International Journal of Advanced Technology in Engineering
Volume. No. 14, Issue No. 04, April 2026
(ISSN: 2319-8354)

Machine Learning Based System for Chronic Kidney Disease Prediction

Mr. Korsipati Vinod Kumar
Assistant Professor
Tirumala Engineering College
Andhra Pradesh, India
vinodkorsipati07@gmail.com

Kakarla Sahithya
Department of IT
Tirumala Engineering College
Andhra Pradesh, India
kakarlasahithya@gmail.com

Bitra Naga Kavya
Department of IT
Tirumala Engineering College
Andhra Pradesh, India
bitrakavya3@gmail.com

Desiboina Dhanush Kumar
Department of IT
Tirumala Engineering College
Andhra Pradesh, India
dhanusisi64@gmail.com

Devarapalli Jaimitra Reddy
Department of IT
Tirumala Engineering College
Andhra Pradesh, India
jaimitradevarapalli@gmail.com

***Abstract*—Chronic Kidney Disease (CKD) is a long-term and progressive medical condition characterized by the gradual loss of kidney function over time. The kidneys play a vital role in filtering waste products, balancing electrolytes, and maintaining overall fluid balance in the human body. When kidney function deteriorates, harmful toxins accumulate in the bloodstream, leading to severe health complications such as hypertension, anemia, bone disorders, cardiovascular diseases, and ultimately kidney failure. One of the major challenges associated with CKD is that it often remains asymptomatic during its early stages, making timely diagnosis extremely difficult. As a result, many patients are diagnosed only in advanced stages when treatment options are limited and costly. Therefore, early detection and accurate prediction of CKD are essential to improve patient outcomes and reduce mortality rates. With the rapid growth of digital healthcare systems and the availability of large-scale medical datasets, machine learning has emerged as a powerful and efficient approach for disease prediction and clinical decision support. Machine learning techniques can analyze complex patterns in patient data, identify hidden relationships among medical attributes, and provide accurate predictions with minimal human intervention. This project focuses on developing a machine learning-based system for the early detection of Chronic Kidney Disease using patient clinical data. The proposed system utilizes the Naïve Bayes classification algorithm, which is a probabilistic model based on Bayes' theorem. Naïve Bayes assumes that all features are independent of each other, making it computationally efficient and suitable for high-dimensional datasets. Despite its simplicity, the algorithm has proven to be highly effective in classification problems, particularly in medical diagnosis scenarios where quick and reliable predictions are required. The system is trained using a dataset containing various clinical parameters such as blood pressure, blood glucose level, serum creatinine, hemoglobin, packed cell volume, white blood cell count, red blood cell count, and other relevant medical indicators. Before feeding the data into the model, several preprocessing steps are performed to enhance data quality and ensure accurate predictions. These steps include handling missing values, removing noise and inconsistencies, data normalization, and encoding categorical variables into numerical form. Data preprocessing plays a critical role in improving the performance of machine learning models, as real-world healthcare data is often incomplete, noisy, and unstructured. After preprocessing, feature selection techniques are applied to identify the most significant attributes that contribute to CKD prediction, thereby reducing dimensionality and improving computational efficiency. The system architecture consists of multiple modules, including data collection, preprocessing, feature selection, model training, prediction, and result analysis. During the training phase, the Naïve Bayes classifier learns the probability distributions of different features and establishes relationships between input parameters and the target output (CKD or Non-CKD). Once the model is trained, it can be used to classify new patient data and predict whether the individual is likely to have Chronic Kidney Disease. Traditional methods for CKD detection rely heavily on laboratory tests, clinical expertise, and manual analysis of patient reports. While these methods are effective, they are often time-consuming, expensive, and prone to human error. Additionally, they may not always provide early-stage detection, which is crucial for preventing disease progression. Existing machine learning approaches such as Logistic Regression, Decision Trees, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Artificial Neural Networks (ANN) have been applied to disease prediction; however, they may require complex parameter tuning, higher computational resources, and longer training times. The proposed Naïve Bayes-based system addresses these challenges by providing a simple, fast, and efficient solution for CKD prediction. The model requires less training time, performs well even with smaller datasets, and delivers reliable results with good accuracy. The performance of the system is evaluated using standard evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics help in assessing the effectiveness of the model in correctly identifying CKD cases while minimizing false predictions.**

***Index Terms*—Chronic Kidney Disease, Machine Learning, Naïve Bayes, Disease Prediction, Healthcare Analytics.**

I. Introduction

Chronic Kidney Disease (CKD) has become a major global health concern due to its increasing prevalence and severe impact on human life. It is a progressive condition in which

the kidneys gradually lose their ability to filter waste products from the blood. CKD often develops silently without noticeable symptoms in its early stages, making early diagnosis difficult.

