International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Survey of Predicting CKD Using Machine Learning

Dr. Ramah Sivakumar[1], R. Vijayalakshmi*[2]

Submitted:14/03/2024 **Revised**: 29/04/2024 **Accepted**: 06/05/2024

Abstract: Chronic kidney disease (CKD) poses a significant global public health challenge, affecting approximately 10% of the worldwide population. Despite recent increases in awareness, understanding of the disease remains limited. Alarmingly, the incidence, morbidity, mortality, and associated healthcare costs of CKD continue to rise, especially in low-income countries. Chronic kidney disease (CKD) represents the most severe stage of kidney damage, where the kidneys gradually lose functionality and may eventually cease to function entirely. Key risk factors for CKD include high blood pressure, cardiovascular disease, diabetes, advanced age, and a family history of kidney failure. Secondary risk factors encompass obesity, autoimmune diseases, systemic infections, urinary tract infections, and other kidney-related issues such as kidney damage, injury, or infection. Treatment strategies for CKD vary based on the patient's physical condition and typically involve lifestyle modifications, medications to manage related health problems, dialysis, and ultimately, kidney transplantation. Early diagnosis is crucial for effective treatment of CKD. The two primary methods for diagnosing CKD are blood and urine tests. However, these manual processes require expert involvement, which can be time-consuming and resource-intensive. To address these challenges, recent research has focused on developing automated, computerized diagnostic approaches using artificial intelligence. In this context, machine learning (ML) has emerged as the preferred choice among researchers due to its efficiency and accuracy.

Keywords: *Chronic kidney disease, Machine Learning, Supervised Learning unsupervised Learning, Kidney transplant*

1. Introduction

Chronic kidney disease (CKD) is a condition where the kidneys become so damaged that they can no longer effectively filter blood. The kidneys are responsible for removing waste and excess water from the blood to produce urine. In CKD, waste accumulates in the body because the damage occurs gradually over a long period. This chronic condition affects individuals globally, leading to various health complications. Diabetes, high blood pressure, and heart disease are significant contributors to CKD, along with factors such as age and gender. Symptoms of CKD include back pain, stomach pain, diarrhea, fever, nosebleeds, rash, and vomiting. Diabetes and high blood pressure are the two primary illnesses that cause long-term kidney damage, making their management crucial in preventing CKD. Unfortunately, CKD often remains asymptomatic until it reaches an advanced stage, leaving many unaware of their condition until it is too late.

Chronic Kidney Disease (CKD) is classified into five stages, based on the glomerular filtration rate (GFR), which measures how well the kidneys are filtering blood.

Stage 1: Kidney Damage with Normal GFR

GFR: 90 or higher

Description: In this stage, there is evidence of kidney damage, but the kidneys still function normally. Symptoms are often absent, and diagnosis typically occurs through routine testing for other conditions.

Management: Focus on addressing underlying conditions and lifestyle changes to prevent progression.

Stage 2: Mild Reduction in Kidney Function

GFR: 60-89

Description: There is a slight decline in kidney function. Like Stage 1, symptoms may not be apparent, and kidney damage is often identified through routine tests.

Management: Regular monitoring, controlling blood pressure and diabetes, and lifestyle modifications.

Stage 3: Moderate Reduction in Kidney Function

GFR: 30-59

Description: This stage is split into 3a (GFR 45-59) and 3b (GFR 30-44). Symptoms may start to appear, such as fatigue, swelling, and changes in urine output.

Management: More intensive monitoring, managing complications, and medications to slow disease progression.

Stage 4: Severe Reduction in Kidney Function

GFR: 15-29

Description: Kidney function is significantly impaired. Symptoms become more pronounced and may include anemia, bone disease, and increased blood pressure.

[1] *Dr. Ramah Sivakumar, Assistant Professor, Department of Computer Science, Bishop Heber College (Autonomous), Trichy -620 017, Tamilnadu, India*
ORCID ID : 0000-0002-5926-6060
[2] *Ms. R. Vijayalakshmi, Research scholar, Department of Computer, Bishop Heber College (Autonomous), Trichy -620 017, Tamilnadu, India*
ORCID ID : 0000-0002-5926-6060
** Corresponding Author Email: giripriya710@gmail.com*

Management: Preparing for potential kidney replacement therapy (dialysis or transplant), and more aggressive treatment of underlying conditions.

