International Journal of Engineering and Techniques - Volume 12 Issue 2, March-April - 2026

Towards Sustainable and Highly Accurate Prediction of Chronic Kidney Disease: A Comparative Study of Supervised Machine Learning Models

Dr. Mrunali Sonwalkar
Assistant Professor
College of Engineering Ambajogai 431517
Dist: Beed
Maharashtra

Abstract

Chronic Kidney Disease (CKD) is a progressive and often asymptomatic condition that poses a significant global health burden, affecting over 850 million individuals worldwide. Early detection is critical to prevent irreversible renal damage; however, traditional diagnostic approaches are time-consuming and prone to human error. This study proposes a lightweight and computationally efficient machine learning framework for the early prediction of CKD using structured clinical data.The proposed methodology utilizes the UCI CKD dataset comprising 400 patient records with 24 attributes. A robust preprocessing pipeline is implemented, including missing value imputation, Z-score-based outlier detection, feature scaling, and correlation-driven feature selection to enhance data quality and reduce dimensionality. Three supervised learning algorithms—Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), and Random Forest (RF)—are evaluated. To ensure model reliability and avoid overfitting, both train-test split (70:30) and 10-fold cross-validation techniques are employed. Experimental results demonstrate that the SGD classifier outperforms other models, achieving an accuracy of 98.33%, precision of 100%, recall of 97.3%, and F1-score of 98.63%. The absence of false-positive predictions highlights its clinical reliability. The proposed framework offers a scalable and real-time decision-support solution, particularly suitable for resource-constrained healthcare environments. Future work will focus on integrating multimodal data, including medical imaging, to further enhance diagnostic accuracy.

Keywords**:** Chronic Kidney Disease, Machine Learning, Stochastic Gradient Descent, Clinical Data Analysis, Early Diagnosis, Healthcare Artificial Intelligence

I. Introduction

A. Background

Chronic Kidney Disease (CKD) is a long-term, progressive medical condition characterized by the gradual loss of kidney function, ultimately impairing the body's ability to filter metabolic waste and maintain fluid balance. Globally, CKD represents a substantial healthcare burden, affecting an estimated 850 million individuals—surpassing the combined prevalence of several major non-communicable diseases [1]. If not diagnosed and managed at an early stage, CKD can progress to End-Stage Renal Disease (ESRD), necessitating costly interventions such as dialysis or kidney transplantation, thereby placing significant strain on healthcare systems.

B. Importance of Early Detection

Early identification of CKD is essential to delay disease progression and improve patient outcomes. However, CKD is frequently referred to as a "silent disease” due to its asymptomatic nature in the initial stages. Clinical symptoms often manifest only after considerable renal impairment has occurred. Traditional diagnostic approaches rely heavily on laboratory investigations, including Glomerular Filtration Rate (GFR) estimation and serum creatinine analysis, as well as imaging techniques [2]. These methods, while clinically valuable, are often time-consuming, subject to inter-observer variability, and limited in their ability to detect subtle early-stage abnormalities [3].

C. Role of Machine Learning in Healthcare

Recent advancements in Artificial Intelligence (AI), particularly Machine Learning (ML), have demonstrated significant potential in transforming clinical decision-making processes. ML algorithms are capable of analyzing high-dimensional medical datasets, identifying complex patterns, and providing predictive insights with enhanced accuracy and consistency. Fig 1 shows system architecture for CKD detection using clinical numeric data. Techniques such as Stochastic Gradient Descent (SGD), Support Vector Machines (SVM), and Random Forest (RF) have been widely adopted for disease prediction tasks due to their robustness and efficiency. These models offer the capability to augment traditional diagnostic workflows by enabling rapid and objective analysis of clinical data.