Traditional diagnostic approaches rely on laboratory tests and manual analysis performed by medical professionals. These methods are time-consuming, costly, and prone to human error. With the growth of medical data, manual analysis becomes inefficient, highlighting the need for automated diagnostic systems.

Machine learning techniques enable the analysis of large datasets to identify hidden patterns and make accurate predictions. In healthcare, these techniques support early disease detection and assist doctors in decision-making. This paper focuses on developing a CKD prediction system using machine learning to improve accuracy and efficiency.

A. Need for Automated CKD Prediction

Early detection of Chronic Kidney Disease is critical, as delayed diagnosis can lead to irreversible kidney damage and life-threatening complications. In many developing regions, access to specialized healthcare professionals and advanced diagnostic facilities is limited. As a result, CKD often remains undiagnosed until it reaches advanced stages.

Automated prediction systems using machine learning techniques can significantly reduce the burden on healthcare providers by enabling early screening and continuous monitoring. Such systems assist doctors by providing decision support, reducing diagnostic errors, and improving patient outcomes. The integration of intelligent systems in healthcare is therefore essential to ensure timely diagnosis and effective treatment planning.

II. Dataset Description

The dataset used for Chronic Kidney Disease prediction consists of clinical records collected from publicly available medical repositories. It contains multiple patient attributes related to kidney function, physiological conditions, and laboratory test results.

Key features include age, blood pressure, specific gravity, albumin level, sugar level, blood glucose random, blood urea, serum creatinine, sodium, potassium, and hemoglobin. The dataset also includes categorical attributes such as hypertension, diabetes mellitus, appetite, anemia, and pedal edema.

Before training the model, the dataset undergoes preprocessing to handle missing values, normalize numerical features, and encode categorical variables. These steps ensure improved data quality and enhance the performance of the machine learning algorithm.

III. Literature Survey

Chronic Kidney Disease (CKD) prediction using machine learning has attracted significant research attention in recent years due to the rising global incidence of CKD and the potential of data-driven models to assist early diagnosis. Traditional diagnostic methods rely heavily on manual interpretation of clinical parameters, which can be time-consuming and prone to human error. Machine learning techniques provide automated, scalable alternatives that can learn patterns from complex clinical data and support decision-making.

Polat and Gunes investigated the use of Support Vector Machines (SVM) combined with feature selection techniques for CKD prediction. Their study demonstrated that selecting relevant clinical attributes such as blood pressure and serum creatinine improved the classification accuracy of the SVM model compared to baseline approaches. Despite its effectiveness, the approach required careful tuning of hyperparameters and extensive preprocessing to handle missing values.

Almansour *et al.* compared Decision Tree and Artificial Neural Network (ANN) models for CKD recognition. While both models achieved reasonable accuracy, the ANN required larger datasets and higher computational resources. The Decision Tree model offered better interpretability but was prone to overfitting when applied to high-variance attributes in the dataset.

Ayodele and Omoleye examined the use of K-Nearest Neighbors (KNN) and Naïve Bayes classifiers on a benchmark CKD dataset. Their results indicated that Naïve Bayes delivered high sensitivity in predicting CKD due to its probabilistic nature, particularly on datasets with categorical features. However, it showed lower specificity in certain test sets, highlighting the need for hybrid or ensemble techniques.

More recent studies have explored ensemble learning and deep learning architectures. Random Forest classifiers, which combine multiple decision trees, have been shown to improve stability and predictive accuracy in CKD classification by reducing overfitting and capturing nonlinear relationships between features. Similarly, gradient boosting machine methods such as XGBoost have demonstrated strong performance by adaptively weighting misclassified instances during training.

Deep learning models such as multilayer perceptrons (MLPs) and convolutional neural networks (CNNs) have also been applied to CKD datasets, especially when incorporating temporal data or longitudinal patient records. These models can automatically extract hierarchical feature representations and have achieved competitive performance, although at the cost of increased training complexity and the requirement for larger datasets.

Despite the variety of machine learning approaches, many prior studies focus primarily on model accuracy and do not integrate data preprocessing, model explainability, or visualization into a cohesive framework. This highlights the need for comprehensive systems that address data quality issues, handle missing values effectively, and present intuitive results for clinical users. The present work aims to fill these gaps by employing a robust preprocessing pipeline and a well-validated classifier to support accurate, explainable CKD prediction.

