*International Journal on Recent and Innovation Trends in Computing and Communication*
***ISSN: 2321-8169 Volume: 11 Issue: 11***
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

# Advancements in Machine Learning for the Diagnosis of Chronic Kidney Disease

**Deepa P. S.**
Research scholar,
Department of Computer Science,
Shri Venkateshwara University, Gajraula, UP, India

**Balaji Venkateswaran**
Research scholar
Department of Computer Science,
Shri Venkateshwara University, Gajraula, UP, India
Email: *balaji.venkateswaran@gmail.com*

**Shuchi Goplani**
Assistant Professor,
Department- AIML
ISBM college of Engineering, Pune, Maharashtra, India
Email: *shuchigoplani@gmail.com*

**Kailash Nath Tripathi**
Assistant Professor
Department of Computer Engineering
ISBM College of Engineering, Pune, Maharashtra, India
Email: *kailash.tripathi@gmail.com*

**B. Murali Krishna**
Sr. Engineer,
Cardinal Health International India pvt Ltd, Bengaluru, Karnataka, India
Email:*banalamurali05@gmail.com*

**Shamshed Ali**
Research scholar
Department of Computer Science
Shri Venkateshwara University, Gajraula, UP, India

**Abstract:** Chronic Kidney Disease (CKD) constitutes a significant global health issue, precipitating damage to the kidneys and stripping many individuals of their most productive years. Alarmingly, 40% of those affected by CKD remain oblivious to their condition, a stark contrast to many other diseases where early detection is more common. Unlike other conditions, CKD eludes cure unless identified promptly in its nascent stages. This research emphasizes the collection of critical indicators such as blood pressure and diabetes status to ascertain the presence of CKD in individuals. It proposes the employment of advanced machine learning techniques, including Random Forest, XGBoost, and Support Vector Machines, aiming to enhance early detection and thereby mitigate the disease's impact. Utilizing a CKD dataset, this study endeavors to predict the likelihood of CKD in individuals, offering a proactive approach to tackle this formidable health challenge.

**Keywords**- Machine Learning (ML), Chronic Kidney Disease(CKD), Random Forest(RFC), XGBoost(XGC), Support Vector Machines(SVM)

## I. INTRODUCTION

The kidneys are vital organs for both humans and animals, performing crucial functions such as osmoregulation and excretion. They play a pivotal role in blood purification, eliminating toxic substances and waste from the body. Chronic Kidney Disease (CKD) poses a significant threat to public health, impairing kidney function and leading to diminished organ performance. In India alone, CKD accounts for approximately one million new cases annually [1]. Regular laboratory tests can detect CKD, allowing for interventions that may halt its progression. Untreated, CKD can progress to permanent kidney failure. Early detection is vital; symptoms of early-stage CKD include high blood pressure, anaemia, poor general health, and weak bones, along with reduced waste elimination due to compromised kidney function. However, some individuals may not exhibit symptoms, making early detection challenging. Machine learning offers a promising solution for predicting CKD presence, leveraging data analysis to identify those at risk. The Glomerular Filtration Rate (GFR) test is paramount for assessing kidney function and determining CKD's stage, with five stages of damage severity categorized based on GFR results.

TABLE 1: STAGES OF CHRONIC KIDNEY DISEASE

| **Stage of Chronic Kidney Disease** | **Description Kidney function** | **e-GFR level** |
|---|---|---|
| I | Normal with Urine symptoms | >90 ml |
| II | Slightly-reduced Urine Symptoms | 60 - 89 ml/min |
| III | Moderately | 30 - 59 ml/min |
| IV | Severely-reduced | 15 - 29 ml/min |
| V | Very severe with kidney failure | < 15 ml/min |

Table 1 illustrates that the awareness of declining kidney functionality typically becomes apparent only after reaching stage II of chronic kidney disease (CKD). Recognizing the crucial role of early detection, the potential to mitigate the progression of CKD becomes evident. The advent of machine learning and artificial intelligence has catalyzed the development of various classifiers and clustering algorithms, which are now instrumental in facilitating earlier identification of CKD. These advanced technologies offer promising avenues for improving patient outcomes by enabling timely interventions before the disease progresses to more advanced stages.

