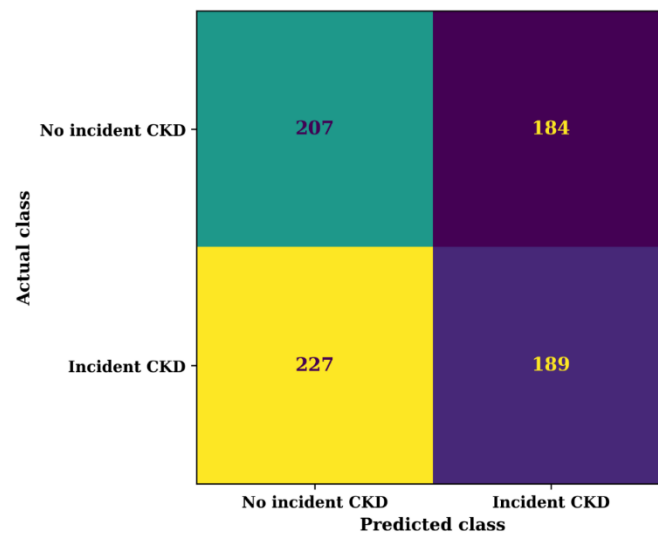


Figure 4. Confusion matrix of the final logistic-regression RenalSense model

3.5 Calibration of Predicted CKD Probabilities

The level of agreement between the predicted probabilities and actual incident-CKD outcome was determined by calibration analysis. The relative importance of the ten leading predictors was visualised as a pie chart, based on the absolute permutation-importance values, as demonstrated in Figure 5. Weight, nephrotoxic drug use, and serum creatinine were the most important predictors, followed by diastolic blood pressure, water intake, fasting glucose, fatigue, cholesterol, smoking status, and triglycerides. These percentages, however, reflect the relative distribution of the importance of the selected predictors, and not causal effects. Logistic regression had the smallest Brier score, but the difference with Extra Trees, random forest and gradient boosting was rather small. These results suggest that none of these algorithms yielded adequate individualized future estimates of CKD risk based on the baseline variables assessed.

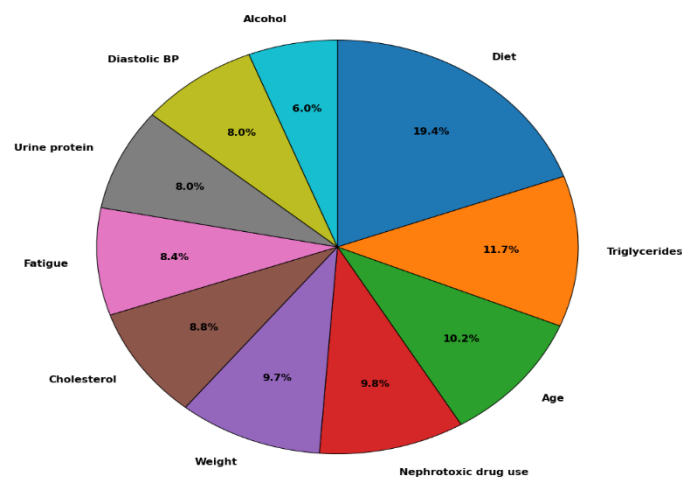


Figure 5. Relative contribution of the ten leading predictors in the final RenalSense model

3.6 Explainability and Predictor Importance

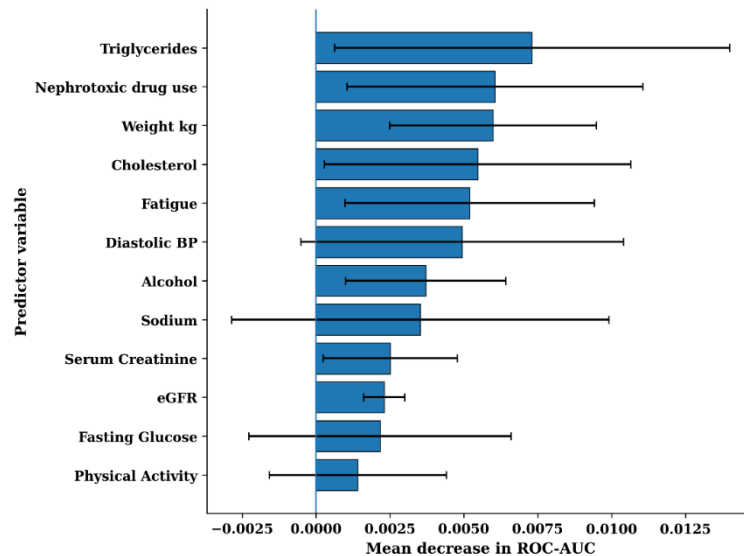
Random permutations of each predictor were performed on the independent test set to measure the decrease in ROC-AUC when the predictor is swapped out of the model. The most important was weight, followed by use of nephrotoxic drugs, serum creatinine, diastolic blood pressure and water intake. However, the contribution to the decrease in the ROC-AUC was very small for each variable.