Stage 5: Kidney Failure

GFR: Less than 15

Description: Also known as end-stage renal disease (ESRD), this stage requires dialysis or a kidney transplant to sustain life. Symptoms are severe and can include nausea, vomiting, muscle cramps, and difficulty breathing.

Management: Dialysis, transplant evaluation, and supportive care to manage symptoms and improve quality of life.

2. Related Works

Machine Learning has revolutionized various fields, including healthcare, by providing powerful tools for data analysis and pattern recognition. In the context of CKD, ML algorithms can analyze large volumes of patient data to identify patterns and predict disease onset with high accuracy. Key ML techniques used in healthcare include supervised learning, unsupervised learning, and deep learning. These methods can handle complex datasets and provide insights that traditional statistical methods may overlook.

In their 2021 study, Pankaj Chittora and Sandeep Chaurasia [1], along with their colleagues, conducted a comprehensive analysis using the chronic kidney disease (CKD) dataset from the UC Irvine Machine Learning Repository. This dataset comprises 400 instances and 24 attributes. The methodology employed in their study encompassed five critical stages: dataset preprocessing, feature selection, classifier application, Synthetic Minority Over-sampling Technique (SMOTE) application, and performance analysis of the classifiers.

The researchers utilized three feature selection methods: Wrapper, Filter, and Embedded techniques. These methods were pivotal in identifying the most relevant attributes for CKD prediction. Following feature selection, the study applied several machine learning classifiers to train the model, including Artificial Neural Network (ANN), C5.0, Logistic Regression, Linear Support Vector Machine (LSVM), K-Nearest Neighbors (KNN), and Random Tree.

In his 2021 study, Khaled Mohamad Almustafa [2] explored the efficacy of various classification algorithms in predicting chronic kidney disease (CKD) using a CKD dataset. The algorithms utilized in this study included Random Tree, Decision Table, K-Nearest Neighbor (K-NN), J48, Stochastic Gradient Descent (SGD), and Naïve Bayes. Almustafa's approach emphasized the importance of feature selection to enhance prediction accuracy and reduce computational complexity.

A significant aspect of the study was the implementation of a feature selection model, which employed different classification algorithms to identify and retain the most relevant attributes for CKD prediction. By reducing the number of attributes, the study aimed to streamline the predictive process and improve model performance. The performance of the classification models was rigorously assessed and compared both with and without the application of feature selection techniques.

In their 2022 study, Qiong Bai [3] and colleagues conducted a thorough investigation into the prediction of end-stage kidney disease (ESKD) among chronic kidney disease (CKD) patients using machine learning (ML) algorithms. The study utilized a retrospective cohort of 748 CKD patients, who were followed for an average duration of 6.3 ± 2.3 years. The research adhered to ethical standards as outlined by the World Medical Association Declaration of Helsinki and received approval from the Peking University Third Hospital Medical Science Research Ethics Committee.

The study employed various ML algorithms, including logistic regression, naïve Bayes, random forest, decision tree, and K-nearest neighbors, to develop predictive models for ESKD. The performance of these models was evaluated using several metrics: accuracy, precision, recall, specificity, F1 score, and area under the curve (AUC). These metrics provided a comprehensive assessment of each model's predictive capabilities.

In 2023, Chamandeep Kaur [4] and colleagues conducted a study focused on predicting chronic kidney disease (CKD) using a dataset containing 400 patient records and 25 features. Their methodology encompassed several key stages: data preprocessing, training of machine learning (ML) models, and model selection. The research team employed a variety of ML algorithms to evaluate their effectiveness in predicting CKD. These algorithms included Random Forest, Decision Tree, Logistic Regression, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). The comparative analysis of these models aimed to identify the most accurate and reliable algorithm for CKD prediction.

A notable aspect of this study was its approach to handling missing data, which is a common challenge in clinical datasets. The researchers implemented techniques to manage and impute missing values, ensuring the integrity and completeness of the dataset. This step was crucial for improving the performance and accuracy of the ML models.

In their 2022 study, Muhammad Minoar Hossain [5] and colleagues conducted an in-depth evaluation of various feature optimization techniques to determine their effectiveness in predicting chronic kidney disease (CKD). The dataset used in this research comprised 400 CKD

patient records, including 25 attributes—24 predictive variables and one decision class. The researchers implemented comprehensive data preprocessing methods, handling both nominal and numerical data. This included filling in missing values and converting nominal values to integer values, ensuring the dataset was suitable for machine learning (ML) analysis.

