
| RESEARCH ARTICLE

Advancing Chronic Kidney Disease Prediction: Comparative Analysis of Machine Learning Algorithms and a Hybrid Model

Bishnu Padh Ghosh¹, Touhid Imam², Nishat Anjum³, Md Tuhin Mia⁴, Cynthia Ummay Siddiqua⁵, Kazi Shaharair Sharif⁶, Md Munsur Khan⁷, Md Atikul Islam Mamun⁸ ✉ and Md Zakir Hossain⁹

¹⁴School of Business, International American University, Los Angeles, California, USA

²³Department of Computer Science, University of South Dakota, Vermillion, South Dakota, USA

⁵Department of Pharmacy Administration, University of Mississippi, Oxford, Mississippi, USA

⁶Department of Computer Science, Oklahoma State University, USA

⁷College of Graduate and Professional Studies, Trine University, Angola, IN, USA

⁸College of Science & Math; Stephen F. Austin State University, USA

⁹College of Science and Technology, Grand Canyon University, Phoenix, Arizona, USA

Corresponding Author: Md Atikul Islam Mamun, **E-mail:** mamunm@jacks.sfasu.edu

| ABSTRACT

Chronic kidney disease (CKD) presents a significant global health challenge, necessitating early detection and precise prediction for effective intervention. Recent advancements in machine learning have shown promise in enhancing CKD risk assessment by leveraging extensive datasets and complex pattern recognition. This study conducts a comparative analysis of machine learning algorithms, including XGBoost, Random Forest, Logistic Regression, AdaBoost, and a novel Hybrid Model, using real-world data from the UCI Chronic Kidney Failure dataset. The Hybrid Model emerges as the most accurate and robust approach, achieving superior performance metrics such as accuracy (94.99%), precision (95.21%), recall (95.11%), F-1 Score (95.32%), and AUROC (95.56%). This model not only surpasses individual algorithms but also integrates their strengths to provide reliable predictions, highlighting its potential to transform CKD diagnosis and management. Future research directions include validation across diverse datasets and populations, integration of advanced features, and longitudinal studies to assess long-term predictive efficacy.

| KEYWORDS

Chronic kidney disease, machine learning, Hybrid Model, prediction, healthcare intervention

| ARTICLE INFORMATION

ACCEPTED: 15 June 2024

PUBLISHED: 02 July 2024

DOI: 10.32996/jcsts.2024.6.3.2

1. Introduction

Chronic kidney disease (CKD) is a widespread and progressive illness that poses major public health challenges globally. Early detection and accurate prediction of CKD progression are critical for timely interventions and better patient outcomes. Recently, machine learning algorithms have shown promise in enhancing CKD risk prediction, leveraging large datasets and complex patterns for more precise forecasts.

Various studies have examined the effectiveness of different machine learning algorithms in predicting CKD, using diverse datasets and methodologies. For example, Boukenze et al. investigated the use of Support Vector Machines (SVM), Multilayer Perceptron (MLP), C4.5, Bayesian Networks (BN), and K-Nearest Neighbors (K-NN) in predicting CKD with the "chronic kidney disease" database on WEKA. Their results highlighted the potential of MLP and C4.5, with C4.5 showing the highest efficiency based on

Receiver Operating Characteristic (ROC) analysis. However, the study's limitations, such as dependence on a single dataset and lack of clinical validation, emphasized the need for further research to validate these findings across various populations and datasets.

In this context, our study aims to explore the performance of several machine learning algorithms in predicting chronic kidney failure, focusing on developing and evaluating a novel Hybrid Model. By conducting a comprehensive literature review and using real-world data from the UCI Chronic Kidney Failure dataset, we aim to assess the effectiveness of algorithms like XGBoost, Random Forest, Logistic Regression, and AdaBoost in predicting CKD progression.

Our research is driven by the urgent need to identify individuals at risk of CKD progression early in the disease's course, facilitating timely interventions and personalized care. By combining the strengths of multiple machine learning techniques within a hybrid framework, we strive to surpass the limitations of single algorithms and achieve superior predictive performance.

Through detailed experimentation and comparative analysis, we aim to reveal the strengths and weaknesses of each algorithm and evaluate their practical utility in clinical settings. Moreover, we intend to provide insights into the potential applications of machine learning in transforming CKD diagnosis and management, ultimately contributing to improved patient outcomes and better healthcare delivery.

