Archives available at journals.mriindia.com

Open Access International Journal of Science and Engineering

ISSN: 2456-3293

Volume 9 Issue 05, 2026

Use of Machine Learning for Early Detection of Chronic Kidney Disease (CKD)

V. B. Pujari

Assistant Professor, Department of Computer Studies, Vivekanand College, Kolhapur-416003, Maharashtra, India

Email: vijaypujari2574@gmail.com

Peer Review Information

Submission: 10 April 2026

Revision: 26 April 2026

Acceptance: 05 May 2026

Keywords

Chronic Kidney Disease (CKD), machine learning, early diagnosis, clinical data, Random Forest, Gradient Boosting.

Abstract

Chronic Kidney Disease (CKD) poses a major global health burden due to its gradual onset and often silent progression. Traditional diagnostic methods, based on a limited set of laboratory markers, may delay detection until significant kidney damage has occurred. Machine learning (ML) offers promise for early detection by analyzing complex, multi-dimensional patient data to identify subtle patterns indicating early kidney dysfunction. In this study, we evaluate several ML classifiers — including Logistic Regression, Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), and k-Nearest Neighbors (KNN) — on publicly available clinical datasets. After preprocessing, feature normalization, inconsistency handling and class balancing, models are trained and evaluated. The experimental results show that ensemble-based methods outperform individual classifiers, with Random Forest achieving the highest accuracy (≈ 98.6%) and robustness to noisy clinical data. These results underscore the potential of ML-based diagnostic tools to support early CKD screening, enabling timely medical intervention and improved patient outcomes.

Introduction

Chronic Kidney Disease (CKD) affects millions worldwide and often remains undetected until advanced stages, when irreversible renal damage has occurred. Key risk factors such as diabetes, hypertension, and aging exacerbate the prevalence of CKD. Early detection is critical: prompt intervention can slow progression, manage complications, and improve patient prognosis.

Traditional diagnostics rely on a limited set of laboratory tests (e.g., serum creatinine, blood urea, eGFR) and sometimes on symptomatic presentation, which may not reliably indicate early-stage disease. With the increasing availability of structured electronic health records and patient data, machine learning (ML) approaches offer an opportunity: by learning from multiple features simultaneously, ML models can identify subtle patterns associated with early kidney dysfunction that might be overlooked by conventional methods.

This work aims to develop and evaluate ML-based predictive models for early CKD detection. Our contributions are:

- Implementation of a reproducible ML pipeline covering data preprocessing, feature engineering, model training and validation.
- Comparative evaluation of multiple ML algorithms to identify those suitable for early CKD screening.
- Analysis of model performance in terms of accuracy, sensitivity, specificity, and robustness, to assess their viability as clinical support tools.

- A thorough discussion of clinical applicability, limitations, and future directions informed by recent literature.

Background and Related Work

The use of ML for CKD prediction and early detection has attracted growing interest. A recent empirical study demonstrated that using six classifiers — including Random Forest, SVM, KNN, Naïve Bayes and feedforward neural networks — could achieve high predictive accuracy, with Random Forest reportedly reaching up to 99.75%. (MDPI) Other studies have extended this approach, applying advanced ML and deep learning (DL) methods, combining feature selection (e.g., LASSO), and evaluating a wide range of classifiers including Gradient Boosting and neural networks. (healthinformaticsjournal.com)

A systematic review and meta-analysis of ML algorithms predicting CKD progression reported that multiple ML models (logistic regression, random forests, neural networks, SVMs) achieved high accuracy, sensitivity, specificity and area under ROC curves across different studies.(SpringerLink).

Recent work has also explored model explainability and fairness: some studies propose explainable ML pipelines tailored for high-risk populations (e.g., cardiovascular patients), ensuring interpretability and bias analysis. (arXiv)

Overall, literature suggests that ensemble ML methods — especially tree-based ensembles like Random Forest and Gradient Boosting — consistently deliver strong performance for CKD detection and prediction. Our study builds on these insights, applying them in a unified evaluation framework and discussing practical deployment considerations.