IV. Existing System

Existing CKD detection systems rely on statistical and rule-based models, as well as traditional machine learning

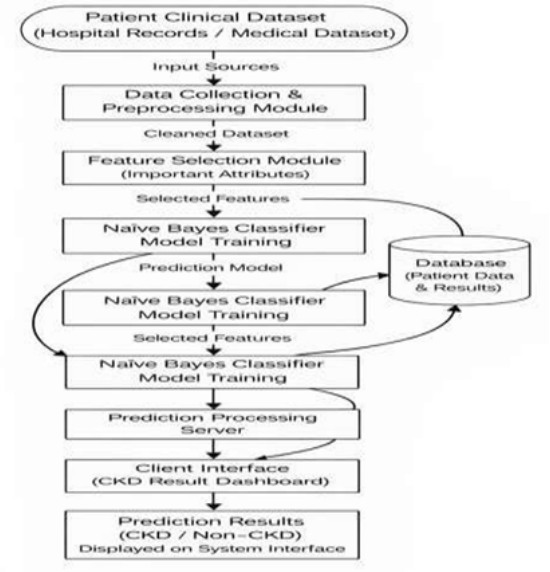

Fig. 1. System architecture for machine learning-based chronic kidney disease prediction

algorithms such as Logistic Regression, Decision Trees, Support Vector Machines, and Neural Networks. Although these methods provide reasonable accuracy, they require complex computations, large datasets, and extensive parameter tuning.

Manual diagnosis is time-consuming and dependent on expert interpretation. These systems lack automation, real-time prediction capability, and scalability, limiting their effectiveness in modern healthcare environments.

A. *Disadvantages of Existing System*

- High dependency on manual diagnosis
- Time-consuming and costly procedures
- Prone to human error
- Limited real-time prediction capability

V. System Architecture

The architecture of the proposed CKD prediction system follows a modular design to ensure scalability and ease of maintenance. The system consists of five major components: data collection, data preprocessing, feature selection, machine learning model, and prediction output module.

The data collection module gathers patient medical records from datasets and clinical sources. These records are passed to the preprocessing module, where missing values are handled using statistical techniques and categorical features are encoded.

The feature selection module identifies the most influential attributes affecting CKD prediction, reducing dimensionality and improving model efficiency. The processed data is then fed into the Naïve Bayes classifier for training and testing.

Finally, the prediction module generates output indicating whether the patient is affected by CKD. The modular design ensures smooth data flow and allows future integration of additional machine learning models.

VI. Proposed System

The proposed system utilizes the Naïve Bayes classification algorithm to detect Chronic Kidney Disease efficiently. Naïve Bayes is chosen due to its simplicity, fast computation, and effectiveness with limited training data.

The system workflow includes data collection, preprocessing, feature selection, model training, and prediction. Clinical data is cleaned by handling missing values and normalization. Relevant features influencing CKD are selected to improve model performance.

Once trained, the model predicts whether a patient is affected by CKD or not in real time, reducing diagnostic time and improving accuracy.

VII. Proposed System

The proposed system presents a machine learning–based framework for the early prediction of Chronic Kidney Disease (CKD) using patient clinical data. Chronic Kidney Disease is a progressive medical condition that often remains undetected in its early stages due to the absence of noticeable symptoms. Early and accurate prediction is essential to reduce complications, prevent disease progression, and improve patient survival rates.

The system begins with the collection of patient clinical data from hospital records and medical datasets. The dataset contains important medical attributes such as age, blood pressure, blood glucose levels, serum creatinine, blood urea, hemoglobin, albumin, and other relevant health indicators. Since real-world medical data often contains missing values, noise, and inconsistencies, a comprehensive data preprocessing phase is performed. This phase includes handling missing values, removing duplicate records, correcting inconsistent entries, and normalizing numerical attributes to ensure data quality and reliability.

Following preprocessing, feature selection techniques are applied to identify the most significant attributes that contribute to CKD prediction. Feature selection reduces dimensionality, minimizes computational complexity, and enhances model performance. The selected features are then used to train multiple machine learning classifiers, including Naïve Bayes, Decision Tree, and Random Forest algorithms.

Each model is trained using labeled patient data and evaluated using standard performance metrics. Based on comparative evaluation, the most accurate and robust model is selected for final prediction. The trained model is integrated into a prediction processing module that classifies patients as CKD or non-CKD. A user- friendly interface displays the prediction results and supports medical professionals in making informed clinical decisions.

