Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

A Review on Utilizing Data Mining Techniques for Chronic Kidney Disease Detection

Shivan Hussein Hassan[1,2], Adnan Mohsin Abdulazeez[3]
shivan.hassan@uod.ac [1], shivan.hussein@auas.edu.krd [1], adnan.mohsin@dpu.edu.krd [2]
[1]University of Duhok / College of Basic Education / Department of Mathematics
[2]Akre University for Applied Sciences/ Technical College of Informatics -Akre/ Department of Information Technology.
[3]Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.

Article Information

Submitted :(date)
Reviewed: (date)
Accepted : (date)

Keywords

Chronic Kidney Disease Diagnosis, Data Mining, CKD Prediction Models, ML Algorithms for CKD, Chronic Kidney Disease Stages Classification.

Abstract

This comprehensive study delves into the application of machine learning (ML) and data mining techniques for the prognosis and diagnosis of Chronic Kidney Disease (CKD), a significant global health concern characterized by the gradual loss of kidney function. Through a detailed examination of various predictive models, the research evaluates the efficacy of different ML algorithms and data mining methodologies in classifying and diagnosing CKD. Utilizing datasets from the UCI machine learning repository and other sources, this study explores a range of ML algorithms-including logistic regression, decision trees, support vector machines, random forest, and deep learning networks-alongside feature selection techniques to enhance prediction accuracy and facilitate early diagnosis. Despite facing challenges such as dataset limitations and the need for external validation, the findings reveal remarkable potential in using ML and data mining to improve CKD diagnosis, with some models achieving accuracy rates exceeding 99%. The research underscores the critical role of technology in advancing CKD diagnosis and management, paving the way for more personalized and effective healthcare solutions.

A. Introduction

CKD, a stealthy assailant lurking within the shadows of public health concerns, has emerged as a formidable adversary against global health. Characterized by the insidious deterioration of kidney function over time, CKD harbors the potential to devastate lives if left unchecked. The kidneys, vital organs for filtering blood, balancing electrolytes, and removing waste, are central to our body's equilibrium. Yet, the gradual nature of CKD means that its presence often remains unnoticed until it reaches an advanced and more perilous stage [1][2][3].

With over 750 million individuals grappling with kidney disease worldwide, the disparity in the prevalence, diagnosis, and treatment of this condition underscores a pressing global challenge [4]. Despite advancements in medical science, early detection of CKD remains a hurdle, compounded by the lack of specific symptoms in its nascent stages [5][6]. This silent epidemic is exacerbated by risk factors such as obesity, diabetes, hypertension, and lifestyle choices, including smoking and excessive alcohol consumption, which further elevate the risk of developing CKD [7][8].

Kidneys regulate essential elements like salt and potassium in the body and purify 5L of blood, but if they fail, toxic fluids increase, leading to CKD. Risk factors include diabetes, hypertension, age, and hereditary conditions [9][10]. Alarmingly, CKD claims more lives annually than some of the most common cancers, placing a significant strain on healthcare systems, particularly in regions where access to adequate treatment is scarce [11][12]. The progression of CKD can culminate in End-Stage Renal Disease (ESRD), necessitating dialysis or kidney transplantation, both of which entail substantial risks and financial implications. Additionally, it creates hormones that regulate electrolyte levels, blood pressure, and acid-base balance in addition to stimulating erythropoiesis and controlling calcium metabolism [13].

Classification methods are widely used in the medical field to enhance the accuracy of disease identification and prediction. This technique falls under the category of data mining [14]. The role of technology and data mining in combatting CKD presents a beacon of hope. By harnessing the power of AI and ML, healthcare professionals can glean insights from vast datasets, potentially paving the way for early detection and more personalized treatment plans [15]. The Nephrotic Syndrome Study Network (NEPTUNE) exemplifies the cutting-edge efforts to apply precision medicine in nephrology, offering a glimpse into a future where CKD can be managed more effectively [16][17].

As CKD ascends the ranks of global health priorities, the collective efforts of the medical community, policymakers, and researchers are crucial in stemming the tide of this growing epidemic [18][19]. The journey toward mitigating the impact of CKD is fraught with challenges, yet it is a path we must tread with determination and innovative thinking to safeguard the health and well-being of millions across the globe [20].

The aim of the paper is to develop and validate predictive models using ML and data mining techniques for the diagnosis and prognosis of CKD. By utilizing various datasets, including those from the UCI Machine Learning Repository, the study seeks to evaluate the effectiveness of different ML algorithms and data mining methodologies in accurately classifying and diagnosing CKD. The

overarching goal is to enhance prediction accuracy, minimize misclassifications, and facilitate early diagnosis, ultimately improving patient outcomes and advancing the management of CKD through technological innovations in healthcare.

B. Background

The studies described present a comprehensive exploration of utilizing ML and data mining techniques for CKD prognosis and diagnosis. CKD is a significant health concern globally, with early detection and management crucial for mitigating its progression and associated complications [21].

The research endeavors primarily revolve around developing predictive models, leveraging diverse ML algorithms, and data mining methodologies to classify and diagnose CKD accurately. These studies utilize various datasets, including those from the UCI Machine Learning Repository [22], with sample sizes ranging from 400 instances to smaller subsets. The aim across these studies is to enhance CKD prediction accuracy, minimize misclassifications, and facilitate early diagnosis to improve patient outcomes [23].

Several ML algorithms, including Logistic Regression, Decision Trees, Support Vector Machines (SVM), Random Forest, Deep Learning networks, AdaBoost, and ensemble techniques, are explored for their efficacy in CKD prediction. Each algorithm demonstrates varying levels of accuracy, sensitivity, and specificity, with some achieving remarkable results exceeding 99% accuracy rates. Additionally, feature selection techniques such as ReliefF, Recursive Feature Elimination (RFE), and correlation-based selection are employed to identify crucial predictors for CKD diagnosis, including physiological parameters like hemoglobin, albumin, and specific gravity [24][25][26].

Despite notable achievements in accuracy, these studies acknowledge several limitations and challenges. These include dataset size constraints, potential biases, interpretability concerns, scalability issues, and the need for external validation and practical implementation considerations. Furthermore, the studies highlight the importance of addressing missing data, dataset balancing, and selecting relevant features to enhance model performance and generalizability [27][28].

