Austin Journal of Clinical Case Reports

Open Access

Austin Publishing Group

Special Article: Case Reports in Hypertension

Prognostic Modeling of Chronic Kidney Disease Progression: Bridging Mild and Severe Stages through a Machine Learning Approach

Karamo Bah[1*]; Amadou Wurry Jallow[2]; Adama Ns Bah[3]

[1]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taiwan

[2]Department of Medical Laboratory Science and Biotechnology, Taipei Medical University, Taiwan

[3]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taiwan

***Corresponding author: Karamo Bah**

Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei 11031, Taiwan.

Email: kamasbah@gmail.com

Received: October 31, 2023

Accepted: November 30, 2023

Published: December 07, 2023

Abstract

Background and Aim: Chronic Kidney Disease (CKD) is a condition where the kidneys gradually lose their ability to function properly over time. It is into stages based on the severity of kidney damage and the level of kidney function. The objective of our study is to employ machine learning models for the prediction of Chronic Kidney Disease (CKD) progression.

Methods: Our study is centered on the prediction of CKD progression from mild (I, II, III) to advanced stages (IV, V, VI). We utilized logistic regression with a lasso-penalized approach and random forest model for our predictive analysis. We assessed the significance of features using the Gini index derived from the random forest model. The performance of our models was evaluated based on the Area Under Receiver Operating Characteristic (AU-ROC), AU-Precision-Recall (PR) curves, recall, precision and accuracy.

Results: Our study showcases remarkable predictive performance of CKD progression from milder (I, II, III) to severe stages (IV, V, VI). Random forest model achieved an accuracy of 85%, a recall rate of 86%, a precision rate of 83%, an AU-ROC score of 92%, and an AU-PR score of 83%. The logistic regression model exhibited an accuracy of 84%, a recall rate of 84%, a precision rate of 85%, an AU-ROC score of 92%, and an AU-PR score of 81%. Regarding variable importance, our model identifies creatinine as the most critical feature, followed by eGFR.

Conclusion: Our findings indicate that machine learning models hold promise in predicting CKD progression with substantial discriminative capabilities, as evidenced by high AUROC curves. This suggests their potential utility in real-world clinical settings for identifying patients at risk of transitioning from mild to severe stages of CKD.

Keywords: Chronic Kidney Disease; Machine Learning; Logistic Regression; Random Forest; Classification Model

Introduction

In 2013, the toll of Chronic Kidney Disease (CKD) claimed the lives of approximately one million individuals [1]. This burden disproportionately afflicts the developing world, where low to middle-income nations bear the weight of 387.5 million CKD cases, comprising 177.4 million male patients and 210.1 million female patients [2]. These statistics underscore the pervasive nature of CKD within developing regions, and the prevalence continues to surge.Chronic Kidney Disease (CKD) stands as a significant medical issue affecting numerous individuals worldwide. This condition entails the gradual deterioration of kidney function, leading to a reduced capacity to efficiently filter waste and excess fluids from the bloodstream, a process vital for urine production [3]. The term "chronic" is applied due to the slow, often extended, progression of this damage. CKD's global impact underscores its status as a pressing concern in healthcare, touching the lives of people across the globe. CKD represents a widespread and severe medical condition [4], characterized by a gradual decline in kidney function, a process that typically unfolds over months to years [3]. One distinguishing aspect of CKD is its silent nature, with symptoms often remaining latent until

Citation: Bah K, Jallow AW, Bah AN. Prognostic Modeling of Chronic Kidney Disease Progression: Bridging Mild and Severe Stages through a Machine Learning Approach. Austin J Clin Case Rep. 2023; 10(9): 1310.