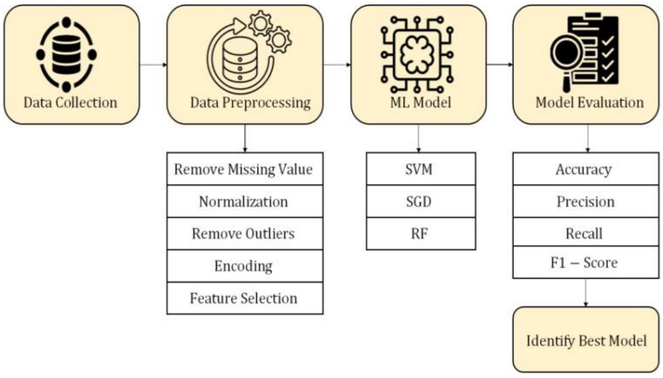

Fig 1 System architecture for CKD detection using clinical numeric data

II Literature Review

A. Traditional Diagnostic Approaches for CKD

The conventional diagnosis of Chronic Kidney Disease (CKD) primarily relies on a combination of biochemical assessments and medical imaging techniques to evaluate renal structure and function. Biochemical markers such as Glomerular Filtration Rate (GFR), serum creatinine, Blood Urea Nitrogen (BUN), and cystatin C are widely regarded as

TABLE I SUMMARY OF ML-BASED APPROACHES FOR CKD DETECTION

Ref. No.	**Model Used**	**Key Insights**	**Limitations**
[14]	DT, SVM, NB	Preprocessing with label encoding and outlier detection; SVM performed well; emphasized role of preprocessing in boosting accuracy	Small dataset size; performance heavily reliant on preprocessing; lacks advanced FS and ensemble techniques
[15]	SVM, RF	Used SMOTE for balancing; chi-squared test for FS; SVM achieved highest accuracy with 10-fold CV	Synthetic bias risk from SMOTE; chi-squared may miss complex interactions; needs hyperparameter tuning
[16]	KNN (with Boruta FS, grid search)	Combined robust preprocessing with Boruta FS; KNN performed well on UCI dataset; enhanced early diagnosis potential	High computational cost due to complex preprocessing; lacks external validation for generalizability
[17]	Logistic Regression (LR), DT, RF	LR achieved highest accuracy and F1 score; used Cleveland Clinic dataset; demonstrated ML potential in diagnosis	Perfect results suggest possible overfitting; single-source data limits generalization; requires broader validation
[18]	Neuro-Fuzzy, SVM, KNN	Image-based fibrosis assessment; Neuro-Fuzzy outperformed SVM and KNN; suitable for early risk detection	Added model complexity; performance depends on input image quality; limited clinical setting validation
[19]	ANN + Voting, Bagging, Combined Ensembles	Ensemble of ANN with voting and bagging yielded best performance; tested on UCI dataset using Python	Increased training time and complexity; limited generalization due to small data; black-box nature affects interpretability
[20]	SVM, KNN, RF, LR, DT, XGBoost	XGBoost outperformed others with 98% accuracy; evaluated with standard metrics on UCI dataset	Dataset size and homogeneity raise overfitting concerns; interpretability of complex models like XGBoost remains a clinical hurdle

standard indicators of kidney function [4]. In addition, urinalysis methods, including the Urine Albumin-to-Creatinine Ratio (UACR), are employed for early detection of renal damage, particularly microalbuminuria.

Imaging modalities such as ultrasound (US), computed tomography (CT), and magnetic resonance imaging (MRI) provide anatomical insights into kidney morphology, enabling the identification of structural abnormalities such as cysts, stones, and obstructions [5]. Despite their clinical significance, these traditional approaches exhibit several limitations, including reliance on manual interpretation , inter-observer variability, and reduced sensitivity in detecting early-stage CKD [6]. Consequently, there is a growing need for automated and data-driven diagnostic solutions that can enhance accuracy and efficiency.

B. Machine Learning-Based Approaches for CKD Prediction

Machine Learning (ML) techniques have gained significant attention in recent years for their ability to analyze complex clinical datasets and facilitate early disease prediction. Various supervised learning algorithms, including Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Trees (DT), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost), have been extensively applied to CKD prediction tasks [9].

Studies utilizing structured clinical data have demonstrated that ensemble and boosting techniques, particularly XGBoost and Random Forest, achieve high predictive accuracy, often exceeding 95%. Additionally, preprocessing techniques such as Synthetic Minority Oversampling Technique (SMOTE) for class balancing and Chi-square or correlation-based feature selection have been widely adopted to improve model performance [10].