Future research directions outlined in these studies emphasize the exploration of advanced ML algorithms, deep learning techniques, and ensemble methods on larger and more diverse datasets. Additionally, there's a call for integrating domain knowledge, addressing interpretability challenges, and validating models in real-world clinical settings to ensure their utility and effectiveness in aiding healthcare professionals in CKD diagnosis and management [29].

The use of gene expression analysis is revolutionizing healthcare by making it more personalized, proactive, and predictive. This approach allows for the development of healthcare strategies that are customized to individual genetic profiles, focusing on prevention and early intervention to improve patient outcomes [30].

Overall, these studies collectively underscore the significant potential of ML and data mining in improving CKD prognosis and diagnosis, offering valuable insights and paving the way for the development of robust computer-aided diagnostic systems to tackle this critical healthcare issue.

C. Literature Review Methodology

In conducting the literature review, the researchers meticulously analyzed studies focusing on machine learning (ML) techniques for predicting chronic kidney disease (CKD). They explored a variety of ML models, including Logistic Regression, Decision Trees, Support Vector Machines (SVM), Random Forest, and deep learning approaches like Artificial Neural Networks (ANN) and Deep Neural Networks (DNN). The methodologies employed in these studies often involved data preprocessing steps such as imputation, balancing with techniques like SMOTE, and feature selection methods including Recursive Feature Elimination (RFE), chi-squared tests, and Principal Component Analysis (PCA). Performance metrics such as accuracy, sensitivity, specificity, and area under the ROC curve were used to evaluate the effectiveness of these models. The findings demonstrated that ensemble methods, particularly those involving bagging and boosting techniques, and advanced ML models like XGBoost and deep learning, often yielded higher accuracy rates, sometimes exceeding 99%. However, several studies highlighted the challenges of small dataset sizes, potential overfitting, and the need for external validation, emphasizing the importance of domain knowledge and the integration of diverse data sources to enhance model robustness.

Despite the promising results, limitations in the reviewed studies pointed to areas requiring further research. Many studies lacked discussions on practical implementation challenges, such as dataset biases, model interpretability, and scalability issues. The dependency on expert pathologists for clinical dataset acquisition and the exclusion of comprehensive feature selection techniques were also noted as constraints. Future research directions suggested in the literature include incorporating more sophisticated ML and deep learning algorithms, using larger and more diverse datasets, and exploring explainable ML techniques for better clinical application. Additionally, integrating additional data on genetics, environmental factors, and lifestyle habits could improve the predictive accuracy and applicability of ML models in CKD prognosis, thereby offering valuable tools for early detection and patient management in healthcare settings.

D. Literature Review

In [31] the authors aims to develop and validate a predictive model for chronic renal disease (CRD) prognosis using a dataset from the UCI Machine Learning Repository comprising 400 samples. Logistic Regression, Decision Trees, and Support Vector Machines were employed as ML classifiers, with the bagging ensemble method utilized to enhance performance. While the Decision Tree classifier yielded the highest accuracy at 95%, the implementation of the bagging ensemble method resulted in a greater accuracy of 97%. Limitations in the study include the relatively small dataset size and the lack of discussion on potential overfitting or external validation of the model. Nevertheless, the research underscores the potential of ML techniques, particularly Decision Trees, for predicting CRD prognosis. The findings emphasize the importance of domain knowledge in interpreting clinical data accurately and suggest future research directions such as applying Random Forest models to manage missing values in

datasets related to various diseases and incorporating additional data on genetics, water consumption, and dietary habits to enhance understanding of CRD.

In [32] the researchers introduced a hybrid approach for diagnosing CKD, combining Support Vector Machine (SVM) classifier optimization with ReliefF feature selection and Principal Component Analysis (PCA) for dimensionality reduction. It achieved impressive prediction accuracies of 92.5% on clinical CKD datasets and 98.5% on benchmarked CKD Datasets, showcasing its effectiveness. However, challenges arose from dependency on expert pathologists for clinical dataset acquisition, leading to delays and potential reluctance in practical implementation. Nonetheless, the approach identified significant CKD risk factors and enhanced diagnostic accuracy by eliminating irrelevant parameters. The study suggested promising prospects for aiding medical professionals in CKD diagnosis and highlighted avenues for future research to further refine the methodology.

In [33] the researchers developed a machine-learning model for forecasting CKD occurrence using publicly available data. SVM and RF achieved high accuracy rates of 99.33% and 98.67%, respectively, with low false negative rates, notably 0.67% for SVM. Data preprocessing steps included imputation, balancing with SMOTE, and feature scaling, while the chi-squared test aided feature selection. The study's real-time diagnostic potential in rural areas, where advanced tests were scarce, highlighted its cost and time efficiency compared to conventional methods like MRI and CT scans. Further research was suggested to address limitations and explore optimizing feature selection techniques for improved model performance.

In [34] the authors assessed various ML models for automating CKD diagnosis, prioritizing accuracy and precision. Employing the JBI scoping review methodology, it identified ensemble-based and deep learning methods as prevalent in this domain, with deep learning achieving an exceptional 99.75% accuracy. The research developed and evaluated an ML model tailored specifically for CKD diagnosis, utilizing tools like Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Despite providing valuable insights, the study lacked detailed discussion on practical implementation challenges and potential biases. Future research avenues may include exploring explainable ML techniques for CKD diagnosis.

In [35] the authors presented a Deep Neural Network model for CKD diagnosis, utilizing the PyTorch library's multi-layer perceptron classifier. It addressed the challenges of diagnosing CKD due to its reliance on multiple features and limited contribution from common symptoms. Despite these obstacles, the model achieved an impressive 100% accuracy, surpassing traditional ML methods like support vector machines and naive Bayes classifiers. The research highlighted neural networks' ability to handle non-linear data, process complex datasets, and autonomously learn essential information, offering promising advancements in medical diagnosis. This achievement not only aided healthcare professionals in addressing diagnostic challenges but also provided a valuable tool for individuals with limited access to medical care. Furthermore, the paper emphasized the importance of advanced libraries in simplifying model implementation and encouraged further exploration into neural network technologies.

