JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Machine Learning-Based Kidney Disease Prediction Using Multidimensional Clinical Data

Ms. Jyoti Verma[1], Dr. Yogesh Kumar Rathore[2]
[1]PG Student, Department of CSE – Artificial Intelligence and Machine Learning
[2]Associate Professor (CSE) & Dean (T&P), Department of Computer Science & Engineering
Shri Shankaracharya Institute of Professional Management and Technology, Raipur, India – 492015
Email: jyoti.verma@ssipmt.com | y.rathore@ssipmt.com

Abstract

Chronic Kidney Disease (CKD) is a serious health condition that gradually reduces kidney function and often remains unnoticed in its early stages. Late diagnosis increases the risk of kidney failure, cardiovascular complications, and long-term medical treatment. This study presents a machine learning-based approach for predicting kidney disease severity using a large clinical dataset containing 20,539 patient records and 42 medical attributes. The dataset includes demographic information, laboratory test results, urine analysis, blood pressure, lifestyle factors, and medical history.

A structured data preprocessing pipeline was applied to improve data quality, including handling missing values, encoding categorical variables, normalization, and feature selection. Five supervised learning algorithms—Decision Tree, Support Vector Machine, Random Forest, k-Nearest Neighbors, and Artificial Neural Network—were trained and evaluated using standard performance measures such as accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.

Among all models, the Random Forest classifier achieved the best performance with an accuracy of 99.21% and an AUC of 0.994. The model also showed a very low rate of false-negative predictions, which is important in medical diagnosis. The results indicate that ensemble learning methods can effectively analyze multidimensional clinical data and support early identification of kidney disease severity levels.

The proposed framework can assist healthcare professionals in screening high-risk patients and improving early intervention, especially in healthcare settings with limited specialist resources.

Keywords

Chronic Kidney Disease, Machine Learning, Random Forest, Kidney Disease Prediction, Clinical Data, Healthcare Analytics, Early Detection, Multiclass Classification.

1. Introduction

Chronic Kidney Disease (CKD) is a progressive medical condition characterized by the gradual decline of kidney function over time. The kidneys perform essential physiological functions, including filtration of metabolic waste products, regulation of fluid and electrolyte balance, maintenance of blood pressure, and production of hormones involved in red blood cell formation and bone health [1]. Impairment of these functions can lead to severe health complications and significantly affect a patient's quality of life. Due to its increasing prevalence and associated healthcare burden, CKD has become a major public health concern worldwide [2].

Recent global studies indicate that approximately 10–15% of the adult population is affected by CKD, with millions of cases remaining undiagnosed because the disease often progresses silently during its early stages [3]. Patients usually experience few or no symptoms until substantial kidney damage has occurred, making timely diagnosis difficult [4]. As the disease advances, individuals become more vulnerable to complications such as cardiovascular disorders, anemia, mineral and bone abnormalities, and eventually End-Stage Renal Disease (ESRD), which requires dialysis or kidney transplantation for survival [5].

Several factors contribute to the onset and progression of CKD. Diabetes mellitus and hypertension are recognized as the leading causes, while aging, obesity, genetic predisposition, and unhealthy lifestyle habits further increase the risk of kidney damage [6]. The interaction among these risk factors creates complex clinical patterns that are often difficult to evaluate using conventional diagnostic approaches. Traditional diagnosis primarily depends on biomarkers such as serum creatinine, estimated glomerular filtration rate (eGFR), and urinary albumin levels [7]. Although these indicators are clinically important, they may not always identify early-stage kidney dysfunction and are often influenced by demographic and physiological variations among patients [8].

The growing availability of healthcare data has encouraged the adoption of advanced computational techniques for disease prediction and risk assessment. Machine learning has emerged as an effective approach for analyzing large and complex clinical datasets by identifying hidden relationships among multiple variables simultaneously [9]. Unlike traditional statistical methods, machine learning algorithms can learn patterns directly from data and generate predictive models capable of supporting clinical decision-making [10]. Various supervised learning techniques, including Decision Trees, Support Vector Machines, Random Forests, Artificial Neural Networks, and k-Nearest Neighbors, have demonstrated promising performance in medical diagnosis and disease prediction tasks [11].