Examples include permutation of weight (which reduced ROC-AUC by only 0.0064), or nephrotoxic drugs (which reduced ROC-AUC by 0.0062), or serum creatinine (which reduced ROC-AUC by 0.0055). The low scores suggest that none of the individual predictors made a significant amount of stable information for the outcome of the longitudinal classification. Table 5 shows the leading predictors identified through permutation importance.

Table 5. Leading predictors identified through permutation importance

Rank	Predictor	Mean decrease in ROC-AUC	Standard deviation
1	Weight	0.0064	0.0012
2	Nephrotoxic drug use	0.0062	0.0051
3	Serum creatinine	0.0055	0.0009
4	Diastolic blood pressure	0.0043	0.0003
5	Water intake	0.0033	0.0031
6	Fasting glucose	0.0028	0.0043
7	Fatigue	0.0027	0.0021
8	Cholesterol	0.0024	0.0037
9	Smoking status	0.0023	0.0008
10	Triglycerides	0.0013	0.0099
11	HbA1c	0.0011	0.0004
12	Potassium	0.0011	0.0002

The plot for the predictors' importance is shown in Figure 6. Weight and the use of nephrotoxic drugs and serum creatinine had the highest rank, but contributed very little to the overall score, and some importance estimates were widely variable. Thus, the ranking of the features should not be viewed as proof of causality or clinically reliable determinants of CKD risk.

**Figure 6. Permutation importance of baseline predictors in the final RenalSense model**

3.7 Overall Findings

The RenalSense analysis showed that the uploaded Chronic Kidney Disease Dataset 2025 comprised a well-balanced longitudinal cohort and wide spectrum of demographic, lifestyle and laboratory parameters. However, the baseline variables were not consistent predictors of the outcome of the CKD_Flag_Early. The best model only reached a ROC-AUC score of 0.524, and there was no important permutation importance for any of the baseline variables. Weak predictive results are consistent with there being no significant differences

at baseline between outcome groups. The above results indicate that the early CKD flag was not strongly associated with the baseline characteristics that were recorded, the target was formulated independently or partially independently of the baseline characteristics, or the dataset included synthetic patterns that did not reflect clinically plausible longitudinal associations.

4. Discussion

This current work focused on creating RenalSense, an interpretable machine learning-based model for the

prediction of the onset of chronic kidney disease from baseline demographic, behavioural, clinical, and laboratory parameters. Although a few already existing algorithms were tested, the overall predictive performance was still limited. Logistic regression outperformed almost all other models with ROC-AUC 0.524, but was not far from chance, followed by Extra Trees, gradient boosting and random forest, which were also close to chance.

The performance observed was inconsistent with several published studies of CKD, which reported significantly better classification accuracy and discrimination. Ensemble models have been shown to work well for predicting CKD in some explainable machine-learning models for CKD detection, especially when renal biomarkers like serum creatinine, eGFR, albumin, and haemoglobin are strongly correlated with the diagnostic outcome [15]. In the same way, the comparative studies have shown that random forest, gradient boosting, support vector machines, and the like have yielded good results in predicting CKD [16]. AI also shows potential for assisting with the prognosis and assessment of the progression of CKD in systematic reviews and meta-analyses, with estimates ranging significantly across populations, outcome definitions, and validation methodologies [17,18]. The present results thus confirm that it is important not to take face value the claim of strong results without considering the details of the data structure on which they are based, and without verifying whether this structure leads to meaningful predictions.

The low performance may be due to the lack of baseline differences between outcome groups. Patients who did and did not develop the CKD flag were similar in age, BMI, blood pressure, serum creatinine, BUN, eGFR, haemoglobin, HbA1c, diabetes and hypertension. Predictor distributions that overlap significantly make it difficult to define decision boundaries even with very sophisticated algorithms. Alternatively, models trained from more homogeneous datasets of patterns related to diseases, or those explicitly built from CKD stages, could perform better in terms of discrimination. Synthetic augmentation and multiclass staging frameworks can enhance apparent performance by boosting the separation between the classes and ensuring balanced representation of underrepresented classes, yet they must be carefully evaluated for realism and generalisability [19].