In [36] the researchers utilized ML methods to predict and classify CKD using a dataset from the Irvine ML Repository. Various algorithms including SVM, KNN, RF,

LR, DT, and XGBoost were employed, with XGBoost achieving the highest accuracy of 98.00%. Preprocessing techniques involved imputation for missing values and feature selection using Recursive Feature Elimination (RFE). The research highlighted the potential for early CKD detection and patient-centered management. Limitations included the absence of deep learning methods and potential dataset constraints. Future work could involve integrating deep learning approaches and addressing raw image datasets related to CKD for further insights.

In [37] the authors focused on predicting CKD using techniques and evaluated classifier accuracy. It explored models such as Gradient Boosting, Gaussian Naïve Bayes, Decision Tree, and Random Forest, achieving notable accuracies. The proposed hybrid model achieved 100% accuracy, outperforming individual classifiers. Feature selection using Pearson correlation and the stacking algorithm enhanced prediction accuracy. Despite potential limitations in dataset bias and generalizability, the research underscored ML's potential in medical diagnosis for its superior accuracy over traditional methods.

In [38] the researchers, CKD prediction was explored using a dataset from the UCI repository. Seven classifier algorithms, including artificial neural network, C5.0, logistic regression, CHAID, linear support vector machine (LSVM), K-Nearest neighbors, and random tree, were employed. Different feature selection techniques, such as correlation-based feature selection, Wrapper method, and LASSO regression, were applied. LSVM consistently achieved the highest accuracy, ranging from 98.46% to 98.86%, while a deep neural network attained the highest accuracy of 99.6%. Synthetic minority over-sampling technique (SMOTE) with full features yielded the best results across all classifiers. It was concluded that SMOTE, particularly with selected features by LASSO regression, enhanced model performance. Notably, logistic regression and KNN were excluded from SMOTE due to inadequate results, highlighting the significance of dataset balancing techniques like SMOTE.

In [39] the researchers focused on developing a model for early CKD detection using ML algorithms like Support Vector Machine, Random Forest, and Artificial Neural Network. Three main stages involved applying baseline classifiers on categorical and non-categorical attributes and then combining them through majority voting. The proposed model demonstrated a 3% accuracy improvement over existing models, with Random Forest achieving the highest accuracy of 92% on combined datasets. These findings offered potential benefits for improving CKD classification accuracy, aiding medical professionals and patients in early diagnosis and intervention. Limitations included reliance on baseline classifiers and specific feature selection, suggesting opportunities for future enhancement through advanced feature selection methods.

In [40] the authors, ML algorithms were employed to diagnose CKD at an early stage, with twelve classifiers tested and the XgBoost classifier yielding the highest performance metrics, including an accuracy, precision, recall, and F1-score of 0.983. Feature selection techniques identified hemoglobin, albumin, and specific gravity as crucial predictors for CKD diagnosis. The research preprocessed the CKD dataset, conducted principal component analysis (PCA) to highlight dominant features, and validated models using selected input parameters. Despite limitations such as reliance on supervised learning and potential variable selection

biases, the study underscored the potential of ML in improving early CKD detection, offering valuable insights for future predictive modeling in kidney disease and related fields.

In [41] the researchers focused on identifying patient characteristics affecting CKD prevalence and extracting actionable insights for medical professionals. Utilizing data mining techniques like Sequential Minimal Optimization and Multilayer Perceptron, the study achieves high accuracy rates, with SMO at 98.31% and Multilayer Perceptron at 98.06%. The extracted action rules serve as valuable guidelines for doctors, aiding in diagnosis and treatment decisions. However, limitations such as reliance on available data and generalizability should be considered. Overall, the study showcases the potential of data mining in supporting medical decision-making processes for CKD.

In [42] the researchers aimed to develop a robust prediction system for CKD using classification and association rule mining techniques through platforms like Weka and SPSS. Five classification algorithms were utilized, with K-nearest neighbor achieving 98.50% accuracy and JRip achieving 96.00%. Real medical datasets were employed, demonstrating significant improvements in CKD prediction accuracy. The study's findings hold potential implications for medical decision-making, patient health, and healthcare cost reduction. Acknowledged limitations include the method's lack of generalization, difficulty with continuous variables, and challenges in handling large datasets. Future plans involve extending the research with different classifier methods or applying the FP-growth algorithm for association rule mining.

In [43] the authors employed data mining techniques, utilizing Weka, Orange, and Python tools, to predict kidney disease. The study focused on preprocessing the CKD dataset and applied five techniques: Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree, K Nearest Neighbor (KNN), and Naïve Bayes (NB). SVM emerged with the highest accuracy, reaching 99.8% according to the Orange tool. While the specific limitations were not outlined, the work underscored the significance of data mining in medical contexts for disease prediction, particularly in the case of CKD.

In [44] the authors proposed an ML-based approach to predict CKD using patient health data and nine ML algorithms, including AdaBoost, achieving 99.17% accuracy. Feature selection methods were applied to handle the relatively small dataset of 400 samples with 24 features. AdaBoost emerged as the top-performing algorithm with 100% sensitivity, specificity, and ROC values when combined with feature selection techniques. However, the study acknowledged potential limitations due to the dataset size. Future work aimed to explore advanced ML and deep learning algorithms on larger datasets to enhance CKD prediction effectiveness.

In [45] the researchers presented a novel algorithm, Probability Weighted AdaBoost (PRAB), aimed at predicting CKD with a focus on reducing misclassifications. ML and data mining techniques were leveraged to analyze physiological parameters from the UCIdataset for disease prediction. The PRAB model achieved an impressive accuracy of 99%, highlighting its effectiveness in early detection. However, specific limitations of the approach were not explicitly addressed. The study suggested further enhancements through the incorporation

of ensemble algorithms and techniques. Overall, the research contributed to advancing healthcare outcomes in CKD management.

