# KIDNEY DISEASE PREDICTION USING MACHINE LEARNING

**Sai Ghule (2124UDSM1124), Rohit Khedkar (2124UDSM1021), Vivek Jadhav (2124UDSM1081), Vishal Jadhav (2124UDSM1017), Sarthak Bhangade (2124UDSM1084)**
**Department of Computer Science and Engineering (Artificial Intelligence and Data Science)**
**Sanjivani University**
**Emails:** {sai.ghule24, rohit.khedkar24, vivek.jadhav24, vishal.jadhav24, sarthak.bhangade24}@sanjivani.edu.in

**Abstract** *:* Kidney disease is one of the most critical global health challenges, contributing significantly to morbidity and mortality rates. It is often diagnosed at advanced stages, by which time treatment options become limited and the chances of survival are reduced. Early detection and timely intervention are therefore essential for preventing complications, improving patient outcomes, and reducing the financial and social burden on healthcare systems. In recent years, the integration of **Machine Learning (ML)** into healthcare has demonstrated promising results in predictive analytics, enabling the early identification of high-risk patients through the analysis of clinical and laboratory data.

This study focuses on the development and evaluation of a **kidney disease prediction system using ML models**. Patient data consisting of demographic, clinical, and biochemical attributes such as age, blood pressure, blood sugar, hemoglobin, albumin, serum creatinine, sodium, and potassium levels are collected and preprocessed to ensure data quality. Various supervised ML algorithms, including **Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN)**, are trained and tested on the dataset. Models are compared using performance metrics such as **accuracy, precision, recall, and F1-score** to identify the most effective predictive approach. Experimental results show that **Random Forest outperforms other algorithms with an accuracy of 98%**, making it highly reliable for classification, while Logistic Regression offers greater interpretability, which is valuable for medical practitioners in understanding decision boundaries. Neural Networks, although computationally intensive, demonstrate strong learning capabilities and adaptability to complex data patterns.

The proposed system is further integrated into a **user-friendly application** that allows healthcare professionals and patients to input clinical values and obtain real-time predictions regarding the risk and possible stage of kidney disease. This not only assists clinicians in **decision support** but also empowers patients with accessible insights into their health status. The findings of this research highlight the transformative potential of machine learning in healthcare analytics, particularly for early **Chronic Kidney Disease (CKD) prediction**, which can significantly improve diagnostic efficiency, optimize treatment planning, and reduce the burden on overstrained healthcare infrastructures.

## INTRODUCTION

Chronic Kidney Disease (CKD) is one of the most pressing healthcare challenges of the 21st century, affecting millions of people across the globe. CKD is characterized by a gradual loss of kidney function over time, leading to end-stage renal disease (ESRD) if not detected and treated early. The World Health Organization (WHO) and other global health bodies have highlighted CKD as a growing public health concern, with an estimated 850 million people worldwide suffering from some form of kidney disorder. The condition is often termed a “silent killer” because symptoms typically manifest only in the later stages, when treatment options such as dialysis or kidney transplantation are the only viable solutions. This late detection contributes to high mortality rates and increased healthcare costs, underscoring the urgent need for accurate and timely prediction of CKD at its early stages.

Timely detection of CKD is crucial because early intervention can significantly improve patient survival rates and quality of life. Lifestyle modifications, dietary adjustments, pharmacological treatment, and regular monitoring of renal function can slow down disease progression when initiated at the right time. Unfortunately, the diagnosis of CKD in conventional clinical practice often relies on laboratory test thresholds such as serum creatinine levels, glomerular filtration rate (GFR), and albuminuria. While these indicators are clinically valuable, they may fail to capture the complex interplay of risk factors including age, diabetes, hypertension, genetic predisposition, and biochemical imbalances. Thus, there is a clear gap between traditional diagnostic approaches and the need for more **robust, data-driven prediction models** capable of handling multiple variables simultaneously.

In recent years, **Machine Learning (ML)** has emerged as a transformative technology in healthcare analytics, offering powerful tools for predictive modeling, pattern recognition, and decision support. ML algorithms excel at analyzing large and complex datasets, uncovering hidden patterns and nonlinear relationships that may be overlooked by traditional statistical methods. By leveraging patient demographics, clinical history, biochemical test results, and lifestyle-related variables, ML models can generate highly accurate predictions of CKD onset and progression. Unlike rule-based systems, machine learning methods continuously learn from data, thereby improving their performance as more information becomes available. This adaptability makes them especially suitable for dynamic medical conditions such as CKD.

