International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

A Comprehensive Review of Data Mining Classification Techniques for Early Detection and Analysis of Chronic Kidney Disease

Anshita Singh[1], Heera Singh Yadav[2]
Department of Computer Science & Engineering,
Babu Sunder Singh Institute of Technology & Management, (AKTU), Uttar Pradesh, India

Abstract— **chronic kidney disease (CKD) poses a growing global health challenge due to its asymptomatic nature in the early stages and the potential for progression to end-stage renal failure. Early detection is vital to ensure timely intervention and reduce mortality rates. Data mining classification techniques have emerged as powerful tools for analyzing complex clinical datasets and improving diagnostic accuracy. This comprehensive review explores various classification methods—including Decision Trees, Support Vector Machines (SVM), Naïve Bayes, k-Nearest Neighbors (k-NN), and ensemble models—used in CKD prediction and analysis. The paper critically examines the strengths, limitations, and performance metrics of these models based on recent studies. It also highlights the role of feature selection, data preprocessing, and imbalanced dataset handling in enhancing model efficiency. The review concludes with insights into current trends, challenges, and future directions for developing robust, interpretable, and clinically relevant classification frameworks to support nephrologists in early CKD diagnosis.**

Keywords— **Chronic Kidney Disease (CKD), Data Mining, Classification Techniques, Machine Learning, Early Detection, Clinical Decision Support, Predictive Modeling, Health Informatics, Imbalanced Data, Feature Selection.**

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a long-term condition characterized by a gradual loss of kidney function over time, often progressing silently until reaching advanced stages. According to the Global Burden of Disease Study, CKD has become one of the leading causes of mortality worldwide, affecting approximately 10% of the global population and resulting in significant healthcare costs and resource utilization [1]. The early stages of CKD are typically asymptomatic, making timely diagnosis and treatment challenging. Consequently, early detection through reliable diagnostic systems is crucial for slowing disease progression, minimizing complications, and improving patient outcomes.

Traditional diagnostic methods rely heavily on clinical expertise and laboratory tests such as serum creatinine levels, estimated glomerular filtration rate (eGFR), and proteinuria. However, these indicators may not always reflect early pathological changes or may be influenced by non-kidney-related factors [2]. To address these limitations, data mining and machine learning techniques have been increasingly employed in the medical domain to discover hidden patterns in large datasets, enabling predictive modeling and decision support systems for various diseases, including CKD [3].

Classification techniques—such as Decision Trees, Support Vector Machines (SVM), Naïve Bayes, k-Nearest Neighbors (k-NN), and Random Forests—have proven effective in identifying high-risk CKD patients and categorizing disease stages based on patient attributes [4]. These models are capable of handling high-dimensional data, capturing non-linear relationships, and providing probabilistic predictions that can assist healthcare professionals in making informed decisions.

Moreover, advancements in feature selection, data preprocessing, and ensemble learning methods have significantly improved model performance, especially in handling noisy, imbalanced, or incomplete healthcare datasets [5]. Recent studies have demonstrated that integrating multiple classifiers and optimizing algorithm parameters can further enhance diagnostic accuracy and robustness [6].

This review aims to present a detailed overview of data mining classification techniques used in the early detection and analysis of CKD. It discusses the theoretical underpinnings, comparative performance, and practical applications of various classifiers, while also highlighting current research trends, limitations, and opportunities for future development in this critical area of healthcare analytics.

II. LITERATURE SURVEY

Over the past decade, the application of data mining and machine learning techniques for the early detection and analysis of Chronic Kidney Disease (CKD) has attracted significant research interest. Several studies have evaluated the effectiveness of classification algorithms in identifying CKD at various stages, using clinical and laboratory data.

A. Decision Tree Algorithms:

Decision Tree classifiers, such as C4.5 and CART, are widely used in CKD diagnosis due to their simplicity and interpretability. Quinlan (1993) introduced C4.5, which has been used in various healthcare datasets for classification purposes [1]. In a study by Shillan et al. (2019), Decision Trees achieved promising accuracy and were favored for their transparent decision rules, making them suitable for clinical use [2].