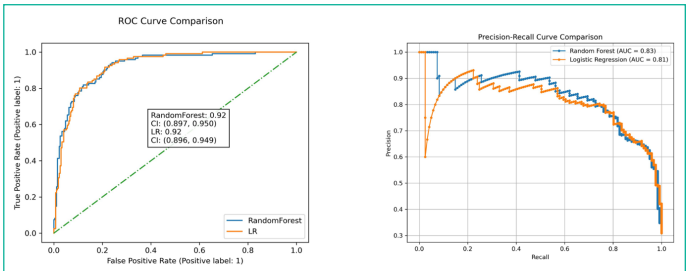
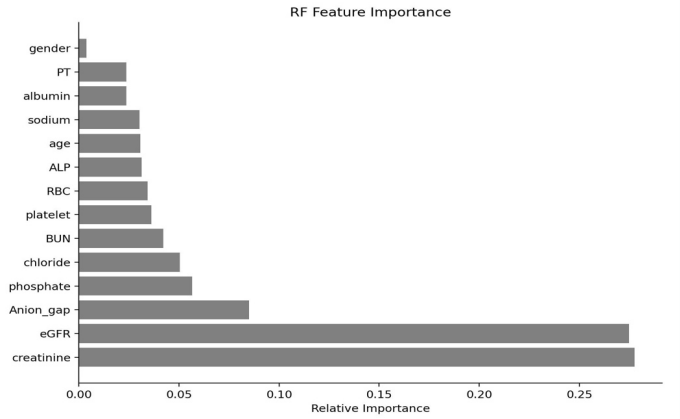


Figure 1: AU-ROC and AU-PR of the machine learning models.



RBC: Red Blood Cell; ALP: Alkaline Phosphate; BUN: Blood Urea Nitrogen; CKD: Chronic Kidney Disease; PT: Prothrombin Time; eGFR: Estimated Glomerular Filtration Rate.

Figure 2: Feature importance from machine learning models.

the disease reaches advanced stages [5]. One distinguishing aspect of CKD is its silent nature, with symptoms often remaining latent until the disease reaches advanced stages.

In recent years, the adoption of Electronic Health Records (EHRs) has witnessed significant growth within healthcare systems [6]. This wealth of electronic health data has ushered in unprecedented opportunities for computational methodologies. These approaches not only serve to enhance our existing understanding of various medical conditions but also enable the development of predictive models for assessing patient risk. For instance, conditions like breast cancer [7] and myocardial infarction [8] have already witnessed successful modelling through the application of machine learning algorithms. Machine learning, a subfield of artificial intelligence, is dedicated to crafting algorithms that can discern patterns or relationships within a set of variables [9]. These algorithms are adept at predicting the value or outcome of an unknown variable based on the information gleaned from historical data. In the realm of healthcare, machine learning models can be effectively harnessed to forecast a patient's susceptibility to a particular disease by analyzing the wealth of information housed within their health records. Furthermore, the output of machine learning algorithms isn't merely a black box; it often provides insights that can be manually scrutinized. This examination aids in deciphering which specific variables play pivotal roles in indicating diverse patient outcomes. Extensive efforts have been dedicated to the early detection of CKD, to initiate treatment in its nascent stages.

The objective of your study is to employ machine learning algorithms, lasso penalized logistic regression [10] and random forests [11] in the context of prediction and risk factor analysis for Chronic Kidney Disease (CKD) progression. Specifically, the focus is on forecasting the transition of CKD from its milder stages (I, II, III) to advanced, severe stages (IV, V, VI). The study seeks to enhance our comprehension of this progression phenomenon across diverse disease stages. The potential ramifications of achieving this goal include the advancement of early intervention strategies and improvements in patient care within the context of CKD management. Through this empirical exploration, we anticipate unravelling deeper insights into the mechanisms governing CKD progression. These insights, in turn, have the potential to equip medical practitioners with tools to refine risk assessment, enabling more timely interventions and tailored patient care strategies. This study aspires to contribute substantively to the enhancement of clinical decision-making, ultimately leading to improved patient outcomes. This model will uncover the salient variables exerting the most significant influence on the transition process.

Related Studies

Chronic Kidney Disease (CKD) is a pervasive and serious global health issue that poses a significant burden on healthcare systems. The condition is characterized by a gradual decline in kidney function over time, with five stages ranging from mild to severe. As CKD advances, it can lead to complications like cardiovascular disease and End-Stage Renal Disease (ESRD), necessitating dialysis or kidney transplantation.

Leveraging machine learning and data mining methods, researchers have embarked on a diverse range of studies aimed at extracting valuable insights from datasets related to Chronic Kidney Disease (CKD) [12]. The adoption of machine learning serves a twofold purpose: to streamline the analytical process, reduce time requirements, and enhance prediction accuracy through data mining categorization techniques [13]. Furthermore, the application of machine learning extends to the realms of disease diagnosis and treatment, encompassing a spectrum of medical conditions. Employing data-gathering techniques, a multitude of endeavors have been undertaken to extract valuable insights from CKD datasets. Numerous studies have been done using machine learning.