In [46] the authors utilized data mining algorithms to predict chronic renal disease, analyzing a dataset of 400 instances with 25 attributes. Preprocessing techniques were employed to handle missing data and select relevant variables for prediction models. Decision Tree, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN) classifiers were compared, with KNN achieving the highest accuracy of 98.3%. Despite the promising results, limitations may have stemmed from the dataset's scope and unaccounted confounding variables. Future research directions included exploring additional parameters like lifestyle and environmental factors, as well as investigating alternative classifiers such as Artificial Neural Networks (ANN) or ensemble techniques to further improve prediction accuracy.

In [47] the authors, Python programming and the CKD dataset from UC Irvine were utilized to develop eight ML models for detecting CKD, with Random Forest achieving the highest accuracy of 99.75%. However, potential biases in the dataset, model interpretability, scalability, and alternative approaches beyond supervised learning were overlooked. Nevertheless, the study highlighted the potential of ML in predictive analysis of CKD, emphasizing the need for robust computer-aided diagnosis systems in healthcare. Future directions included collecting local datasets for specific diagnostic models and exploring advanced techniques like deep learning and neural networks for further enhancement.

In [48] the researchers combined feature selection and ensemble techniques to enhance CKD prediction accuracy, focusing on various classifiers like KNN, J48, ANN, NB, and SVM. Notable achievements included accuracies of 97% to 98% for individual classifiers and 98.85% for ensemble techniques on the condensed CKD dataset. Limitations might have included a narrow focus on classifier precision and possible oversight of other factors influencing CKD prediction. Nevertheless, the study underscored the effectiveness of the approach and suggested further exploration of ensemble methods and feature selection techniques for improving CKD prediction models and their application in medical diagnoses.

In [49] the authors introduced a new deep learning model for early CKD detection, utilizing Recursive Feature Elimination (RFE) to identify crucial predictive features. Key features identified included Hemoglobin, Specific Gravity, Serum Creatinine, Red Blood Cell Count, Albumin, Packed Cell Volume, and Hypertension. Comparing with other ML techniques, the proposed deep neural network achieved 100% accuracy in CKD prediction, suggesting its potential utility for nephrologists. However, a limitation of the study was the use of small datasets, highlighting the need for larger and more diverse datasets to improve model performance. Future directions involved evaluating the model on larger datasets with additional clinical parameters to enhance prediction accuracy.

In [50] the researchers reviewed state-of-the-art data mining algorithms for predicting diseases, utilizing techniques such as Naive Bayes, J48, REFTree, SMO, Multilayer Perceptron, and Vote on datasets obtained from UCI respiratory, employing the WEKA tool for experimental setup and evaluation. Limitations included the complexity of certain algorithms like KNN, SMO, and Random Forest, as well as the inadequacy of statistical models in handling big data. It concluded

that SVM and Naive Bayes algorithms were commonly used and achieved high accuracy, with Random Forest performing best for heart, liver, and kidney disease prediction, and SMO exhibiting the highest accuracy (77.34%) for diabetes prediction. Future work could involve applying big data techniques with Spark and exploring deep learning approaches.

In [51] the researcher proposes a diagnostic prediction model for CKD using IoT sensor technology and classification methods, aiming to improve CKD prediction and severity determination, particularly in developing countries with limited healthcare access. Achieving 97% accuracy, 99% sensitivity, and 95% specificity using decision tree (J48) classifier on dataset 3, the model shows promise but faces limitations in feature selection and scalability for real-world IoT implementation. Future work includes validating the model with larger datasets and exploring ensemble ML methods for improved performance, highlighting the potential impact on CKD management in resource-constrained settings.

In [52] the authors explored the use of ML for early CKD diagnosis in developing nations, aiming to address healthcare limitations. It assessed various algorithms' efficacy through qualitative and quantitative analyses, utilizing Weka software and CKD data. Findings revealed the J48 decision tree as suitable for CKD screening, achieving 95.00% accuracy and aligning closely with nephrologists' opinions. Although random forest showed promise, it posed interpretation challenges. The study emphasized attribute count's impact on cost and classifier performance, vital in resource-limited settings. Overall, it highlighted the need for accessible and accurate ML tools to tackle healthcare obstacles in CKD diagnosis within developing countries.

In [53] the researchers employed ensemble methods and three base learners to enhance CKD diagnosis accuracy. Data preprocessing preceded classification to normalize independent variables. Evaluation metrics included accuracy, specificity, sensitivity, kappa, and ROC criteria. Ensemble techniques outperformed individual base learners, with Random Subspace achieving 100% accuracy using KNN. The study underscored the importance of accuracy in medical diagnosis. Implemented in Python, the ensemble approach demonstrated superior performance, making it suitable for CKD prediction.

In [54] the authors focused lay on achieving a solid diagnosis for CKD rather than seeking the ideal solution, utilizing data mining techniques. Specifically, two algorithms, Random Forest algorithm, and Back Propagation Neural Network were employed for diagnosis. The important results included the Random Forest Algorithm achieving a certainty of 88.7% and a Receiver Operating Characteristics Area under the Curve (ROC AUC) of 99.2%, while the Back Propagation Neural Network demonstrated a higher efficiency with a certainty of 98.40%. However, the text did not explicitly outline the limitations of the research beyond its focus on solid diagnosis.

In [55] the authors utilized ML classification algorithms, specifically Random Forest and J48 decision tree algorithm, to predict the various stages of CKD. It highlighted the limitation of Random Forest in terms of accuracy and speed compared to J48, with J48 achieving an accuracy of 85.5% in predicting CKD stages while demonstrating improved performance over Random Forest. Additionally, the time taken by J48 for prediction was significantly lower at 0.03 seconds compared

to 0.28 seconds for Random Forest. The findings suggested that J48 was more accurate and efficient, making it a preferable choice for building an automated system for CKD detection, potentially aiding physicians in generating a decision support system for diagnosing CKD.

In [56] the researchers employed Decision Tree methods and Identification of Pattern Mining (IPM) techniques to classify data concerning CKD patients. It achieved an accuracy of 91.75% with Decision Tree algorithms and 96.75% with IPM techniques, indicating their potential in diagnosing CKD patients accurately. Decision Tree algorithms were commended for their ability to construct comprehensive trees using all dataset features, streamlining the prediction process and facilitating early treatment initiation for CKD patients. Nonetheless, the study acknowledged limitations, such as dataset size and missing attribute values, hindering the attainment of higher accuracy, like the desired 99.99%. Addressing these limitations would have required a larger dataset with zero missing values for future research.