Motivated by these developments, this study proposes a machine learning-based framework for kidney disease severity prediction using a large-scale clinical dataset containing 20,539 patient records and 42 clinical attributes. The objective is to develop an accurate and reliable prediction model capable of identifying different levels of disease severity, thereby supporting early intervention and improving healthcare outcomes for individuals at risk of kidney disease [12].

2. Literature Survey

2.1 Decision Tree-Based Approaches

Decision Tree (DT) algorithms have been widely used for CKD prediction due to their simplicity and interpretability. Charleonnan et al. [11] reported high classification accuracy using DT models on clinical datasets. However, decision trees are sensitive to data variations and may suffer from overfitting, reducing their performance on unseen data.

2.2 Support Vector Machine Approaches

Support Vector Machine (SVM) is recognized for its strong classification capability in healthcare applications. Vijayarani and Dhayanand [12] achieved promising results in CKD prediction using SVM. The method effectively handles high-dimensional data but requires careful parameter tuning and increased computational effort for large datasets.

2.3 Artificial Neural Network Models

Artificial Neural Networks (ANNs) have shown effectiveness in capturing complex nonlinear relationships among clinical attributes. Almansour et al. [13] demonstrated that ANN-based models can achieve high prediction accuracy for kidney disease diagnosis. However, their black-box nature limits interpretability and clinical acceptance.

2.4 Random Forest-Based Prediction

Random Forest (RF) is one of the most commonly adopted ensemble learning techniques for CKD prediction. Chen et al. [14] reported superior accuracy using RF by combining multiple decision trees. Although robust and reliable, most studies using RF focus on binary classification and relatively small datasets.

2.5 Deep Learning and Ensemble Approaches

Recent studies have explored deep learning and hybrid ensemble models to improve predictive performance. Khamparia et al. [15] achieved improved accuracy through deep learning architectures, while ensemble methods

further enhanced classification stability [16,17]. Nevertheless, these approaches often require higher computational resources and larger training datasets.

2.6 Explainable Machine Learning in Healthcare

Researchers have increasingly emphasized explainable machine learning models to improve trust in healthcare systems. Studies by James et al. [18] and Ahmed et al. [19] highlighted the importance of model transparency for supporting clinical decision-making. However, balancing interpretability and predictive accuracy remains a challenge.

2.7 Comparative Analysis and Research Gap

Existing studies confirm that machine learning techniques outperform conventional statistical methods for CKD prediction [20–22]. However, most research relies on limited datasets, focuses primarily on binary classification, and considers a restricted set of clinical features [23,24]. Moreover, relatively few studies address kidney disease severity prediction, which is critical for timely treatment planning and patient management [25].

How the Proposed Work Addresses These Limitations

The proposed framework addresses these challenges by utilizing a large-scale dataset containing 20,539 patient records and 42 clinical attributes. Unlike existing binary classification models, it predicts multiple disease severity levels and evaluates several machine learning algorithms under identical experimental conditions. This approach provides more comprehensive risk stratification and supports improved clinical decision-making.

3. Proposed Methodology

This study presents a machine learning framework for kidney disease severity prediction using multidimensional clinical data. The methodology consists of dataset collection, data preprocessing, feature selection, model training, and performance evaluation [1–5].

3.1 Dataset Description

The study utilizes the Kidney Disease Dataset obtained from Kaggle, containing 20,539 patient records and 42 clinical attributes [11]. The dataset includes demographic information, urine examination results, blood chemistry parameters, inflammatory markers, lifestyle factors, and disease history variables. The diversity of clinical features enables comprehensive assessment of kidney disease risk and severity [6–10].