Bemando et al. delved into an exploration of the intricate relationship between blood-related diseases and their distinctive characteristics. Employing a range of classifier methods including Gaussian Naive Bayes, Bernoulli Naive Bayes, and Random Forest, these researchers brought forth compelling insights. Notably, in their investigation, Naive Bayes exhibited remarkable accuracy, surpassing other algorithms [14]. In a distinct avenue of medical research, Kumar and Polepaka crafted an innovative approach to predict illnesses. Their arsenal included powerful tools like Random Forest and Convolutional Neural Networks (CNN), alongside other machine learning methodologies. These algorithms demonstrated notable prowess in classifying illness datasets, delivering precision, recall, and F1-score metrics of excellence. Intriguingly, Random Forest stood out, showcasing superior accuracy and statistical performance [15]. The pursuit of enhanced statistical analysis outcomes led Acharya et al. to navigate the landscape of medical-linked illness datasets. Employing a multifaceted approach that included Convolutional Neural Networks (CNN) and an array of machine learning algorithms, they ventured into the realm of ECG datasets. Here, they achieved a commendable classification accuracy rate of 94% [16]. In the domain of medical illness prediction, Desai et al. devised a sophisticated methodology. The author harnessed the capabilities of both back-propagation Neural Networks (NN) and Logistic Regression (LR) classification algorithms. These strategic choices yielded distinctive outcomes, with a comprehensive statistical analysis concluding that logistic regression outperformed other algorithms in terms of accuracy and predictive capabilities [17]. Patil et al. undertook the creation of a comprehensive database dedicated to ECG arrhythmia-related

medical conditions. Within this endeavour, the researchers harnessed the potential of machine learning approaches, including Support Vector Machine (SVM) and the ingenious Cuckoo Search-Optimized Neural Network. The results were impressive, with the support vector machine yielding an enhanced accuracy rate of 94.44% [18].

Methods

Data Source

In this retrospective study, we conducted a comprehensive analysis using data sourced from the Medical Information Mart for Intensive Care (MIMIC) repositories. These repositories house a vast collection of de-identified health-related information about critically ill patients admitted to the Beth Israel Deaconess Medical Center, a leading tertiary medical institution located in Boston, USA [19].

The dataset at our disposal encompasses a diverse range of variables, including demographic details, vital signs, laboratory results, prescription records, and clinical notes. These data sources offer invaluable insights into the profiles of critically ill patients.

For this investigation, we focused specifically on the latest iteration of the MIMIC databases, namely MIMIC-III v1.4. This clinical database spans a timeframe from 2001 to 2012, incorporating data recorded through two distinct systems: MetaVision (iMDSoft, Wakefield, MA, USA) and CareVue (Philips Healthcare, Cambridge, MA, USA). It's noteworthy that the initial Philips CareVue system, which archived data from 2001 to 2008, was subsequently succeeded by the more advanced MetaVision data management system. The MetaVision system continues to be actively employed for data management and analysis to this day.

Table 1: Description of the patient populations.

	Mild stages			Severe stages		
	stage I	stage II	stage III	stage IV	stage V	VI (ESRD)
ICD-9 Code (s)	585.1	585.2	585.3	585.4	585.5	585.6
# Cases	13	104	557	225	59	1,002
Total Patients	Mild (n=674)			Severe (n=1,286)		

ESRD: End-Stage Renal Disease

Table 2: Baseline characteristics.

CKD patients characteristic	n=1,960
Age in years *median (min. – max.)*	69(20–88)
Gender (Male) n (%)	1,185(60)
Platelet (K/uL) *median (IQR)*	205(152.2–270.4)
RBC (m/uL) *median (IQR)*	3.27(3.00–3.59)
Albumin (g/dL) median *(IQR)*	3.10(2.82–3.35)
ALP (IU/L) *median (IQR)*	96.5(78.0–124.5)
Anion gap *median (IQR)*	15.76(13.7–18.0)
Chloride (mEq/L) median *(IQR)*	102(98.5–105.7)
Creatinine (mg/dL) *median (IQR)*	3.14(1.86–5.2)
Phosphate (mg/dL) *median (IQR)*	4.05(3.40–4.95)
Sodium (mEq/L) *median (IQR)*	138.6(136.2–140.7)
BUN (mg/dL) *median (IQR)*	43.24(30.7–59.4)
PT (s) *median (IQR)*	14.5(13.13–17.4)
eGFR (mL/min/1.73 m²) *median (IQR)*	17.93(10.2–32.8)