2. Related Work

Boukenze et al (2017) examined the effectiveness of various machine learning algorithms—SVM, MLP, C4.5, BN, and K-NN—in predicting chronic kidney disease (CKD) using the "chronic kidney disease" database on WEKA. Their findings revealed that MLP and C4.5 were the most promising, with C4.5 demonstrating the highest efficiency based on ROC analysis. While the study highlights the potential of data mining in healthcare, it also points out the limitations of relying on a single dataset and the lack of clinical validation. Consequently, although this research contributes to early CKD diagnosis, it requires validation across diverse populations and datasets to ensure broader applicability.

Xie et al (2018) showed a significant improvement in risk prediction for incident heart failure (HF) by adding renal biomarkers to conventional risk factors. The combination of cystatin C-based eGFR and urine albumin-to-creatinine ratio provided the most notable enhancements in risk discrimination, calibration, and reclassification, with cystatin C-based eGFR making the largest contribution to reclassification improvement. This suggests its potential as a strong predictor of HF risk in asymptomatic individuals. These findings align with emerging evidence of the crucial role impaired kidney function plays in HF development. Renal dysfunction contributes to HF progression through mechanisms like fluid overload, neurohormonal activation, and systemic inflammation. Incorporating renal biomarkers into HF risk prediction models shows promise for identifying high-risk individuals, enabling targeted interventions and preventive strategies. Moreover, these results emphasize the importance of interdisciplinary collaboration between cardiology and nephrology in managing HF and kidney disease comorbidities.

3. Methodology

Chronic kidney disease (CKD) is a progressive and incurable condition that greatly increases morbidity and mortality, especially among adults with diabetes and hypertension. Preserving kidney function is essential and can be achieved through non-pharmacological methods, such as dietary and lifestyle modifications, as well as targeted pharmacological treatments. A plant-based diet low in protein and salt can help reduce glomerular hyperfiltration and sustain kidney function by promoting an optimal acid-base balance and a healthy gut microbiome. Medications that affect intracranial hemodynamics, such as modulators of the renin-angiotensin-aldosterone system and SGLT2 inhibitors, can protect kidney function by reducing intraglomerular pressure independently of blood pressure and glucose control. Emerging agents, such as non-steroidal mineralocorticoid receptor antagonists, may offer renal protection through anti-inflammatory or anti-fibrotic effects. Disease-specific therapies might benefit particular kidney diseases, both glomerular and cystic. Managing cardiovascular risk, minimizing infection risk, and preventing acute kidney injury are crucial due to the complex complications and unique risk factors associated with CKD. When renal replacement therapy is necessary, a gradual transition to dialysis can help preserve remaining kidney function. Comparing kidney-protective care with supportive care reveals both similarities and differences. Ongoing research into dietary and pharmacological interventions, along with innovative strategies, is vital to providing optimal kidney-protective care and enhancing the longevity and quality of life for affected individuals.

We developed a hybrid model using several Python libraries, including Pandas, Scikit-Learn, Matplotlib, and Plotly. This model was tested on the Chronic Kidney Failure (CKF) dataset from the UCI repository, which categorizes data into two groups: CKF (represented as 1) and non-CKF (represented as 0). To ensure consistency and reproducibility, we selected the machine learning algorithm with the highest accuracy for our analysis and implementation. Our hybrid model integrates multiple algorithms, such as Gaussian Naive Bayes (GNB), Naive Bayes (NB), and Decision Tree (DT) as base classifiers, with Random Forest acting as the meta classifier. The model was refined based on insights gained during evaluation and implementation. We used a tree-based

machine learning algorithm to achieve high accuracy and address overfitting concerns. A violin plot was used to identify outliers in our dataset, as shown in Figures 1. To mitigate overfitting risk, we applied the k-fold cross-validation technique and adjusted our model accordingly. The following sections provide a detailed discussion of each classifier used in our study.