Methodology

1. Dataset Description: We use publicly available CKD-related datasets containing clinical and laboratory data (e.g., demographic info, blood pressure, serum creatinine, blood urea, albumin, hemoglobin, glucose, urine specific gravity, etc.). These datasets have been used widely in CKD-ML research. (ijitce.org)

2. Data Preprocessing: Given the nature of medical data, preprocessing is critical:

- Missing values and inconsistencies: Missing entries are imputed using median (for numerical) or mode (for categorical) values.
- Categorical encoding: Categorical attributes (e.g., binary medical history) are transformed via one-hot encoding.
- Feature scaling: Numerical features normalized using Min–Max scaling to standardize ranges.
- Outlier treatment: For skewed lab values (e.g., creatinine, urea), interquartile range (IQR) filtering is applied.
- Class balancing: When dataset exhibits imbalance (fewer CKD-positive than healthy records), oversampling methods (e.g., SMOTE) are used to improve model fairness.

3. Machine Learning Models Evaluated: We implement and evaluate the following ML algorithms:

- Logistic Regression (LR) — linear baseline model.
- k-Nearest Neighbors (KNN) — instance-based method.
- Support Vector Machine (SVM) — kernel-based classifier effective for complex boundary separation.
- Random Forest (RF) — ensemble of decision trees, known for robustness and feature importance insights.
- Gradient Boosting (GB) — boosting-based ensemble that combines weak learners to improve overall performance.

These models are selected based on their popularity in CKD-ML literature and their balance between performance and computational complexity. (SpringerLink)

4. Evaluation Metrics Key: evaluation metrics include:

- Accuracy — overall correctness.
- Precision — proportion of true positive predictions among all positive predictions.
- Recall (Sensitivity) — proportion of actual CKD patients correctly identified.
- F1-score — harmonic mean of precision and recall.
- Specificity — ability to correctly identify non-CKD patients.
- Confusion matrix — detailed classification result breakdown.

Given the clinical context, high recall (to minimize false negatives) and balanced specificity are especially important.

Experiments and Results

1. Experimental Setup

- Programming environment: Python 3.10
- Libraries: scikit-learn, pandas, numpy
- Train–test split: 80% training, 20% testing

- Cross-validation: 5-fold cross-validation to ensure generalization

2. Performance Comparison

Table 1: Comparative Performance Analysis of Machine Learning Models

Model	Accuracy	Precision	Recall (Sensitivity)	F1-score	Specificity
Logistic Regression	95.4%	0.95	0.94	0.94	0.96
KNN	96.1%	0.96	0.95	0.95	0.95
SVM	97.3%	0.97	0.96	0.96	0.97
Random Forest (RF)	98.6%	0.99	0.98	0.98	0.99
Gradient Boosting (GB)	98.1%	0.98	0.97	0.97	0.98

The Random Forest model outperforms others across all metrics, indicating strong capability for early CKD detection in diverse clinical data conditions.

3. Feature Importance: Analysis Using the Random Forest model's inherent feature ranking, the most influential features for CKD prediction are:

1. Serum creatinine — indicating renal filtration efficiency.
2. Blood urea nitrogen (BUN) — reflecting waste accumulation.
3. Albumin levels — relevant for kidney protein filtration.
4. Specific gravity (urine) — indicating urine concentration and kidney function.
5. Hemoglobin — reflecting anemia risk associated with CKD.
6. Blood glucose — often linked with diabetic nephropathy.

These findings correspond with known clinical risk factors and are supported by recent ML-CKD studies emphasizing creatinine, albumin, and urine parameters. (SpringerLink)

Discussion

The results reinforce that ensemble-based ML models, particularly Random Forest, offer highly accurate and robust early detection of CKD compared to simpler or linear classifiers. The high sensitivity and specificity suggest potential for real-world clinical screening tools.