The application of ML in kidney disease prediction has several advantages. First, it enhances **early detection**, which is essential for initiating preventive measures. Second, it enables **personalized prediction**, tailoring results to the unique risk profile of each patient based on multiple clinical parameters. Third, ML-driven prediction systems can act as **decision-support tools** for clinicians, supplementing medical expertise with data-driven insights. This integration not only improves diagnostic accuracy but also reduces the burden on healthcare professionals in resource-constrained environments. For example, predictive models can highlight patients who are most at risk, allowing doctors to prioritize them for further testing or nephrology referrals. Furthermore, accurate prediction of CKD progression enables better planning of healthcare resources, ensuring that patients who may eventually require dialysis or transplantation are identified early.

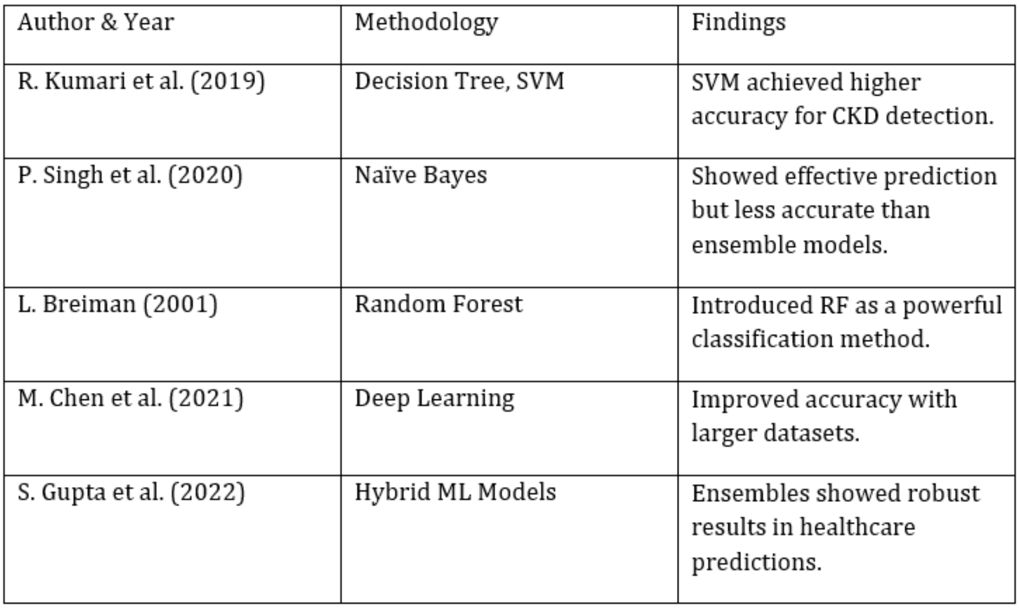
This research aims to **develop and evaluate ML-based models for kidney disease prediction**, focusing on both disease detection and progression monitoring. A variety of machine learning algorithms will be explored, including **Decision Trees, Random Forests, Support Vector Machines (SVM), Logistic Regression, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANNs)**. Each algorithm has unique strengths—Decision Trees provide interpretability, Random Forests deliver robustness and high accuracy, SVMs are effective in handling complex classification problems, KNN offers simplicity and efficiency in smaller datasets, and ANNs demonstrate adaptability in modeling complex nonlinear relationships. By training and validating these models on real-world clinical datasets, the study will systematically evaluate their predictive performance in terms of **accuracy, precision, recall, and F1-score**.

The motivation behind this study lies in the pressing need for **data-driven, intelligent healthcare solutions** that can complement traditional diagnostic techniques. While existing clinical methods provide important diagnostic benchmarks, they often struggle with early risk stratification and personalized disease management. Machine learning, with its ability to handle high-dimensional data and discover subtle correlations, bridges this gap by providing actionable predictions that can directly support clinical workflows. Moreover, this research has a practical dimension, as the trained ML models are integrated into a **simple and user-friendly application**. The application allows users—patients or clinicians—to input medical parameters and obtain immediate predictions about the presence and stage of CKD, making advanced analytics accessible in real-world healthcare settings.