In [57] the authors evaluated seven ML techniques for classifying a kidney patient dataset into CKD or NOTCKD. Techniques included NBTree, J48, SVM, LR, MLP, Naïve Bayes, and CHIRP. CHIRP demonstrated superior performance with minimal error rates and 99.75% accuracy, suggesting its promise for CKD prognosis. The study lacked explicit discussion on limitations, interpretability, scalability, and real-world applicability. Tools used for implementation and evaluation were not specified but likely included popular ML libraries. Overall, CHIRP's effectiveness suggested it as a promising technique for CKD diagnosis compared to other evaluated methods.

Table1. Summarizing The Literature Review Mentioned Along With Mentioned Along With Their Key Details

Ref.	Year	Based Model	Accuracy	Advantage	Limitations
[31]	2023	Bagging Ensemble	97%	Bagging ensemble method increased accuracy to 97%.	Relatively small dataset, Lack of discussion on potential overfitting or external validation
[32]	2023	Support Vector Machine (SVM), ReliefF, PCA	92.5% - 98.5%	Hybrid approach with SVM optimization and PCA enhanced diagnostic accuracy.	Dependency on expert pathologists for dataset acquisition
[33]	2020	Random Forest, Machine Learning (ML)	99.75%	High accuracy and low false negative rates, effective in rural areas.	Variability in the effectiveness of data mining techniques
[34]	2023	Ensemble-based, Deep Learning	99.75%	Deep learning achieved exceptional accuracy of 99.75%.	Lack of detailed discussion on practical implementation challenges and potential biases
[35]	2023	Deep Neural Network, PyTorch	100%	Achieved 100% accuracy using a multi-layer perceptron classifier.	Reliance on multiple features, Limited contribution from

					common symptoms
[36]	2023	XGBoost, KNN, RF, LR, DT, XGBoost	98.00%	XGBoost achieved the highest accuracy of 98.00%.	Absence of deep learning methods and potential dataset constraints
[37]	2023	Gradient Boosting, Gaussian Naïve Bayes, Decision Tree, Random Forest	100%	Hybrid model achieved 100% accuracy with Pearson correlation and stacking algorithm.	Potential dataset bias and generalizability limitations
[38]	2021	Linear Support Vector Machine (LSVM), Deep Neural Network	98.46% - 99.6%	LSVM and a deep neural network achieved high accuracies with SMOTE.	Inadequate results for logistic regression and KNN with SMOTE, dataset balancing significance highlighted
[39]	2023	Support Vector Machine, Random Forest, Artificial Neural Network	92%	3% accuracy improvement over existing models with majority voting.	Reliance on baseline classifiers and specific feature selection
[40]	2023	XgBoost	0.983	XgBoost classifier achieved high performance metrics.	Reliance on supervised learning and potential variable selection biases
[41]	2022	Sequential Minimal Optimization (SMO), Multilayer Perceptron	98.06% - 98.31%	High accuracy rates with Sequential Minimal Optimization and Multilayer Perceptron.	Reliance on available data and generalizability
[42]	2020	K-nearest neighbor (KNN), JRip	96.00% - 98.50%	Achieved 98.50% accuracy with K-nearest neighbor.	Method's lack of generalization, difficulty with continuous variables, and challenges in handling large datasets
[43]	2022	Support Vector Machine (SVM)	99.8%	SVM reached 99.8% accuracy.	Not explicitly outlined limitations
[44]	2023	AdaBoost, Feature selection	99.17%	AdaBoost with feature selection achieved high sensitivity, specificity, and ROC values.	Acknowledged dataset size limitations
[45]	2024	Probability Weighted AdaBoost (PRAB)	99%	PRAB model achieved 99% accuracy in predicting CKD.	Specific limitations not addressed
[46]	2021	Decision Tree, Support Vector Machine (SVM), Random Forest, K-Nearest	98.3%	KNN achieved the highest accuracy of 98.3%.	Limitations from dataset scope and unaccounted confounding variables

		Neighbors (KNN)			
[47]	2021	Random Forest	99.75%	Random Forest achieved the highest accuracy of 99.75%.	Potential biases in the dataset, model interpretability, scalability, and alternative approaches beyond supervised learning
[48]	2023	Ensemble techniques	97% - 98.85%	Ensemble techniques achieved up to 98.85% accuracy.	Narrow focus on classifier precision and possible oversight of other factors influencing prediction
[49]	2022	Deep Neural Network	100%	Deep learning model achieved 100% accuracy in CKD prediction.	Use of small datasets
[50]	2021	SVM, Naive Bayes, Random Forest	Various	Random Forest performed best for heart, liver, and kidney disease prediction using various data mining algorithms.	Complexity of certain algorithms, inadequacy of statistical models
[51]	2021	Decision tree (J48)	97%	Achieved high accuracy, sensitivity, and specificity with J48 classifier in a resource-constrained setting.	Limitations in feature selection and scalability
[52]	2020	J48 decision tree, Random Forest	95.00%	J48 decision tree aligned closely with nephrologists' opinions with 95.00% accuracy.	Interpretation challenges with Random Forest
[53]	2020	Random Subspace, KNN	100%	Ensemble methods with Random Subspace and KNN achieved 100% accuracy.	Not specified
[54]	2020	Random Forest, Back Propagation Neural Network	88.7% - 99.2%, 98.40%	Back Propagation Neural Network demonstrated higher efficiency with a certainty of 98.40%.	Lack of explicit limitations beyond focus on solid diagnosis
[55]	2021	Random Forest, J48	85.5%	J48 decision tree was more accurate and efficient for predicting CKD stages.	Random Forest accuracy and speed limitations
[56]	2021	Decision Tree, Identification of Pattern Mining (IPM)	91.75% - 96.75%	IPM techniques achieved higher accuracy in diagnosing CKD.	Dataset size and missing attribute values hindering higher accuracy
[57]	2020	CHIRP	99.75%	CHIRP showed minimal error rates and 99.75% accuracy in CKD prognosis.	No explicit discussion on limitations

The table summarizes various predictive models used in medical diagnostics, detailing their accuracy, benefits, and drawbacks. Ensemble methods and deep learning techniques are prominent for their high accuracy but are limited by small

dataset sizes and implementation challenges. Support Vector Machines and hybrid models show effectiveness with broad accuracy ranges, though they face issues with dataset acquisition and generalization. Random Forest and Decision Trees are highlighted for their precision, yet scalability and bias remain problematic. Overall, while the models achieve high accuracies, they are commonly challenged by dataset limitations, potential overfitting, and the need for external validation, underscoring the need for careful evaluation of model applicability and improvement areas.