The study focused on six feature optimization techniques, encompassing feature importance, feature selection, and feature reduction methods. The goal was to assess how these techniques impacted the accuracy of CKD diagnosis models. The researchers meticulously analyzed the performance of each technique, comparing their approach with existing methods to identify the most effective optimization strategy.

In their 2022 study, Manal A. Abdel-Fattah [6] and colleagues conducted a comprehensive analysis to predict chronic kidney disease (CKD) using machine learning (ML) techniques. The research followed a six-step methodology: data collection, data preprocessing, feature selection methods, model optimization and training, application of ML models, and evaluation of the models. The dataset used in the study was sourced from the UCI Machine Learning Repository and consisted of 400 samples, each with 24 features and one class label. The initial data preprocessing stage involved handling missing values and outliers to ensure the quality and integrity of the dataset. Feature selection was performed using two methods: Relief-F and chi-squared tests. These methods were employed to identify and select the most important features from the dataset, aiming to improve the performance of the predictive models. The models were optimized using a grid search strategy combined with stratified K-Fold cross-validation, ensuring robust and unbiased model training. Six ML classification algorithms were applied: Decision Tree (DT), Logistic Regression, Random Forest (RF), Support Vector Machine (SVM), Naive Bayes, and Gradient-Boosted Trees (GBT). These algorithms were chosen for their diverse strengths and complementary capabilities in handling different types of data and classification challenges. The performance of the models was evaluated using standard metrics, including accuracy, precision, recall, and F1-score. The study's results demonstrated that the SVM, DT, and GBT classifiers, when applied with the selected features, achieved the best performance, attaining 100% accuracy. This finding underscores the effectiveness of these classifiers in predicting CKD when coupled with appropriate feature selection.

Furthermore, the study compared the performance of the Relief-F feature selection method with the chi-squared feature selection method and the full feature set. The results indicated that the Relief-F method outperformed the chi-squared method and the full feature set, highlighting its superior ability to enhance model performance.

In their 2022 study, Rahul Sawhney [7] and colleagues focused on diagnosing chronic kidney disease (CKD) using a deep neural network-based Multi-Layer Perceptron (MLP) classifier. The dataset used in this research included attributes from 400 Indian patients, sourced from Managiri, Karaikudi, and Apollo Hospitals in Tamil Nadu, India. The study employed a combination of two feature extraction techniques and three feature selection methods to enhance the prediction accuracy of CKD. The research demonstrated the effectiveness of the deep neural network-based MLP classifier in diagnosing CKD. By integrating advanced feature extraction and selection techniques, the study aimed to enhance the predictive capabilities of the model. The use of supervised learning and batch training principles further strengthened the model's accuracy and reliability.

In their 2022 study, Á Sobrinho [8] and colleagues investigated the application of various machine learning techniques to predict chronic kidney disease (CKD) using medical records of Brazilian patients. The research methodology encompassed several critical components: data preprocessing, model implementation, validation methods, data augmentation, and multi-class classification metrics. The study began with data preprocessing, which involved handling missing values, normalizing data, and preparing it for model training. This step ensured the integrity and consistency of the dataset. To address the challenges of imbalanced and limited-size datasets, the researchers applied data augmentation techniques and evaluated the effectiveness of different oversampling methods. This approach aimed to improve the representativeness of minority classes and enhance model performance. The research utilized both ensemble and non-ensemble models, specifically decision tree (DT), random forest (RF), and multi-class AdaBoosted DTs algorithms. The implementation of these models was followed by a thorough validation process. Various validation methods were employed, including hold-out validation, multiple stratified cross-validation (CV), and nested CV, to ensure robust and unbiased evaluation of model performance. To further refine the models, the study explored dynamic selection methods and applied linear regression to verify the statistical significance of the results. This comprehensive approach provided deeper insights into the models' predictive capabilities and their robustness under different conditions. One of the key contributions of this research was the development of a decision support system (DSS) based on the machine learning model with the highest performance. This DSS was designed to assist healthcare providers in the identification and monitoring of CKD within Brazilian communities, leveraging the predictive power of the optimized machine learning models.

In their 2022 study, Sarah A. Ebiaredoh-Mienye[9] and

colleagues presented an advanced approach to improving chronic kidney disease (CKD) detection by combining filter-based feature selection techniques with a cost-sensitive variant of the AdaBoost algorithm. The methodology included a detailed comparison between the traditional AdaBoost algorithm and the proposed cost-sensitive AdaBoost, emphasizing their differences and the benefits of the latter in handling imbalanced datasets. The study provided a comprehensive overview of the experimental setup and the results obtained from applying the proposed approach. Various performance evaluation metrics, including accuracy, precision, recall, and F1-score, were used to assess the effectiveness of the proposed model. The results demonstrated that the combination of information gain and cost-sensitive AdaBoost significantly improved CKD detection compared to traditional methods.