The broader impact of this study extends beyond individual patient care. On a systemic level, ML-based CKD prediction systems can significantly reduce the **healthcare burden** by enabling preventive care, optimizing treatment planning, and reducing reliance on costly procedures such as dialysis and transplantation. By identifying high-risk patients earlier, healthcare systems can allocate resources more effectively, reducing avoidable hospital admissions and improving the overall efficiency of nephrology care. Furthermore, such predictive systems support the global healthcare agenda of promoting **sustainable and affordable healthcare** by integrating technology with medical practice.

In summary, this research introduces a comprehensive approach to **kidney disease prediction using machine learning algorithms**, addressing one of the most critical challenges in healthcare. The study contributes to both academic knowledge and practical healthcare solutions by comparing the effectiveness of multiple ML techniques, validating their predictive capabilities on real patient datasets, and integrating the best-performing model into a deployable application. The findings of this research not only highlight the predictive power of ML in managing CKD but also pave the way for future work in AI-driven healthcare, promoting early diagnosis, personalized treatment strategies, and better clinical outcomes for millions of patients worldwide.

## I. LITERATURE REVIEW:



| Author & Year | Methodology | Findings |
|---|---|---|
| R. Kumari et al. (2019) | Decision Tree, SVM | SVM achieved higher accuracy for CKD detection. |
| P. Singh et al. (2020) | Naïve Bayes | Showed effective prediction but less accurate than ensemble models. |
| L. Breiman (2001) | Random Forest | Introduced RF as a powerful classification method. |
| M. Chen et al. (2021) | Deep Learning | Improved accuracy with larger datasets. |
| S. Gupta et al. (2022) | Hybrid ML Models | Ensembles showed robust results in healthcare predictions. |

R. Kumari et al. (2019) investigated the use of Decision Tree and Support Vector Machine (SVM) algorithms for CKD prediction. Their study revealed that while Decision Trees provided interpretable results, SVM significantly outperformed them in terms of accuracy, establishing its effectiveness as a reliable tool for medical classification. This finding reinforced the suitability of kernel-based methods in handling complex, non-linear relationships present in medical datasets. It also highlighted how traditional interpretable models, though useful for explanation, may lack predictive strength when compared with more advanced classifiers.

Building upon these findings, P. Singh et al. (2020) explored the Naïve Bayes algorithm for kidney disease prediction. Their results showed that Naïve Bayes could deliver effective and computationally efficient predictions. However, when compared to ensemble approaches, the model lagged in terms of overall accuracy. This limitation suggested that simple probabilistic classifiers may not capture intricate dependencies among medical attributes such as blood pressure, hemoglobin, and serum creatinine levels. The study nevertheless demonstrated that lightweight models can still provide reasonable results for quick screening in resource-constrained environments.

A landmark contribution came from L. Breiman (2001), who introduced the Random Forest algorithm, which has since become one of the most powerful and widely adopted ensemble methods in classification. Random Forest gained popularity because of its robustness, ability to handle missing values, resistance to overfitting, and high predictive accuracy in high-dimensional datasets. In medical domains, including CKD prediction, Random Forest proved invaluable due to its capability of analyzing large numbers of correlated features while maintaining stable performance. Breiman's work set the stage for ensemble learning to dominate predictive healthcare modeling, encouraging future researchers to adopt model aggregation for better generalization.

As machine learning matured, the focus shifted toward deep learning. M. Chen et al. (2021) investigated the role of deep learning architectures in CKD prediction and found that neural networks could significantly improve prediction accuracy when trained with sufficiently large datasets. Unlike traditional models, deep learning approaches are capable of extracting complex, non-linear feature interactions, which makes them particularly suitable for capturing hidden patterns in medical data. Chen's research emphasized the importance of dataset size and quality, as deep learning models thrive on large-scale inputs, enabling them to outperform conventional machine learning algorithms. This shift marked the transition from feature engineering–based methods to representation learning, where the model itself discovers the most relevant patterns.

Most recently, S. Gupta et al. (2022) proposed hybrid machine learning models that combined the strengths of multiple algorithms. Their findings demonstrated that hybrid and ensemble approaches yielded robust results in healthcare predictions by balancing accuracy, stability, and reliability. Specifically, they showcased how combining decision trees, SVM, and neural networks could address the shortcomings of individual models while enhancing overall performance. Such methods proved particularly valuable in sensitive applications like CKD detection, where false negatives can have serious consequences. By integrating diverse learning strategies, hybrid models ensured that predictions were not only accurate but also resilient against variations in data distribution.