E. Discussion

The discussion around the utilization of machine learning (ML) and data mining techniques for chronic kidney disease (CKD) prognosis and diagnosis unveils a promising avenue toward revolutionizing healthcare approaches in nephrology. Through extensive analysis, it is evident that ML algorithms—ranging from logistic regression to advanced deep learning networks—coupled with sophisticated data mining methodologies, hold the potential to significantly improve the accuracy of CKD diagnosis and the prediction of its progression. This review not only showcases the high accuracy rates achieved by various models but also highlights the critical role of feature selection and dataset optimization in enhancing model performance. Despite the challenges posed by dataset limitations and the necessity for real-world validation, moving forward, it becomes imperative to focus on the integration of more comprehensive and diversified datasets, alongside the exploration of new ML techniques, to overcome existing hurdles and fully harness the potential of ML and data mining in CKD management.

This discussion lays the groundwork for future endeavors in the field, emphasizing the necessity for continued research and collaboration among healthcare professionals, data scientists, and policymakers. Such collaborative efforts are essential to optimize CKD diagnosis and treatment strategies, ultimately leading to better patient outcomes and a deeper understanding of kidney disease dynamics. Future research should prioritize the development of standardized protocols for data collection and sharing, ensuring data quality and facilitating the integration of diverse data sources. Additionally, exploring novel ML techniques, such as transfer learning and federated learning, could further enhance the robustness and generalizability of predictive models. By addressing these aspects, the healthcare community can make significant strides in the early detection, prevention, and management of CKD, improving the overall quality of care for patients worldwide.

F. Conclusions

The investigation into the use of ML and data mining for CKD prognosis and diagnosis has demonstrated significant potential in enhancing patient outcomes. This review explored a wide array of ML algorithms and data mining techniques, successfully employing them to accurately classify and diagnose CKD stages. Despite encountering challenges like limited dataset sizes and the necessity for comprehensive external validation, the research achieved notable successes, with certain models surpassing 99% accuracy rates. These accomplishments highlight the transformative impact of technology in healthcare, particularly in the early

detection and management of CKD. Future directions should focus on overcoming existing limitations through the integration of larger, more diverse datasets, and the implementation of advanced ML techniques, ensuring the models' applicability and effectiveness in real-world clinical settings. Ultimately, this study contributes to the growing body of evidence supporting the use of ML and data mining in the fight against CKD, offering new avenues for improving diagnosis, treatment, and patient care in the field of nephrology.

G. References

[1] A. Pathak, M. Asma Gani, A. H. Tasin, S. N. Sania, M. Adil, and S. Akter, "Chronic Kidney Disease (CKD) Prediction Using Data Mining Techniques," vol. 10, no. 5, pp. 976–988, 2021, doi: 10.1007/978-3-030-68154-8_82.

[2] M. G et al., "Fuzzy Logic-Based Systems for the Diagnosis of Chronic Kidney Disease," Biomed Res. Int., vol. 2022, p. 2653665, 2022, doi: 10.1155/2022/2653665.

[3] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," J. Big Data, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00657-5.

[4] R. C. Poonia et al., "Intelligent Diagnostic Prediction and Classification Models for Detection of Kidney Disease," Healthc., vol. 10, no. 2, 2022, doi: 10.3390/healthcare10020371.

[5] S. Eddy, L. H. Mariani, and M. Kretzler, "Integrated multi-omics approaches to improve classification of chronic kidney disease," Nat. Rev. Nephrol., vol. 16, no. 11, pp. 657–668, 2020, doi: 10.1038/s41581-020-0286-5.

[6] A. Prabowo, S. Wardani, A. Muis, R. Gea, N. Atan, and B. Tarigan, "Journal of Computer Networks , Architecture and High Performance Computing Diagnosis and Prediction of Chronic Kidney Disease Using a Stacked Generalization Approach Journal of Computer Networks , Architecture and High Performance Computing," vol. 6, no. 1, pp. 467–471, 2024.

[7] S. Danish, S. Aamer, S. P. Kharde, and S. S. Gadekar, "Review on Chronic Kidney Disease Using Naïve Bayes Algorithm," no. 08, pp. 1520–1522, 2020.

[8] K. Dharmarajan, "Prediction of Chronic Kidney Disease using Classification techniques ISSN NO : 2347-6648 Prediction of Chronic Kidney Disease using Classification techniques Page No : 1420," no. March, 2020.

[9] W. Wang, G. Chakraborty, and B. Chakraborty, "Predicting the risk of chronic kidney disease (Ckd) using machine learning algorithm," Appl. Sci., vol. 11, no. 1, pp. 1–17, 2021, doi: 10.3390/app11010202.

[10] P. A. Moreno-Sanchez, "Data-Driven Early Diagnosis of Chronic Kidney Disease: Development and Evaluation of an Explainable AI Model," IEEE Access, vol. 11, no. April, pp. 38359–38369, 2023, doi: 10.1109/ACCESS.2023.3264270.

[11] J. Qezelbash-Chamak, S. Badamchizadeh, K. Eshghi, and Y. Asadi, "A survey of machine learning in kidney disease diagnosis," Mach. Learn. with Appl., vol. 10, no. May, p. 100418, 2022, doi: 10.1016/j.mlwa.2022.100418.