RBC: Red Blood Cell; ALP: Alkaline Phosphate; BUN: Blood Urea Nitrogen; CKD: Chronic Kidney Disease; PT: Prothrombin Time; eGFR: Estimated Glomerular Filtration Rate; IQR: Interquartile Range; mg/dL: Milligrams per Deciliter; IU/L: International Units per Litre; K/uL: Thousand per Microliter; m/uL: Million per Microliter, %: Percentage; min: Minimum; max: Maximum; mEq/L: Milliequivalents per Liter; s: Seconds. Continuous values were recorded as median (IQR), and categorical values (absolute numbers and percentages).

Table 3: Classification Metrics from the Machine Learning Models.

Models	Accuracy	Recall	Precision	AU-ROC	AU-PR
Random Forest Model	85	86	83	92	83
Logistic Regression Model	84	84	85	92	81

AU-ROC: Area Under Receiver Operating Characteristic; AU-PR: Area Under Precision-Recall

Patients Population

The patients in this study were selected based on their ICD-9 codes, which are a standardized way of categorizing medical conditions and diagnoses. In the context of Chronic Kidney Disease (CKD), the ICD-9 codes used in this study represented different stages of the disease. Among the patient cohort, 674 individuals exhibited mild stages (I, II, III) of CKD, indicative of milder manifestations of this condition. Furthermore, a group of 1,286 patients received diagnoses reflecting severe stages (IV, V, VI) of CKD. The distribution of patients within these distinct categories is detailed in Table 1.

Model Construction Methods

To assess the predictive capabilities of our models, we adopted two distinct machine learning approaches: random forests and logistic regression with a lasso-penalized approach.

Random forest models are well-regarded for their exceptional accuracy and robustness in handling high-dimensional data [20]. Moreover, they excel at capturing intricate nonlinear relationships within the data [21]. However, one challenge with random forests lies in their interpretability, which is often a critical factor in healthcare contexts. To address this concern, we extended our exploration to logistic regression models.

Logistic regression, although not naturally suited for high-dimensional data, can be enhanced with a lasso-penalized approach. This lasso, or L1-regularization, incorporates a penalty term into the model's objective function. Its purpose is twofold: first, it penalizes features that provide only marginal information, and second, it encourages the selection of a concise set of highly predictive features for the final model [22]. This strategy is essential for achieving both comprehensible and accurate models, especially when dealing with patient datasets containing numerous unique features. Logistic regression without dimensionality reduction can yield suboptimal results in such complex scenarios.

The L1 or "lasso" penalty has gained widespread acceptance as an effective method for dimensionality reduction in regression. To validate the performance of our models, we employed leave-one-out cross-validation for both random forests and logistic regression.

In the case of random forests, the models were constructed using the remaining $k-1$ folds of training data. We created 500 decision trees for each forest and determined the number of features to consider at each split as the square root of the total number of features. Trees were grown until they reached leaf purity whenever possible. Additionally, we used a balanced approach when calculating the Gini gain for split decisions. This involved assigning greater weight to the lower class to balance their influence with the majority class.

In logistic regression analysis, we introduced an additional layer of internal cross-validation aimed at fine-tuning the pen-

alty coefficient for L1 regularization. This involved a multi-step process. We performed $k-1$ round of internal cross-validation. Each round consisted of 1 tuning fold and $k-2$ training folds. During this process, we explored 10 different penalty values distributed logarithmically between 10^-4 and 10^4. Evaluation of these penalty values occurred within an internal cross-validation layer. The optimal penalty value was chosen based on model performance, and assessed using a weighted accuracy metric. This metric ensured equal consideration of two outcome classes. For each external cross-validation fold, the remaining folds were utilized to determine the most suitable penalty value. A new predictive model was then trained using these selected folds. Finally, this model was evaluated on the original held-out fold.

It's important to note that, similar to our random forest models, logistic regression models also employed a balanced class weight approach to ensure fair treatment of both the two outcome classes.