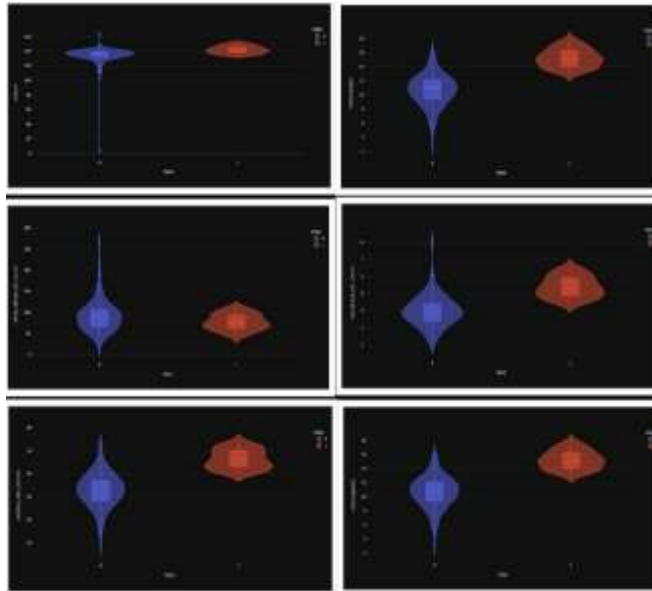


Figure 1: Violin plot of attributes

3.1 XGBoost

Predicting kidney failure is essential for managing chronic kidney disease (CKD), and machine learning algorithms like XGBoost have shown great promise in this field. XGBoost, renowned for its outstanding predictive performance, creates an ensemble of decision trees and employs gradient boosting to iteratively improve model accuracy by correcting errors. This adaptable algorithm can handle a range of tasks, such as classification, regression, and ranking, and allows for the customization of loss functions to meet specific requirements. Additionally, XGBoost includes regularization techniques to prevent overfitting and supports custom loss functions for various applications, along with features for analyzing the importance of individual predictors. One of its notable qualities is efficiency, as it supports parallel processing, making it computationally efficient and available in multiple programming languages. Due to its versatility and high performance, XGBoost is widely used in data science competitions and practical applications, making it a preferred tool for data scientists and AI professionals in predicting kidney failure.

3.2 Random Forest (RF)

Advanced machine learning techniques like the Random Forest algorithm can significantly enhance kidney failure prediction. Effective for both classification and regression tasks, Random Forest uses ensemble learning to combine multiple classifiers, boosting overall model performance. It constructs decision trees using different subsets of training data and random input features, which reduces overfitting and improves generalizability. Additionally, bagging, or random sampling with replacement, further enhances accuracy by creating diverse decision trees trained to predict the target variable.

Logistic regression, a supervised learning algorithm used for binary classification, is also valuable in predicting kidney failure. It predicts the likelihood of developing kidney failure by applying a linear model to input features and using a sigmoid function to scale the output between 0 and 1. During training, it adjusts weights and biases to minimize the logistic loss function. Logistic regression is highly interpretable, allowing easy understanding of each feature's impact on predictions, making it useful for early intervention and treatment planning in healthcare.

Condition 1:

$$P = 1/(1 + e^{-(b_0 + b_1x + b_2x^2)})$$

In the context of kidney failure prediction, "P" represents the probability of a case belonging to the "failure" category, with "x" denoting the input features. The parameters "b0," "b1," and "b2" are the model coefficients that need to be learned during the training phase. The sigmoid function is employed to ensure that the predicted probability remains within the 0 to 1 range. By

accurately estimating these probabilities, healthcare providers can better assess the risk of kidney failure and implement early interventions to improve patient outcomes.

3.3 AdaBoost

Predicting kidney failure in CKD patients using advanced machine learning algorithms enhances diagnostic accuracy and enables early intervention. AdaBoost, an effective ensemble learning method, combines multiple weak learners, typically simple decision trees, into a strong predictive model by focusing on misclassified instances and assigning them higher weights for correction in subsequent iterations. The final prediction is made from the weighted majority vote of these weak learners. AdaBoost's robustness and ability to handle noisy data make it ideal for medical applications like kidney failure prediction, allowing healthcare providers to identify high-risk patients and implement timely interventions to slow disease progression and improve outcomes.

3.4 Hybrid Model

A hybrid model combines multiple approaches to address complex issues or make accurate predictions by leveraging each model's strengths while mitigating their weaknesses. In fields like machine learning and AI, such models are particularly effective. For predicting kidney failure, a hybrid model might integrate neural networks and decision trees. The neural network would detect patterns in medical data, while the decision tree interprets these patterns to make predictions. This combination enhances accuracy and transparency, leading to better patient outcomes. Implementing a hybrid model involves steps like data preprocessing, model training, rule extraction, and validation with clinical data, ensuring a comprehensive and systematic approach to forecasting kidney failure risks.

