

Proportional Performance Evaluation of Naïve Bayes, Decision Tree, and Random Forest Algorithms for Chronic Kidney Disease Prediction

Dr. Rajesh S Walse¹

¹Assistant Professor (Computer Science),
Department of Dairy Business Management,
College of Dairy Technology, Warud (Pusad),
M.A.F.S.U, Nagpur, Maharashtra State, India
rajeshwalse@gmail.com

Abstract

Chronic Kidney Disease (CKD) is a progressive medical disorder that poses significant challenges to healthcare systems worldwide. Early diagnosis and timely intervention are essential to reduce disease progression and improve patient outcomes. Machine learning techniques have demonstrated considerable potential in developing intelligent prediction systems capable of assisting clinicians in disease diagnosis and management. This study presents a comparative evaluation of three widely used classification algorithms, namely Naïve Bayes, Decision Tree (J48), and Random Forest, for the prediction of chronic kidney disease. Experimental analysis was conducted using the Chronic Kidney Disease dataset obtained from the UCI Machine Learning Repository, consisting of 400 patient records and 25 clinical attributes. Data preprocessing and classification experiments were performed using the WEKA data mining platform. The performance of the classifiers was assessed using a 70% training and 30% testing strategy. Experimental results revealed classification accuracies of 97.50%, 98.33%, and 100% for Naïve Bayes, Decision Tree (J48), and Random Forest algorithms, respectively. Comparative analysis based on confusion matrices, kappa statistics, error measures, and receiver operating characteristic (ROC) curves demonstrated the superior performance of the Random Forest classifier. The findings indicate that ensemble learning approaches provide enhanced predictive capability and can effectively support early diagnosis of chronic kidney disease. The proposed framework may assist healthcare professionals in clinical decision-making and contribute to the development of intelligent disease prediction systems.

Keywords— Chronic Kidney Disease, Machine Learning, Naïve Bayes, Decision Tree, Random Forest, Classification, WEKA, Healthcare Analytics.

I. Introduction

Chronic Kidney Disease (CKD) is one of the most prevalent chronic illnesses affecting millions of individuals worldwide. The gradual deterioration of kidney function can result in severe complications, including end-stage renal disease and cardiovascular disorders. Owing to increasing incidences of diabetes, hypertension, unhealthy lifestyles, and aging populations, the burden of chronic kidney disease has risen considerably over recent decades. Consequently, early diagnosis and effective disease management have become critical requirements in modern healthcare systems.

The rapid advancement of machine learning and data mining techniques has provided powerful tools for extracting valuable knowledge from medical datasets. These intelligent approaches enable the identification of hidden patterns and relationships among clinical parameters, thereby supporting healthcare professionals in making accurate diagnostic decisions. The application of computational intelligence in healthcare has attracted significant attention because of its ability to improve prediction accuracy and reduce the dependence on costly laboratory investigations.

Among the various classification techniques available in machine learning, Naïve Bayes, Decision Tree, and Random Forest algorithms are widely employed due to their simplicity, interpretability, and

predictive capabilities. Naïve Bayes utilizes probabilistic principles to perform classification tasks, whereas Decision Tree algorithms provide rule-based structures that facilitate understanding of decision-making processes. Random Forest, an ensemble learning technique, combines multiple decision trees to enhance prediction accuracy and minimize over fitting. Several studies have explored machine learning approaches for chronic kidney disease prediction. However, comparative investigations involving probabilistic, tree-based, and ensemble classifiers remain relatively limited. Furthermore, identifying the most suitable algorithm for CKD diagnosis is essential for developing reliable clinical decision-support systems.

In this study, a comprehensive comparative analysis of Naïve Bayes, Decision Tree (J48), and Random Forest algorithms is presented using the Chronic Kidney Disease dataset obtained from the UCI Machine Learning Repository. Performance evaluation is conducted using various statistical measures, including classification accuracy, confusion matrices, kappa statistics, error measures, and receiver operating characteristic curves. The objective of this research is to determine the most effective classification model for chronic kidney disease prediction and to demonstrate the applicability of machine learning techniques in healthcare analytics.