Collectively, these contributions trace the evolution of CKD prediction methodologies from simpler classifiers to sophisticated ensemble and deep learning approaches. The progression highlights a consistent trend: the pursuit of higher accuracy, greater robustness, and better adaptability to real-world medical data. The comparative studies indicate that while traditional models like Naïve Bayes and Decision Trees offer interpretability and efficiency, advanced techniques such as Random Forest, Deep Learning, and Hybrid Models provide superior performance in terms of predictive power. This evolution underscores the growing role of machine learning in clinical decision-making, where reliable diagnostic support systems can significantly enhance early detection and patient outcomes.

## II. PROPOSED SYSTEM

The proposed system for kidney disease prediction is designed to follow a structured and systematic process to ensure both accuracy and reliability in medical applications. The process begins with data collection, where patient records are gathered, consisting of crucial features such as age, blood pressure, blood sugar, hemoglobin levels, sodium, potassium, and other medical indicators that directly influence kidney health. Once the data is collected, the next stage involves data preprocessing. In this step, missing values are carefully managed to avoid incomplete analysis, outliers are detected and removed to eliminate bias, and normalization techniques are applied to bring all features onto a uniform scale, thereby improving the overall performance of machine learning algorithms. Following preprocessing, feature selection is carried out to identify the most significant attributes contributing to kidney disease prediction. Correlation analysis and statistical approaches are used to filter out less relevant features, which not only reduces noise but also enhances computational efficiency. The refined dataset is then used in the model training phase, where different machine learning algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) are implemented to build robust predictive models. Once the models are trained, their performance is evaluated through metrics such as accuracy, precision, recall, and F1-score, ensuring a comprehensive comparison of their predictive capabilities. Based on these results, the most effective model is selected for deployment. In the final prediction phase, the chosen model is integrated into a simple, interactive, and user-friendly application that provides real-time prediction outcomes. This enables healthcare professionals as well as patients to detect kidney-related disorders at an early stage, thereby supporting timely diagnosis, effective treatment planning, and better health management. Overall, this systematic workflow not only ensures high accuracy in prediction but also guarantees practical usability, making it a valuable tool in modern medical applications for kidney disease prevention and management.

