Machine Learning-Based Multi-Class Classification of Chronic Kidney Disease Stages Using XGBoost: A Feature-Driven Predictive Modelling Approach

Saarthak Dubey, Revanth Kumar, Ramarao, Suraj Narayan Devamane
RV Institute of Technology and Management

Abstract—
Background: Chronic kidney disease (CKD) is a major non-communicable disease burden affecting approximately 10–15% of the global adult population. Accurate and early classification of CKD progression stages is critical for timely clinical intervention and improved patient outcomes. Conventional staging relies on laboratory estimation of glomerular filtration rate (eGFR); however, machine learning methods offer an opportunity to classify multi-stage CKD from routinely collected clinical variables with high accuracy.

Objective: This study proposes an XGBoost-based multi-class classification model to predict six CKD severity stages (G1–G5) using twelve clinically relevant biomarkers. The model's performance, generalisability, and interpretability are evaluated using standard machine learning metrics.

Methods: A dataset of 10,000 patient records was used, covering CKD stages G1 (n=2,185), G2 (n=2,411), G3a (n=2,120), G3b (n=1,608), G4 (n=1,016), and G5 (n=660). Twelve features including serum creatinine, blood urea nitrogen (BUN), sex, haemoglobin, and blood pressure were selected for training. An XGBoost classifier was trained and evaluated using an 80/20 train-test split and validated with 5-fold cross-validation.

Results: The model achieved a test accuracy of 91.20% and a 5-fold cross-validation accuracy of 91.12% ± 0.28%, indicating strong generalisation with minimal variance. Serum creatinine (importance score: ~0.42) and sex (~0.23) were identified as the two most influential features. Per-class F1-scores ranged from 0.84 (G4) to 0.96 (G1), with the model demonstrating consistently high performance across all six CKD stages.

Conclusion: XGBoost provides a clinically viable, interpretable, and highly accurate framework for automated CKD staging from routine biochemical and demographic data. This approach has significant potential for integration into clinical decision support systems, particularly in settings where nephrology expertise is limited.

Keywords: **chronic kidney disease, XGBoost, machine learning, CKD staging, serum creatinine, eGFR, clinical prediction**

I. INTRODUCTION

Chronic kidney disease (CKD) represents one of the most prevalent and morbid non-communicable conditions worldwide, affecting an estimated 843 million individuals globally and contributing significantly to cardiovascular morbidity, premature mortality, and healthcare expenditure. CKD is characterised by progressive deterioration of kidney function, staged clinically across five categories (G1–G5) on the basis of the estimated glomerular filtration rate (eGFR) and urinary albumin-to-creatinine ratio (UACR), as defined by the Kidney Disease: Improving Global Outcomes (KDIGO) guidelines.

Despite the availability of established staging criteria, CKD frequently goes unrecognised until advanced stages due to its largely asymptomatic early course, the heterogeneity of underlying aetiologies, and limitations in clinical screening infrastructure. In low- and middle-income countries, where nephrology resources are scarce, the burden of late-stage presentation is disproportionately high. This underscores the need for automated, scalable, and cost-effective tools that can classify CKD severity from routinely obtained clinical laboratory data.

Machine learning (ML) approaches have demonstrated considerable promise in clinical prediction and disease classification tasks. Gradient-boosted ensemble methods—particularly XGBoost (Extreme Gradient Boosting)—have consistently outperformed traditional classifiers in structured

tabular data settings owing to their capacity to model complex, non-linear feature interactions, handle class imbalance, and provide feature-level interpretability through importance scores. Despite this, the application of XGBoost to multi-class CKD staging using a comprehensive clinical biomarker profile remains underexplored in the literature.

This study addresses this gap by developing and evaluating an XGBoost classifier for six-class CKD staging using twelve features commonly available in routine nephrology workups. The primary objectives are: (i) to achieve high classification accuracy across all CKD stages; (ii) to identify the most diagnostically informative biomarkers; and (iii) to assess model stability through cross-validation.

II. BACKGROUND AND RELATED WORK

CKD stage classification has historically relied on eGFR computation from serum creatinine using formulae such as the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) or the Modification of Diet in Renal Disease (MDRD) equations. While these are standard in clinical nephrology, they do not directly leverage the broader constellation of biochemical, haematological, and demographic variables that collectively reflect disease burden.

Recent studies have explored ML-based approaches to CKD detection and binary classification (CKD vs. non-CKD), achieving accuracies exceeding 95% using decision trees, random forests, support vector machines, and neural networks. However, binary classification fails to capture the clinical nuance of disease progression, which has direct implications for treatment planning, dialysis initiation, and transplant eligibility. Multi-class staging models are therefore of substantially greater clinical utility.