The imperative to leverage machine learning (ML) for diagnosing chronic kidney disease (CKD) stems from the intricate and often asymptomatic nature of this disease, coupled with its profound impact on public health. CKD remains a stealthy adversary; its early stages frequently go unnoticed due to the absence of symptoms, leading to delayed treatment and, consequently, severe complications including irreversible kidney failure. The traditional methods of diagnosis, reliant on symptom observation and standard laboratory tests, face limitations in early detection and risk stratification, underscoring the need for more advanced, predictive approaches.

Machine learning, with its ability to sift through and analyze vast datasets to identify patterns and correlations beyond human discernment, presents a transformative solution. By harnessing ML algorithms, healthcare professionals can integrate and evaluate diverse data points—from demographic information and genetic predispositions to subtle variations in lab results and vital signs. This holistic analysis can predict CKD presence and progression more accurately and at an earlier stage than ever before, offering a crucial window for intervention that can drastically alter the disease's trajectory. Furthermore, the predictive power of ML can personalize patient care, tailoring treatment plans to individual risk profiles and thereby enhancing outcomes. It also holds promise for unraveling the complex etiologies of CKD, potentially unveiling novel risk factors and therapeutic targets.

In essence, the motivation to employ ML in CKD diagnosis lies in its potential to revolutionize early detection and management, transitioning from a reactive to a proactive and personalized healthcare paradigm. This shift not only promises to improve patient outcomes and quality of life but also to alleviate the societal and economic burdens posed by CKD, paving the way for a healthier future.

The structure of this paper is designed to provide a comprehensive understanding of the study; Section 3 offers an overview of the ML algorithms deployed. Section 4 outlines the methodology adopted for developing the predictive models. Section 5 presents the experimental outcomes derived from these models. Lastly, Section 6 concludes the study and discusses potential directions for future research.

## II. II.LITERATURE SURVEY

This compilation of research highlights various approaches to chronic kidney disease (CKD) classification through machine learning, categorizing the studies based on the techniques and algorithms employed.

## III. Neural Networks and Related Algorithms:

S.Ramya and Dr.N.Radha[4] explored enhancing diagnosis time and accuracy using machine learning classification algorithms, focusing on the classification of CKD stages. They analyzed algorithms including the Basic Propagation Neural Network, RBF, and RF, with findings indicating the RBF algorithm outperformed others, achieving 85.3% accuracy.

J. Snegha[10] proposed a system employing data mining techniques like the Random Forest algorithm and the Back Propagation neural Network. Their comparison revealed superior performance from the Back Propagation algorithm, utilizing a feedforward neural network.

### *A. Random Forest and Decision Trees*

Gunarathne W.H.S.D et.al.[3] compared the results of various models and concluded that the Multiclass Decision Forest algorithm exhibited greater accuracy over others for a dataset reduced to 14 attributes.

Baisakhi Chakraborty [9] developed a CKD prediction system using several machines learning techniques, including K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine, and Multi-Layer Perceptron Algorithm. Among these, Random Forest was chosen for its superior accuracy, precision, and recall.

### *B. Innovative Approaches and Feature Selection*

Asif Salekin and John Stankovic applied a novel approach to detect CKD using machine learning algorithms on a dataset with 400 records and 25 attributes. They utilized K-nearest neighbors, Random Forest, and Neural Networks, employing a wrapper method for feature reduction, which proved highly accurate in CKD detection.

Mohammed Elhoseny (2019) described a system for CKD detection employing Density-based feature selection with Ant Colony Optimization (ACO), using wrapper methods for feature selection.

### *C. Handling Missing Value*

S.Dilli Arasu and Dr. R. Thirumalaiselvi[5] addressed the challenge of missing values in CKD datasets, which can compromise model accuracy and prediction results. They implemented a recalculation process for CKD stages, replacing missing values with recalculated ones to enhance data integrity.