The major contributions of this work are summarized as follows:

1. Development of a machine learning framework for chronic kidney disease prediction.
2. Comparative evaluation of Naïve Bayes, Decision Tree (J48), and Random Forest algorithms.
3. Assessment of classification performance using multiple statistical indicators.
4. Identification of the most effective predictive model for CKD diagnosis.
5. Demonstration of the potential of intelligent systems in supporting healthcare decision-making.

The remainder of this paper is organized as follows. Section II presents a review of related studies. Section III describes the dataset and methodology employed in the study. Section IV

discusses the experimental results and comparative performance analysis. Finally, Section V summarizes the conclusions and outlines future research directions.

II. Related Work

The integration of data mining and machine learning techniques into healthcare applications has significantly improved disease diagnosis and prediction capabilities. Researchers have employed various computational approaches to extract hidden patterns from medical datasets and develop intelligent decision-support systems for healthcare management.

Bharara et al. [1] presented a review of knowledge extraction techniques for business operations using data mining methodologies. Their study emphasized the role of intelligent algorithms in analyzing large datasets and supporting decision-making processes. Similarly, Ariff et al. [2] proposed an RFID-based health management framework for livestock monitoring, demonstrating the effectiveness of automated systems in health-related applications.

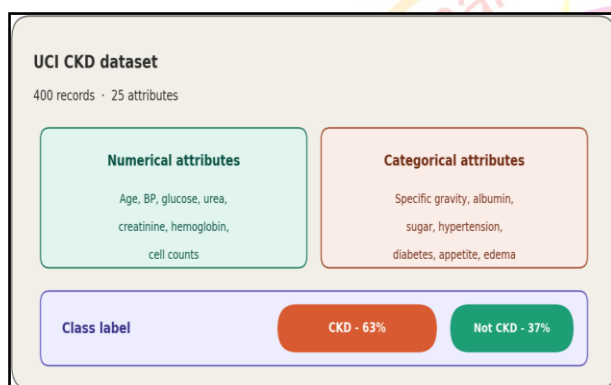
Chen et al. [3] introduced a fast clustering algorithm for determining cluster centers efficiently. Their work improved clustering performance by reducing computational complexity and enhancing cluster formation. Chuan et al. [5] applied clustering techniques to investigate disease progression among peritoneal dialysis patients and reported that clustering approaches are capable of discovering meaningful relationships in clinical data.

Arasu and Thirumalaiselvi [4] developed an imputation-based framework for predicting kidney-related disorders using data mining methods. Their study highlighted the importance of preprocessing techniques in improving the quality of predictive models. Likewise, Kunwar et al. [6] investigated several classification algorithms for chronic kidney disease diagnosis and demonstrated the applicability of machine learning techniques in medical data analysis.

Luong [7] employed K-Means clustering to analyze disease progression patterns and illustrated the usefulness of clustering algorithms for understanding disease characteristics. Güllüoğlu [8] utilized data mining methodologies for customer segmentation and showed that classification and

clustering techniques can effectively identify hidden patterns within large datasets.

Kumar and Khatri [9] implemented the WEKA environment for disease prediction and reported that machine learning tools provide efficient mechanisms for medical data analysis. Khanna [10] investigated the economic burden associated with dialysis treatment in India and emphasized the importance of early diagnosis to minimize healthcare expenditure. Similarly, Varma [11] discussed the increasing prevalence of chronic kidney disease in India and highlighted the necessity for developing intelligent predictive frameworks.



Uboltham et al. [12] proposed an acute kidney injury diagnostic system based on KDIGO guidelines and demonstrated that computational intelligence techniques can significantly improve clinical decision-making processes.

Although previous studies have extensively investigated various machine learning approaches for disease prediction, comparatively limited research has focused on performing a comprehensive comparative evaluation of probabilistic, tree-based, and ensemble learning techniques for chronic kidney disease diagnosis. Moreover, the relative performance of Naïve Bayes, Decision Tree (J48), and Random Forest algorithms requires further investigation to identify the most reliable predictive framework. Therefore, the present study aims to address this research gap by conducting a detailed comparative analysis of these three classification techniques using the Chronic Kidney Disease dataset. Performance assessment is carried out using accuracy measures, confusion matrices, error statistics, kappa coefficients, and receiver operating characteristic curves to determine the most effective algorithm for CKD prediction.