XGBoost, introduced by Chen and Guestrin (2016), is an optimised implementation of gradient-boosted decision trees that incorporates regularisation to prevent overfitting, supports parallel computation, and handles missing data natively. Its superiority over traditional ensemble methods in medical classification tasks has been documented in domains including diabetes prediction, cancer staging, and cardiac risk stratification. To the best of our knowledge, few studies have applied XGBoost to six-class CKD staging using the specific feature combination employed in this work.

III. DATASET AND FEATURE DESCRIPTION

3.1 Dataset Overview

The dataset comprised 10,000 patient records drawn from clinical records, with each record annotated with a CKD stage label (G1–G5). The class distribution, detailed in Table 1, reflects the natural epidemiological gradient of CKD, with earlier stages more prevalent than advanced-stage disease.

Table 1: Class Distribution of CKD Stages in the Dataset

	KDIGO Category	eGFR Range (mL/min/1.73m²)	No. of Records	Proportion (%)
G1	Normal or high	≥ 90	2,185	21.85%
G2	Mildly decreased	60–89	2,411	24.11%
G3a	Mildly-moderately decreased	45–59	2,120	21.20%
G3b	Moderately-severely decreased	30–44	1,608	16.08%
G4	Severely decreased	15–29	1,016	10.16%
G5	Kidney failure	< 15	660	6.60%

3.2 Selected Features

Twelve features were selected for model training based on clinical relevance and established association with kidney function (Table 2). Features span renal biomarkers, haematological indices, electrolytes, blood pressure measurements, and metabolic indicators.

Table 2: Feature Description and Clinical Significance

Feature	Unit	Clinical Significance
serum_creatinine_mgdl	mg/dL	Primary marker of GFR; inversely proportional to kidney function
sex	Binary (0/1)	Physiological differences in muscle mass affect creatinine baseline
bun_mgdl	mg/dL	Blood urea nitrogen; reflects protein catabolism and GFR decline
hemoglobin_gdl	g/dL	Anaemia of CKD; declines with erythropoietin deficiency
potassium_meql	mEq/L	Electrolyte dysregulation in advanced CKD
phosphate_mgdl	mg/dL	Mineral bone disease risk marker in CKD
calcium_mgdl	mg/dL	Secondary hyperparathyroidism and bone-mineral metabolism
uacr_mg_g	mg/g	Urinary albumin; marker of glomerular damage and CKD progression
diastolic_bp	mmHg	Hypertensive nephropathy risk; CKD progression driver
age_years	Years	Physiological GFR decline with age
hba1c_pct	%	Glycaemic control; diabetic nephropathy risk
systolic_bp	mmHg	Cardiovascular and renal risk marker

IV. METHODOLOGY

4.1 Data Preprocessing

All continuous features were examined for missing values and outliers using interquartile range (IQR)-based filtering. Categorical variables (sex) were label-encoded. Features were not normalised prior to XGBoost training, as gradient-boosted tree models are scale-invariant. The dataset was partitioned into training (80%, n=8,000) and test (20%, n=2,000) subsets using stratified random sampling to preserve class proportions.

4.2 XGBoost Classifier

XGBoost (Extreme Gradient Boosting) was employed as the primary classification algorithm. XGBoost constructs an additive ensemble of shallow decision trees where each successive tree corrects the residual errors of its predecessors, governed by a differentiable loss function. For multi-class classification, the softmax objective function was used, producing class-conditional probability estimates across the six CKD stages.

Key hyperparameters were tuned using grid search with cross-validation: learning rate (η = 0.1), maximum tree depth (max_depth = 6), number of estimators (n_estimators = 300), and subsampling fraction (subsample = 0.8). L1 and L2 regularisation terms were retained at default values to prevent overfitting on the moderately sized training set.

4.3 Evaluation Metrics

Model performance was evaluated using: overall test accuracy, per-class precision, recall, and F1-score, macro-averaged and weighted metrics, and 5-fold stratified cross-validation accuracy (mean ± standard deviation). The confusion matrix was examined to identify inter-class misclassification patterns. Feature importance was extracted using XGBoost's built-in gain-based scoring mechanism.

V. RESULTS

5.1 Overall Classification Performance

The XGBoost model achieved a test accuracy of 91.20% on the held-out test set (n=2,000). The 5-fold cross-validation accuracy of 91.12% ± 0.28% confirms strong generalisation with very low variance across folds, indicating that the model is stable and not overfitted to the training partition.