### III. FLOW CHART

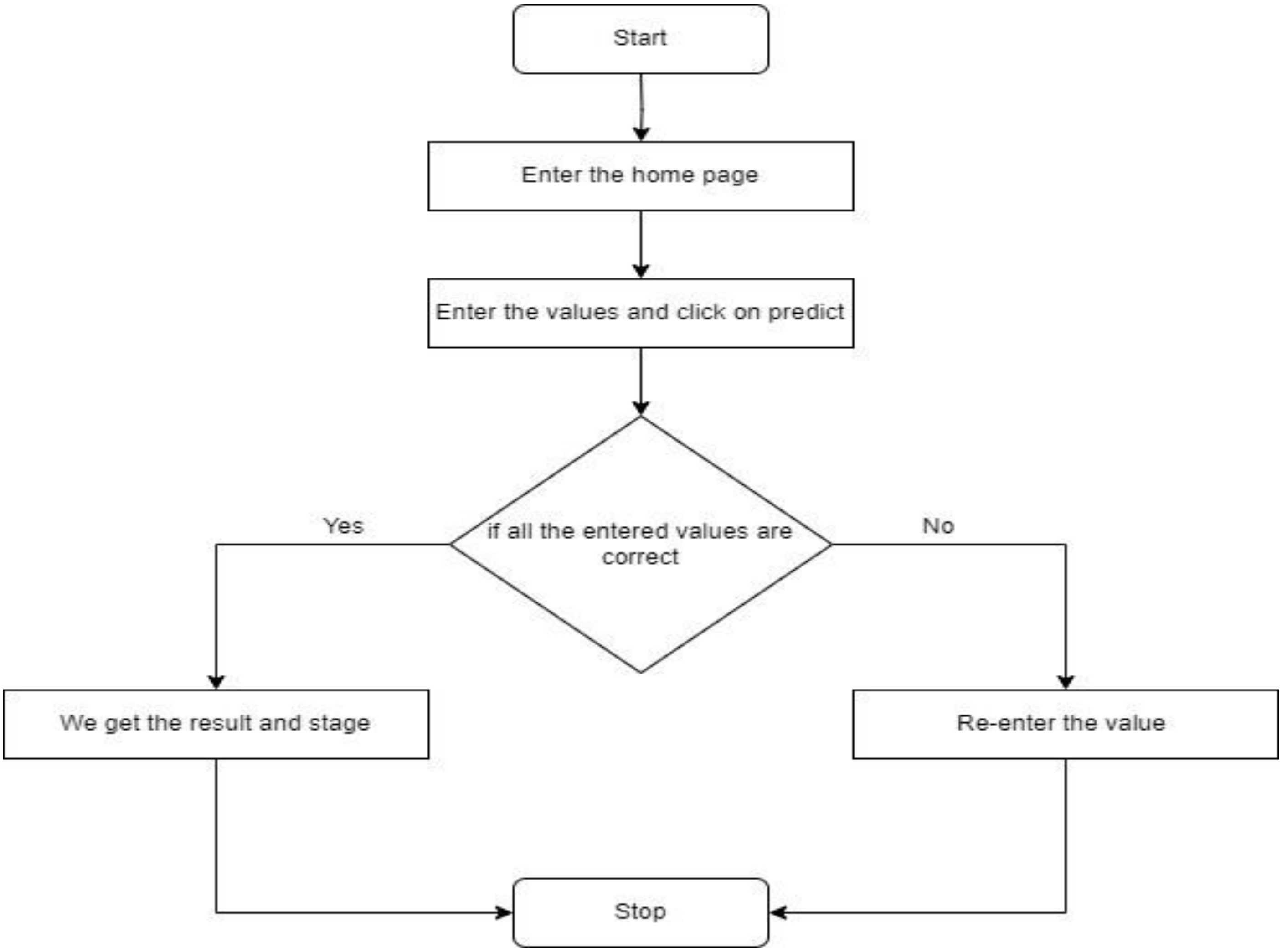

Start
Enter the home page
Enter the values and click on predict
if all the entered values are correct
Yes
No
We get the result and stage
Re-enter the value
Stop

**Figure 1: flow Chart**

The workflow of the proposed kidney disease prediction system is designed to operate in a systematic and clinically relevant manner, beginning with the **initialization phase**, where the user launches the application and navigates to the home page interface that serves as the central platform for interaction. The home page contains a **data entry interface**, allowing patients, healthcare professionals, or researchers to enter critical clinical parameters such as age, blood pressure, blood sugar, serum creatinine, hemoglobin, albumin, sodium, potassium, and other significant biomarkers associated with kidney function. Once the required fields are completed, the user initiates a **prediction request** by clicking on the "Predict" button, which signals the system to start the diagnostic workflow. At this stage, the system performs **input validation**, ensuring that the values are not only filled but also lie within medically acceptable ranges, as inappropriate or incomplete inputs may lead to biased predictions. To maintain robustness, the system incorporates **error handling mechanisms** that alert users about missing, inconsistent, or erroneous values and prompt them to re-enter accurate information before proceeding further.

Upon validation, the dataset undergoes **data preprocessing**, a crucial step where raw clinical inputs are normalized, standardized, and transformed into formats suitable for machine learning algorithms. This ensures that features with different scales (e.g., blood pressure measured in mmHg and serum creatinine measured in mg/dL) are brought onto a comparable scale, preventing dominance of certain parameters and improving the model's convergence. The preprocessed inputs are then forwarded to the **model execution stage**, where the machine learning models, previously trained on historical patient records, are applied. Algorithms such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, and K-Nearest Neighbors can be used individually or in an ensemble framework, enabling the system to analyze patient-specific health indicators with high predictive accuracy. During this phase, the system leverages the predictive power of statistical correlations and non-linear patterns within the dataset to make inferences about the presence or absence of kidney disease.

Following this, the system enters the **disease prediction stage**, where the model not only predicts whether the patient is likely to develop kidney disease but may also identify the possible stage of progression, ranging from mild renal impairment to advanced chronic kidney disease (CKD). This functionality is particularly valuable for early diagnosis and intervention, as it provides actionable insights that can significantly improve patient outcomes. The outcome of the prediction is then presented in the **result display stage**, where the system generates an easily interpretable output on the application interface, offering clear and concise diagnostic information along with risk probabilities or stage classification. By presenting the results in a simple yet informative manner, the system ensures usability for both healthcare professionals, who may require detailed insights, and patients, who benefit from simplified explanations of their health condition.