III. Materials And Methods

A. Dataset Description:

The experimental investigation conducted in this study utilized the Chronic Kidney Disease (CKD) dataset obtained from the UCI Machine Learning Repository. The dataset consists of 400 patient records described by 25 clinical attributes associated with kidney function and related physiological conditions. These attributes include both numerical and categorical variables, providing comprehensive information required for disease prediction. The numerical parameters include age, blood pressure, blood glucose random level, blood urea, serum creatinine, sodium concentration, potassium concentration, hemoglobin level, packed cell volume, white blood cell count, and red blood cell count. Categorical attributes comprise specific gravity, albumin, sugar, red blood cell condition, pus cell condition, bacteria, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia, and the corresponding class label indicating CKD or non-CKD status.

Among the total observations, approximately 63% correspond to chronic kidney disease cases, while the remaining 37% represent healthy individuals. The dataset provides sufficient diversity and complexity for evaluating machine learning algorithms used for disease prediction.

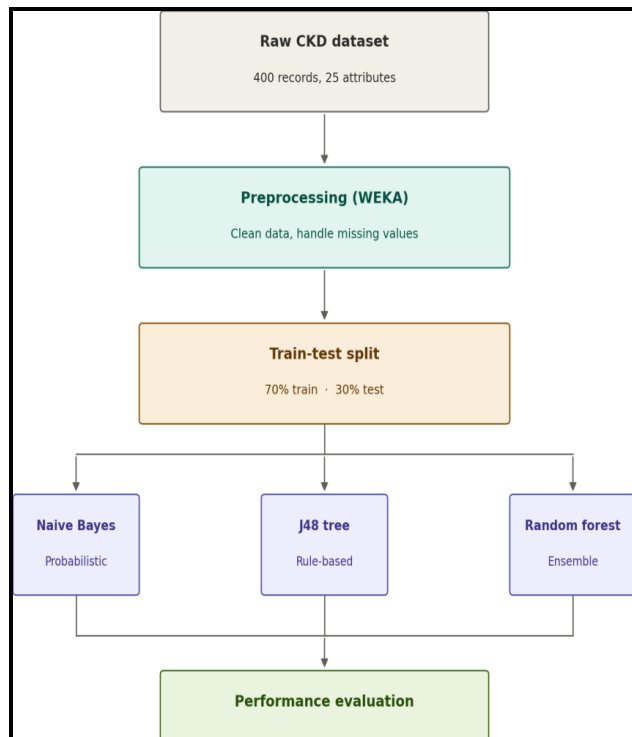
B. Data Preprocessing

Data preprocessing plays a vital role in improving the quality and reliability of predictive models. Prior to classification, the dataset was subjected to cleaning and transformation procedures to eliminate inconsistencies and manage missing values. The preprocessing operations were performed using the WEKA data mining platform.

Irrelevant and incomplete records were appropriately handled to ensure data consistency. Numerical and categorical variables were arranged in a suitable format for classifier training and testing. The resulting processed dataset was subsequently utilized for model development and performance evaluation.

C. Naïve Bayes Classifier

Experimental Methodology Pipe Line



The full pipeline: raw data is cleaned in WEKA, split 70/30 into train and test sets, then fed in parallel to three classifiers (Naive Bayes, J48, Random Forest) before a shared performance evaluation stage.

Naïve Bayes is a probabilistic classification algorithm based on Bayes' theorem. The algorithm assumes that all predictor variables are conditionally independent with respect to the target class. Despite this simplifying assumption, Naïve Bayes has demonstrated remarkable effectiveness in various classification problems because of its computational simplicity and efficiency.