Furthermore, to enhance reliable generalization and reduce the risk of overfitting, we adopted a hybrid method that blends both holdout and cross-validation techniques. Our dataset was divided into separate segments: 80% for training purposes and 20% for autonomous testing.

Model Evaluation Methods

In our model evaluation, we utilize established quantitative metrics such as Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) curves. These metrics provide a comprehensive assessment of model performance [23,24]. To obtain predicted probabilities for each patient, we employ k-fold cross-validation. The final probability vector is constructed by combining these predictions while ensuring each patient's inclusion only once.

Our primary metrics of interest are the Area Under the ROC curve (AUC-ROC) and the area under the PR curve (AUC-PR). While both metrics are reported, we focus primarily on AUC-ROC due to its robustness, particularly when dealing with imbalanced class distributions [23].

For a qualitative evaluation, we examine the most influential features identified by our models. In the case of random forests, we apply the Gini importance method outlined by Breiman in 1984 to assess feature importance [11]. This approach ensures a comprehensive evaluation of our models, maintaining methodological integrity in alignment with established practices in the field.

Equation

The Glomerular Filtration Rate (GFR) was determined using a widely recognized formula known as the 4-variable Modification of Diet in Renal Disease (MDRD-4) formula. This formula is employed to estimate the GFR, a crucial measure of kidney function. The MDRD-4 formula is expressed as follows:

MDRD-4 = $175 \times (\text{Scr})^{-1.154} \times (\text{Age})^{-0.203} \times (0.742 \text{ if female}) \times (1.178 \text{ if black})$ [25] where:

MDRD−4: The abbreviation for the 4-variable MDRD formula.

Age: Age of the individual in years.

Scr: Serum creatinine level in milligrams per deciliter (mg/dL).

Gender: Gender of the individual (0.742 for females).

Ethnicity: Ethnicity of the individual (1.178 for individuals of black ethnicity).

This formula is employed to estimate GFR, a crucial measure of kidney function, based on these variables. It is particularly valuable in clinical settings for assessing kidney health and is widely accepted for its accuracy and reliability.

Results

This study revolves around a dataset comprising 1,960 patients who received diagnoses of CKD. Among these individuals, 674 were classified within the mild stages (I, II, III) of CKD, while the remaining 1,286 exhibited the severe stages (IV, V, VI) of the condition. In the study population, 1,185 (60%) were male participants with a median age of 69 years, spanning an age range of 20 to 88 years, Table 2.

Machine Learning Models

Overall, our study demonstrates exceptional predictive performance in forecasting CKD progression from mild (I, II, III) to severe stages (IV, V, VI). Table 3 and Figure 1 provide detailed insights into the classification metric scores. In the case of the random forest model, we achieved an accuracy rate of 85%, a recall rate of 86%, a precision rate of 83%, an AU-ROC score of 92%, and an AU-PR score of 83%. Conversely, the logistic regression model exhibited an accuracy rate of 84%, a recall rate of 84%, a precision rate of 85%, an AU-ROC score of 92%, and an AU-PR score of 81%. Notably, random forests slightly outperform logistic regression across many of these metrics.

Variable Importance

Random Forest feature importance is a valuable technique employed in Random Forest machine learning models to gauge the significance of input features. Its primary aim is to discern which features exert the most influence on the model's predictive outcomes. The applications of feature importance are multifaceted, encompassing tasks like gaining insights into the underlying dataset, making informed decisions about feature selection, and pinpointing potential determinants of the target variable. In this study, features are selected based on Gini impurity. This technique relies on the Gini impurity index to quantify feature importance. It operates by scrutinizing each decision tree in the forest and noting the reduction in Gini impurity achieved when specific features are used to split nodes. Features that lead to a notable reduction in impurity during node splitting are accorded higher importance. These techniques collectively offer a comprehensive understanding of feature importance, aiding in model interpretation, selection, and overall comprehension of data dynamics. The most important feature in our model is creatinine followed by eGFR, anion gap, phosphate, chloride, Blood Urea Nitrogen (BUN), platelet, Red Blood Cell (RBC) etc, Figure 2.

Discussion

In our study, we concentrated on predicting Chronic Kidney Disease (CKD), classifying stages I, II, and III as mild and stages IV, V, and VI as severe. Across various tasks and types of data, our observed impressive AU-ROC scores, often surpassing the 0.8 threshold. These high scores attest to the exceptional quality of our models, highlighting their efficacy in predicting CKD. Considering the high mortality rate associated with CKD, early prediction stands as a crucial asset for healthcare professionals.

