**EARLY PREDICTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING: A COMPREHENSIVE REVIEW**

**Konda Akhila1, B. Sandhya Rani2, J. Chaithanya3, Remalli Rohan4**

1M. Sc-Data Science-PG Student,

2,3 Degree Lecturer, Department of Computer Science,

Telangana Social Welfare Residential Degree College for Women, Jagathgirigutta,

Hyderabad, India.

4Researcher, Computer Science Educator, Hyderabad, Telangana, India.

**ABSTRACT**

The chronic disease known as chronic kidney disease (CKD) is defined by a slow decrease in kidney function. Dialysis or a kidney transplant may be necessary if CKD is not treated. Early identification is essential because it provides immediate actions that may lower healthcare costs, extend the life, and slow down the development of disease. Interest in creating predictive models for CKD detection has increased as a result of recent developments in data mining and Machine Learning (ML). By identifying CKD risk factors early through the analysis of clinical and demographic data, these models improve early treatment. The development of CKD, the early diagnosis, and the possible use of predictive modelling to improve patient outcomes by identifying individuals at higher risk are all examined in this study. Healthcare providers can more accurately predict the growth of diseases with the help of ML algorithms.

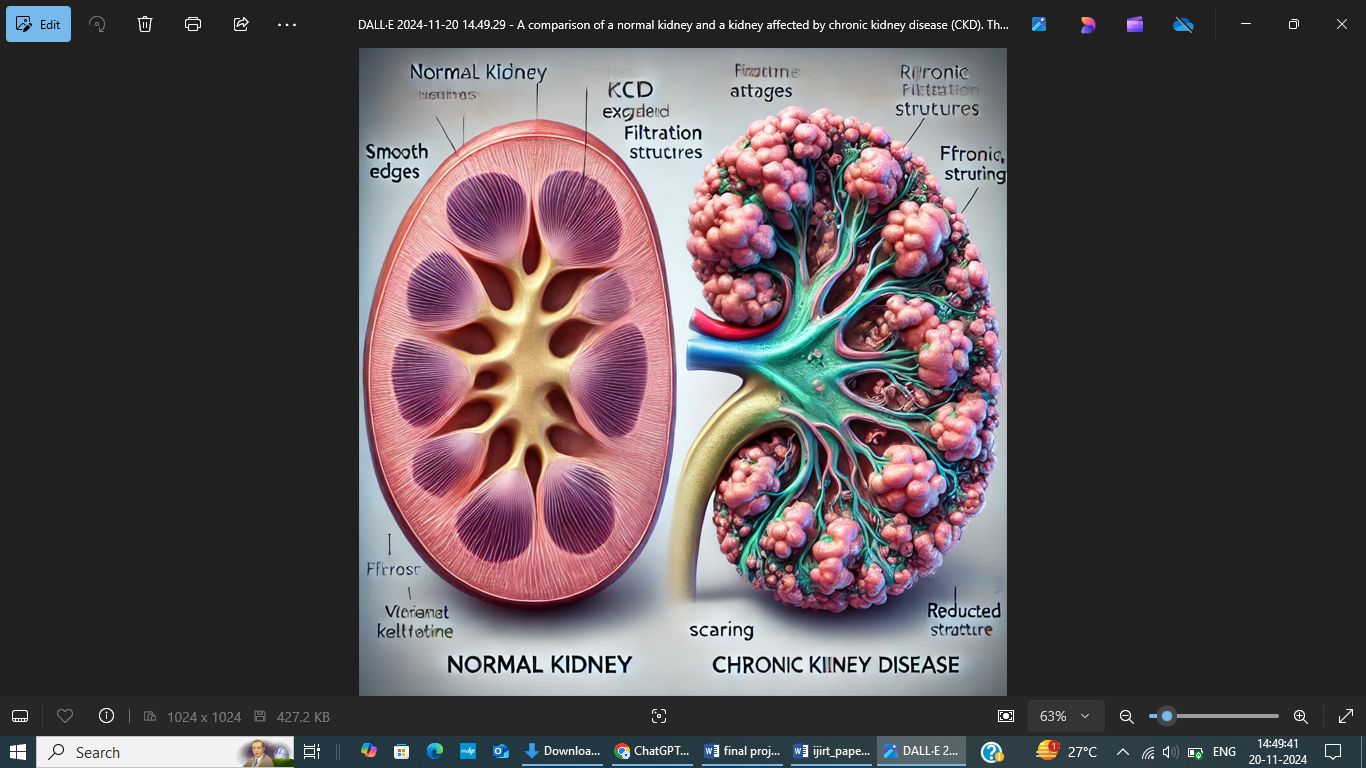
**Keywords:** Chronic Kidney Disease (CKD), Early Detection, Machine Learning, Predictive Modelling, Risk Assessment**.**