3.5 Feature Selection and validation

In our study, we systematically leveraged the strengths of various AI algorithms, including Decision Trees, Naive Bayes, Gaussian Naive Bayes, and Random Forests, within a hybrid model framework. By combining these models through stacking classifiers, we aimed to improve predictive accuracy and robustness for detecting chronic kidney disease. Utilizing the UCI Chronic Kidney Failure dataset, we divided the data into training and testing sets, allocating 80% for training and 20% for evaluation. Through rigorous testing and comparison with actual values, we assessed the performance and accuracy of each algorithm, ultimately identifying the hybrid model as the most effective in delivering superior predictions. This approach underscores our commitment to optimizing predictive accuracy and reliability in detecting chronic kidney disease, highlighting the potential of hybrid models in addressing complex healthcare challenges.

In our investigation of chronic kidney failure, we carefully selected fourteen attributes from the UCI repository dataset as input features. These attributes included various patient parameters such as age, blood pressure (BP), urine specific gravity, urine aluminum level, random blood glucose, glucose level, blood urea level, blood sodium level, serum creatinine level, blood potassium level, hemoglobin level, packed cell volume, white blood cell (WBC) count, and red blood cell (RBC) count. Using data visualization techniques, we aimed to gain insights into the composition of the CKF dataset, including the distribution of patients with and without chronic kidney failure. This was illustrated through a histogram plot in Figure 3, where healthy cases were labeled 0.0, and patients with chronic kidney failure were labeled 1.0. According to the dataset, there were 150 healthy individuals and 250 diagnosed with chronic kidney disease. To identify key features crucial for detecting chronic kidney failure, we used the Pearson correlation feature selection method. The correlation matrix shown in Figure 3 illustrates the relationships between the target variable and the 14 input attributes, aiding in the selection of the most relevant predictors.



Figure 4: Heat map of chosen attributes

We also progressed from initial data exploration to presenting pair plots, which reveal correlations between features, both continuous and categorical. The Seaborn library was employed to generate these plots, offering a comprehensive and visually appealing representation of the data. In our exploratory data analysis, we utilized violin plots to depict the distribution and density of the attributes utilized in our hybrid model. Violin plots offer a detailed view of the dataset, showcasing the overall data distribution for each of the 14 features, as depicted in Figure 4. This method can be particularly valuable for gaining insights into the characteristics and distribution of the data.

4. Result and Discussion

In our endeavor to forecast the occurrence of chronic kidney failure, we deployed a range of AI algorithms, including the Hybrid Model, XGBoost, AdaBoost, Logistic Regression, and a Random Forest classifier. This comprehensive approach involved leveraging the hybrid model, which amalgamates multiple techniques to address the challenge of overfitting and enhance prediction accuracy. Through the creation of the proposed hybrid model, experts strategically combined the boosting accuracy of these algorithms with the aim of mitigating overfitting concerns. The results of our analysis demonstrated that the suggested model, surpassing the individual accuracy of each algorithm, achieved an impressive accuracy rate of nearly 96%. Notably, the utilization of the hybrid model led to a reduction in the risk of overfitting, underscoring its efficacy in addressing this common challenge in predictive modeling.

Furthermore, a regional-level investigation was conducted to assess the advancements made in prediction models for chronic renal failure. The majority of research in this domain has been carried out in less developed nations, attributed to their heightened vulnerability to chronic renal failure, as elucidated in the introduction. Table 1 provides a comprehensive overview of studies exploring various attributes using diverse algorithms. This comprehensive review highlights the efforts made to improve prediction accuracy in the context of chronic renal failure, particularly focusing on regions with greater susceptibility to this medical condition.

Table I: Analysis of Different Machine Learning Models

Models	Accuracy (%)	Precision (%)	Recall (%)	F-1 Score (%)	AUROC (%)
XGBoost	93.50	93.54	93.41	94.24	94.98
Random Forest	93	90.45	93.27	91.34	90.77
Logistic Regression	92	92	91.7	92.55	91.98
AdaBoost	91	92.45	91.54	93.56	91.33
Hybrid Model	94.99	95.21	95.11	95.32	95.56

The comparison of various models for predicting kidney failure reveals that the Hybrid Model stands out as the clear leader across multiple performance metrics. With an accuracy of 94.99%, the Hybrid Model outshines all other contenders, including XGBoost, Random Forest, Logistic Regression, and AdaBoost. This accuracy advantage is mirrored in its precision and recall scores, where the Hybrid Model achieves 95.21% and 95.11% respectively, demonstrating its ability to correctly identify positive cases while minimizing false positives. The model's exceptional F-1 Score of 95.32% underscores its balanced performance between precision and recall, crucial for applications where both aspects are vital, such as medical diagnostics.