This body of work underscores the diversity and potential of machine learning in advancing CKD classification and prediction, showcasing a range of algorithms from neural networks to decision trees and innovative feature selection techniques.

## IV. MACHINE LEARNING ALGORITHMS

### *A. Random Forest Classifier:*

Random Forest (RF) [6], an ensemble learning technique, leverages multiple decision trees during the training phase to output the mean prediction of individual trees, enhancing prediction accuracy and robustness. This method utilizes random sampling with replacement from the training dataset to construct each sub-tree model, allowing these models to operate in parallel independently. The aggregation of results from all sub-models leads to a final prediction that benefits from the diverse perspectives of each tree.

Distinctively, Random Forest introduces variations in the construction of decision trees compared to traditional methods [7]. While standard decision trees aim for the optimal branching decision at each node to minimize entropy, thereby creating a highly specific path based on the entire set of variables, Random Forest selects split points at each node from a random subset of predictors. This strategic randomness in choosing split points from among the best available options for a subset of predictors rather than the entire set effectively reduces the risk of overfitting. Overfitting, a common pitfall of using a single, deep decision tree where the model becomes too tailored to the training data and performs poorly on unseen data, is mitigated through this diversified approach. By building a forest of trees where each is slightly different, Random Forest achieves a balance between detail and generalization, making it a powerful tool for predictive modeling (Figure 1).

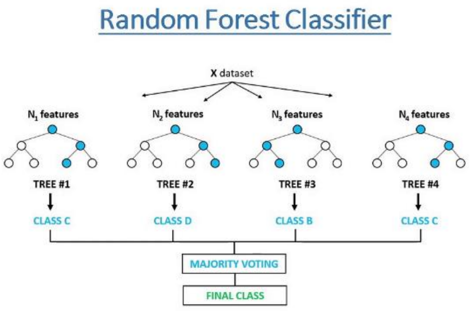

Figure 1. Simplified structure of Random Forest

### *B. XGBoost:*

XGBoost stands as a highly efficient and optimized gradient boosting library, enhancing the suite of machine learning algorithms within the Gradient Boosting Decision Tree (GBDT) framework. This advanced system distinguishes itself by focusing on the sequential improvement of models through the careful minimization of residuals. Unlike the approach taken by Random Forest, where each tree is built independently, XGBoost crafts each subsequent tree to specifically address and reduce the residuals left by its predecessor, thereby refining the model's accuracy progressively with each step.

A key innovation of XGBoost lies in its analytical depth. While traditional GBDT methodologies rely solely on the first derivative of error information to guide decision tree growth, XGBoost employs a second-order Taylor expansion of the cost function. This allows the algorithm to consider both the first and second derivatives of the loss function, offering a more nuanced understanding of the direction and curvature of error gradients. As a result, XGBoost can navigate the path to minimization with greater precision, significantly enhancing model performance.

Furthermore, XGBoost's flexibility extends to its support for custom cost functions. This adaptability enables practitioners to tailor the learning process to specific objectives and constraints, making XGBoost a versatile tool capable of tackling a wide array of predictive modeling challenges. The combination of sequential residual minimization, advanced analytical techniques, and customization options positions XGBoost as a formidable force in the field of machine learning (Figure 2).