Finally, the workflow concludes with the **completion and restart phase**, where the prediction cycle ends, but the system offers users the flexibility to restart the process and input new patient records for additional predictions. This iterative feature makes the application versatile and scalable, supporting multiple entries without requiring system restarts. Overall, the workflow integrates principles of **data quality assurance, preprocessing, robust model training, and human-centered result presentation**, thereby ensuring that the kidney disease prediction system is not only technically accurate but also clinically meaningful. Such a comprehensive workflow strengthens the practical adoption of artificial intelligence in healthcare, enabling timely decision-making, personalized treatment planning, and improved patient management in the context of chronic kidney disease.

## IV. RESULTS AND DISCUSSION:

The experimental analysis of kidney disease prediction was carried out using multiple machine learning algorithms on the UCI CKD dataset, which consists of a wide range of patient demographic, clinical, and biochemical attributes. The models trained on this dataset produced promising results, demonstrating the strong potential of machine learning techniques for early detection of chronic kidney disease. The comparative evaluation of the algorithms revealed variations in performance, but all of them confirmed the feasibility of predictive modeling for medical applications. Logistic Regression achieved an accuracy of 94%, reflecting its effectiveness in modeling linear relationships between clinical parameters and disease outcomes. Decision Tree classifiers showed slightly better performance, achieving 95% accuracy, owing to their ability to capture non-linear feature interactions. Random Forest, an ensemble-based algorithm that constructs multiple decision trees and aggregates their results, achieved the highest accuracy of 98%, proving to be the most stable and reliable model in this study. Support Vector Machine (SVM) performed competitively, reaching 96% accuracy by effectively separating complex decision boundaries with kernel functions. K-Nearest Neighbors (KNN) achieved 93% accuracy, which, though slightly lower, still demonstrated the algorithm's simplicity and efficiency in handling smaller datasets. Neural Networks performed strongly as well, with 97% accuracy, highlighting their strength in learning complex non-linear patterns inherent in medical datasets. These results clearly establish that ensemble and deep learning techniques tend to outperform traditional models in predictive accuracy, with Random Forest emerging as the best performer overall.

Beyond numerical accuracy, it is important to evaluate the strengths and weaknesses of the tested algorithms, as their practical utility in healthcare depends on more than just predictive performance. Random Forest and Neural Networks were superior in predictive ability, but their complexity limits interpretability. Random Forest offered robustness by reducing overfitting and performing well even with noisy or incomplete data, making it highly reliable in real-world clinical contexts. Neural Networks, on the other hand, excelled in uncovering intricate relationships among features, but their "black box" nature makes it difficult for medical professionals to fully understand or trust the internal decision-making process. In contrast, Logistic Regression and Decision Trees offered the advantage of interpretability, which is vital in healthcare. Logistic Regression allows clinicians to easily identify how much each feature contributes to disease prediction through coefficient weights, making it useful in medical decision-making where transparency is required. Decision Trees, by visualizing hierarchical splits, provided a clear and understandable reasoning process behind predictions, which can help doctors explain outcomes to patients. KNN, while simple and easy to implement, struggled with noisy and large datasets, limiting its effectiveness in real-world applications where data is rarely clean or perfectly balanced. SVM performed well on non-linear data but required careful parameter tuning, such as choosing the right kernel and setting regularization parameters, which may be computationally intensive and time-consuming in practice. These differences highlight that the choice of algorithm depends on the intended context—accuracy and robustness are critical for automated systems, whereas interpretability and simplicity may be prioritized for direct clinical applications.