1. **INTRODUCTION**

According to WHO's 2021 Global Health Estimates, Chronic Kidney Disease (CKD) is still the ninth biggest reason for death globally, making it a serious global health concern. Deaths from CKD increased by about 95% between 2000 and 2021, indicating the disease's rising incidence, particularly in low- and middle-income nations. It is defined by an ongoing decrease in kidney function, which affects the organ’s capacity to filter waste, maintain electrolyte balance, and control blood pressure. As waste adds up, it leads to serious health issues. Interest in predictive modelling for early identification and intervention has increased due to increasing rates and effects of chronic kidney disease.

* 1. **Early Detection**

CKD must be identified early for a number of reasons. It can lower medical expenses, improve patient’s quality of life, and avoid serious problems from advanced renal disease. Early detection of CKD enables early treatments that can improve patient life span and delay the course of the disease.



**Figure 1:** Normal kidney and chronic kidney disease kidney

* 1. **Symptoms**

Early identification is made more difficult by the fact that CKD symptoms can be unclear and can develop later. Typical indicators include vomiting and feeling sick, appetite loss, weakness and sleepiness, disturbances in sleep, variations in the frequency of urination, reduced mental capacity, tightness in the muscles, Legs and ankles swelling, itchy and dry skin, high blood pressure, breathlessness, pain in the chest. Many people lack any symptoms until the disease grows to a more severe stage because of the kidney’s making up processes.

* 1. **Causes and Risk Factors**

Multiple diseases and conditions that affect kidney function, such as diabetes, high blood pressure, kidney disease, and urinary system blockage, can lead to CKD. The following risk factors could raise your chance of getting CKD: hypertension and diabetes, heart-related problems, obesity from smoking, ethnic background (such as Asian American, Native American, or Black), kidney disease's family history, and Older age.

* 1. **Complications**

Multiple problems affecting different body systems might result from CKD. Heart disease, anemia, fluid retention, and decreased immunity are a few examples of these results. In addition, CKD can have a major effect on the reproductive system. The development of strong prediction models that combine ensemble methods, understanding methods like Shapley Additive Explanations (SHAP), and efficient handling of missing data have been the focus of recent developments as the field of CKD management changes. According to these studies, machine learning algorithms—such as ensemble techniques and Gaussian Processes—show interest in accurately classifying patients and detecting risk factors.

1. **LITERATURE SURVEY**

The review of the literature looks at the Support Vector Machine (SVM) classifier and new developments in data analysis and Machine Learning (ML) techniques for early CKD prediction. We aim to determine the efficiency, accuracy, and flexibility of SVM and other algorithms in detecting CKD early on by examining research from 2019 to 2023. Data set, feature extraction, algorithm, and results to improve CKD prediction using ML methods is presented in Table1.

Revathy [1] created a machine learning framework for early CKD diagnosis using Decision Tree, Random Forest, and SVM. Although complete dataset preparation was missing, the framework showed positive outcomes for early prediction. Reducing CKD predictors to basic features was the main goal of Ward & Almasoud [2]. Gradient Boosting achieved 99.1% accuracy, but they indicated that focus on a small number of features could leave out important indications. Using big data and machine learning, Naik et al. [3] were able to predict chronic disorders like stroke in central China with 94.8% accuracy, indicating that this could also apply to other chronic condition. On the UCI dataset, Ghosh et al. [4] tested SVM, ADABOOST, LDA, and Gradient Boosting for CKD; Gradient Boosting proved to be the most accurate, achieving 99.8%, and was suggested using additional algorithms.

Ekanayake et al. [5] compared 11 machine learning models for CKD with a focus on bias reduction. They found that Random Forest and Extra Tree were the most accurate, and they indicated expert input for increased security.  
In a review of CKD prediction models, Bhattacharyya [6] emphasized the necessity of thorough model comparisons for understanding, which is limited by dataset variation. Using SVM and Ant Colony Efficiency, Reshma et al. [7] created a minimal-feature CKD prediction model and suggested testing several feature selection techniques. According to Rob Quinn et al.'s evaluation of prediction models for end-stage CKD in Canada [8], the Fine-Gray model is the most successful in predicting long-term risk, particularly when it comes to risk of death. With models like K-Means and Separation Forest, Antony Linta et al. [9] used unsupervised machine learning to predict CKD with 99% accuracy, which could be helpful for clinics with few experts.