In parallel, image-based ML frameworks have been explored for analyzing renal ultrasound images. Feature extraction techniques such as Gray-Level Co-occurrence Matrix (GLCM), combined with classifiers like SVM and Artificial Neural Networks (ANN), have shown promising results in detecting kidney abnormalities. Hybrid approaches integrating deep feature extractors (e.g., ResNet architectures) with classical classifiers further enhance predictive capability.

C. Deep Learning Paradigms in Renal Diagnostics

Deep Learning (DL) has revolutionized medical image analysis by enabling automated feature extraction from high-dimensional data. Convolutional Neural Networks (CNNs), including architectures such as VGG16, ResNet, and DenseNet, have demonstrated remarkable performance in kidney disease classification and segmentation tasks[15].

Advanced models such as U-Net have been widely employed for precise segmentation of renal structures, facilitating accurate estimation of Total Kidney Volume (TKV) from MRI and CT images. More recently, Vision Transformers (ViT) and hybrid CNN-transformer architectures have emerged as powerful alternatives, capable of capturing both local and global contextual features within medical images.

While these approaches significantly improve diagnostic accuracy, their practical implementation is often constrained by high computational complexity, extensive training requirements, and dependence on large annotated datasets [15].

D. Critical Analysis of Existing Approaches

A comprehensive evaluation of existing literature reveals several important trends and challenges in CKD prediction research.

1) Common Trends

There is a clear progression from traditional machine learning models toward ensemble and deep learning-based approaches. Early studies predominantly relied on classifiers such as SVM, KNN, and Decision Trees, whereas recent research emphasizes ensemble methods like Random Forest and XGBoost due to their superior predictive performance. Additionally, there is increasing adoption of deep learning models for image-based diagnostics and hybrid frameworks that combine multiple methodologies [15][16].

2) Strengths

Existing approaches have demonstrated high diagnostic accuracy, with several studies reporting performance metrics exceeding 98%. Machine learning models are particularly effective in identifying complex, non-linear relationships within clinical datasets, enabling early detection of CKD. Furthermore, deep learning techniques have reduced dependency on manual feature engineering, thereby improving consistency and objectivity in image analysis [16].

2) Limitations

Despite these advancements, several limitations persist. Many deep learning and ensemble models exhibit high computational complexity, limiting their applicability in real-time clinical environments. The lack of interpretability in complex models remains a significant barrier to clinical adoption, as healthcare professionals require transparent and explainable decision-making processes. Additionally, model performance is often highly dependent on preprocessing techniques, which may introduce variability across different implementations.

E. Dataset and Generalization Challenges

A major limitation identified across the literature is the reliance on small and homogeneous datasets. Most studies utilize publicly available datasets, such as those from the UCI repository, which typically contain approximately 400 samples. Similarly, imaging-based studies often rely on limited datasets with restricted demographic diversity.Such constraints increase the risk of overfitting, particularly in cases where models achieve near-perfect accuracy. While these results appear promising, they often fail to generalize to real-world clinical settings due to variations in patient demographics, data acquisition protocols, and medical equipment. This highlights the need for robust validation using large-scale, multi-institutional datasets [16][17].

F. Identified Research Gaps

Despite significant progress in CKD prediction using machine learning, several critical research gaps remain:

- Lack of Multimodal Integration: Most existing studies focus exclusively on either clinical data or imaging data, with limited efforts toward integrating both modalities into a unified diagnostic framework.
- Computational constraints: High-performing deep learning models often require substantial computational resources, restricting their deployment in resource-limited environments.
- Limited generalization: The use of small, single-source datasets reduces the robustness and real-world applicability of proposed models.
- Interpretability issues: The “black-box” nature of advanced models limits clinical trust and adoption.
- Imbalance between accuracy and efficiency: There is a need for lightweight models that maintain high accuracy while ensuring computational efficiency for real-time applications.