In their 2023 study, A. Farjana [10] and colleagues proposed and evaluated nine machine learning (ML) approaches to predict chronic kidney disease (CKD). The dataset used in this research consisted of 400 records with 14 attributes and was sourced from Kaggle.com. The aim was to identify the most effective classifier for predicting CKD by comparing the performance of various ML models. The study implemented and compared the following nine ML models: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression (LR), Naïve Bayes, Extra Tree Classifiers, AdaBoost, XGBoost, LightGBM. CKD dataset used in this study included 14 attributes relevant to kidney function and patient health. The data preprocessing steps ensured that the dataset was clean and suitable for training and testing the models. The research involved a comprehensive evaluation of each model's performance. Standard metrics such as accuracy, precision, recall, and F1-score were used to assess the effectiveness of the models. The study's comparative analysis provided insights into the strengths and weaknesses of each classifier in predicting CKD.

3. Results and Discussions

Results and discussion are describing the table 1.

4. Research Gap

Machine learning (ML) shows promise in predicting chronic kidney disease (CKD), but several research gaps remain. High-quality, diverse datasets are lacking, limiting model generalizability. Effective feature selection and engineering are needed to capture CKD's multifactorial nature. ML models must be interpretable for clinical use, but current methods often lack transparency. Class imbalance in CKD datasets biases predictions, necessitating advanced handling techniques. Temporal dynamics of CKD require longitudinal data analysis for accurate predictions. Integrating ML models into clinical workflows is crucial for real-world impact. Ethical and legal concerns around patient privacy and algorithmic bias must be addressed to ensure equitable outcomes. Collaborative efforts across disciplines are essential to advance ML-based CKD prediction.

Table 1. Results and Discussions

S.No	Author and Year	Technique Applied	Dataset	Output	Drawback
1	Pankaj Chittora, Sandeep Chaurasia et al.	ANN, C5.0, Logistic Regression, CHAID, KNN, Random Tree LSVM	UCI	Full Features: C5.0 - highest accuracy -96.10% Feature Selection: CFS: LSVM: Accuracy - 95.12%	The efficacy of machine learning models is dataset-dependent, showing comparable results to prior research with no significant differences;
2	Chamandeep Kaur et al.	Random Forest, Decision Tree, Logistic Regression, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM)	UCI	RF: 96%	Managing missing data, which could result in false positives, as the data was lost at random and perfect accuracy is unachievable without a collaborative imputer.

3	Qiong Bai et al.	Logistic regression, Naïve Bayes, Fandom Forest, Decision tree, and K-nearest neighbors	Cohort Study	Navie Bayes: 86%	A cohort of fewer than 1000 subjects and a rare occurrence of ESKD, affecting model performance, and the lack of initial quantitative urine tests, leading to an ESKD prediction model without urine variables.
4	Manal A Abdel-Fattah et al.	Decision tree, Logistic regression, Naive Bayes, Random Forest, Support Vector Machine, and Gradient-Boosted Trees	UCI	LDA: 99.5%	It includes the challenges associated with big data analytics, particularly in handling enormous and heterogeneous data storage in real-time.
5	Andressa C. M. da Silveira et al.	Decision tree (DT), Random forest (RF), and Multi-class AdaBoosted DTs algorithms.	Brazilian Community	DT: 94.44%	It includes high computational costs from using gridSearchCV for parameter optimization, especially for ensemble models, and a reduced number of manually augmented instances for the high-risk class, affecting performance evaluation despite partial mitigation through nested cross-validation.
6	Sarah A. Ebiaredoh-Mienye et al.	Traditional adaboost algorithm and the proposed cost-sensitive adaboost	UCI	Proposed AdaBoost: 96.7%	Dataset size, biases in data collection, generalizability of findings
7	Minhaz Uddin Emon et al.	Logistic Regression, Naive Bayes, Multilayer Perceptron, Stochastic Gradient Descent, Adaptive Boosting, Bagging, Decision Tree, and Random Forest	UCI	Random Forest: 99%	Data Quality and Generalizability
8	Hamida Ilyas et al.	J48 and Random Forest	UCI	J48:85.5%	Relatively small dataset used for analysis, which may impact the generalizability of the findings.
9	Md. Ariful Islam et al.	SVM, KNN, LGBM, XgBoost, CatBoost, Ada, and hybrid models	UCI	Xgboost with PCA:99.2%	Relatively small dataset used for analysis, which may impact the generalizability of the findings.