B. Support Vector Machines (SVM):

SVM has demonstrated high classification accuracy in CKD prediction, particularly in high-dimensional spaces. Kusiak et al. (2005) applied SVM to medical diagnosis problems and

found it effective for binary classification [3]. Similarly, Rashid et al. (2020) achieved over 95% accuracy in CKD detection using an SVM model on the UCI CKD dataset, proving its robustness and generalization capability [4].

C. Naïve Bayes Classifier:
Naïve Bayes, despite its simplistic assumption of feature independence, has been employed successfully in CKD prediction. Patil and Kumaraswamy (2009) compared Naïve Bayes with other classifiers and found it to be computationally efficient with acceptable predictive performance [5]. However, its performance often suffers when the independence assumption is violated.

D. k-Nearest Neighbors (k-NN):
k-NN is a non-parametric classifier that has shown reasonable accuracy in diagnosing CKD when trained on normalized and balanced datasets. A study by Kora and Kalva (2015) revealed that k-NN achieved high sensitivity and specificity, especially when an optimal value of 'k' was selected through cross-validation [6].

E. Ensemble Methods (Random Forest, AdaBoost, XGBoost):
Ensemble techniques such as Random Forest and Gradient Boosting have gained traction due to their superior accuracy and stability. Random Forest, introduced by Breiman (2001), is an ensemble of decision trees that reduces overfitting by averaging predictions [7]. Sarker et al. (2020) applied Random Forest to CKD datasets and achieved over 98% accuracy [8]. Similarly, XGBoost has shown excellent performance in dealing with imbalanced datasets, which are common in medical data.

F. Deep Learning Approaches:
Though less explored in CKD prediction compared to other chronic diseases, deep learning models like artificial neural networks (ANN) have started gaining popularity. Thomas and Jayapradha (2021) reported that ANN outperformed traditional classifiers in terms of precision and recall in detecting early-stage CKD [9].

G. Feature Selection and Preprocessing:
Effective feature selection significantly influences classification performance. Algorithms such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and correlation-based feature selection (CFS) have been used to identify the most relevant attributes. Kumar and Ramana (2016) emphasized that feature reduction not only boosts accuracy but also improves computational efficiency [10].

H. Handling Imbalanced Data:
CKD datasets are often imbalanced, with more non-CKD than CKD instances. Techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and cost-sensitive learning have been adopted to mitigate this issue. Studies by Ahmed et al. (2021) have shown that oversampling combined with ensemble classifiers enhances minority class prediction without compromising overall performance [11].

In summary, while various classification techniques have shown promise in CKD prediction, their effectiveness depends largely on the quality of data, preprocessing methods, and the choice of performance metrics. The growing interest in explainable AI also underscores the need for interpretable models in clinical applications.

TABLE 1: LITERATURE REVIEW TABLE FOR PREVIOUS YEAR RESEARCH PAPER COMPARISON

S. No	Title of Research Paper	Authors	Year	Method/Technique Used	Dataset	Key Findings
1	Prediction of Chronic Kidney Disease Using Machine Learning Algorithms	Vijayarani & Dhayanand	2015	SVM, Decision Tree, Naïve Bayes	UCI CKD	SVM outperformed others with 91% accuracy.
2	Classification of CKD Using Data Mining Techniques	Soni et al.	2011	J48, Naïve Bayes, Random Forest	UCI CKD	J48 achieved best accuracy (84%).
3	Early Detection of CKD Using Decision Tree Algorithms	Hassan et al.	2017	ID3, C4.5	Clinical Dataset	C4.5 provided better performance in accuracy and interpretability.
4	Diagnosis of Chronic Kidney Disease Using Artificial Neural Network	Jayapradha & Thomas	2021	ANN	UCI CKD	ANN showed 97% accuracy and good generalization.
5	Chronic Kidney Disease Prediction Using Random	Yadav et al.	2020	Random Forest	UCI CKD	Random Forest gave over 96% accuracy.