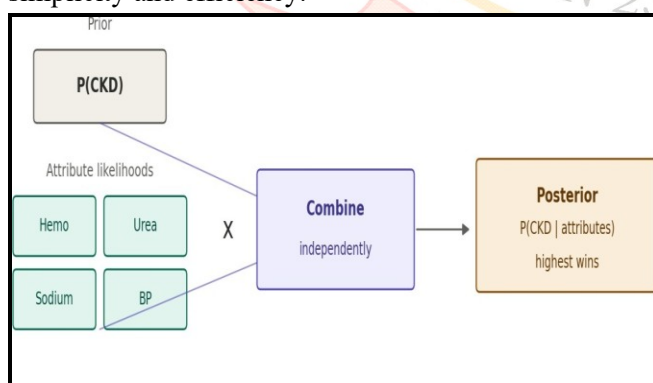


Fig : Naïve Bayes combines a prior probability $P(\text{CKD})$ with each attribute's likelihood independently (the “naive” conditional-independence assumption), producing a posterior probability ; the class with the high estposterioris selected.

conditional-independence assumption), producing a posterior probability ; the class with the high estposterioris selected

The probability of a class variable given a set of predictor attributes is calculated using Bayes' theorem as follows:

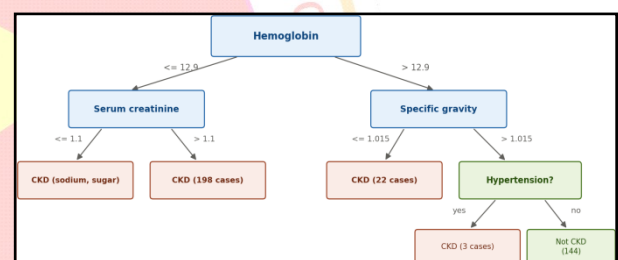
$$[P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}]$$

where:

- $P(C|X)$ denotes the posterior probability of class (C),
- $P(X|C)$ represents the likelihood of predictor variables,
- $P(C)$ is the prior probability of the class,
- $P(X)$ denotes the probability of observing the predictor variables.

The classifier assigns an instance to the class possessing the highest posterior probability.

D. Decision Tree (J48) Algorithm

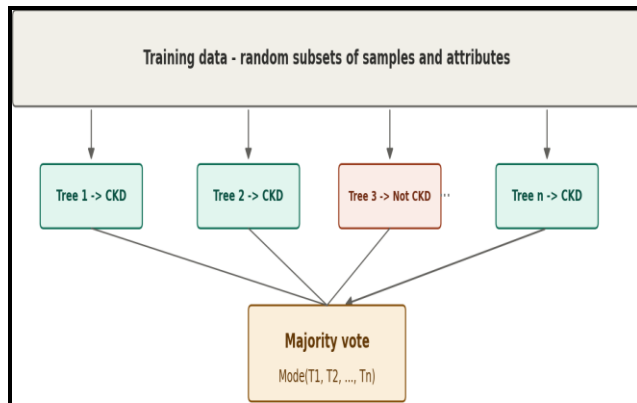


Decision Tree (J48) is a supervised learning algorithm used for constructing hierarchical tree structures based on attribute values. The algorithm recursively partitions the dataset into homogeneous subsets by selecting the most informative attributes. Internal nodes represent decision criteria, whereas leaf nodes correspond to class labels. J48 employs information gain as the splitting criterion and generates interpretable classification rules. The tree structure facilitates understanding of relationships among clinical variables and enables visualization of the decision-making process. Because of its simplicity and transparency, J48 is widely employed in healthcare applications.

E. Random Forest Classifier

Random Forest is an ensemble learning technique that combines multiple decision trees to improve classification accuracy and reduce overfitting. Each tree is constructed using randomly selected subsets of training samples and predictor variables. The final prediction is obtained through

majority voting among individual trees. The ensemble approach enhances model robustness and minimizes variance associated with a single decision tree. Random Forest has gained significant popularity because of its capability to handle high-dimensional datasets and its resistance to noise and missing values.



The classification output is determined according to $\hat{Y} = \text{Mode}(T_1, T_2, \dots, T_n)$

where:

- $(T_1, T_2, \dots, T_n)$ denote the individual decision trees,
- $(\hat{Y})$ represents the final predicted class obtained by majority voting.