In this regard, both random forest and logistic regression models demonstrated comparable performance in identifying mild and severe CKD cases.

Furthermore, our study delved into the broader landscape of CKD progression. Leveraging machine learning algorithms, such as lasso penalized logistic regression and random forests, we embarked on a comprehensive exploration. It's noteworthy that random forests consistently outperformed logistic regression, mainly due to their superior ability to capture intricate data relationships [26]. This attribute makes them particularly valuable for CKD prediction. Our primary focus revolved around predicting CKD advancement and, notably, forecasting the transition from milder stages to severe phases. This research venture unveiled valuable insights into the factors contributing to CKD progression, offering a pioneering approach to early detection and intervention. These findings hold significant clinical relevance, equipping healthcare practitioners with the means to identify high-risk patients at an earlier stage. This, in turn, facilitates targeted interventions, encourages lifestyle modifications, and tailors therapies to mitigate CKD progression. By embracing a multifaceted approach that combines machine learning prowess with clinical acumen, our study marks a substantial stride forward in the quest to unravel and address the intricacies of CKD progression, ultimately striving to elevate the quality of care and the well-being of affected individuals.

In assessing variable importance, the top five (5) most influential features in our model to predict CKD progression from mild to severe is creatinine followed by eGFR, anion gap, phosphate, chloride, and blood urea nitrogen (BUN).

The measurement of serum creatinine levels is a fundamental diagnostic tool in monitoring kidney health. As CKD progresses, creatinine levels tend to rise, signaling a decline in kidney function [27]. This elevation in creatinine is indicative of impaired Glomerular Filtration Rate (GFR) [28], a key parameter for assessing kidney function. Monitoring creatinine levels over time helps healthcare providers classify CKD into different stages, ranging from mild (Stage 1) to severe (Stage 5 or end-stage renal disease).

eGFR is a critical parameter in the evaluation and management of CKD [29]. It aids in diagnosis, staging, and monitoring, guiding treatment decisions and promoting better outcomes for individuals living with CKD [30]. Regular measurement of eGFR over time helps track the progression of CKD. A declining eGFR indicates worsening kidney function and may prompt more aggressive management strategies to slow down disease progression.

The anion gap is a valuable parameter that can provide insights into acid-base balance and the presence of metabolic acidosis, which are relevant considerations in the management of CKD. It is used to assess the body's acid-base balance by comparing the concentrations of positively charged ions (sodium and potassium) with negatively charged ions (chloride and bicarbonate) in the blood. In CKD, particularly in advanced stages, the kidneys may have difficulty maintaining the body's acid-base balance. An elevated anion gap can be an indicator of metabolic acidosis, a condition where there is an excess of acid in the body [31]. Regular monitoring of the anion gap is important for healthcare providers to make informed treatment decisions and optimize care for CKD patients. Furthermore, Blood Urea Nitrogen (BUN) is a clinically important parameter in the management of CKD. It provides valuable information about renal function, uremia, fluid and electrolyte balance, and disease progression [32]. Regular monitoring of BUN levels is an essential part of CKD care to ensure timely interventions and optimize patient outcomes.

Clinical relevance for Combining the Stages (I, II, III) and Stages (IV, V, VI)

The decision to group CKD Stages I to III and CKD Stages IV to VI in this study was based on a recognized clinical approach. The rationale behind this grouping is to distinguish between lower-risk stages (I, II, III) and higher-risk stages (IV, V, VI) of CKD. This division helps in studying the progression of CKD from relatively mild to more severe stages and understanding the factors that contribute to this transition. Furthermore, from a clinical perspective, the management and progression of CKD often involve the monitoring and intervention strategies that apply across all stages. For instance, interventions related to blood pressure control, medication management, and lifestyle modifications can be relevant to patients across various stages of CKD. Therefore, it makes sense to consider them collectively. CKD is characterized by a progressive decline in kidney function. Stages I, II, and III represent milder forms of kidney dysfunction, while stages VI, V, and VI signify more severe impairment. However, the progression from milder to severe stages is continuous and can be influenced by a variety of factors. Analyzing the entire spectrum together allows for a more comprehensive understanding of the disease's progression. In clinical practice, healthcare providers often need to assess a patient's risk of progressing to more severe stages of CKD. Combining the stages can help in developing predictive models that assist clinicians in identifying patients at higher risk of disease progression. Depending on the research objectives, combining stages can be practical. For instance, we aim to develop a predictive model for CKD progression, having a broader dataset that encompasses all stages can provide a more robust and generalizable model. Conducting separate analyses for each CKD stage can be resource-intensive, particularly if the dataset is limited. Combining stages can streamline the analysis process and make more efficient use of available resources. CKD is a complex condition influenced by various clinical, genetic, and lifestyle factors. In the real world, patients may transition between stages due to disease management and treatment. Therefore, analyzing the disease continuum captures this complexity more accurately.

