

Enhancing Chronic Kidney Disease Prediction Using Machine Learning Algorithms

By: hizam ahmed alsaeedi

Abstract:

Chronic Kidney Disease (CKD) is a serious health problem because it usually develops quietly and is often diagnosed too late. Catching it early can help reduce complications and improve how patients are treated. In this study, we plan to explore different machine learning techniques—such as Decision Trees, Support Vector Machines, Random Forests, and Deep Belief Networks—to create a model that predicts CKD with both high accuracy and ease of understanding. Our research will compare these models and show that using combined methods (like ensemble techniques and carefully choosing which features to include) can help doctors spot CKD sooner and plan treatment more effectively.

Chapter 1:

Introduction:

CKD affects millions worldwide, often leading to End-Stage Kidney Disease (ESKD) if left untreated. The challenge in early diagnosis stems from the absence of specific symptoms and reliance on traditional clinical tests. ML techniques present a unique opportunity to enhance predictive accuracy, leveraging patient records to identify risk factors associated with CKD progression.

Background:

Traditional CKD diagnosis relies on glomerular filtration rate (GFR), albumin levels, and biochemical markers. However, statistical models lack generalizability across diverse patient groups. Recent studies, including those by Debal & Sitote (2022) and Elkholy et al. (2021), have demonstrated ML's potential in CKD classification and stage prediction using advanced methods like DBN. Additionally, Tekale et al. (2018) and Kethineni et al. (2025) have explored ML-based CKD classification, emphasizing Decision Trees and Random Forests for clinical application.

Problem Statement:

Detecting CKD progression remains challenging due to its asymptomatic nature and reliance on conventional models. Existing risk equations show limited adaptability, prompting the need for advanced ML models that utilize clinical data for high-accuracy CKD prediction. This research aims to refine ML-based CKD models, improving sensitivity, interpretability, and clinical relevance.

Research Objectives:

1. Assess the predictive capability of ML algorithms for CKD detection.
2. Compare ML models to traditional statistical approaches for reliability.
3. Optimize feature selection methods to improve diagnostic accuracy.
4. Evaluate the clinical applicability of ML-based CKD screening tools.

Research Scope:

This study focuses on ML-driven CKD prediction using patient data from published research. Feature selection techniques and comparative model performance analyses are emphasized, ensuring results are clinically interpretable.

Chapter 2 (Literature review):

Literature Review:

Chronic Kidney Disease (CKD) and the Need for Predictive Models

CKD is a growing global health concern, affecting millions worldwide and often leading to End-Stage Kidney Disease (ESKD) if left untreated. Early detection is critical for implementing preventive strategies and improving patient outcomes. Traditional CKD diagnostic models rely on factors such as glomerular filtration rate (GFR) and albuminuria, but their effectiveness varies across populations due to demographic and genetic differences. As a result, machine learning (ML) techniques have emerged as an alternative, leveraging large datasets to uncover subtle patterns in patient health indicators.

Machine Learning for CKD Prediction

Recent advances in artificial intelligence and machine learning have facilitated the development of predictive models for CKD and ESKD. Several studies have explored different ML algorithms, each demonstrating varying levels of accuracy, sensitivity, and specificity:

- (Debal & Sitote, 2022) evaluated multiple ML models for CKD prediction, identifying Random Forest (RF) as the most effective algorithm when combined with recursive feature elimination and tenfold cross-validation. Their findings suggest that RF surpasses Decision Trees (DT) and Support Vector Machines (SVM) in predictive performance, reinforcing its potential for clinical applications.
- (Elkholy et al., 2021) introduced a Deep Belief Network (DBN) for early CKD prediction, using Softmax activation and Categorical Cross-Entropy as a loss function. Their approach demonstrated superior accuracy compared to traditional statistical models, emphasizing the benefits of deep learning architectures in disease classification.
- (Kethineni et al., 2025) examined CKD prediction models using Decision Trees and ensemble learning methods, reinforcing the clinical utility of ML-based prediction tools. Their study suggests that ensemble models improve interpretability, making them suitable for healthcare decision-making.
- (Tekale et al., 2018) analyzed machine learning applications in CKD classification, comparing SVM, Random Forest, and Decision Trees. Their findings indicate that ML models can significantly enhance CKD detection when trained on structured patient data, supporting their integration into clinical workflows.
- (Bai et al., 2022) investigated the feasibility of ML models in predicting CKD progression to ESKD, comparing logistic regression, naïve Bayes, Random Forest, Decision Tree, and K-Nearest Neighbors (KNN) against the Kidney Failure Risk Equation (KFRE). Their study found that ML models provide higher sensitivity, making them valuable for early-stage patient screening.