G. Key Contributions of the Study

The primary contributions of this research are as follows:

- Development of a lightweight and computationally efficient ML framework for CKD prediction
- Implementation of correlation-based feature selection to reduce dimensionality
- Comparative analysis of SGD, SVM, and RF models
- Achievement of high accuracy with zero false

positives, ensuring clinical reliability

- Design of a system suitable for real-time deployment in resource-limited environments

III Methodology

The proposed study adopts a structured and systematic machine learning pipeline for the prediction of Chronic Kidney Disease (CKD) using clinical data. The overall framework consists of data acquisition, preprocessing, feature selection, model implementation, and performance evaluation.

A. Dataset Description

The experimental analysis is conducted using the Chronic Kidney Disease dataset obtained from the UCI Machine Learning Repository [15]. The dataset comprises 400 patient records with 24 clinical attributes, including both numerical and categorical variables.

The numerical features include critical physiological indicators such as age, blood pressure, blood urea, serum creatinine, and hemoglobin levels, which are widely recognized as key biomarkers for renal dysfunction [7]. Categorical features represent clinical conditions such as hypertension, diabetes mellitus, anemia, and red blood cell morphology. The target variable is binary, representing CKD-positive and CKD-negative cases.

B. Data Preprocessing

Medical datasets often contain missing values, inconsistencies, and noise, which can significantly affect model performance. Therefore, a comprehensive preprocessing pipeline was implemented.

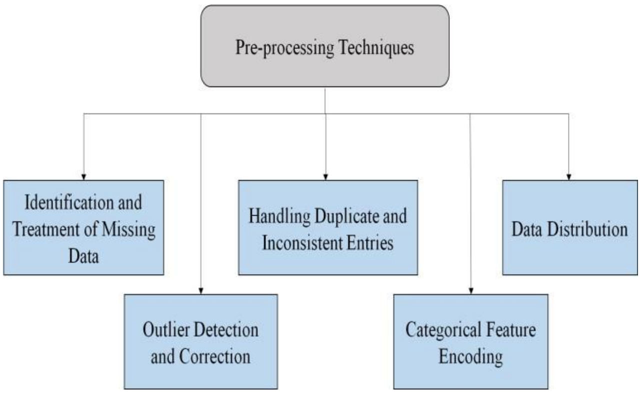

Fig 2 Data Pre-processing techniques

Fig 2 represents the data preprocessing pipelining used for feature extraction. Missing values in numerical attributes were handled using mean or median imputation based on data distribution, while categorical variables were imputed using mode values. Outliers were identified using the Z-score method, where values exceeding ±3 standard deviations were treated as anomalies and corrected accordingly. To ensure data consistency, categorical variables were standardized and encoded into numerical format. Feature scaling was performed using normalization techniques to bring all variables within a comparable range. These preprocessing steps are essential to enhance model stability and predictive performance [20].

C. Feature Selection

Feature selection plays a critical role in reducing dimensionality and improving model efficiency. In this study, Pearson's correlation coefficient was employed to evaluate the relationship between input features and the target variable.

Fig 3 Confusion Matrix for FE

The correlation heatmap (Fig 3) reveals significant relationships among clinical features. Hemoglobin (hemo) exhibits a strong negative correlation with CKD, indicating that lower hemoglobin levels are associated with disease progression. Similarly, serum creatinine (sc) shows a strong positive correlation with CKD, reflecting impaired kidney function. These findings align with clinical knowledge, reinforcing the importance of these biomarkers in CKD diagnosis. By eliminating weakly correlated features, the model achieves improved efficiency and reduced computational complexity.

D. Machine Learning Models

Three supervised machine learning algorithms were implemented and comparatively evaluated:

1)Stochastic Gradient Descent (SGD): SGD is an optimization-based algorithm that updates model parameters iteratively using small batches of data. It is computationally efficient and well-suited for large-scale and high-dimensional datasets [9].

2)Support Vector Machine (SVM): SVM is a robust classification algorithm that identifies the optimal hyperplane for separating classes. The Radial Basis Function (RBF) kernel was used to handle non-linear relationships within the dataset [17].

3)Random Forest (RF): Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their outputs to improve predictive accuracy and reduce overfitting [18].