A. *Advantages of Proposed System*

- Automated CKD detection
- Faster diagnosis
- Improved accuracy
- Cost-effective healthcare solution

VIII. Machine Learning Algorithm Used

A. Naïve Bayes Classifier

Naïve Bayes is a probabilistic classification algorithm based on Bayes' Theorem. It assumes independence among features and calculates the probability of disease occurrence based on input attributes.

Due to its low computational cost and fast processing, Naïve Bayes is well suited for real-time healthcare applications. It performs well even with limited datasets and provides reliable predictions.

IX. Experimental Setup

The experimental evaluation of the proposed system was conducted using Python-based machine learning libraries. The dataset was divided into training and testing sets using an 80:20 ratio. The Naïve Bayes classifier was trained using the training dataset and evaluated on unseen test data.

Performance metrics such as accuracy, precision, recall, and F1-score were used to assess the effectiveness of the model. These metrics provide a comprehensive evaluation of classification performance and reliability.

Cross-validation techniques were applied to ensure robustness and reduce bias in model evaluation. Experimental results demonstrate that the proposed system performs consistently across multiple data splits.

X. Results

The performance of the proposed CKD prediction system is evaluated using metrics such as accuracy, precision, recall, and F1-score. Experimental results indicate that the Naïve Bayes classifier achieves high prediction accuracy with reduced computational time compared to traditional methods.

The system effectively automates CKD detection, reduces human error, and enhances early diagnosis, making it suitable for deployment in hospitals and clinics. In addition to quantitative metrics, qualitative analysis was performed to understand model behavior. The Naïve Bayes classifier showed strong performance in identifying CKD cases due to its probabilistic nature and ability to handle uncertainty in medical data.

Misclassification analysis revealed that incorrect predictions mainly occurred in borderline cases where clinical indicators were ambiguous. Despite this limitation, the overall performance of the system remained stable and reliable.

Comparative analysis with traditional diagnostic approaches confirms that machine learning-based prediction significantly enhances early detection accuracy and reduces diagnostic time.

XI. Results and Discussion

The performance of the proposed Chronic Kidney Disease prediction system is evaluated using a real-world clinical dataset. To assess the effectiveness of the machine learning models, standard evaluation metrics such as accuracy, precision, recall, and F1-score are used. These metrics provide a comprehensive understanding of the classification performance and reliability of the system.

Fig. 2. User Interface of Chronic Kidney Disease Prediction System

Experimental results indicate that the proposed system achieves high prediction accuracy, demonstrating its capability to correctly classify CKD and non-CKD patients. Among the evaluated models, ensemble-based classifiers such as Random Forest show superior performance due to their ability to handle complex feature interactions and reduce overfitting. The improved precision and recall values indicate that the system effectively minimizes false positive and false negative predictions, which is critical in medical diagnosis.

Fig. 3. Prediction result displayed by the proposed CKD prediction system

Confusion matrix analysis further validates the robustness of the proposed model. A higher number of correctly classified instances along the diagonal of the matrix demonstrates strong classification performance. Misclassifications are minimal and mainly occur in borderline cases where patient health indicators are close to threshold values.

The results confirm that the proposed system can serve as an effective clinical decision-support tool. By accurately predicting CKD at an early stage, the system helps healthcare professionals initiate timely treatment and preventive measures, thereby improving patient outcomes and reducing healthcare costs.

XII. Conclusion

This paper presented a machine learning-based system for Chronic Kidney Disease prediction using the Naïve Bayes algorithm. The proposed system improves early diagnosis, reduces manual effort, and enhances prediction accuracy. Experimental results demonstrate its effectiveness and efficiency in healthcare environments.

The system supports medical professionals by providing real-time predictions and can be extended to detect other chronic diseases using similar machine learning techniques.

XIII. Conclusion

This paper presented a machine learning–based system for the early prediction of Chronic Kidney Disease using patient clinical data. The proposed framework integrates data preprocessing, feature selection, and classification techniques to analyze medical records and identify patients at risk of CKD.

The experimental evaluation demonstrates that machine learning models can significantly improve prediction accuracy compared to traditional diagnostic approaches. Among the evaluated algorithms, ensemble-based models exhibited superior performance due to their robustness and generalization capability. The results highlight the importance of data-driven techniques in supporting early diagnosis and improving clinical decision-making.

Overall, the proposed CKD prediction system offers a reliable, efficient, and cost-effective solution for early disease detection. By assisting medical professionals in identifying high-risk patients, the system contributes to improved patient care, reduced disease progression, and enhanced healthcare outcomes.