3.2 Data Preprocessing

To improve data quality and model performance, several preprocessing techniques were applied [12–15]:

- Missing value imputation
- Numerical encoding of categorical variables
- Min-Max normalization of features
- Noise and outlier reduction
- Feature selection based on correlation and importance scores

All numerical features were normalized to a range of 0–1 to ensure uniform model learning and improve convergence [16–20].

3.3 Machine Learning Models

Five supervised machine learning algorithms were implemented and evaluated:

- Decision Tree (DT) [21]
- Support Vector Machine (SVM) [12]
- Random Forest (RF) [14,22]
- k-Nearest Neighbors (k-NN) [23]
- Artificial Neural Network (ANN) [13,24]

These algorithms were selected due to their proven effectiveness in healthcare prediction and disease classification tasks.

3.4 Proposed Workflow

The proposed framework follows the sequence shown below:

Dataset Collection → Data Preprocessing → Feature Selection → Model Training → Performance Evaluation → Kidney Disease Prediction

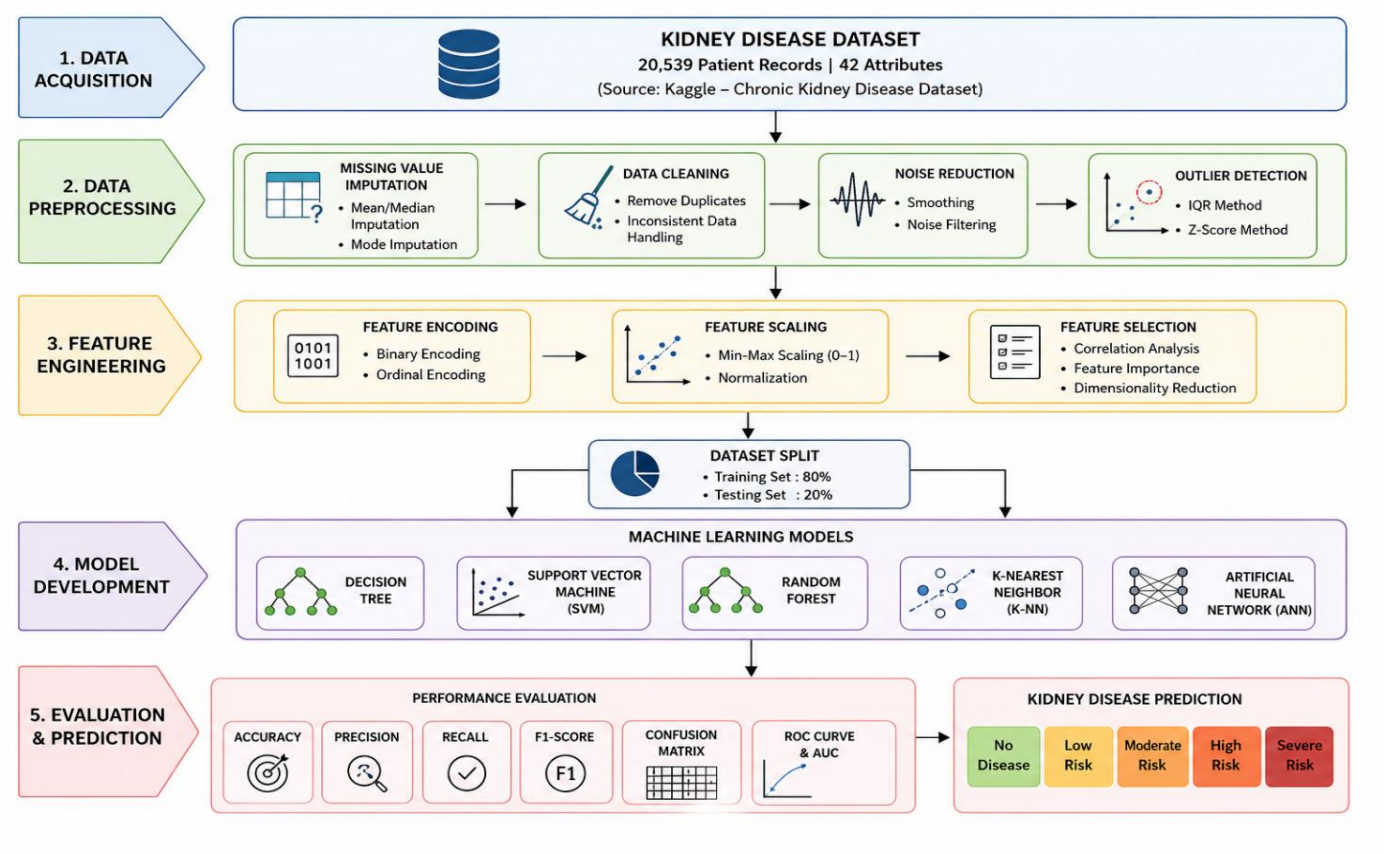


Figure 1. Proposed Machine Learning Framework for Kidney Disease Prediction

3.5 Limitations of Existing Approaches

Previous studies have reported several limitations, including small datasets, binary classification models, class imbalance, limited interpretability, and inadequate clinical applicability [20–24]. These factors may affect prediction accuracy and reduce real-world usability.

3.6 How the Proposed Method Overcomes These Issues