[12] M. E. Barakat, C. G. Chin, and L. I. Ee, "Performance Analysis of Chronic Kidney Disease Detection Based on K-Nearest Neighbors Data Mining," Int. J. Intell. Syst. Appl. Eng., vol. 11, no. 8s, pp. 393–400, 2023.
[13] V. KR, M. S. Maharajan, B. K, and N. Sivakumar, "Classification of adaptive back propagation neural network along with fuzzy logic in chronic kidney disease," e-Prime - Adv. Electr. Eng. Electron. Energy, vol. 7, no. February, p. 100463, 2024, doi: 10.1016/j.prime.2024.100463.
[14] S. Muawanah, U. Muzayanah, M. G. R. Pandin, M. D. S. Alam, and J. P. N. Trisnaningtyas, "Stress and Coping Strategies of Madrasah's Teachers on Applying Distance Learning During COVID-19 Pandemic in Indonesia," Qubahan Acad. J., vol. 3, no. 4, pp. 206–218, 2023, doi: 10.48161/Issn.2709-8206.
[15] S. Rama Sree and S. N. S. V. S. C. Ramesh, "Artificial intelligence aided diagnosis of Chronic Kidney Disease," J. Crit. Rev., vol. 7, no. 12, pp. 909–923, 2020, doi: 10.31838/jcr.07.12.159.
[16] C. Delrue, S. De Bruyne, and M. M. Speeckaert, "Application of Machine Learning in Chronic Kidney Disease: Current Status and Future Prospects," Biomedicines, vol. 12, no. 3, p. 568, 2024, doi: 10.3390/biomedicines12030568.
[17] G. U. John, Y. M. Malgwi, I. Adamu, and E. Peter, "Hybrid Diagnostic Model for Kidney Disease Prediction Using Data Mining Techniques," African J. Adv. Sci. Technol. Res., vol. 11, no. 1, pp. 75–89, 2023.
[18] A. Pinto, D. Ferreira, C. Neto, A. Abelha, and J. Machado, "Data mining to predict early stage chronic kidney disease," Procedia Comput. Sci., vol. 177, no. 2018, pp. 562–567, 2020, doi: 10.1016/j.procs.2020.10.079.
[19] H. Iftikhar, M. Khan, Z. Khan, F. Khan, H. M. Alshanbari, and Z. Ahmad, "A Comparative Analysis of Machine Learning Models: A Case Study in Predicting Chronic Kidney Disease," Sustain., vol. 15, no. 3, pp. 1–13, 2023, doi: 10.3390/su15032754.
[20] Asra Fatima, Shireen Fatima, and Ayesha Kiran, "Big Data and Machine Learning Based Early Chronic Kidney Disease Prediction," J. Sci. Res. Technol., no. 3, pp. 22–28, 2024, doi: 10.61808/jsrt88.
[21] F. Sanmarchi, C. Fanconi, D. Golinelli, D. Gori, T. Hernandez-Boussard, and A. Capodici, "Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review," J. Nephrol., vol. 36, no. 4, pp. 1101–1117, 2023, doi: 10.1007/s40620-023-01573-4.
[22] H. Herle and K. V. Padmaja, "Relative Merits of Data Mining Algorithms of Chronic Kidney Diseases," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 6, pp. 575–583, 2021, doi: 10.14569/IJACSA.2021.0120667.
[23] R. Alanazi, "Identification and Prediction of Chronic Diseases Using Machine Learning Approach," J. Healthc. Eng., vol. 2022, 2022, doi: 10.1155/2022/2826127.
[24] M. M. Hassan et al., "A Comparative Study, Prediction and Development of Chronic Kidney Disease Using Machine Learning on Patients Clinical Records," Human-Centric Intell. Syst., vol. 3, no. 2, pp. 92–104, 2023, doi: 10.1007/s44230-023-00017-3.

[25] M. Rashed-Al-Mahfuz, A. Haque, A. Azad, S. A. Alyami, J. M. W. Quinn, and M. A. Moni, "Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening," IEEE J. Transl. Eng. Heal. Med., vol. 9, no. April, pp. 1–11, 2021, doi: 10.1109/JTEHM.2021.3073629.

[26] Y. Kale, S. Rathkanthiwar, P. Fulzele, and N. J. Bankar, "XGBoost Learning for Detection and Forecasting of Chronic Kidney Disease (CKD)," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 17s, pp. 137–150, 2024.

[27] P. A. Moreno-sanchez, "An Explainable Classification Model for Chronic Kidney Disease Patients," no. June, 2021.

[28] V. Gulati and N. Raheja, "Comparative analysis of machine learning techniques based on chronic kidney disease dataset," IOP Conf. Ser. Mater. Sci. Eng., vol. 1131, no. 1, p. 012010, 2021, doi: 10.1088/1757-899x/1131/1/012010.

[29] S. Akter et al., "Comprehensive Performance Assessment of Deep Learning Models in Early Prediction and Risk Identification of Chronic Kidney Disease," IEEE Access, vol. 9, no. iii, pp. 165184–165206, 2021, doi: 10.1109/ACCESS.2021.3129491.

[30] N. M. Farhan and B. Setiaji, "Indonesian Journal of Computer Science," Indones. J. Comput. Sci., vol. 12, no. 2, pp. 284–301, 2023, [Online]. Available: http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135

[31] C. Kaur, M. S. Kumar, A. Anjum, M. B. Binda, M. R. Mallu, and M. S. Al Ansari, "Chronic Kidney Disease Prediction Using Machine Learning," J. Adv. Inf. Technol., vol. 14, no. 2, pp. 384–391, 2023, doi: 10.12720/jait.14.2.384-391.

[32] V. Singh and D. Jain, "A Hybrid Parallel Classification Model for the Diagnosis of Chronic Kidney Disease," Int. J. Interact. Multimed. Artif. Intell., vol. 8, no. 2, pp. 14–28, 2023, doi: 10.9781/ijimai.2021.10.008.

[33] A. Vijayalakshmi and V. Sumalatha, "Survey on diagnosis of chronic kidney disease usingmachine learning algorithms," Proc. 3rd Int. Conf. Intell. Sustain. Syst. ICISS 2020, pp. 590–595, 2020, doi: 10.1109/ICISS49785.2020.9315880.

[34] Tsehay Admassu Assegie and Yenework Belayneh Chekol, "The Performance of Machine Learning for Chronic Kidney Disease Diagnosis," Emerg. Sci. Innov., vol. 1, pp. 33–40, 2023, doi: 10.46604/emsi.2023.11285.