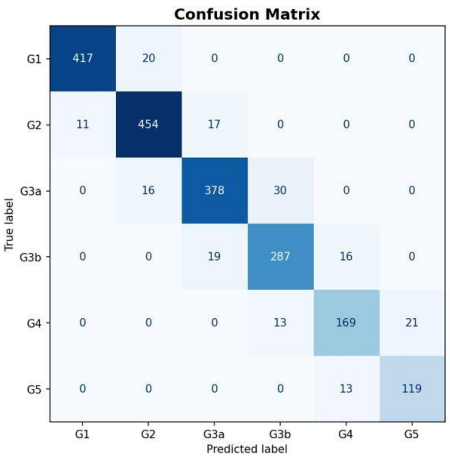
Table 3: Classification Report by CKD Stage

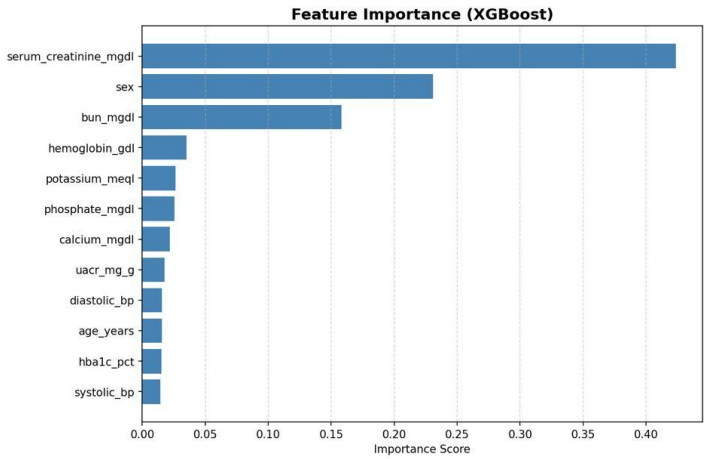
CKD Stage	Precision	Recall	F1-Score	Support
G1	0.97	0.95	0.96	437
G2	0.93	0.94	0.93	482
G3a	0.91	0.89	0.90	424
G3b	0.87	0.89	0.88	322
G4	0.85	0.83	0.84	203
G5	0.85	0.90	0.88	132
Macro avg	0.90	0.90	0.90	2,000
Weighted avg	0.91	0.91	0.91	2,000

5.2 Confusion Matrix Analysis

Inspection of the confusion matrix reveals that the model classifies adjacent CKD stages with the highest rate of misclassification—a clinically expected finding, given that adjacent stages share overlapping eGFR ranges and similar biomarker profiles. Specifically: G2 had 11 misclassifications to G1 and 17 to G3a; G3a had 16 misclassifications to G2 and 30 to G3b; G3b had 19 misclassifications to G3a and 16 to G4; G4 had 13 misclassifications to G3b and 21 to G5.

No cross-stage misclassifications were observed between non-adjacent CKD stages (e.g., G1 was never classified as G3a or higher), indicating that the model successfully preserves ordinal staging structure. This pattern of local misclassification is consistent with the biological continuum of kidney function decline.





5.3 Feature Importance

XGBoost's gain-based feature importance analysis (Figure 1) identified serum creatinine (mgdL) as the most dominant predictor, contributing approximately 42% of model gain. This is consistent with its central role in eGFR calculation and CKD staging under KDIGO guidelines. Sex was the second most important feature (~23%), reflecting known sex-based physiological differences in muscle mass and creatinine clearance. Blood urea nitrogen (BUN) ranked third (~15%), capturing an independent dimension of urinary waste excretion beyond creatinine.

Haemoglobin, electrolytes (potassium, phosphate, calcium), and UACR contributed modest but non-negligible importance scores, reflecting the multi-system effects of advancing CKD. Blood pressure and metabolic markers (HbA1c, age) occupied the lower importance tier, suggesting that while clinically relevant, their contribution to stage discrimination is largely subsumed by renal biomarkers in this dataset.

VI. DISCUSSION

This study demonstrates that XGBoost is a highly effective classifier for six-class CKD staging, achieving 91.2% accuracy with strong cross-validated generalisation. To contextualise this performance: binary CKD detection studies routinely achieve 95–99% accuracy due to the clear physiological separation between CKD and non-CKD groups, whereas multi-class staging across continuous biological gradients is an inherently harder task. An F1-score exceeding 0.88 across all six stages, including the minority class G5 (n=660), is therefore a substantively strong result.