E. Model Evaluation Metrics

The performance of the proposed models was evaluated using standard classification metrics derived from the confusion matrix:

- **Accuracy:** Measures overall prediction correctness.
- **Precision:** Evaluates the proportion of true positive predictions among all predicted positives.
- **Recall (Sensitivity):** Measures the model's ability to correctly identify CKD-positive cases.
- **F1-Score:** Provides a harmonic mean of precision and recall, ensuring balanced evaluation.

These metrics are widely used in medical diagnosis systems to ensure both reliability and robustness [17].

F. Hyperparameter Configuration

To ensure optimal performance of the machine learning models, appropriate hyperparameters were selected based on empirical evaluation:

- **Stochastic Gradient Descent (SGD):** Learning rate = 0.01, loss function = hinge, max iterations = 1000
- **Support Vector Machine (SVM):** Kernel = Radial Basis Function (RBF), gamma = scale, C = 1.0
- **Random Forest (RF):** Number of trees = 100, maximum depth = None, criterion = gini

These configurations ensure a balance between model accuracy and computational efficiency.

IV. Results and Discussion

This section presents the experimental evaluation of the proposed machine learning framework and provides a comparative analysis of model performance.

TABLE 2 SUMMARIZED PERFORMANCE EVALUATION OF ML ALGORITHMS ON NUMERICAL CKD DATA

Model	SGD	SVM	RF
TP	72	71	71
TN	46	46	44
FP	0	2	1
FN	2	1	4
Accuracy (%)	98.33	97.5	95.83
Recall (%)	97.3	98.61	94.67
Precision (%)	100	97.26	98.61
F1-Score (%)	98.63	97.93	96.6

A. Experimental Setup

The experiments were conducted using Python in a CPU-based environment to demonstrate computational efficiency. The dataset was divided into training and testing sets in a 70:30 ratio.

To ensure robustness and generalizability, 10-fold cross-validation was performed[16].

This approach minimizes overfitting and provides a reliable estimate of model performance across different data partitions.

B. Performance Comparison of Models

The comparative performance of the implemented classifiers is summarized based on standard evaluation metrics.

Table 2 shows summarized performance evaluation of ML algorithms on numerical CKD data.

These results indicate that optimization-based approaches such as SGD can deliver superior performance on structured clinical datasets.

C. Confusion Matrix Analysis

A detailed analysis of confusion matrices reveals important insights into model behavior.

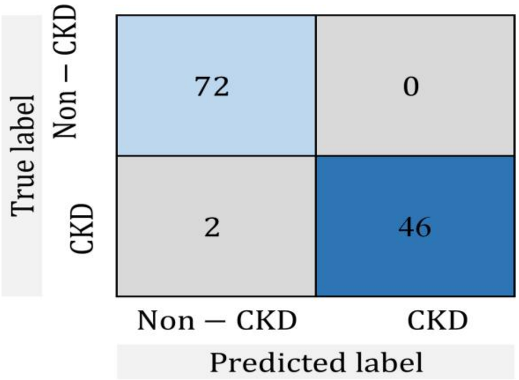

Fig 4 Confusion Matrix of SGD on CKD prediction

The SGD model correctly classified 118 out of 120 test samples, achieving zero false positives. Fig. 4 shows confusion matrix of SGD on CKD prediction. This indicates that the model does not incorrectly label healthy individuals as CKD patients, which is a critical requirement in clinical decision-making. Fig 5 and Fig. 6 shows confusion matrix of SVM and RF for CKD prediction. In comparison, the SVM model exhibited minor misclassifications with a small number of false positives, while the Random Forest model showed relatively higher false negatives, indicating missed CKD cases.