F. Experimental Configuration

The classification experiments were performed using the WEKA software environment. Three machine learning algorithms, namely Naïve Bayes, Decision Tree (J48), and Random Forest, were implemented for comparative analysis. A train-test split strategy was adopted, where 70% of the data was used for model training and the remaining 30% was utilized for testing purposes. Performance evaluation was conducted using several statistical indicators, including:

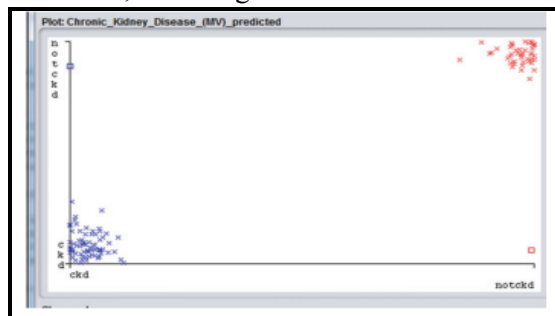


Fig: Decision Tree J48 algorithm to visualize classifier error

1. Classification Accuracy
2. Confusion Matrix
3. Kappa Statistics
4. Mean Absolute Error (MAE)
5. Root Mean Squared Error (RMSE)
6. Relative Absolute Error
7. Root Relative Squared Error
8. Receiver Operating Characteristic (ROC) Area
9. Precision
10. Recall
11. F-Measure

The comparative analysis of these performance measures enables identification of the most effective classification algorithm for chronic kidney disease prediction. Furthermore, the study investigates the ability of probabilistic, tree-based, and ensemble learning approaches to provide reliable support for clinical decision-making.

IV. Results And Performance Analysis

A. Experimental Evaluation:

The performance of the proposed classification framework was evaluated using the WEKA data mining platform. Three machine learning algorithms, namely Naïve Bayes, Decision Tree (J48), and Random Forest, were implemented and tested on the Chronic Kidney Disease dataset containing 400 instances and 25 attributes. A 70% training and 30% testing strategy was adopted to assess the predictive capabilities of the classifiers. The effectiveness of the models was investigated using several performance indicators, including classification accuracy, kappa statistics, error measures, precision, recall, F-measure, and receiver operating characteristic (ROC) curves. These measures provide a comprehensive assessment of the classification models and facilitate comparison among the algorithms.

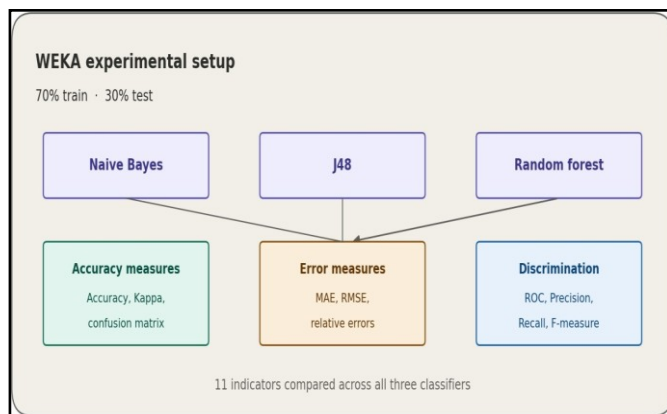


Fig: Accuracy measures (accuracy, Kappa, confusion matrix), error measures (MAE, RMSE, relative errors), and discrimination measures (ROC area, precision, recall, F-measure).

B. Performance of Naïve Bayes Classifier

The Naïve Bayes classifier demonstrated satisfactory predictive capability for chronic kidney disease diagnosis. The algorithm correctly classified 117 out of 120 testing instances, resulting in an overall classification accuracy of 97.50%. Only three instances were incorrectly classified, corresponding to an error rate of 2.50%. The classifier exhibited high precision and recall values for both CKD and non-CKD classes, indicating its effectiveness in identifying disease patterns. Furthermore, the receiver operating characteristic curve showed an area close to unity, confirming the reliability of the probabilistic classification approach. The results indicate that Naïve Bayes provides an efficient and computationally inexpensive framework for chronic kidney disease prediction.

C. Performance of Decision Tree (J48)

The Decision Tree (J48) algorithm achieved improved classification performance compared with the Naïve Bayes classifier. The model correctly identified 118 out of 120 observations, corresponding to an accuracy of 98.33%, while only two instances were misclassified. The tree-based structure generated interpretable decision rules based on clinical parameters such as hemoglobin level, serum creatinine concentration, sodium level, and sugar content. These decision rules facilitate understanding of the relationships among predictor variables and provide valuable insights into disease diagnosis. The confusion matrix analysis revealed balanced classification performance across both disease

categories. Moreover, the ROC curve and kappa statistic indicated a high level of agreement between predicted and actual class labels.