In order to improve patient care, Wainstein et al. [10] analyzed the prediction of CKD development focusing on ML function in developing customized monitoring and treatment plans. Sumitha et al. [11] highlighted the value of early screening for better patient outcomes by using ensemble classifiers to identify early CKD by combining Random Forest and Naïve Bayes. Sitote & Debal [12] used Random Forest, SVM, and Decision Tree to predict the stages of CKD, with Random Forest showing the best results. They underlined how important earlier detection is. Pal [13]-Bagging with models such as SVM and Decision Tree improved the accuracy of CKD prediction by up to 97.23%, Ebiaredoh-Mienye et al. [14] developed a 99.8% accurate cost-sensitive AdaBoost CKD model and proposed applying it to bigger, imbalanced datasets for better illness identification.

The Kidney Failure Risk Equation was shown to be the best accurate machine learning model for predicting end-stage kidney disease (ESKD) by Bai et al. [15]; they suggested more research into indicators. Su et al. [16] used Random Forest to predict the course of CKD, highlighting urea as a critical predictor and offering suggestions for improving prediction models for better results. Bhutani [17] used Artificial Neural Networks (ANNs) to predict CKD, arguing deep learning research for improved outcomes. A CKD prediction system was developed by Ahmed et al. [18], which compared models such as Random Forest and Logistic Regression and suggested larger datasets to increase model accuracy Using classifiers such as KNN, Decision Tree, and Naïve Bayes, Barot [19] proposed a CKD prediction model and suggested model merging for improved accuracy. The efficiency of ensemble methods was highlighted by Chaurasia and Pal [13], who compared ensemble algorithms for CKD prediction. However, they noted that complexity was a hurdle for practical use. Using models like Random Forest and AdaBoost, Gowtham et al. [21] tested ensemble learning for CKD detection and discovered that understanding was a barrier to clinical application.

Hosena [22] emphasized the value of early diagnosis while pointing out that clinical applicability is limited by complexity. The study focused on high-risk CKD prediction using SVM, Random Forest, and ANN Li et al. [23] used machine learning to analyze CKD risk variables in Bangladesh, finding socioeconomic predictors while pointing out the drawbacks of regional datasets for universal use.Shelake et al. [24] indicated new indicators to increase model robustness and used PCA to improve CKD prediction with Random Forest and SVM Using algorithms like Decision Tree, SVM, and Random Forest, Reddy et al. [25] created a CKD prediction model, highlighting the importance of larger datasets for validating results. Islam in [26] XGBOOST has a 98.3% accuracy rate in testing 12 ML classifiers for CKD, indicating greater use of ML for early CKD identification in regardless of feature limitations. [27] ElMalah et al. (2023) studied the use of ML models for predicting CKD. They compared models like SVM, Random Forest, and Logistic Regression using metrics such as accuracy, precision, and recall. Their findings show that SVM performed the best with an accuracy of 98%, proving it to be effective for CKD prediction.