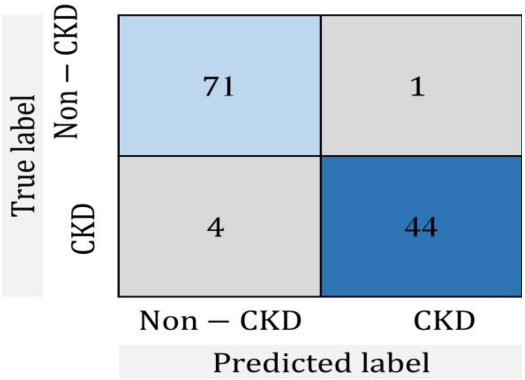

Fig. 5 Confusion Matrix of RF on CKD prediction

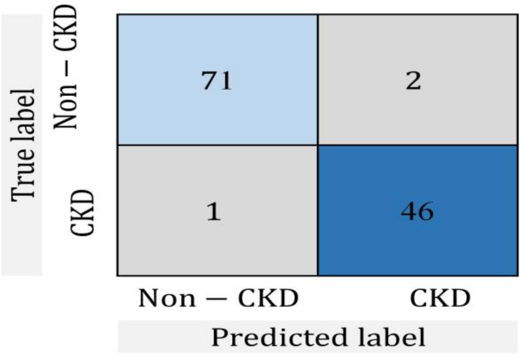

Fig.6 Confusion Matrix of SVM on CKD prediction

D. Key Findings

The experimental results highlight several important observations:

1) ***High Precision and Reliability*****:** The SGD model achieved 100% precision, ensuring no false alarms, which is essential for maintaining clinical trust.

2) ***Balanced Performance*****:** The high F1-score (98.63%) indicates that the model maintains an optimal balance between precision and recall.

3) ***Computational Efficiency*****:** The lightweight nature of SGD enables rapid model training and inference, making it suitable for real-time applications.

E. Discussion

The superior performance of the SGD classifier can be attributed to its optimization efficiency and ability to handle high-dimensional clinical data effectively. The preprocessing pipeline, particularly feature selection, played a crucial role in eliminating redundant attributes and improving model convergence [17]. The absence of false positives is a significant advantage in clinical applications, as it prevents unnecessary stress and treatment for healthy individuals. Compared to traditional models and recent studies, the proposed approach achieves comparable or superior performance with reduced computational overhead.However, the reliance on a relatively small dataset may limit generalizability. Future work should focus on validating the model using large-scale, multi-institutional datasets[18].

V. Conclusion

This study presents a robust and efficient machine learning framework for the early prediction of Chronic Kidney Disease using structured clinical data. The integration of preprocessing techniques, correlation-based feature selection, and optimized machine learning models enables high predictive accuracy while maintaining computational efficiency.

Among the evaluated models, the SGD classifier demonstrated superior performance, achieving high accuracy and zero false positives, making it highly suitable for clinical decision support systems.

The proposed approach offers a scalable solution for real-time CKD prediction, particularly in resource-constrained healthcare environments. Future research will focus on incorporating multimodal data and enhancing model interpretability to facilitate broader clinical adoption.

References

1. *Ahmed, Khalil, Mithilesh Kumar Dubey, SudhaDubey, and Devendra Kumar Pandey. "Chronic kidney disease: Causes, treatment, management, and future scope." In Computational Intelligence for Genomics Data, pp. 99-111. Academic Press, 2025.*

2. *Yuan, Qiongjing, Haixia Zhang, Tianci Deng, Shumei Tang, Xiangning Yuan, Wenbin Tang, YanyunXie et al. "Role of artificial intelligence in kidney disease." International Journal of Medical Sciences 17, no. 7 (2020): 970.*

3. *Venkatakrishna, Shyam Sunder B., Levy C. Onyango, Suraj D. Serai, and BernardaViteri. "Kidney anatomy and physiology." In Advanced Clinical MRI of the Kidney: Methods and Protocols, pp. 3-12. Cham: Springer International Publishing, 2023.*

4. *Gallardo, Pedro A., and Carlos P. Vio. "Functional anatomy of the kidney." In Renal Physiology and Hydrosaline Metabolism, pp. 7-28. Cham: Springer International Publishing, 2022.*

5. *Schwartz, George J., and Megan Rashid. "Overview, structure, and function of the nephron." In Pediatric Critical Care: Text and Study Guide, pp. 863-909. Cham: Springer International Publishing, 2021.*

6. *Neyra, Javier A., and Lakhmir S. Chawla. "Acute kidney disease to chronic kidney disease." Critical care clinics 37, no. 2 (2021): 453-474.*