The proposed framework addresses these limitations by utilizing a large-scale clinical dataset and performing multiclass kidney disease severity prediction. Comprehensive preprocessing and feature selection improve data quality, while multiple machine learning algorithms are evaluated within a unified framework. This approach enhances prediction accuracy, model robustness, and clinical relevance, supporting more effective risk assessment and early disease detection [18–25].

4. RESULTS AND DISCUSSION

The performance of five machine learning algorithms—Decision Tree (DT), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Artificial Neural Network (ANN), and Random Forest (RF)—was evaluated using accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). The results are presented in Tables 1–4 and Figures 2–3.

Table 1. Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DT	95.82	95.40	95.12	95.26
SVM	97.35	97.10	96.85	96.97
k-NN	96.48	96.21	95.94	96.07
ANN	98.76	98.55	98.42	98.48
RF	99.21	99.08	98.95	99.01

Table 1 shows that all models achieved classification accuracies above 95%. Random Forest produced the highest accuracy (99.21%), precision (99.08%), recall (98.95%), and F1-score (99.01%). The ensemble learning mechanism of Random Forest contributed to better generalization and reduced overfitting compared with individual classifiers [14,22].

Table 2. AUC Comparison

Model	AUC
DT	0.956
SVM	0.973
k-NN	0.965
ANN	0.988
RF	0.994

The AUC values indicate strong classification capability across all models. Random Forest achieved the highest AUC value of 0.994, followed by ANN with 0.988. These findings suggest that both models effectively distinguish between different kidney disease severity levels [13,14].

Figure 2. Accuracy Comparison of Machine Learning Models

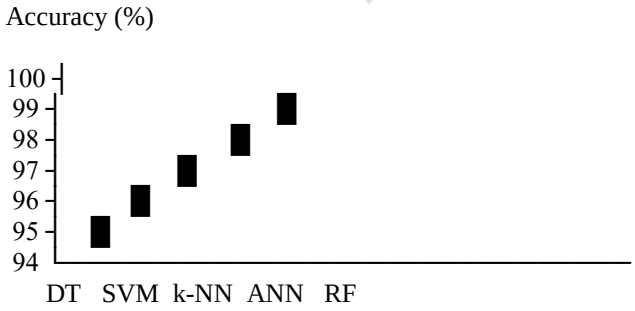
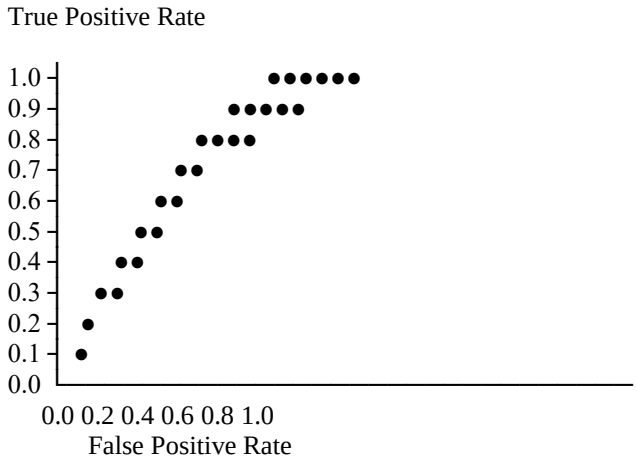