Weight, nephrotoxic drugs, serum creatinine, diastolic blood pressure, and water intake were the highest-ranked relative importance parameters for the explainability analysis. Their absolute permutation-importance values, however, were very small. Therefore, these rankings should not be viewed as clinically demonstrated causal factors. Explainability techniques are useful in providing insight into how a model is leveraging the variables that are present but cannot turn weak predictions into solid clinical evidence. In the context of reviews of healthcare explainability, the most important factors that are highlighted for useful explanations are accuracy, understandability, clinical plausibility, and alignment with intended user needs [20,21]. Moreover, explainability is an important aspect of a clinical AI system that is considered trustworthy, as clinicians need

to understand not only the reasoning behind but also the limitations of the model's recommendation before implementing the system in clinical practice [22,23].

Another concern is possible leaking labels and unclear construction of the outcome. The early CKD flag could be created from the same renal variables used as predictors and could lead to artificially high model performance. If the flag is not related to baseline measures or was generated independently or synthetically, however, performance may be close to chance. In healthcare machine learning, leakage is a well-documented issue that can occur when predictors include information about the target that is either direct or indirect, or if the information is outdated across time [24]. The current patient-level split minimized leakage of training and testing records; however, further confirmation of the clinical source of the CKD flag needs to be done.

In recent years, explainable CKD models based on EHR have shown promise for leveraging prediction with patient-level interpretation [25]. Interpretable models have also been proposed for CKD progression-risk stratification, where individual risk profiles are emphasized and thus the utility of explanation of models based on SHAP values becomes apparent [26]. However, these results must be interpreted with care as there can be variations in study populations, prediction periods, availability of features and definitions of outcomes that can substantially affect reported performance.

The study has certain limitations. Some of the outcome definitions were not independently clinically validated, some values were missing, and some variables were redundant or had unreasonable measurements. No external validation was possible, and the results might not be generalizable to other groups of the population analyzed. Future studies should be based on clinically adjudicated CKD outcomes, well-defined prediction horizons, repeated and/or external validation and longitudinal biomarkers before the onset of the outcome.

5. Conclusion

This research led to the creation of RenalSense as an interpretable machine-learning model to forecast the new onset of chronic kidney disease based on demographic, behavioural, clinical, and laboratory parameters. Various supervised learning algorithms were considered and tested, such as logistic regression, random forest, Extra Trees and gradient boosting. Of these models, logistic regression performed best with an ROC-AUC; however, the overall discrimination was weak and just better than chance. Other measures of predictive value, such as sensitivity, specificity, F1-score, Matthews correlation coefficient, average precision, and calibration measures, also showed low predictive value. Weight, Use of nephrotoxic drugs, Serum creatinine, Diastolic blood pressure, and Water intake were the most influential factors identified in the explainability analysis. However, their absolute permutation-importance values were quite small, reflecting the weak and unstable contribution of each variable to the prediction of CKD. These findings were similar to the lack of significant baseline differences

between those who developed early CKD-flagged and non-CKD-flagged patients. The study underscores that strong predictor-outcome relationships, clear target construction and good quality data are essential for sophisticated machine-learning algorithms to succeed in the domain of weak relationships, uncertain target construction, or limited data quality. Explainability enhanced transparency in presenting how existing variables were used by the model, but failed to provide clinical reliability or causality. RenalSense should therefore be considered a methodological proof of concept, not a deployable clinical decision supporting system.