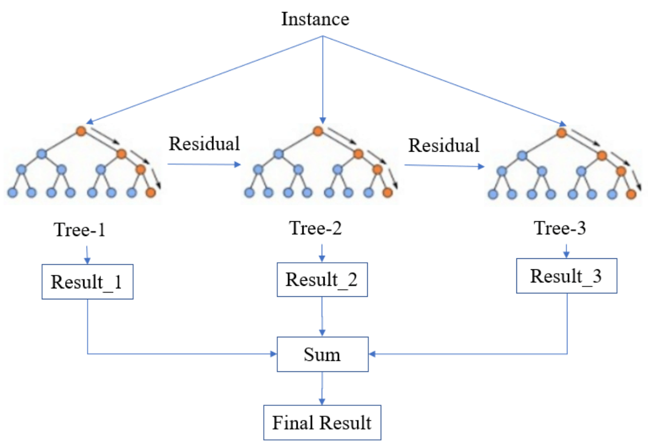

Figure 2. Simplified Structure of XGBoost

### *C. Support Vector Machines(SVM)*

Support Vector Machines (SVM) are a pivotal component in the suite of supervised learning models utilized within machine learning for both classification and regression tasks. The essence of SVM lies in its capacity to categorize training examples distinctly into one of two groups, thereby establishing a foundation for a model that adeptly assigns new instances to either category, functioning as a non-probabilistic binary linear classifier.

The core mechanism that empowers SVM to perform with remarkable efficiency is its strategy of transforming data into a high-dimensional feature space. This transformation facilitates the classification of data points by constructing a hyperplane or set of hyperplanes in this elevated feature space. Notably, this approach is especially beneficial when dealing with datasets that are not linearly separable in their original space. By projecting the data into a higher dimension, SVM makes it feasible to delineate classes that were otherwise intertwined, using linear decision boundaries.

This capability to effectively handle non-linear separability without direct manipulation of the original data space, through the use of kernel functions, underscores the robustness and versatility of SVM as a classification tool. Whether applied to simple linearly separable problems or complex datasets requiring intricate separation, SVM continues to be a cornerstone algorithm for tackling diverse challenges in classification and regression analysis within the realm of machine learning.

## V. PROPOSED METHOD

The proposed method for building the predictive models for the CKD(Figure 3) is as follows:

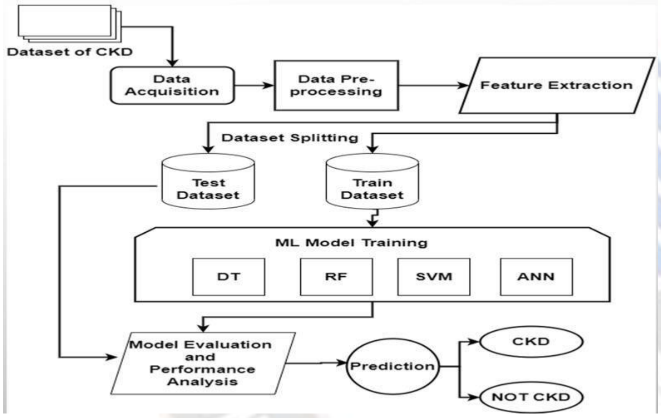

Figure 3. Steps involved in training and testing the model

### A. *Dataset:*

For this project, the CKD Dataset from the UCI repository is utilized, encompassing 400 samples across two distinct classes with 25 attributes, including 11 numeric, 13 categorical, and one class attribute. The dataset suffers from some missing data values, which adds to the challenge of analysis. It encompasses a range of patient data, such as age, blood pressure, red and white blood cell counts, hemoglobin levels, and more, offering a comprehensive view of factors relevant to CKD. Attributes covered in this dataset include Age, Blood Pressure, Specific Gravity, Albumin, Sugar, Red Blood Cells, Pus Cell, Pus Cell clumps, Bacteria, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, Red Blood Cell count, White Blood Cell count, Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema, Anemia, and a Classification attribute that distinguishes between the classes. This rich dataset provides a solid foundation for exploring the predictive capabilities of machine learning models in diagnosing and understanding the nuances of CKD.

### B. *Data Pre-processing:*

Data pre-processing is a crucial step in preparing a dataset for machine learning analysis, ensuring that the data is in a format that algorithms can efficiently process. This stage involves handling missing values, which are either filled with the mean, median, mode, or a constant value for numeric attributes, or replaced with the most frequent value for categorical attributes. Additionally, categorical data must be converted from object type to a numerical format, typically float64, to facilitate analysis. This conversion is often achieved through label encoding, which assigns an integer value to each unique category, effectively transforming categorical attributes into numerical ones. The pandas library proves invaluable for these pre-processing tasks, streamlining the process of preparing data for subsequent analysis.