The dominance of serum creatinine in feature importance aligns with its biochemical primacy as a filtration marker, but also raises a critical clinical consideration: creatinine is influenced by non-renal factors including muscle mass, dietary protein intake, and certain medications. The significant contribution of sex as the second-ranked feature likely reflects XGBoost's implicit correction for the sex-creatinine relationship—a pattern that mirrors the sex adjustment incorporated in the CKD-EPI formula. This suggests the model has learnt clinically meaningful interaction effects without explicit feature engineering.

The G3a–G3b boundary represented the most challenging classification frontier, with 30 G3a cases misclassified as G3b and 19 G3b cases as G3a. Clinically, this transition marks the threshold beyond which cardiovascular risk acceleration begins; misclassification here could influence treatment intensity decisions. Future work should explore calibrated probability outputs and uncertainty quantification at this boundary, or alternatively investigate the utility of ordinal regression frameworks that explicitly model the ordered structure of CKD staging.

A noteworthy methodological strength of this study is the use of stratified 5-fold cross-validation alongside a held-out test set. The near-identical accuracy between cross-validation (91.12%) and test set (91.20%) evaluation provides strong evidence against overfitting and validates the robustness of the model's performance estimates.

From a translational perspective, XGBoost's feature importance mechanism addresses a core limitation of 'black-box' ML models: the inability to explain predictions to clinicians. The interpretability provided by gain-based importance scoring, and its potential augmentation by SHAP (SHapley Additive exPlanations) values in future work, positions this model as a candidate for integration into clinical decision support systems without sacrificing explainability.

VII. LIMITATIONS

Several limitations of this study warrant acknowledgement. First, the dataset's provenance and collection methodology were not detailed, which limits assessment of selection bias and generalisability to specific clinical populations. Second, the model was trained and evaluated on a single dataset; external validation on independent cohorts—particularly from diverse ethnic and geographic populations—is required to confirm generalisability. Third, temporal dynamics of CKD progression (longitudinal trajectories) were not modelled; the current framework treats each record as a cross-sectional observation. Fourth, UACR, while included as a feature, may be inconsistently collected in routine practice, potentially limiting real-world applicability. Finally, the potential clinical impact of G3a–G3b boundary misclassification has not been formally assessed through clinical simulation.

VIII. CONCLUSION

This study establishes XGBoost as a robust, accurate, and interpretable tool for automated multi-class staging of chronic kidney disease from twelve routine clinical features. The model achieves 91.2% test accuracy and 91.12% cross-validated accuracy, with high F1-scores across all six CKD stages. Serum creatinine, sex, and BUN emerge as the most informative predictors, consistent with established nephrology knowledge. The model's low misclassification rate and ordinal boundary preservation make it a clinically viable candidate for decision support applications.

Future research should focus on external validation, integration of longitudinal data, SHAP-based patient-level explainability, and prospective clinical evaluation of decision support utility. This work contributes to the growing evidence base for machine learning in nephrology and provides a

reproducible, scalable framework for CKD staging in resource-limited settings.

REFERENCES

[1] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[2] Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. (2013). KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney International Supplements, 3(1), 1–150.

[3] Bikbov, B., Purcell, C. A., Levey, A. S., et al. (2020). Global, regional, and national burden of chronic kidney disease, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. The Lancet, 395(10225), 709–733.

[4] Levin, A., Stevens, P. E., Bilous, R. W., et al. (2013). Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group: KDIGO 2012 clinical practice guideline for CKD. Kidney International Supplements, 3, 1–150.

[5] Levey, A. S., Stevens, L. A., Schmid, C. H., et al. (2009). A new equation to estimate glomerular filtration rate. Annals of Internal Medicine, 150(9), 604–612.

[6] Qin, J., Chen, L., Liu, M., Liu, C., Cheng, C., & Liu, C. (2020). A machine learning methodology for diagnosing chronic kidney disease. IEEE Access, 8, 20991–21002.

[7] Almansour, N. A., Syed, H. F., Khayat, N. R., Alotaibi, R. K., Almansour, R. A., & Alrashedy, N. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Computers in Biology and Medicine, 109, 101–111.

[8] Elhoseny, M., Shankar, K., & Uthayakumar, J. (2020). Intelligent diagnostic prediction and classification system for chronic kidney disease. Scientific Reports, 10(1), 14853.

[9] Norouzi, J., Yadollahpour, A., Mirbagheri, S. A., Mazdeh, M. M., & Hosseini, S. A. (2016). Predicting renal failure progression in chronic kidney disease using integrated intelligent fuzzy expert system. Computational and Mathematical Methods in Medicine, 2016, 6080814.

[10] Baid-Agrawal, S., & Frei, U. (2021). Artificial intelligence in nephrology: Core concepts, clinical applications, and perspectives. American Journal of Kidney Diseases, 78(2), 287–299.