[35] R. Sawhney, A. Malik, S. Sharma, and V. Narayan, "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease," Decis. Anal. J., vol. 6, no. January, p. 100169, 2023, doi: 10.1016/j.dajour.2023.100169.

[36] V. K. Venkatesan, M. T. Ramakrishna, I. Izonin, R. Tkachenko, and M. Havryliuk, "Efficient Data Preprocessing with Ensemble Machine Learning Technique for the Early Detection of Chronic Kidney Disease," Appl. Sci., vol. 13, no. 5, 2023, doi: 10.3390/app13052885.

[37] H. Khalid, A. Khan, M. Zahid Khan, G. Mehmood, and M. Shuaib Qureshi, "Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease," Comput. Intell. Neurosci., vol. 2023, pp. 1–14, 2023, doi: 10.1155/2023/9266889.

[38] P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," IEEE Access, vol. 9, pp. 17312–17334, 2021, doi: 10.1109/ACCESS.2021.3053763.
[39] S. Pal, "Prediction for chronic kidney disease by categorical and non_categorical attributes using different machine learning algorithms," Multimed. Tools Appl., vol. 82, no. 26, pp. 41253–41266, 2023, doi: 10.1007/s11042-023-15188-1.
[40] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," J. Pathol. Inform., vol. 14, no. September 2022, p. 100189, 2023, doi: 10.1016/j.jpi.2023.100189.
[41] A. S. Turiac and M. Zdrodowska, "Data Mining Approach in Diagnosis and Treatment of Chronic Kidney Disease," Acta Mech. Autom., vol. 16, no. 3, pp. 180–188, 2022, doi: 10.2478/ama-2022-0022.
[42] A. Alaiad, H. Najadat, B. Mohsen, and K. Balhaf, "Classification and Association Rule Mining Technique for Predicting Chronic Kidney Disease," J. Inf. Knowl. Manag., vol. 19, no. 1, 2020, doi: 10.1142/S0219649220400158.
[43] I. Saha, M. K. Gourisaria, and G. M. Harshvardhan, "Classification System for Prediction of Chronic Kidney Disease Using Data Mining Techniques," Lect. Notes Networks Syst., vol. 318, no. May, pp. 429–443, 2022, doi: 10.1007/978-981-16-5689-7_38.
[44] M. M. El Sherbiny, E. Abdelhalim, H. El-Din Mostafa, and M. M. El-Seddik, "Classification of chronic kidney disease based on machine learning techniques," Indones. J. Electr. Eng. Comput. Sci., vol. 32, no. 2, pp. 945–955, 2023, doi: 10.11591/ijeecs.v32.i2.pp945-955.
[45] S. M. Imran, N. Prakash, K. Hazeena, and S. Sivasubramanian, "Robust Chronic Kidney Impact Identification System Using Prab Algorithm," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 7s, pp. 404–411, 2024.
[46] D. C. Akalya, "International Journal of Informatics , Information System and Computer Engineering Risks of Chronic Kidney Disease Prediction using various Data Mining Algorithms," no. December 2021, 2023.
[47] M. M. Nishat et al., "A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms," EAI Endorsed Trans. Pervasive Heal. Technol., vol. 7, no. 29, pp. 1–12, 2021, doi: 10.4108/eai.13-8-2021.170671.
[48] M. Majid et al., "Using Ensemble Learning and Advanced Data Mining Techniques to Improve the Diagnosis of Chronic Kidney Disease," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 10, pp. 470–480, 2023, doi: 10.14569/IJACSA.2023.0141050.
[49] V. Singh, V. K. Asari, and R. Rajasekaran, "A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease," Diagnostics, vol. 12, no. 1, pp. 1–22, 2022, doi: 10.3390/diagnostics12010116.
[50] M. Nabeel, S. Majeed, M. J. Awan, H. Muslih-Ud-din, M. Wasique, and R. Nasir, "Review on effective disease prediction through data mining techniques," Int. J. Electr. Eng. Informatics, vol. 13, no. 3, pp. 717–733, 2021, doi: 10.15676/IJEEI.2021.13.3.13.

[51] M. Hosseinzadeh et al., "A diagnostic prediction model for chronic kidney disease in internet of things platform," Multimed. Tools Appl., vol. 80, no. 11, pp. 16933–16950, 2021, doi: 10.1007/s11042-020-09049-4.
[52] A. Sobrinho, A. C. M. D. S. Queiroz, L. Dias Da Silva, E. De Barros Costa, M. Eliete Pinheiro, and A. Perkusich, "Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries: A Comparative Analysis of Machine Learning Techniques," IEEE Access, vol. 8, pp. 25407–25419, 2020, doi: 10.1109/ACCESS.2020.2971208.
[53] O. A. Jongbo, A. O. Adetunmbi, R. B. Ogunrinde, and B. Badeji-Ajisafe, "Development of an ensemble approach to chronic kidney disease diagnosis," Sci. African, vol. 8, p. e00456, 2020, doi: 10.1016/j.sciaf.2020.e00456.
[54] D. K. Srivastava and P. K. Tiwari, "Chronic kidney disease prediction using data mining algorithms," Handb. Res. Dis. Predict. Through Data Anal. Mach. Learn., no. July, pp. 92–111, 2020, doi: 10.4018/978-1-7998-2742-9.ch006.
[55] H. Ilyas et al., "Chronic kidney disease diagnosis using decision tree algorithms," BMC Nephrol., vol. 22, no. 1, pp. 1–11, 2021, doi: 10.1186/s12882-021-02474-z.
[56] S. Rajarajeswari and T. Tamilarasi, "Risks of Chronic Kidney Disease Prediction using various Data Mining Algorithm," Int. J. Adv. Netw. Appl., vol. 12, no. 06, pp. 4776–4780, 2021, doi: 10.35444/ijana.2021.12607.
[57] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An empirical evaluation of machine learning techniques for chronic kidney disease prophecy," IEEE Access, vol. 8, pp. 55012–55022, 2020, doi: 10.1109/ACCESS.2020.2981689.