### C. *Feature selection*

Feature selection is then employed to identify the most impactful attributes for the prediction task, enhancing model performance by eliminating irrelevant or redundant features. This step is essential for improving the efficiency of machine learning models, as it reduces computational complexity and execution time. Effective feature selection not only streamlines the model but also can significantly enhance predictive accuracy.

Following feature selection, the dataset is divided into two segments: 80% for training and 20% for testing. This split allows for the application of proposed models—such as Random Forest, XGBoost, and Support Vector Machines—to the training set, with their performance evaluated based on prediction accuracy using the test set. Through this process, the most effective model can be determined, guiding the selection of the optimal approach for addressing the specific machine learning task at hand.

## VI. RESULTS AND DISCUSSION

The table 2 presents the performance metrics of three classifiers—Random Forest Classifier (RFC), XGBoost (XGB), and Support Vector Machine (SVM)—evaluated based on four key metrics: Accuracy, Recall, Precision, and F1-score. Here's a breakdown of what each metric represents and what the values signify for each classifier:

### A. *Accuracy*

This metric measures the overall correctness of the classifier, i.e., the ratio of correct predictions to the total number of predictions. A higher accuracy indicates better overall performance. RFC has the highest accuracy at 0.98, indicating it correctly predicts 98% of the outcomes. XGB follows with an accuracy of 0.96, while SVM has the lowest accuracy at 0.78. SVM achieves a perfect recall of 1, suggesting it identifies all true positives correctly but at the cost of more false positives, as indicated by its lower precision.

TABLE 2: RESULTS OF THE CLASSIFIERS

| **Classifiers** | **Accuracy** | **Recall** | **Precision** | **F1-score** |
|---|---|---|---|---|

Article Received: 25 July 2023 Revised: 12 September 2023 Accepted: 30 September 2023

| | | | | |
|---|---|---|---|---|
| RFC | 0.98 | 0.95 | 1 | 0.97 |
| XGB | 0.96 | 0.93 | 0.87 | 0.90 |
| SVM | 0.78 | 1 | 0.63 | 0.77 |

RFC and XGB have recall scores of 0.95 and 0.93, respectively, indicating they are also proficient at identifying true positives. This metric assesses the proportion of true positive predictions in the total predicted positives. Higher precision indicates fewer false positives.

- RFC achieves perfect precision at 1, meaning all its positive predictions are correct.
- XGB and SVM have precision scores of 0.87 and 0.63, respectively, indicating a varying degree of false positives in their predictions.

### B. *F1-score*

The F1-score is the harmonic mean of precision and recall, providing a single metric to assess a balance between them. An F1-score closer to 1 indicates a balanced classifier with both high recall and precision.

- RFC leads with an F1-score of 0.97, showcasing an excellent balance between precision and recall.
- XGB has an F1-score of 0.90, indicating a good balance, whereas SVM's F1-score of 0.77 suggests it is less effective at balancing recall and precision compared to the others.

In summary, RFC emerges as the most effective classifier across all metrics, showcasing high accuracy, recall, precision, and F1-score. XGB also performs well but slightly lags behind RFC. SVM, while achieving perfect recall, struggles with precision, leading to a lower overall F1-score and accuracy. This table effectively compares the classifiers' abilities to predict outcomes accurately while maintaining a balance between identifying all relevant instances and ensuring the correctness of those identifications.