10	Khaled Mohamad Almustafa	Random tree, decision table, K-nearest neighbor (K-NN), J48, stochastic gradient descent (SGD), and Naïve Bayes	UCI	Naïve Bayes:99%	Relatively small dataset used for analysis, which may impact the generalizability of the findings

5. Conclusion

In conclusion, the application of machine learning (ML) methods for predicting chronic kidney disease (CKD) holds significant potential to revolutionize early diagnosis and personalized treatment strategies. Our survey has highlighted the various ML techniques employed in CKD prediction, including supervised learning, unsupervised learning, and ensemble methods. While these approaches have demonstrated promising results, several challenges and research gaps persist. Key issues such as data quality and availability, feature selection and engineering, model interpretability, handling imbalanced data, temporal dynamics, integration with clinical workflows, and ethical considerations remain critical areas for further research. Addressing these gaps will require the development of high-quality, comprehensive datasets and advanced algorithms that can accurately and transparently predict CKD across diverse populations and healthcare settings.

Future research should focus on creating robust, explainable ML models that are seamlessly integrated into clinical practice, ensuring they are both user-friendly and actionable for healthcare providers. Additionally, ethical frameworks must be established to safeguard patient privacy and mitigate algorithmic biases, promoting equitable healthcare outcomes. By tackling these challenges through interdisciplinary collaboration, the potential of ML in CKD prediction can be fully realized, leading to improved patient outcomes and more efficient healthcare delivery. As the field progresses, continuous advancements in ML techniques and their application to CKD will undoubtedly contribute to a deeper understanding and better management of this critical public health issue.

Funding: This research received no external funding.

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of

their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Pankaj Chittora, Sandeep Chaurasiai et al [2021]. Prediction of Chronic Kidney Disease - A Machine Learning Perspective, IEEE

[2] Khaled Mohamad Almustafa [2021], Prediction of chronic kidney disease using different classification algorithms, Elsevier

[3] Q Bai, C Su, W Tang, Y Li et al [2022]. Machine learning to predict end stage kidney disease in chronic kidney disease, Scientific Reports

[4] Chamandeep Kaur, M. Sunil Kumar et al [2023]. Chronic Kidney Disease Prediction Using Machine Learning. Journal of Advances in Information Technology, Vol. 14, No. 2, 2023

[5] Muhammad Minoar Hossain et al. 2022, Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease, Elsevier

[6] MA Abdel-Fattah, NA Othman, N Goher et al [2022]. Predicting chronic kidney disease using hybrid machine learning based on apache spark, Computational Intelligence and Neuroscience

[7] Rahul Sawhney, Aabha Malik, Shilpi Sharma, Vipul Narayan. A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease, Decision Analytics Journal, Volume 6, 2023, 100169, ISSN 2772-6622,

[8] Á Sobrinho, LD Silva, EB Costa et al [2022], Exploring Early Prediction of Chronic Kidney Disease Using Machine Learning Algorithms for Small and Imbalanced Datasets. Appl. Sci. 2022, 12(7), 3673

[9] Ebiaredoh-Mienye, Sarah A., Theo G. Swart, Ebenezer Esenogho, and Ibomoiye Domor Mienye. 2022. "A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease"

[10] A Farjana, FT Liza, PP Pandit, MC Das, M Hasan, F Tabassum, MH Hossen, et al [2023]. Predicting Chronic Kidney Disease Using Machine Learning Algorithms, IEEE

[11] Francesco Sanmarchi, Claudio Fanconi et al [2023]. Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review. Journal of Nephrology (2023) 36:1101–1117

[12] Mirza Muntasir Nishat1, Fahim Faisal, et al [2021]. A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms. EAI Endorsed Transactions on Pervasive

Health and Technology 10 2021 - 11 2021 | Volume 7 | Issue 29 | e1

[13] Md. Ariful Islam Md. Ziaul Hasan Majumder, Md. Alomgeer Hussein [2023], chronic kidney disease prediction based on machine learning algorithms. Journal of Pathology Informatics, Volume 14, 2023, 100189

[14] Minhaz Uddin Emon et al [2021], Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches, IEEE