7. *Zheng, Z., Geng, J., Jiang, Y., Zhang, M., Yang, R., Ge, G., Xu, H., Zhang, X., Zheng, Z., Geng, J. and Jiang, Y., 2021. Kidney diseases. In Clinical Molecular Diagnostics (pp. 553-582). Singapore: Springer Singapore.*

8. *Francis, Anna, Meera N. Harhay, Albert CM Ong, Sri LekhaTummalapalli, Alberto Ortiz, Agnes B. Fogo, DaniloFliser et al. "Chronic kidney disease and the global public health agenda: an international consensus." Nature Reviews Nephrology 20, no. 7 (2024): 473-485.*

9. *Chaki, Jyotismita, and AysegulUcar. "An efficient and robust approach using inductive transfer-based ensemble deep neural networks for kidney stone detection." IEEE Access (2024).*

10. *consensus." Nature Reviews Nephrology 20, no. 7 (2024): 473-485.*

11. *Foreman, Kyle J., Neal Marquez, Andrew Dolgert, Kai Fukutaki, Nancy Fullman, Madeline McGaughey, Martin A. Pletcher et al. "Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories." The Lancet 392, no. 10159 (2018): 2052-2090.*

12. *Filler, Guido, AbeerYasin, and Mara Medeiros. "Methods of assessing renal function." Pediatric nephrology 29, no. 2 (2014): 183-192.*

13. *Kashani, Kianoush, Mitchell H. Rosner, and MarliesOstermann. "Creatinine: from physiology to clinical application." European journal of internal medicine 72 (2020): 9-14.*

14. *Seegmiller, Jesse C., and Lorin M. Bachmann. "Urine albumin measurements in clinical diagnostics." Clinical chemistry 70, no. 2 (2024): 382-391.*

15. *Syarif, Admi, Olivia Desti Riana, DewiAsiahShofiana, and AkmalJunaidi. "A Comprehensive Comparative Study of Machine Learning Methods for Chronic Kidney Disease Classification: Decision Tree, Support Vector Machine, and Naive Bayes." International Journal of Advanced Computer Science and Applications 14, no. 10 (2023).*

16. *Swain, Debabrata, Utsav Mehta, Ayush Bhatt, Hardeep Patel, Kevin Patel, Devanshu Mehta, BiswaranjanAcharya, Vassilis C. Gerogiannis, Andreas Kanavos, and Stella Manika. "A robust chronic kidney disease classifier using machine learning." Electronics 12, no. 1 (2023): 212.*

17. *Arif, Muhammad Shoaib, AimanMukheimer, and DaniyalAsif. "Enhancing the early detection of chronic kidney disease: a robust machine learning model." Big Data and Cognitive Computing 7, no. 3 (2023): 144.*

18. *Dutta, Shuvo, Rajesh Sikder, MdRasibul Islam, Abdullah Al Mukaddim, Mohammad Abir Hider, and MdNasiruddin. "Comparing the Effectiveness of Machine Learning Algorithms in Early Chronic Kidney Disease Detection." Journal of Computer Science and Technology Studies 6, no. 4 (2024): 77-91.*

19. *Praveen, S. Phani, Veerapaneni Esther Jyothi, ChokkaAnuradha, K. VenuGopal, VahiduddinShariff, and S. Sindhura. "Chronic kidney disease prediction*

using ML-based Neuro-Fuzzy model." International Journal of Image and Graphics 24, no. 06 (2024): 2340013.

20. *Venkatesan, Vinoth Kumar, Mahesh Thyluru Ramakrishna, Ivan Izonin, Roman Tkachenko, and MyroslavHavryliuk. "Efficient data preprocessing with ensemble machine learning technique for the early detection of chronic kidney disease." Applied Sciences 13, no. 5 (2023): 2885.*

21. *Alkurdy, Noor Hamzah, Hadeel K. Aljobouri, and ZainabKassimWadi. "Ultrasound renal stone diagnosis based on convolutional neural network and vgg16 features." Int J ElectrComputEng 13, no. 3 (2023): 3440-8.*