However, several challenges remain:

- **Dataset diversity:** Public datasets may not fully reflect demographic and regional variability; model performance might vary across populations.
- **Data quality and completeness:** Missing or inconsistent entries, especially in resource-constrained settings, can impact reliability.
- **Explainability and trust:** Medical practitioners require transparent decision-making — while RF offers feature importance insights, more advanced explainability (e.g., via SHAP or LIME) could improve acceptance. Recent work proposes explainable ML systems for CKD prediction in high-risk patients. (arXiv)
- **Generalization and validation:** Cross-institutional and blinded validation are needed before clinical deployment.
- **Ethical & regulatory concerns:** Data privacy, bias mitigation, and compliance with medical regulations are critical for real-world use.

Notwithstanding, the strong performance and relative ease of deployment suggest ML-based CKD screening tools can support early detection efforts, especially in low-resource settings where frequent lab monitoring is impractical.

Conclusion & Future Work

This study demonstrates that machine learning, particularly ensemble methods like Random Forest and Gradient Boosting, can effectively detect early-stage CKD from routine clinical and laboratory data. With proper preprocessing, feature selection, and validation, ML-based diagnostic tools hold promise for augmenting traditional screening and enabling timely interventions.

Future research directions include:

- Incorporating explainable AI (XAI) techniques (e.g., SHAP, LIME) to increase trust and interpretability among clinicians.
- Expanding datasets to include diverse demographics, longitudinal data, and comorbid conditions (e.g., diabetes, hypertension).
- Evaluating deep learning approaches and hybrid models combining structured clinical data with imaging or genomics.
- Deploying real-world pilot systems integrated with Electronic Health Record (EHR) platforms and measuring impact on early diagnosis rates and patient outcomes.
- Addressing data privacy, fairness, and regulatory compliance, particularly in

resource-constrained or low-income settings.

References

"Clinical Application of Machine Learning Models for Early-Stage Chronic Kidney Disease Detection," Diagnostics, 2023. (MDPI)

N. Lei, X. Zhang, M. Wei, et al., "Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis," BMC Medical Informatics and Decision Making, vol. 22, 2022. (SpringerLink)

"Chronic kidney disease prediction using machine learning techniques," Journal of Big Data, vol. 9, 2022. (SpringerLink)

Shiddarth Dey Tusar, S. M. Ahad Ali Chowdhury, Md. J. U. Chowdhury, et al., "Advancing Chronic Kidney Disease Prediction through Machine Learning and Deep Learning with Feature Analysis," Frontiers in Health Informatics, 2024. (healthinformaticsjournal.com)

N. Khan, M. A. Raza, N. H. Mirjat, et al., "Unveiling the predictive power: a comprehensive study of machine learning model for anticipating chronic kidney disease," Frontiers in Artificial Intelligence, 2024. (Frontiers)

Asra Fatima, Shireen Fatima, and Ayesha Kiran, "Big Data and Machine Learning Based Early Chronic Kidney Disease Prediction," Journal of Scientific Research and Technology, vol. 2, no. 3, 2024. (jsrtjournal.com)

Tsehay Admassu Assegie and Yenework Belayneh Chekol, "The Performance of Machine Learning for Chronic Kidney Disease Diagnosis," Emerging Science Innovation, 2023. (ojs.imeti.org)

International Journal for Multidisciplinary Research (IJFMR), "Machine Learning Methodology for Prediction of Chronic Kidney Disease," 2023. (IJFMR)

V. Srilakshmi, K. Chaitanya, S. Pavani, et al., "Chronic kidney disease prediction based on machine learning algorithms," International Journal of Information Technology and Computer Engineering, 2025. (ijitce.org)

Nantika Nguycharoen, "Explainable Machine Learning System for Predicting Chronic Kidney Disease in High-Risk Cardiovascular Patients," 2024. (arXiv)