	Forest					
6	Prediction of CKD Using SVM and PCA	Rashid et al.	2020	SVM + PCA	UCI CKD	Feature reduction improved accuracy to 95%.
7	Classification of CKD with Naïve Bayes	Mishra & Jaiswal	2018	Naïve Bayes	UCI CKD	Achieved 88% accuracy, efficient for small datasets.
8	Machine Learning Approach for Early Prediction of CKD	Rani & Raj	2019	SVM, k-NN, Decision Tree	UCI CKD	k-NN provided 92% accuracy with optimal k.
9	Comparative Analysis of ML Algorithms for CKD Diagnosis	Patel & Doshi	2020	SVM, RF, ANN	UCI CKD	Ensemble of SVM and RF yielded best result.
10	Deep Learning Model for Early Detection of CKD	Khan et al.	2021	Deep Neural Network	Local Hospital Data	DNN achieved 97.5% accuracy; better with more layers.
11	Predicting CKD Using Hybrid Classification Models	Sarker et al.	2020	XGBoost, RF, AdaBoost	UCI CKD	XGBoost delivered highest F1 score (0.98).
12	CKD Diagnosis Using Logistic Regression and Feature Selection	Roy et al.	2022	Logistic Regression, RFE	UCI CKD	Reduced features increased accuracy to 90%.
13	Imbalanced Data Handling in CKD Detection	Ahmed et al.	2021	SMOTE + Ensemble Learning	UCI CKD	Oversampling improved minority class recall.
14	CKD Prediction Using Ensemble Classifiers	Kumar & Singh	2019	Voting Classifier	UCI CKD	Ensemble outperformed individual models.
15	Data Mining for Medical Diagnosis of CKD	Uddin et al.	2018	J48, RF, Naïve Bayes	UCI CKD	J48 provided best performance with balanced data.
16	Classification of Kidney Disease Using Machine Learning	Palaniappan & Awang	2008	Decision Tree	Malaysian Dataset	Proved feasibility of DT for small datasets.
17	CKD Analysis Using Bayesian and Lazy Classifiers	Sharma et al.	2017	Naïve Bayes, k-NN	UCI CKD	Naïve Bayes faster; k-NN more accurate.
18	Automated Diagnosis System for CKD	Ramya & Vijay	2020	ANN, SVM	UCI CKD	ANN gave 96.2% accuracy.
19	Feature Selection for CKD Using PSO	Maheswari et al.	2019	Particle Swarm Optimization	UCI CKD	PSO improved classifier accuracy.
20	Analyzing CKD Using Decision Tree and	Meena & Anand	2017	J48, NB	UCI CKD	J48 outperformed Naïve Bayes with 94%

	Naïve Bayes					accuracy.
21	Enhanced CKD Diagnosis Using Fuzzy Logic	Adebayo et al.	2020	Fuzzy Logic	Clinical Data	Fuzzy logic improved interpretability.
22	Comparative Study of Classifiers for CKD Prediction	Chauhan & Jaiswal	2019	SVM, k-NN, ANN	UCI CKD	ANN provided best accuracy and stability.
23	A Predictive Model for CKD Using XGBoost	Kapoor et al.	2022	XGBoost	UCI CKD	XGBoost achieved 98.2% accuracy.
24	Hybrid Deep Learning Model for CKD	Fatima et al.	2021	CNN + RNN	Local Hospital Data	Hybrid model showed better sensitivity.
25	Performance Evaluation of ML Techniques for CKD Detection	Singh & Tiwari	2019	SVM, RF, Logistic Regression	UCI CKD	RF achieved highest AUC score.

III. ALGORITHMS

In the domain of Chronic Kidney Disease (CKD) prediction and classification, numerous data mining and machine learning algorithms have been applied to enhance the accuracy, sensitivity, and interpretability of diagnostic models. The commonly used algorithms can be broadly categorized into traditional machine learning, ensemble methods, and advanced models like deep learning.

A. Decision Tree (DT)
Description: A tree-like structure where nodes represent features, branches represent decisions, and leaves represent outcomes.

Advantage: Easy to interpret and implement.

Common Variants: ID3, C4.5, CART.

B. Support Vector Machine (SVM)
Description: A supervised learning model that identifies an optimal hyperplane for separating classes.

Advantage: Works well with high-dimensional data and provides robust classification performance.

Kernel Types: Linear, Polynomial, Radial Basis Function (RBF).

C. Naïve Bayes (NB)
Description: A probabilistic classifier based on Bayes' theorem assuming feature independence.

Advantage: Fast, simple, and efficient on small datasets.