D. Performance of Random Forest Classifier

Among the investigated algorithms, the Random Forest classifier produced the best performance. The model correctly classified all 120 testing instances and achieved a classification accuracy of 100%, with zero misclassified observations. The ensemble learning mechanism employed by Random Forest contributed to improved robustness and minimized classification errors. The kappa statistic reached its maximum value of one, indicating perfect agreement between predicted and actual class labels. Similarly, the ROC area achieved a value of one, reflecting excellent discrimination capability. The obtained results demonstrate the superiority of ensemble learning techniques for chronic kidney disease prediction and highlight the effectiveness of Random Forest in handling complex medical datasets.

E. Comparative Performance Analysis

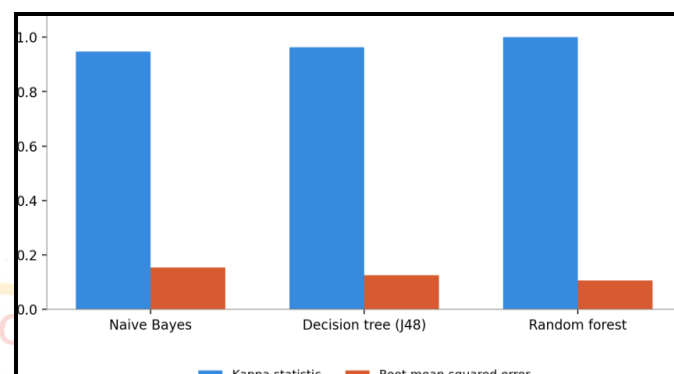
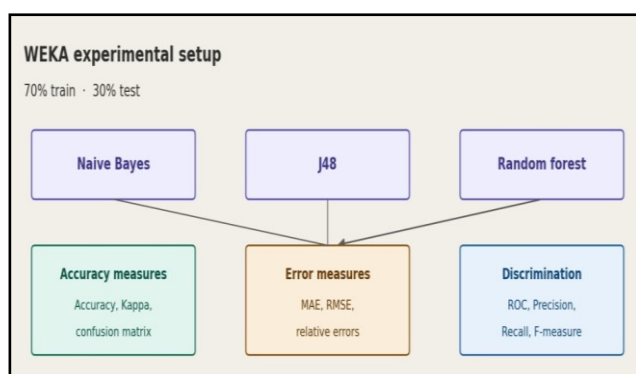
A comparative analysis of the three classification algorithms revealed notable differences in predictive performance. Naïve Bayes achieved an accuracy of 97.50%, whereas Decision Tree (J48) improved the accuracy to 98.33%. Random Forest outperformed both approaches by achieving 100% classification accuracy. In addition to accuracy, Random Forest produced the highest values of precision, recall, and F-measure while maintaining the lowest error rates. The ensemble classifier also exhibited the strongest agreement according to kappa statistics and provided the largest area under the ROC curve. These findings suggest that Random Forest offers superior predictive capability and greater reliability than probabilistic and single-tree classifiers. The ensemble strategy effectively reduces overfitting and enhances model generalization, thereby improving classification performance.

F. Confusion Matrix and ROC Curve Analysis

Confusion matrix analysis provided additional insights into the predictive behavior of the classification models. Naïve Bayes and Decision Tree (J48) exhibited a small number of misclassifications, whereas Random Forest correctly identified all CKD and non-CKD instances without error. Receiver operating characteristic analysis further confirmed

the effectiveness of the classifiers. The ROC area values obtained for all models approached unity, indicating excellent discrimination between CKD and non-CKD classes. However, Random Forest

efficient decision-support system for assisting healthcare professionals in early disease detection and treatment planning.



Achieved the highest ROC performance and demonstrated the most reliable prediction capability.