Limitations and Future Research Recommendations

One limitation of this study is its exclusive focus on comparing two specific machine learning algorithms, random forests, and logistic regression, for predicting CKD progression. While these models exhibit promising results, their performance may exhibit variations across different datasets and clinical contexts. To enhance the robustness and generalizability of predictive models, future research should consider the evaluation of additional machine learning algorithms. Moreover, this study did not include some common risk factors, including genetic factors, lifestyle and environmental factors that could enhance predictive accuracy. Expanding the scope of features considered in subsequent studies can address this limitation. Another limitation stems from the retrospective nature of the study, relying on data from the MIMIC-III database. Although MIMIC-III provides a valuable research resource, it presents inherent limitations, including potential data entry errors, missing data, and limited applicability to populations beyond the database. Consequently, the external validity of the models and their relevance to diverse patient cohorts and healthcare settings may

be affected. To mitigate this limitation, future research should incorporate data from multiple sources to bolster the generalizability of predictive models and validate their performance in real-world clinical scenarios.

Conclusion

Our findings showed that random forests tend to slightly outperform logistic regression, demonstrating a higher capability for accurate CKD progression prediction. Furthermore, our models have leveraged common established risk factors associated with CKD progression, shedding light on the relationships between various stages of CKD.

These discoveries highlight the potential utility of our models in clinical settings, where they could serve as valuable tools for identifying CKD patients at risk of transitioning from mild to severe stages. Our study lays the foundation for further exploration and implementation of these models in real-world healthcare applications.

Author Statements

Author Contributions

K Bah, Ns Bah and AW Jallow conceived the study. K Bah was responsible for the methodology; K Bah, Ns Bah, & AW Jallow managed the software; K Bah, Ns Bah and AW Jallow were responsible for validation; K Bah, Ns Bah, and AW Jallow conducted the formal analysis; K Bah, Ns Bah & AW Jallow conducted the investigation; K Bah was responsible for data curation; K Bah wrote the original draft; K Bah, Ns Bah and AW Jallow reviewed and edited the draft. All authors have read and agreed to the published version of the manuscript.

Financial Support

This study was conducted without the involvement of external funding sources.

Statement on Informed Consent

Given that our study exclusively employed de-identified data from the MIMIC III database, patient consent requirements were waived because of the anonymous nature of the data.

Availability of Data

This study adheres to the data policy and regulations of MIMIC III, but it may be made available by the corresponding author upon reasonable request.

Declaration of Conflicts of Interest

The authors declared no conflicts of interest.

References

1. Yashfi SY, Islam MA, Pritilata, Sakib N, Islam T, Shahbaaz M, et al. Risk prediction of chronic kidney disease using machine learning algorithms. In: 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2020. IEEE Publications; 2020; 2020.

2. Khalid H, Khan A, Zahid Khan M, Mehmood G, Shuaib Qureshi M. Machine learning hybrid model for the prediction of chronic kidney disease. Comp Intell Neurosci. 2023; 2023: 9266889.

3. Chen Z, Zhang X, Zhang Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. Int Urol Nephrol. 2016; 48: 2069-75.

4. Charleonnan A, et al. Predictive analytics for chronic kidney disease using machine learning techniques, Management and Innovation Technology International Conference (MITicon). IEEE Publications. 2016; MIT–80.

5. Levey AS, Coresh J. Chronic kidney disease. Lancet. 2012; 379: 165-80.

6. Hsiao CJ, Hing E, Ashman J. Trends in electronic health record system use among office-based physicians, United States, 2007-2012. Natl Health Stat Report. 2014: US Department of Health and Human Services, Centers for Disease Control and. 1-18.

7. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989; 81: 1879-86.