Limitation: Assumption of feature independence may reduce accuracy.

D. k-Nearest Neighbors (k-NN)
Description: A non-parametric method where classification is based on the majority vote of the nearest neighbors.

Advantage: Simple and effective with proper distance metrics.

Parameter: The choice of 'k' is critical for performance.

E. Random Forest (RF)
Description: An ensemble of decision trees using bagging technique.

Advantage: Reduces overfitting, improves generalization, and handles missing data well.

F. Logistic Regression (LR)
Description: A statistical model used for binary classification problems.

Advantage: Effective for linearly separable data and interpretable coefficients.

G. Gradient Boosting Machines (GBM) / XGBoost
Description: An ensemble method that builds models sequentially to reduce bias.

Advantage: High accuracy, handles missing data, and robust to overfitting.

Popular Tools: XGBoost, LightGBM, CatBoost.

H. Artificial Neural Networks (ANN)
Description: Inspired by the human brain, ANN uses interconnected nodes (neurons) in layers.

Advantage: Capable of modeling complex non-linear relationships.

Limitation: Requires large datasets and computational resources.

I. Deep Neural Networks (DNN)
Description: A multi-layer ANN that can automatically extract hierarchical features.

Application: Useful in modeling subtle CKD patterns not evident in shallow models.

J. Fuzzy Logic Systems
Description: Uses degrees of truth rather than binary decisions, suitable for medical uncertainty.

Advantage: Mimics human reasoning in decision-making processes.

K. Ensemble Voting Classifier
Description: Combines predictions from multiple classifiers (e.g., SVM, RF, NB) to improve accuracy.

Types: Hard voting (majority), soft voting (average probabilities).

L. Feature Selection Algorithms
Examples: Recursive Feature Elimination (RFE), Correlation-Based Feature Selection (CFS), Particle Swarm Optimization (PSO).

Purpose: To reduce dimensionality, eliminate irrelevant features, and improve model performance.

IV. CONCLUSION

Chronic Kidney Disease (CKD) remains a critical public health concern that demands timely diagnosis and intervention. With the rise of data availability in the healthcare sector, data mining and machine learning classification techniques have shown great promise in enhancing the early detection and analysis of CKD. This comprehensive review highlights the strengths, limitations, and comparative performance of various algorithms including Decision Trees, Support Vector Machines, Naïve Bayes, k-Nearest Neighbors, Logistic Regression, ensemble models like Random Forest and XGBoost, and advanced neural network-based approaches.

Each classification algorithm offers unique advantages—ranging from interpretability and speed to accuracy and robustness—depending on the nature of the dataset and application. Ensemble and hybrid models generally outperform single classifiers in terms of precision and recall, especially when combined with appropriate feature selection and data preprocessing techniques. Moreover, the increasing use of deep learning and fuzzy systems further broadens the scope of intelligent diagnostic tools.

However, challenges remain in handling imbalanced datasets, ensuring model interpretability, and translating machine learning predictions into actionable clinical insights. Future research should focus on developing explainable AI models, integrating real-time clinical data, and validating these techniques through clinical trials to build trustworthy, scalable, and patient-centric decision support systems.

Ultimately, integrating advanced data mining techniques into clinical practice has the potential to revolutionize the early detection and personalized treatment of CKD, thereby improving patient outcomes and optimizing healthcare resources.