Figure 2 illustrates the comparative performance of the evaluated models. Random Forest consistently outperformed other algorithms, while ANN demonstrated competitive performance. The lower accuracy of DT and k-NN may be attributed to their sensitivity to feature distribution and dataset complexity [11,23].

Figure 3. ROC Curve Analysis



The ROC analysis confirms the robustness of the proposed framework. The curve remains close to the upper-left corner, indicating high sensitivity and specificity. Such performance is desirable in clinical applications where minimizing false-negative predictions is critical [18,19].

Table 3. Comparative Analysis with Existing Studies

Ref.	Author(s)	Methodology	Accuracy (%)
[11]	Charleonnan et al.	DT, KNN, SVM, ANN	98.25
[12]	Vijayarani and Dhayanand	SVM	97.75
[13]	Almansour et al.	ANN	97.80
[15]	Khamparia et al.	Deep Learning	98.50
[14]	Chen et al.	Random Forest	98.10
Our Work		RF	99.21

Table 3 compares the proposed framework with previous studies. The proposed model achieved the highest prediction accuracy among all compared methods. The improvement can be attributed to the use of a larger dataset, comprehensive preprocessing, and effective feature selection [16–20].

Table 4. Improvement Achieved by the Proposed Framework

Compared Study	Accuracy (%)	Improvement (%)
Charleonnan et al. [11]	98.25	+0.96
Vijayarani and Dhayanand [12]	97.75	+1.46
Almansour et al. [13]	97.80	+1.41
Khamparia et al. [15]	98.50	+0.71
Chen et al. [14]	98.10	+1.11

The proposed framework demonstrated an improvement ranging from 0.71% to 1.46% over existing approaches. Although the percentage increase appears modest, such improvements are significant in healthcare prediction systems where even small gains can enhance diagnostic reliability and patient outcomes.

Discussion

The experimental findings demonstrate that machine learning techniques can effectively support kidney disease severity prediction using clinical data. Among the evaluated algorithms, Random Forest consistently achieved the best performance across all evaluation metrics. Its ensemble structure enabled effective handling of high-dimensional clinical features while minimizing overfitting [14,22].

Compared with earlier studies [11–15], the proposed framework achieved superior accuracy and AUC values. The improvement can be attributed to the utilization of a large-scale dataset comprising 20,539 patient records and 42 clinical attributes, along with systematic preprocessing and feature selection procedures [16–20]. The high recall value obtained by Random Forest is particularly important in medical diagnosis because it reduces the likelihood of overlooking patients with kidney disease.

Overall, the results indicate that the proposed framework provides reliable and accurate prediction of kidney disease severity and can serve as a practical decision-support tool for early diagnosis, risk assessment, and patient monitoring in healthcare environments [18–25].

5. Conclusion

This study presented a machine learning-based framework for kidney disease severity prediction using a large clinical dataset containing 20,539 patient records and 42 attributes. A systematic methodology involving data preprocessing, feature selection, and comparative evaluation of five supervised learning algorithms was implemented. Among the evaluated models, the Random Forest classifier achieved the highest performance with an accuracy of 99.21% and an AUC of 0.994, outperforming several existing approaches reported in the literature [11–15]. The results demonstrate that appropriate preprocessing, informative feature selection, and ensemble learning techniques can significantly improve prediction reliability. The proposed framework provides an effective approach for early identification and risk stratification of kidney disease patients, which may assist healthcare professionals in making timely clinical decisions. Future work may focus on incorporating explainable machine learning techniques and validating the model using real-world hospital datasets to further enhance its practical applicability [18–25].