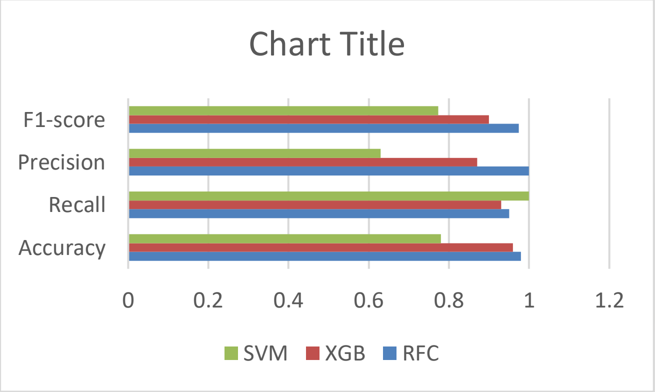

Figure.4 Performance metrics scores for dataset

The figure 4 shows the values of precision, Recall and F1 -score Performance metrics for three classifiers for our dataset.

## VII. CONCLUSION

The table presents a comprehensive evaluation of three classifiers—Random Forest Classifier (RFC), XGBoost (XGB), and Support Vector Machine (SVM)—based on metrics of accuracy, recall, precision, and F1-score. The Random Forest Classifier emerges as the most proficient, showcasing an exceptional balance of accuracy (0.98), precision (1.00), and an F1-score (0.97), with a slightly lower recall of 0.95. This indicates that RFC is not only accurate overall but also maintains a high level of precision, rarely misclassifying negative instances as positive.

XGBoost, while slightly trailing behind RFC, still demonstrates strong performance with an accuracy of 0.96 and an F1-score of 0.90. Its precision (0.87) and recall (0.93) indicate that it is quite reliable in identifying positive instances, though it experiences a minor drop in precision compared to RFC.

The Support Vector Machine, however, shows a marked difference in its performance metrics. While it achieves a perfect recall score of 1, indicating it identifies all positive instances, its precision is considerably lower at 0.63. This disparity results in the lowest accuracy (0.78) and F1-score (0.77) among the classifiers, suggesting that while SVM excels at detecting positive cases, it does so at the expense of misclassifying a significant number of negative cases as positive.

In conclusion, the Random Forest Classifier stands out as the most effective model for this application, offering a superior blend of accuracy, recall, precision, and F1-score. Its performance suggests that it is the most suitable choice for scenarios where both the identification of positive instances and the avoidance of false positives are critical. XGBoost also presents itself as a robust alternative, especially in contexts where a slight compromise on precision is acceptable. On the other hand, SVM, despite its unmatched recall, might be best reserved for cases where identifying every positive instance is paramount, and the cost of false positives is less consequential. These findings underscore the importance of selecting the right classifier based on the specific requirements and constraints of the task at hand.

## REFERENCES

[1] CDC, A. W. (2020). Centers for disease control and prevention. Available: https://www.cdc.gov/kidneydisease/publicationsresources/2019national-facts.html [Accessed: 1-feb2020].

[2] UCI Machine Learning Repository [Online] https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney _ Disease [Accessed: 24-Sep-2019].

[3 Islam, M. A., Majumder, M. Z. H., & Hussein, M. A. (2023). Chronic kidney disease prediction based on machine learning algorithms. *Journal of pathology informatics*, *14*, 100189.

[4]. S.Ramya, Dr. N.Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative Research in Computer

and Communication Engineering,Vol. 4, Issue 1, January 2016.

[5]. S.Dilli Arasu and Dr. R.Thirumalaiselvi, "Review of Chronic Kidney Disease based on Data Mining IJISRT (International Journal of Innovative Science and Research Technology)

[6] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

[7]. Denil, M., Matheson, D., & De Freitas, N. (2014, January). Narrowing the gap: Random forests in theory and in practice. In *International conference on machine learning* (pp. 665-673). PMLR.

[8]. Himanshu Sharma,M A Rizvi,"Prediction of Heart Disease using Machine Learning Algorithms: A Survey",International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169,Volume: 5 Issue: 8

[9] Baisakhi Chakraborty, "Development of Chronic Kidney Disease Prediction Using Machine Learning", International Conference on Intelligent Data Communication Technologies, 2019.

[10] J. Snegha, "Chronic Kidney Disease Prediction using Data Mining", International Conference on Emerging Trends, 2020