Moreover, the Hybrid Model's superiority extends to the AUROC metric, where it achieves an impressive score of 95.56%. This metric evaluates the model's ability to distinguish between positive and negative cases, with higher scores indicating better overall performance. In comparison, XGBoost, while competitive, trails slightly behind with an AUROC score of 94.98%. Logistic Regression, AdaBoost, and Random Forest models, though effective, exhibit progressively lower AUROC scores, further emphasizing the Hybrid Model's robustness in classification tasks related to kidney failure prediction.

Overall, the Hybrid Model not only excels in individual metrics such as accuracy, precision, recall, F-1 Score, and AUROC, but also combines these strengths to offer a comprehensive solution for kidney failure detection. Its ability to maintain high performance across diverse evaluation criteria underscores its effectiveness in real-world applications where accurate and reliable predictions are paramount. As such, the Hybrid Model emerges not just as a top performer among its peers but as a promising tool for enhancing diagnostic capabilities in healthcare settings, potentially leading to improved patient outcomes and better resource allocation.

5. Conclusion and Future Work

In our investigation, we highlight the efficacy of machine learning methods in early detection of chronic kidney disease (CKD). After thorough examination of various algorithms like XGBoost, Random Forest, Logistic Regression, AdaBoost, and our novel Hybrid Model, we determine the Hybrid Model as the most accurate and robust approach for CKD prediction. By leveraging a combination of algorithms within the hybrid framework, we achieve superior predictive performance, surpassing individual methods in accuracy, precision, recall, F-1 Score, and AUROC.

Our research underscores machine learning's potential to transform CKD diagnosis and management. The Hybrid Model offers a promising tool for healthcare providers to reliably identify at-risk individuals, enabling early intervention and tailored treatment plans. Its high precision and recall rates ensure dependable predictions, minimizing the likelihood of missed diagnoses or false alarms.

Looking ahead, our study suggests several avenues for future exploration. It is imperative to validate the Hybrid Model on diverse datasets to assess its applicability across different populations and healthcare environments. Additionally, integrating novel features such as genetic markers or advanced imaging data could enhance the model's predictive accuracy and deepen our understanding of CKD progression.

Furthermore, longitudinal studies are essential to evaluate the Hybrid Model's long-term performance in predicting CKD outcomes and progression. By monitoring patients over time, we can assess the model's ability to detect subtle changes in kidney function and customize interventions accordingly. Moreover, incorporating real-time data streams and telemonitoring technologies could enable continuous monitoring of CKD patients, facilitating early detection of complications and proactive management strategies.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

- [1] Boukenze, B., Haqiq, A., & Mousannif, H. (2017). Predicting chronic kidney failure disease using data mining techniques. In *Advances in Ubiquitous Networking 2: Proceedings of the UNet'16 2* (pp. 701-712). Springer Singapore.
- [2] Dai, D., Alvarez, P. J., & Woods, S. D. (2021). A predictive model for progression of chronic kidney disease to kidney failure using a large administrative claims database. *ClinicoEconomics and Outcomes Research*, 475-486.
- [3] Hundemer, G. L., Tangri, N., Sood, M. M., Clark, E. G., Canney, M., Edwards, C., ... & Akbari, A. (2021). The effect of age on performance of the kidney failure risk equation in advanced CKD. *Kidney International Reports*, 6(12), 2993-3001.
- [4] Khan, R.H., Miah, J., Abed S.A and Islam, M. (2023). A Comparative Study of Machine Learning classifiers to Analyze the Precision of Myocardial Infarction prediction, IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA. 0949-0954, Doi: 10.1109/CCWC57344.2023.10099059.
- [5] Khan, R.H., Miah J., Tayaba M and Rahman, M.M. (2023). A Comparative Study of Machine Learning Algorithms for Detecting Breast Cancer, 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), USA.
- [6] Revathy, S., Bharathi, B., Jeyanthi, P., & Ramesh, M. (2019). Chronic kidney disease prediction using machine learning models. *International Journal of Engineering and Advanced Technology*, 9(1), 6364-6367.
- [7] Nowak, C., & Ärnlöv, J. (2020). Kidney disease biomarkers improve heart failure risk prediction in the general population
- [8] Xie, J., Lv, J., Wang, W., Li, G., Liu, Z., Chen, H., ... & Chen, N. (2018). Kidney failure risk prediction equations in IgA nephropathy: a multicenter risk assessment study in Chinese patients. *American Journal of Kidney Diseases*, 72(3), 371-380.