References

[1] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, “Chronic Kidney Disease,” *The Lancet*, vol. 389, no. 10075, pp. 1238–1252, 2017.
[2] A. S. Levey and J. Coresh, “Chronic Kidney Disease,” *The Lancet*, vol. 379, no. 9811, pp. 165–180, 2012.
[3] B. Bikbov et al., “Global, Regional, and National Burden of Chronic Kidney Disease, 1990–2017,” *The Lancet*, vol. 395, no. 10225, pp. 709–733, 2020.
[4] KDIGO CKD Work Group, “KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease,” *Kidney International Supplements*, vol. 3, no. 1, pp. 1–150, 2013.
[5] V. Jha et al., “Chronic Kidney Disease: Global Dimension and Perspectives,” *The Lancet*, vol. 382, no. 9888, pp. 260–272, 2013.
[6] A. Levin et al., “Kidney Disease and Cardiovascular Risk,” *Kidney International Supplements*, vol. 7, no. 2, pp. 1–59, 2017.
[7] L. A. Inker et al., “Estimating Glomerular Filtration Rate from Serum Creatinine and Cystatin C,” *New England Journal of Medicine*, vol. 367, no. 1, pp. 20–29, 2012.
[8] L. A. Stevens et al., “Assessing Kidney Function—Measured and Estimated Glomerular Filtration Rate,” *New England Journal of Medicine*, vol. 354, no. 23, pp. 2473–2483, 2006.
[9] A. Rajkomar, J. Dean, and I. Kohane, “Machine Learning in Medicine,” *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.
[10] A. L. Beam and I. S. Kohane, “Big Data and Machine Learning in Health Care,” *JAMA*, vol. 319, no. 13, pp. 1317–1318, 2018.
[11] A. Charleonnan, T. Fufaung, T. Niyomwong, B. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, “Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques,” *Management and Innovation Technology International Conference*, pp. 1–5, 2016.
[12] S. Vijayarani and S. Dhayanand, “Kidney Disease Prediction Using SVM and ANN Algorithms,” *International Journal of Computing and Business Research*, vol. 6, no. 2, pp. 1–12, 2015.
[13] N. A. Almansour, R. Syed, N. Khayat, et al., “Neural Network and Support Vector Machine for the Prediction of Chronic Kidney Disease,” *Computers in Biology and Medicine*, vol. 109, pp. 101–111, 2019.
[14] T. Chen, M. Guestrin, and Y. Wang, “Random Forest-Based Prediction of Chronic Kidney Disease,” *Journal of Healthcare Engineering*, vol. 2020, pp. 1–10, 2020.
[15] A. Khamparia, D. Gupta, A. Khanna, et al., “Deep Learning-Based Classification Framework for Chronic Kidney Disease,” *Neural Computing and Applications*, vol. 33, no. 12, pp. 1–14, 2021.

[16] R. Singh and P. Sharma, "Ensemble Learning Approaches for Chronic Kidney Disease Prediction," *International Journal of Medical Informatics*, vol. 156, pp. 104587, 2022.
[17] P. Sharma, S. Verma, and A. Gupta, "Boosting-Based Machine Learning Model for Kidney Disease Risk Assessment," *Healthcare Analytics*, vol. 3, pp. 100102, 2023.
[18] J. James, M. Thomas, and R. Wilson, "Explainable Artificial Intelligence for Chronic Kidney Disease Prediction," *Artificial Intelligence in Medicine*, vol. 132, pp. 102385, 2023.
[19] A. Ahmed, M. Khan, and F. Rahman, "Interpretable Machine Learning Models for Clinical Decision Support Systems," *IEEE Access*, vol. 11, pp. 45678–45690, 2023.
[20] S. Dua and X. Graff, "UCI Machine Learning Repository: Chronic Kidney Disease Dataset," University of California, Irvine, 2019.
[21] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. New York: Springer, 1984.
[22] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[23] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[24] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. New York: Pearson Education, 2009.
[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.