Feature Selection and Optimization in CKD Prediction Models

Feature selection plays a crucial role in enhancing ML model performance. Studies suggest that recursive feature elimination and correlation-based analysis improve prediction accuracy by identifying the most relevant health indicators:

- (Xiao et al., 2019) examined ML tools for CKD prediction and validated feature selection techniques that improve model efficiency. Their work highlights the importance of removing redundant variables to enhance interpretability and reduce computational complexity.
- (Tangri, 2011) developed the Kidney Failure Risk Equation (KFRE), widely used for CKD prediction. However, its applicability outside Western populations has been debated, leading researchers to explore ML alternatives.
- (Peeters et al., 2013) validated the KFRE in European CKD patients but identified limitations in non-Western cohorts. This underscores the need for adaptable models trained on diverse patient demographics.

Challenges and Future Directions

Despite ML's promising applications in CKD prediction, several challenges remain. One of the major limitations is the reliance on structured datasets, which may exclude real-world complexities such as missing values, unstructured clinical notes, and patient-specific variables. Additionally, external validation remains a critical step in ensuring ML models generalize effectively across diverse populations.

Future studies should focus on:

1. **Expanding datasets** to include more diverse populations.
2. **Incorporating additional biomarkers** such as inflammatory markers and genetic indicators.
3. **Enhancing model interpretability** to ensure medical practitioners can confidently use ML-based tools in clinical settings.

Chapter 3 (Research methodology)

Research Methodology:

1. Data Collection: Patient records from existing CKD studies.
2. Feature Selection: Recursive elimination and variance analysis.
3. Model Training: DT, RF, SVM, and DBN trained using fivefold cross-validation.
4. Performance Evaluation: Accuracy, precision, recall, and AUC scores compared to conventional models.

Expected Outcomes:

- ML models demonstrate higher sensitivity than traditional risk equations.
- Random Forest and Deep Belief Networks emerge as top-performing models.
- ML models provide real-world feasibility for clinical decision support.

Gantt Chart:

A structured timeline outlines key research milestones: data preprocessing, model training, validation, and analysis.

Key	Duration
data preprocessing	3 weeks
model training	3 weeks
validation	3 weeks
analysis	4 weeks

References

- Bai, Q., Su, C., Tang, W., & Li, Y. (2022). Machine learning to predict end stage kidney disease in chronic kidney disease. *Scientific Reports*, 12(1). <https://doi.org/10.1038/s41598-022-12316-z>
- Debal, D. A., & Sitote, T. M. (2022). Chronic kidney disease prediction using machine learning techniques. *Journal of Big Data*, 9(1). <https://doi.org/10.1186/s40537-022-00657-5>
- Elkholy, S. M. M., Rezk, A., & Saleh, A. A. E. F. (2021). Early Prediction of Chronic Kidney Disease Using Deep Belief Network. *IEEE Access*, 9, 135542–135549. <https://doi.org/10.1109/ACCESS.2021.3114306>
- Kethineni, Likitha., Nithinchandra, Kumar, N., & Sk, S. Sultana. (2025). Chronic Kidney Disease Prediction Based On Machine Learning Algorithms. *ITM Web of Conferences*, 74, 01004. <https://doi.org/10.1051/itmconf/20257401004>
- Peeters, M. J., van Zuilen, A. D., van den Brand, J. A. J. G., Bots, M. L., Blankestijn, P. J., Wetzels, J. F. M., Group, for the M. S., Blankestijn, P. J., Bots, M. L., van Buren, M., ten Dam, M. A. G. J., Kaasjager, K. A. H., Ligtenberg, G., Sijpkens, Y. W. J., Henk E., S., van de Ven, P. J. G., Vervoort, G., Vleming, L.-J., Wetzels, J. F. M., & van Zuilen, A. D. (2013). Validation of the kidney failure risk equation in European CKD patients. *Nephrology Dialysis Transplantation*, 28(7), 1773–1779. <https://doi.org/10.1093/ndt/gft063>
- Tangri, N. (2011). A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure. *JAMA*, 305(15), 1553. <https://doi.org/10.1001/jama.2011.451>
- Tekale, S., Shingavi, P., & Wandhekar, S. (2018). Prediction of Chronic Kidney Disease Using Machine Learning Algorithm. *IJARCCCE*, 7(10), 92–96. <https://doi.org/10.17148/IJARCCCE.2018.71021>
- Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., Zhu, S., & Ye, Z. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *Journal of Translational Medicine*, 17(1). <https://doi.org/10.1186/s12967-019-1860-0>