8. Weiss JC, Natarajan S, Peissig PL, McCarty CA, Page D. Machine learning for personalized medicine: predicting primary myocardial infarction from electronic health records. Ai Mag. 2012; 33: 33-45.

9. Zhou Z-H. Machine learning. 2021. Springer nature.

10. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B. 1996; 58: 267-88.

11. Livingston F. Implementation of Breiman's random forest machine learning algorithm. ECE591Q machine learning Journal Paper. 2005; 1-13.

12. Qin J, Chen L, Liu Y, Liu C, Feng C, Chen B. A machine learning methodology for diagnosing chronic kidney disease. IEEE Access. 2019; 8: 20991-1002.

13. Arora M, Sharma EA. Chronic kidney disease detection by analyzing medical datasets in Weka. Int J Comput Appl. 2016; 6: 20-6.

14. Bemando C, Miranda E, Aryuni M. Machine-learning-based prediction models of coronary heart disease using naïve bayes and random forest algorithms. In: International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM). IEEE Publications; 2021: 2021.

15. Ram Kumar RP, Polepaka S. Performance comparison of random forest classifier and convolution neural network in predicting heart diseases. In: Proceedings of the third international conference on computational intelligence and informatics. ICCII. Springer. 2020: 683-91.

16. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adam M, Gertych A, et al. A deep convolutional neural network model to classify heartbeats. Comput Biol Med. 2017; 89: 389-96.

17. Desai SD, Giraddi S, Narayankar P, Pudakalakatti NR, Sulegaon S. Back-propagation neural network versus logistic regression in heart disease classification. In: Advanced computing and communication technologies. Proceedings of the 11th ICACCT 2018. Springer; 2019: 133-44.

18. Patil DD, P. Singh R, M. Thakare V, K. Gulve A. Analysis of ECG arrhythmia for heart disease detection using SVM and cuckoo search optimized neural network. Int J Eng Technol. 2018; 7: 27-33.

19. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi

M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016; 3: 160035.

20. Biau G. Analysis of a random forests model. J Mach Learn Res. 2012; 13: 1063-95.

21. Rigatti SJ. Random forest. J Insur Med. 2017; 47: 31-9.

22. Mullah MAS, Hanley JA, Benedetti A. Lasso type penalized spline regression for binary data. BMC Med Res Methodol. 2021; 21: 83.

23. Fan J, Upadhye S, Worster A. Understanding Receiver Operating Characteristic (ROC) curves. Can J Emerg Med. 2006; 8: 19-20.

24. Boyd K, Eng KH, Page CD. Area under the precision-recall curve: point estimates and confidence intervals. In: Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2013, Prague, Czech Republic. Proceedings, Part III. Vol. 13. Springer; September 23-27, 2013; 451-66.

25. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. Ann Intern Med. 1999; 130: 461-70.

26. Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. 2018; 19: 270.

27. Rule AD, Larson TS, Bergstralh EJ, Slezak JM, Jacobsen SJ, Cosio FG. Using serum creatinine to estimate glomerular filtration rate: accuracy in good health and in chronic kidney disease. Ann Intern Med. 2004; 141: 929-37.

28. Carrero JJ, Elinder CG. The Stockholm CREAtinine Measurements (SCREAM) project: fostering improvements in chronic kidney disease care. J Intern Med. 2022; 291: 254-68.

29. Ku E, Xie D, Shlipak M, Hyre Anderson A, Chen J, Go AS, et al. Change in measured GFR versus eGFR and CKD outcomes. J Am Soc Nephrol. 2016; 27: 2196-204.

30. Chertow GM, Beddhu S. Modification of eGFR-Based CKD definitions: perfect, or enemy of the good? J Am Soc Nephrol. 2019; 30: 1807-9.

31. Asahina Y, Sakaguchi Y, Kajimoto S, Hattori K, Doi Y, Oka T, et al. Association of time-updated anion gap with risk of kidney failure in advanced CKD: a cohort study. Am J Kidney Dis. 2022; 79: 374-82.

32. Seki M, Nakayama M, Sakoh T, Yoshitomi R, Fukui A, Katafuchi E, et al. Blood urea nitrogen is independently associated with renal outcomes in Japanese patients with stage 3-5 chronic kidney disease: a prospective observational study. BMC Nephrol. 2019; 20: 115.