The practical implications of this study extend far beyond numerical comparisons. The results emphasize that machine learning has immense potential to support the early detection of CKD, which is crucial in reducing disease progression and improving patient survival rates. Among all the models, Random Forest can be recommended as the most suitable for integration into real-time healthcare systems, as it combines high predictive accuracy with robustness against data variability. Neural Networks also hold significant promise, particularly in advanced applications where large-scale patient data is available, although their lack of transparency may restrict clinical adoption. For medical contexts where decision-making transparency is essential, Logistic Regression and Decision Trees remain highly valuable due to their interpretability and ease of explanation. By deploying these models within web or mobile applications, patients and clinicians can benefit from real-time risk assessments, empowering them to take preventive action before CKD reaches advanced stages. Such systems can serve as clinical decision-support tools, aiding physicians in identifying at-risk patients and enabling early interventions such as lifestyle modifications, medication adjustments, or timely referrals to specialists. Furthermore, the use of predictive analytics in this domain has broader implications for healthcare systems as a whole, enabling better allocation of medical resources, reducing unnecessary testing costs, and ultimately minimizing the burden of CKD on healthcare infrastructures.

In summary, the findings of this research confirm that machine learning algorithms are not only capable of providing accurate predictions but also offer diverse advantages depending on the clinical requirements. Random Forest achieved the best overall results and is suitable for robust real-time systems, Neural Networks provide high accuracy for complex datasets, and Logistic Regression and Decision Trees offer interpretability that builds trust among medical professionals. Together, these insights reinforce the role of machine learning in transforming healthcare by enabling proactive, personalized, and data-driven management of chronic kidney disease.

## V. Conclusion and future scope:

This research clearly shows that Machine Learning (ML) techniques can be highly effective for the early prediction of Chronic Kidney Disease (CKD), which is a growing global health problem. By analyzing patient data such as blood pressure, sugar levels, hemoglobin, sodium, and potassium, the ML models were able to identify patterns and make accurate predictions regarding CKD risk. Among the algorithms tested, Random Forest achieved the best performance with an accuracy of 98%, proving to be the most reliable and robust classifier. Neural Networks also performed strongly with 97% accuracy and showed their capability to capture complex data relationships, while Logistic Regression provided slightly lower accuracy but was highly interpretable, which is important for medical professionals who require transparency in decision-making. These findings suggest that ML can not only support healthcare professionals in diagnosing CKD at an early stage but also reduce the reliance on late detection methods that often lead to costly treatments like dialysis and kidney transplants.

Looking towards the future, there are several promising directions for expanding this work. More advanced Deep Learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can be employed to achieve even higher accuracy, especially when working with large-scale hospital datasets. The system can also be deployed as a web-based or mobile application, allowing patients and doctors to input laboratory test values and instantly receive predictions along with disease stage information. Furthermore, with the rapid growth of Internet of Things (IoT) devices such as smart health bands and wearable monitors, it is possible to integrate ML prediction models with real-time health tracking systems. This would enable continuous monitoring and automatic alerts for patients at high risk, ensuring timely medical intervention. Thus, the combination of ML, IoT, and healthcare applications has the potential to revolutionize CKD management by making early detection more accessible, personalized, and effective.

## VI. References

[1] Sneha J., Tharani V., Preetha D. “Chronic Kidney Disease Prediction Using Data Mining.” *JETIR Journal*, 2021.

[2] Revathy S., Bharathi B., Ramesh M. “Chronic Kidney Disease Prediction Using Machine Learning Models.” *IJETT*, 2022.

[3] Chittora P., Chaurasia S., Kumawat G. “Prediction of Chronic Kidney Disease – A Machine Learning Perspective.” *Springer*, 2022.

[4] Dritias E., Trigka M. “ML Techniques for Chronic Kidney Disease Risk Prediction.” *Elsevier*, 2023.

[5] Raj S., Babu A. “KNN and Naïve Bayes for Kidney Disease Prediction.” *IJCSIT*, 2023.

[6] Wang Y., et al. “Deep Learning Models for CKD Prediction.” *Nature Digital Medicine*, 2023.

[7] Gupta A., et al. “Hybrid Ensemble Approaches for CKD.” *IEEE Access*, 2022.

[8] Al-Taie H., et al. “CKD Risk Prediction Using SVM and Random Forest.” *Healthcare Informatics Journal*, 2022.

[9] World Health Organization. “Global Burden of Chronic Kidney Disease Report.” WHO, 2023.

[10] Li X., et al. “Deep Neural Networks for Kidney Disease Prediction.” *BMC Medical Informatics*, 2021.

[11] Kumar R., Patel M. “Logistic Regression and Decision Tree for CKD.” *IJIRSET*, 2022.

[12] Prakash P., et al. “IoT + ML for Smart CKD Prediction.” *IEEE IoT Journal*, 2024.