**Table.1: Summary of methodologies**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author** | **Dataset** | **Feature extraction** | **Algorithm** | **Result** |
| Revathy [1] | UCI repository | Focuses on key health metrics like blood composition and patient history to improve predictive accuracy | Decision tree, SVM and Random forest | 94%, 98% and  99% accuracy respectively. |
| Marwa Almasoud, Tomas E. Ward [2] | Custom dataset | ANOVA, Pearson's correlation, Cramer's V test | Logistic SVM, Random Forest, Gradient Boosting Regression | Gradient Boosting achieved 99.1% accuracy |
| Shreekanth Jogar et al. [3] | Central China hospital data | Latent factor model for missing data reconstruction | Decision tree, Random forest, Naïve Bayes | 94.8% accuracy with faster convergence |
| Pronab Ghosh  et al. [4] | UCI ML Repository | Imputation of missing values using KNN splitting dataset into training (80%) &  Testing (20%) | SVM, Ada Boost, LDA, Gradient Boosting | Gradient Boosting achieved 99.8% accuracy. |
| Debnath Bhattacharya [6] | Health records from India | CKD classification outcomes Patient health data attributes | Logistic regression,  Decision tree, KNN, SVM and RF | 75%,72%,69% 75%,  and 74% accuracy  Respectively. |
| Reshma S et al. [7] | UCI repository. | Minimum feature selection | Ant Colony Optimization, Support Vector Machine | Predicted CKD with minimal feature use. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Linta Antony et al. [9] | Clinical data from various sources, unspecified in the abstract. | To enhance classification accuracy, specifically through feature reduction techniques with k-means cluster | K-Means Clustering, DB-Scan, Isolation Forest, Auto encoder. | Achieved 99% accuracy in classifying CKD and Non-CKD using feature reduction with K-Means. |
| Sumitha et al. [11] | UCI ML repository CKD dataset | Key features: blood pressure, blood sugar levels, e GFR, age, hemoglobin etc. | Decision tree, Naive Bayes, Random forest and Proposed model | 94% ,95% ,96% and 98% accuracy respectively. |
| Saurabh Pal  [13] | UCI Machine Learning Repository (25 features) | Bagging ensemble method | Logistic Regression, Decision Tree, Support Vector Machine, Bagging ensemble | Best accuracy with Decision Tree: 95.92%, |
| Sarah A. Ebiaredoh-Mienye et al. [14] | Apollo    hospitals | Information gain-based feature selection | Logistic regression,  Decision tree,  Random forest and SVM | 94%,90%,95% and  93% accuracy  respectively. |
| Qiong Bai, Chunyan Su, et al. [15] | 748 CKD patients from a longitudinal cohort (2006-2008) | Baseline characteristics, blood test result | Logistic regression,  Naive Bayes,  Random forest and KNN | 75% ,86% ,82% and  84% accuracy respectively. |
| Chuan-Tsung Su, et al.  [16] | 858 CKD patients from a veteran's hospital in Taiwan | Logistic regression  Random forest, XGBOOST, SVM and GNB | Patient demographics, laboratory data | 74%,90% ,79% ,92% and 63% accuracy  respectively. |
| Dibaba Adeba Debal et al. [17] | Patient records from st.pawols hospital ,ehiopia , | Analysis of variance and recursive feature elimination | Random Forest, Support Vector Machine, Decision Tree | 78.3%,63% and  77.5% accuracy respectively. |
| Iftekhar Ahmed et al. [18] | CKD dataset from UCI | Filled missing values with mean/mode and random sampling | Random Forest, SVM, Naïve Bayes, Logistic Regression, KNN, boost, Decision Tree, Ada Boost | Random Forest and Logistic Regression achieved 99% accuracy. |
| Mitisha Barot [19] | CKD dataset from UCI | Risk factors related to CKD | KNN, Naive Bayes, Decision tree and Supreme boosting classifier | 94%,96% ,97% and  99% accuracy respectively. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Vikas Chaurasia, Dr. Saurabh Pal [20] | CKD dataset from UCI | Age, blood pressure,  serum | Ada boost, GBM, Random forest and ET | 98%,98%,99%, and  98% accuracy  respectively. |
| Md Nayeem Hosena [22] | CKD dataset from UCI repository | Data pre-processing, filtering | SVM, Random Forest, ANN | Random Forest provided reliable accuracy. |
| Shravani Shelake, et al. [24] | CKD dataset from UCI repository | PCA for dimensionality reduction | Random Forest, Support Vector Machines, Decision Tree, PCA | Random Forest and SVM showed high accuracy. |
| O.Nikhilesh Reddy, et al.[25] | Data on 400 patients with 24 health attributes | Selected attributes based on health indicators | Decision Tree, KNN, Random Forest, SVM, Naive Bayes | Random Forest and SVM showed high accuracy. |
| Md. Ariful Islam [26] | 25 variables initially, reduced to top 30% | Feature selection reduced variables to top predictors | XG BOOST, other machine learning classifiers | XG Boost achieved an accuracy of 98.3%. |
| Aya ElMalah et al. [27] | Kaggle dataset;401 records 24 features | Preprocessed (mean strategy for missing Min-Max scaling) | Logistic regression, decision tree, random forest, support vector classifier | 96.66%,100%,100%  100% accuracy. |

**3. CONCLUSION AND FUTURE SCOPE**

Machine learning (ML) based CKD prediction has great potential for early diagnosis and specific therapy plans. The overview examines a number of ML methods that have shown value in the prediction of CKD, including supervised, unsupervised, and ensemble learning. The ML methods are limited by issues such as imbalanced datasets, feature engineering, model understanding, data quality, and historical patient data variation. To make CKD prediction systems dependable and useful in healthcare, it is essential to include ML models into clinical processes and solve moral issues. ML models must be included into medical processes and moral issues must be resolved for CKD prediction systems to be accurate and useful in healthcare.