References

1. Bello AK, Okpechi IG, Levin A, Ye F, Damster S, Arruebo S, et al. An update on the global disparities in kidney disease burden and care across world countries and regions. *The Lancet Global Health*. 2024;12(3):e382–95.
2. Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. *Kidney international supplements*. 2022;12(1):7–11.
3. Mark PB, Stafford LK, Grams ME, Aalruz H, Abd ElHafeez S, Abdelgalil AA, et al. Global, regional, and national burden of chronic kidney disease in adults, 1990–2023, and its attributable risk factors: a systematic analysis for the Global Burden of Disease Study 2023. *The Lancet*. 2025;406(10518):2461–82.
4. Ma X, Liu R, Xi X, Zhuo H, Gu Y. Global burden of chronic kidney disease due to diabetes mellitus, 1990–2021, and projections to 2050. *Frontiers in endocrinology*. 2025;16:1513008.
5. Christofides EA, Desai N. Optimal Early Diagnosis and Monitoring of Diabetic Kidney Disease in Type 2 Diabetes Mellitus: Addressing the Barriers to Albuminuria Testing. *J Prim Care Community Health*. 2021 Jan;12:21501327211003683. doi:10.1177/21501327211003683
6. ElSayed NA, Aleppo G, Bannuru RR, Bruemmer D, Collins BS, Ekhlaspour L, et al. 11. Chronic kidney disease and risk management: Standards of care in diabetes—2024. *Diabetes Care*. 2024;47:S219.
7. Stevens PE, Ahmed SB, Carrero JJ, Foster B, Francis A, Hall RK, et al. KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney international*. 2024;105(4):S117–314.
8. Kalantar-Zadeh K, Jafar TH, Nitsch D, Neuen BL, Perkovic V. Chronic kidney disease. *The lancet*. 2021;398(10302):786–802.
9. Sanmarchi F, Fanconi C, Golinelli D, Gori D, Hernandez-Boussard T, Capodici A. Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review. *Journal of nephrology*. 2023;36(4):1101–17.
10. Segal Z, Kalifa D, Radinsky K, Ehrenberg B, Elad G, Maor G, et al. Machine learning algorithm for early detection of end-stage renal disease. *BMC Nephrol*. 2020 Dec;21(1):518. doi:10.1186/s12882-020-02093-0
11. Chen L, Shao X, Yu P. Machine learning prediction models for diabetic kidney disease: systematic review and meta-analysis. *Endocrine*. 2023 Dec 23;84(3):890–902. doi:10.1007/s12020-023-03637-8
12. Dholariya S, Dutta S, Sonagra A, Kaliya M, Singh R, Parchwani D, et al. Unveiling the utility of artificial intelligence for prediction, diagnosis, and progression of diabetic kidney disease: an evidence-based systematic review and meta-analysis. *Current Medical Research and Opinion*. 2024 Dec;40(12):2025–55. doi:10.1080/03007995.2024.2423737
13. Tan Y, Ma Y, Rao S, Sun X. Performance of deep learning for detection of chronic kidney disease from retinal fundus photographs: A systematic review and meta-analysis. *European Journal of Ophthalmology*. 2024 Mar;34(2):502–9. doi:10.1177/11206721231199848
14. Hassan MT. Chronic Kidney Disease dataset 2025 [Internet]. [cited 2026 Aug 6]. Available from: <https://www.kaggle.com/datasets/topu101/chronic-kidney-disease-dataset-2025>
15. Ghosh SK, Khandoker AH. Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific Reports*. 2024;14(1):3687.
16. Khan N, Raza MA, Mirjat NH, Balouch N, Abbas G, Yousef A, et al. Unveiling the predictive power: a comprehensive study of machine learning model for anticipating chronic kidney disease. *Frontiers in Artificial Intelligence*. 2024;6:1339988.
17. Khalid F, Alsadoun L, Khilji F, Mushtaq M, Eze-Odurukwe A, Mushtaq MM, et al. Predicting the progression of chronic kidney disease: a systematic review of artificial intelligence and machine learning approaches. *Cureus [Internet]*. 2024 [cited 2026 Aug 6];16(5). Available from: <https://www.cureus.com/articles/243965-predicting-the-progression-of-chronic-kidney-disease-a-systematic-review-of-artificial-intelligence-and-machine-learning-approaches.pdf>
18. Pan Q, Tong M. Artificial intelligence in predicting chronic kidney disease prognosis. A systematic review and meta-analysis. *Renal Failure*. 2024 Dec 31;46(2):2435483. doi:10.1080/0886022X.2024.2435483
19. Gogoi P, Valan JA. A multiclass machine learning framework for chronic kidney disease staging using CTGAN-based synthetic data augmentation and explainable AI. *Computer Methods in Biomechanics and Biomedical Engineering*. 2026 Jan 12;1–18. doi:10.1080/10255842.2025.2610677
20. Allgaier J, Mulansky L, Draelos RL, Pryss R. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artificial Intelligence in Medicine*. 2023;143:102616.
21. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon [Internet]*. 2023 [cited 2026 Aug 6];9(5). Available from: [https://www.cell.com/heliyon/fulltext/S2405-8440\(23\)03317-0](https://www.cell.com/heliyon/fulltext/S2405-8440(23)03317-0)
22. Kundu S. AI in medicine must be explainable. *Nature medicine*. 2021;27(8):1328–1328.
23. Markus AF, Kors JA, Rijnbeek PR. The role of

- explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*. 2021;113:103655.
24. Davis SE, Matheny ME, Balu S, Sendak MP. A framework for understanding label leakage in machine learning for health care. *Journal of the American Medical Informatics Association*. 2024;31(1):274–80.
25. Rizwan M, Naseem R, Khan MA, Ahmad A, Fozan M, Muhammad N. Explainable AI machine learning framework for chronic kidney disease prediction utilizing electronic health records. *BMC Med Inform Decis Mak*. 2026 Jul 3;26(1):244. doi:10.1186/s12911-026-03602-1
26. Zheng JX, Li X, Zhu J, Guan SY, Zhang SX, Wang WM. Interpretable machine learning for predicting chronic kidney disease progression risk. *DIGITAL HEALTH*. 2024 Jan;10:20552076231224225. doi:10.1177/20552076231224225