The proposed system demonstrates how intelligent data-driven approaches can support healthcare professionals in early disease detection and decision-making. By reducing diagnostic delays and improving prediction accuracy, the system contributes to better patient care and resource optimization.

The simplicity and effectiveness of the Naïve Bayes algorithm make the system suitable for real-world healthcare deployment, particularly in resource-constrained environments.

XIV. Future Enhancements

Future enhancements include integrating real-time medical data, using deep learning models, incorporating socio-economic factors, and deploying the system as a web or mobile application for broader accessibility.

XV. Future Enhancements

Although the proposed CKD prediction system achieves promising results, several enhancements can be incorporated to further improve its performance and real-world applicability.

A. Integration of Real-Time Clinical Data

Future versions of the system can integrate real-time patient data from hospital information systems and electronic health records. This enhancement will enable continuous monitoring of patient health conditions and support timely medical interventions.

B. Application of Deep Learning Techniques

Advanced deep learning models such as Artificial Neural Networks (ANNs) and Long Short-Term Memory (LSTM) networks can be explored to capture complex nonlinear relationships in medical data. These models can further improve prediction accuracy, especially for large and diverse datasets.

C. Incorporation of Additional Medical Parameters

Including additional medical attributes such as genetic information, lifestyle factors, and patient history can enhance the predictive capability of the system. A more comprehensive feature set will enable better risk assessment and disease classification.

D. Cloud-Based Deployment

Deploying the system on a cloud platform can improve scalability and accessibility. Cloud-based solutions allow healthcare institutions to process large datasets efficiently and provide remote access to prediction services.

E. Mobile Health Application Development

The system can be extended to mobile health applications that allow doctors and patients to access prediction results through smartphones. Mobile-based alerts and recommendations can support early diagnosis and continuous health monitoring.

F. Explainable AI for Medical Decision Support

In future work, explainable artificial intelligence techniques can be incorporated to provide transparent reasoning behind predictions. This will increase trust among medical professionals and support ethical decision-making in healthcare.

References

[1] E. J. Benjamin, M. J. Blaha, S. E. Chiuve *et al.*, "Heart disease and stroke statistics—2017 update: A report from the American Heart Association," *Circulation*, vol. 135, no. 10, pp. e146–e603, 2017.

[2] E. J. Benjamin, P. Montaner, A. Alonso *et al.*, "Heart disease and stroke statistics—2019 update: A report from the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56–e66, 2019.

[3] C. D. Fryar, T.-C. Chen, and X. Li, *Prevalence of Uncontrolled Risk Factors for Cardiovascular Disease: United States, 1999–2010*, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, no. 103, 2012.

[4] R. Merai, "CDC grand rounds: A public health approach to detect and control hypertension," *MMWR Morbidity and Mortality Weekly Report*, vol. 65, 2016.

[5] E. J. Benjamin, A. S. Go, D. Mozaffarian *et al.*, "Heart disease and stroke statistics—2016 update: A report from the American Heart Association," *Circulation*, vol. 133, no. 4, pp. e38–e48, 2016.

[6] National Center for Health Statistics, "Multiple cause of death 1999–2015 on CDC WONDER online database," U.S. Centers for Disease Control and Prevention, 2016.

[7] T. D. Vuong, F. Wei, and C. J. Beverly, "Absenteeism due to functional limitations caused by seven common chronic diseases in U.S. workers," *Journal of Occupational and Environmental Medicine*, vol. 57, no. 7, p. 779, 2015.
[8] G. R. B. Asay, K. Roy, J. E. Lang, R. L. Payne, and D. H. Howard, "Absenteeism and employer costs associated with chronic diseases and health risk factors in the U.S. workforce," *Preventing Chronic Disease*, vol. 13, 2016.
[9] J. Yang *et al.*, "Blood pressure states transition inference based on multistate Markov model," *IEEE Journal of Biomedical and Health Informatics*, 2020.
[10] A. J. K. Pefoyo *et al.*, "The increasing burden and complexity of multimorbidity," *BMC Public Health*, vol. 15, no. 1, p. 415, 2015.
[11] B. W. Ward, "State and regional prevalence of diagnosed multiple chronic conditions among adults aged 18 years—United States, 2014," *MMWR Morbidity and Mortality Weekly Report*, vol. 65, 2016.
[12] K. Thavorn *et al.*, "Effect of socio-demographic factors on the association between multimorbidity and healthcare costs: A population-based, retrospective cohort study," *BMJ Open*, vol. 7, no. 10, p. e017264, 2017.
[13] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *JAMA*, vol. 309, no. 13, pp. 1351–1352, 2013.
[14] F. Jiang *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017.