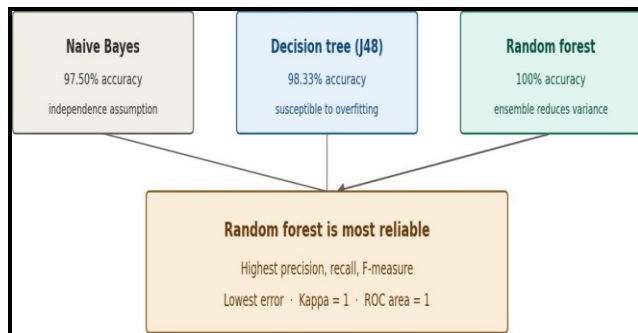
Naive Bayes	Decision tree (J48)	Random forest																											
<table> <tr> <th>Predicted</th><th>CKD</th><th>Not CKD</th></tr> <tr> <th>CKD</th> <td>73 true CKD</td><td>3 missed CKD</td></tr> <tr> <th>Not CKD</th> <td>0 false alarm</td><td>44 true not CKD</td></tr> </table>	Predicted	CKD	Not CKD	CKD	73 true CKD	3 missed CKD	Not CKD	0 false alarm	44 true not CKD	<table> <tr> <th>Predicted</th><th>CKD</th><th>Not CKD</th></tr> <tr> <th>CKD</th> <td>75 true CKD</td><td>1 missed CKD</td></tr> <tr> <th>Not CKD</th> <td>1 false alarm</td><td>43 true not CKD</td></tr> </table>	Predicted	CKD	Not CKD	CKD	75 true CKD	1 missed CKD	Not CKD	1 false alarm	43 true not CKD	<table> <tr> <th>Predicted</th><th>CKD</th><th>Not CKD</th></tr> <tr> <th>CKD</th> <td>76 true CKD</td><td>0 missed CKD</td></tr> <tr> <th>Not CKD</th> <td>0 false alarm</td><td>44 true not CKD</td></tr> </table>	Predicted	CKD	Not CKD	CKD	76 true CKD	0 missed CKD	Not CKD	0 false alarm	44 true not CKD
Predicted	CKD	Not CKD																											
CKD	73 true CKD	3 missed CKD																											
Not CKD	0 false alarm	44 true not CKD																											
Predicted	CKD	Not CKD																											
CKD	75 true CKD	1 missed CKD																											
Not CKD	1 false alarm	43 true not CKD																											
Predicted	CKD	Not CKD																											
CKD	76 true CKD	0 missed CKD																											
Not CKD	0 false alarm	44 true not CKD																											
3 misclassified (2.50%)	2 misclassified (1.66%)	0 misclassified (0%)																											

G. Discussion of Findings

The experimental investigation demonstrates that machine learning algorithms can effectively support chronic kidney disease diagnosis using clinical attributes. Although all three classifiers provided satisfactory performance, Random Forest consistently produced the most accurate and robust predictions. The superior performance of Random Forest can be attributed to its ensemble learning mechanism, which combines the outputs of multiple decision trees and minimizes classification variance. In contrast, Naïve Bayes relies on independence assumptions that may not always hold for medical datasets, while Decision Tree models are susceptible to over fitting. Overall, the findings indicate that Random Forest represents the most suitable classification approach for chronic kidney disease prediction among the algorithms considered in this study. The proposed framework may serve as an

V. Conclusion And Future Scope

This study presented a comparative analysis of three machine learning algorithms, namely Naïve Bayes, Decision Tree (J48), and Random Forest, for the prediction of chronic kidney disease using the Chronic Kidney Disease dataset obtained from the UCI Machine Learning Repository. Experimental analysis was conducted using the WEKA platform with a 70% training and 30% testing strategy. The performance of the classification models was evaluated through several statistical measures, including classification accuracy, confusion matrices, kappa statistics, error measures, precision, recall, F-measure, and receiver operating characteristic curves. The experimental findings demonstrated that all three classification techniques were capable of identifying chronic kidney disease with high accuracy. The Naïve Bayes classifier achieved an accuracy of 97.50%, whereas the Decision Tree (J48) algorithm improved the classification performance to 98.33%. Among the investigated models, the Random Forest classifier exhibited superior predictive capability and achieved a classification accuracy of 100%, with no misclassified instances. Furthermore, the Random Forest model provided the highest values of precision, recall, F-measure, and ROC area, indicating excellent discrimination between CKD and non-CKD classes.