**4. REFERENCES**

1. Revathy, S. “Chronic Kidney Disease Prediction Using Machine Learning Algorithms”. The International Conference on Computer Science and Applied Mathematics, IOP Conf. Series: Journal of Physics: Conf. Series, 1255, 012024, (2019).
2. Almasoud, M., & Ward, T. E. “Detection of Chronic Kidney Disease Using Machine Learning Algorithms with Least Number of Predictors”. International Journal of Advanced Computer Science and Applications, vol 10(8) pg. 89–95, (2019).
3. Jogar, S et al. “Chronic Diseases Prediction over Big Data by Using Machine Learning”. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol 5(3), pg. 246-250, (2019).
4. Ghosh, et al. “Chronic kidney disease prediction using machine learning techniques”. *International Journal of Computer Applications*, vol 175(13), pg. 1-7, (2020).
5. Ekanayake, I. U., & Herath, D. “Predicting chronic kidney disease status using machine learning methods”, (2020).
6. Bhattacharyya D. “Comparative study of classification techniques for chronic kidney disease prediction”. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*,vol 5(3), pg. 246-251, (2020).
7. Reshma, S., et al. “A predictive model for chronic kidney disease using machine learning techniques”. *International Journal of Advanced Research in Computer Science and Software Engineering*, vol 10(5), pg.123-130, (2020).
8. Al Wahsh, H., et al. “Kidney failure risk prediction has implications for disease management”, (2021).
9. Antony, L., et al. “A Comprehensive Unsupervised Framework for Chronic Kidney Disease Prediction”. Journal Title, (2021).
10. Wainstein, M, et al. Machine learning and chronic kidney disease risk prediction, (2021).
11. Sumitha, T., et al. “Prediction of Early Symptoms for Chronic Kidney Disease by Ensembled classifiers”. *Turkish Online Journal of Qualitative Inquiry*, vol 12(4), pg.2321-2335, (2021).
12. Dibaba, A, et al. “Chronic Kidney Disease Prediction Using Machine Learning Techniques”. *Journal Name*, *Volume* (Issue), Page Numbers. (2022).
13. Pal, S. “Chronic Kidney Disease Prediction Using Machine Learning Techniques”. (2022).
14. Ebiaredoh-Mienye, et al. “Cost-sensitive Learning for Chronic Kidney Disease Detection”, (2022).
15. Qiong Bai, et al. “Machine Learning Models for the Prediction of Renal Failure in Chronic Kidney Disease: A Retrospective Cohort Study”. *Department of Healthcare Information and Management, Ming Chuan University, Taoyuan, Taiwan*, (2022).
16. Chuan-Tsung Su, et al. “Machine Learning Models for the Prediction of Renal Failure in Chronic Kidney Disease: A Retrospective Cohort Study”. *Department of Healthcare Information and Management, Ming Chuan University, Taoyuan, Taiwan*. (2022).
17. Naveya Bhutani. “Predictive Analysis of Chronic Kidney Disease Using Machine Learning Techniques”. *Department of Information Technology, Maharaja Agrasen Institute of Technology*, (2022).
18. Iftekhar Ahmed, et al. “Computer-Aided Predictive Model for Chronic Kidney Disease Diagnosis”. *Department of CSE, Leading University, Sylhet, Bangladesh*, (2022).
19. Mitisha Barot. “Predicting Chronic Kidney Disease Using Machine Learning Classifiers”. *Department of Computer Science*, (2022)
20. Vikas Chaurasia & Dr. Saurabh Pal. “Application of Artificial Intelligence Techniques in Predicting Chronic Kidney Disease”. *Dept. of Computer Applications, VBS Purvanchal University, Jaunp,* (2022).
21. Gowtham A., et al. “Ensemble Learning Techniques for Chronic Kidney Disease Detection”. *EPRA International Journal of Research and Development*, vol 8(4), (2023).
22. MHosena. N. “Machine Learning Approaches for Early Prediction of Chronic Kidney Disease” *Unpublished Manuscript*, (2023).
23. Li, Y., et al. “Risk Factor Prediction for Chronic Kidney Disease: A Statistical Approach”. *BLSME*. Warwick Evans Publishing. (2023).
24. Shelake, S., et al. “Chronic Kidney Disease Prediction Using Machine Learning”. *International Research Journal of Modernization in Engineering, Technology and Science*, vol 6(5), pg. 89-75, (2023).
25. Reddy, O. N., et al. “Chronic Kidney Disease Prediction Using Machine Learning Algorithms”. *International Journal of Modernization in Engineering, Technology and Science*,vol 6(5), (2023).
26. slam, Md. A. “Predictive Modelling for Early Detection of Chronic Kidney Disease Using Machine Learning Classifiers”, (2023).
27. ElMalah, A., et al. “Chronic Kidney Disease Classification Using ML Algorithms”. Kafrelsheikh Journal of Information Science, (2023).
28. <https://www.verywellhealth.com/kidney-disease-4014698>