REFERENCES

[1] GBD Chronic Kidney Disease Collaboration. (2020). Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet, 395(10225), 709–733.
[2] Levey, A. S., & Coresh, J. (2012). Chronic kidney disease. The Lancet, 379(9811), 165–180.
[3] Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920–1930.
[4] Dua, S., & Du, X. (2016). Data mining and machine learning in cybersecurity. CRC Press.
[5] Choubey, D. S., Paul, S., & Srivastava, R. (2020). A machine learning based approach for the diagnosis of chronic kidney disease. Procedia Computer Science, 167, 674–681.
[6] Kaur, H., & Kumari, V. (2022). Predictive modeling and classification for chronic kidney disease using ensemble techniques. Health and Technology, 12(4), 809–822.
[7] Khatri, P., & Srivastava, S. (2020). Prediction of chronic kidney disease using data mining techniques. International Journal of Advanced Science and Technology, 29(4), 8840–8847.
[8] Hemanth, D. J., & Anitha, J. (2018). Machine learning-based diagnosis of chronic kidney disease using feature selection and classification algorithms. Procedia Computer Science, 132, 1232–1240.
[9] Kalaiselvi, T., & Saranya, R. (2021). Detection of chronic kidney disease using classification algorithms. Journal of Physics: Conference Series, 1916(1), 012019.
[10] Bhatia, S., & Soni, S. (2016). A predictive model for the diagnosis of chronic kidney disease using machine learning techniques. International Journal of Computer Applications, 160(7), 22–26.
[11] Bhattacharjee, A., & Chaki, R. (2021). Ensemble learning techniques for predicting chronic kidney disease. Advances in Intelligent Systems and Computing, 1183, 27–35.
[12] Agarwal, A., & Sinha, M. (2022). Comparative study of machine learning classifiers for CKD prediction. International Journal of Engineering Trends and Technology, 70(2), 200–205.
[13] Kandasamy, R., & Chandrasekar, C. (2017). CKD prediction using classification algorithms. International Journal of Pure and Applied Mathematics, 117(22), 525–531.
[14] Patel, H., & Prajapati, P. (2018). Study and analysis of the prediction of chronic kidney disease using XGBoost. International Journal of Scientific Research in Computer Science, 6(2), 67–71.
[15] Bharati, S., Podder, P., & Mondal, M. R. H. (2021). Hybrid deep learning for diabetes and CKD prediction. Healthcare Analytics, 1, 100016.
[16] Anurag et. al., “Load Forecasting by using ANFIS”, International Journal of Research and Development in

Applied Science and Engineering, Volume 20, Issue 1, 2020
[17] Raghawend, Anurag, "Detect Skin Defects by Modern Image Segmentation Approach, Volume 20, Issue 1, 2020.
[18] Sunil, M. M., & Patil, A. (2016). Performance analysis of classification algorithms for chronic kidney disease prediction. International Journal of Computer Science and Information Technologies, 7(3), 1141–1146.
[19] & Choudhury, D. (2021). Chronic kidney disease prediction using ensemble learning. International Journal of Engineering Research & Technology, 10(3), 285–291.
[20] Murugesan, B., & Srinivasan, K. (2022). Classification and prediction of CKD using supervised learning algorithms. Measurement: Sensors, 24, 100478.
[21] Rao, N. V., & Rani, B. U. (2019). Machine learning models for chronic kidney disease prediction. Materials Today: Proceedings, 37, 3360–3364.
[22] Wagh, M. P., & Mahajan, R. S. (2021). Detection of CKD using logistic regression and decision tree. International Journal of Scientific & Engineering Research, 12(8), 1132–1136.
[23] Ahmed, M., & Pathan, M. (2019). Early detection of chronic kidney disease by hybrid model. Procedia Computer Science, 165, 751–759.
[24] Prakash, J., & John, A. (2020). Predicting CKD using ML techniques: A case study. Journal of Applied Science and Computations, 7(6), 89–95.
[25] Saxena, S., & Singh, D. (2022). Optimizing feature sets for early CKD detection. Machine Learning with Applications, 6, 100165.
[26] Saeed, M. H., & Abdelrahman, A. M. (2021). Comparison of classifiers for CKD prediction. Computer Methods and Programs in Biomedicine, 208, 106280.
[27] Qureshi, S. A., & Usman, M. (2021). CKD prediction system using ensemble classifiers. International Journal of Advanced Computer Science and Applications, 12(4), 221–227.
[28] Verma, S., & Khanna, R. (2020). Medical diagnosis of CKD using machine learning algorithms. International Journal of Scientific & Technology Research, 9(4), 345–349.
[29] Naik, G. R., & Nayak, P. (2022). Hybrid machine learning model for CKD prediction. Biomedical Signal Processing and Control, 68, 102709.
[30] Yassin, A. A., & Elhoseny, M. (2018). An intelligent system for CKD classification using ML algorithms. Journal of Ambient Intelligence and Humanized Computing, 9(6), 1805–1814.
[31] Yadav, S., & Mishra, S. K. (2021). Decision support system for early prediction of CKD using data mining. ICT Express, 7(2), 205–211.