The comparative investigation revealed that ensemble learning approaches offer greater robustness and predictive reliability than individual classifiers. The results obtained in this study indicate that Random Forest is a highly effective technique for chronic kidney disease diagnosis and can be utilized as a reliable decision-support tool in healthcare applications. The proposed framework has the potential to assist medical practitioners in early disease detection, treatment planning, and the prevention of further kidney-related complications. Despite the promising results obtained in this work, certain limitations remain. The study was conducted using a relatively small dataset consisting of 400 patient records, which may limit the generalization capability of the developed models. In addition, the analysis focused only on three classification algorithms and did not consider advanced ensemble and deep learning approaches. Future research may involve the investigation of additional machine learning techniques such as Support Vector Machines, Extreme Gradient Boosting, AdaBoost, Artificial Neural Networks, and Deep Learning architectures. Feature selection and dimensionality reduction methods may also be incorporated to enhance computational efficiency and predictive performance. Furthermore, larger and more diverse clinical datasets can be utilized to improve model generalization and robustness. The integration of machine learning models with real-time clinical decision-support systems and Internet of Medical Things (IoMT) platforms represents another promising direction for future healthcare applications.

In summary, the results of this study demonstrate that machine learning techniques provide an effective framework for chronic kidney disease prediction. Among the evaluated algorithms, Random Forest emerged as the most reliable and

accurate model, making it a suitable candidate for the development of intelligent healthcare systems aimed at supporting early diagnosis and improving patient outcomes.

References

1. S. Bharara, A. Sai Sabitha, and A. Bansal, "A review on knowledge extraction for business operations using data mining," in *Proc. 7th Int. Conf. Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, Jan. 2017, pp. 512–518.
2. M. H. Ariff, I. Ismarani, and N. Shamsuddin, "RFID-based systematic livestock health management system," in *Proc. IEEE Conf. Systems, Process and Control (ICSPC)*, Kuala Lumpur, Malaysia, Dec. 2014, pp. 111–116.
3. J. Chen et al., "A novel cluster center fast determination clustering algorithm," *Applied Soft Computing*, vol. 57, pp. 539–555, Aug. 2017.
4. S. D. Arasu and R. Thirumalaiselvi, "A novel imputation method for effective prediction of coronary kidney disease," in *Proc. 2nd Int. Conf. Computing and Communication Technologies (ICCCT)*, Chennai, India, Feb. 2017, pp. 127–136.
5. Z. Chuan, T. Ying, B. Li, Z. Yuqun, and L. Fuhua, "Application of clustering analysis to explore syndrome evolution law of peritoneal dialysis patients," in *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine*, Shanghai, China, Dec. 2013, pp. 23–26.
6. V. Kunwar et al., "Chronic kidney disease analysis using data mining classification techniques," in *Proc. 6th Int. Conf. Cloud System and Big Data Engineering (Confluence)*, Noida, India, Jan. 2016.
7. T. A. Duc Luong, "A K-Means approach to clustering disease progressions," Dept. Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA, Tech. Rep., 2017.
8. S. S. Güllüoğlu, "Segmenting customers with data mining techniques," in *Proc. 3rd Int. Conf. Digital Information, Networking and Wireless Communications (DINWC)*, Moscow, Russia, Feb. 2015, pp. 154–159.

9. N. Kumar and S. Khatri, "Implementing WEKA for medical data classification and early disease prediction," in *Proc. 3rd IEEE Int. Conf. Computational Intelligence and Communication Technology (CICT)*, Ghaziabad, India, Feb. 2017.
10. U. Khanna, "The economics of dialysis in India," *Indian Journal of Nephrology*, vol. 19, no. 1, pp. 1–4, Jan. 2009.
11. P. P. Varma, "Prevalence of chronic kidney disease in India: Where are we heading?," *Indian Journal of Nephrology*, vol. 25, no. 3, pp. 133–135, May 2015.
12. I. Uboltham, N. Prompoon, and W. Pan-Ngum, "AKIHelper: Acute kidney injury diagnostic tool using KDIGO guideline approach," in *Proc. IEEE/ACIS 15th Int. Conf. Computer and Information Science (ICIS)*, Okayama, Japan, Jun. 2016, pp. 1–6.

