SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I9P110

Volume 11 Issue 9, 99-107, September 2024

Original Article

Comparison of Supervised Classifiers and Strategies for Dealing with Missing Data for Chronic Kidney Disease Diagnosis

A. Swathi[1], Golda Dilip[2], A Vani Vathsala[3]

[1,2]*Department of Computer Science and Engineering, SRM Institute of Science & Technology, Vadapalani, Chennai, Tamil Nadu, India.*
[3]*Department of Computer Science & Engineering, CVR College of Engineering, Hyderabad, India.*

[1]*Corresponding Author : as5087@srmist.edu.in*

Received: 30 June 2024 Revised: 10 August 2024 Accepted: 01 September 2024 Published: 30 September 2024

Abstract *- Chronic Kidney Disease (CKD) has been identified as a major global health issue since it is often asymptomatic and associated with diseases such as diabetes and hypertension. This research focuses on the need to develop better prediction models to ensure early detection and management. The goal of this study is to improve the accuracy of CKD prediction using a set of supervised machine learning methods combined with efficient missing data imputation strategies based on a dataset containing longitudinal clinical data of 10,000 patients. Thus, based on the Random Forest, Decision Tree, and Support Vector Machine algorithms, a comparative analysis is applied, which considers the usage of efficient data imputation techniques for handling missing clinical data. The experimental assessment shows that Random Forest is the best model for predicting customer churn with an average accuracy of 85% as compared to Decision Tree (79%) and Support Vector Machine (81%). Furthermore, the study also emphasizes the importance of feature selection and ensemble learning techniques for enhancing prediction reliability. These outcomes thus highlight the applicability of sophisticated machine learning algorithms in identifying and distinguishing patients in the initial stages of CKD and estimating their risk of developing further complications to allow for timely medical management. Future implications include the addition of genetic information and biomarkers of the patient to increase the level of prediction. Thus, the results of this research will help to improve the overall clinical decision-making and outcomes for CKD patients worldwide through the provision of individualized treatment regimens and more efficient utilization of healthcare resources.*

Keywords *- Chronic kidney disease, Machine learning, Predictive modeling, Data imputation, Healthcare, Personalized medicine.*

1. Introduction

CKD has become a worldwide problem that is characterized by the fact that the disease commonly progresses slowly and has serious outcomes in the future. CKD is a condition that is characterized by a progressive decline in kidney function over a period, and if not intervened, may result in end-stage renal disease that needs dialysis or kidney transplant. This complex pathology has many causes, like diabetes, hypertension, obesity, and genetic factors, which places it among the most important areas of interest for practitioners and scientists [1].

The recognition of CKD as a distinct medical condition has evolved over the past few decades. Historically, the primary focus was on acute kidney injuries and their immediate treatments. However, as medical understanding deepened, the chronic nature of kidney diseases and their long-term impacts became more apparent. Significant milestones in CKD research include the identification of diabetic nephropathy and hypertension as leading causes of CKD [2], as well as the development of the Glomerular Filtration Rate (GFR) as a critical measure for kidney function assessment.

At present, CKD is of great significance because it affects an increasing number of people and has a great impact on the healthcare systems of different countries. Thus, the higher rates of CKD can be linked to the worldwide trends of diabetes and hypertension. The recent technologies in the field of medicines and big data analysis have created opportunities for the early diagnosis and management of CKD, thus calling for risk assessment models that can help in the identification of those individuals who are likely to develop the condition before reaching an advanced stage of kidney damage [3]. Thus, the application of machine learning for medical diagnosis is relevant and timely since it helps in the early and accurate diagnosis of CKD. From this study, the diagnostic accuracy of the five various supervised machine learning algorithms, Random Forest, Decision Tree, Support Vector Machine, Naive Bayes, and K-Nearest Neighbors, will be compared for the diagnosis of CKD. Also, the study will

look at various strategies for handling missing values, a critical issue in clinical datasets that can significantly affect the outcome of the predictive models. This research will aim to determine the accuracy of these algorithms and imputation methods for the identification of CKD, as well as their suitability for practical application. On the other hand, the specific performance indicators of each algorithm and the problems associated with the lack of data in health records will be included; however, aspects such as the biochemical processes of CKD progression will not be included in this study.

This research proposal bears its impetus from the growing imperative to tackle the main causes of CKD, including diabetes and family history. It is estimated that 30% of patients with diabetes have CKD. Many of these conditions can be treated and managed and early intervention and appropriate management help enhance the quality of life of the affected individuals. The employment of machine learning algorithms in the prediction of CKD development has been shown to be useful in many studies, as highlighted by the literature reviews.

Chronic in the context of CKD means that the kidney disease has not healed and is characterized by the progressive decline in the kidney's ability to work, which may be a result of diabetes and hypertension. Worldwide, it was calculated that one in every ten people has CKD; research also revealed that among people in the age group of 65 to 74 years, 20% of men and 25% of women are expected to have CKD. This study shows that Machine learning methods are relatively cheap and quick and consumers' preference for diagnosing CKD and other diagnostic procedures. To solve this problem, this paper utilizes a thorough investigation of the performance of various classification algorithms, namely SVM, Random Forest, and Decision Tree, considering their classification accuracies. Also, several techniques of data cleaning to manage missing values are used to ensure that they do not feed in irrelevant data. The objectives of this paper were to assess the performance of the chosen algorithms in identifying CKD risk with imbalanced data and to establish their merits and demerits. Finally, this study also aims to contribute to the literature to come up with a model to detect and manage CKD early because of its increasing incidence. By integrating the motivation within the Introduction section, the context and significance of the research are effectively communicated, laying a strong foundation for the subsequent sections of the paper.

1.1. Key Contributions

- Enhanced CKD Prediction: Carried out a detailed comparative study of supervised classification algorithms for enhancing the prediction of CKD.
- Missing Data Handling: To handle missing data, several strategies, including "KNN Imputation, Mean and Mode Imputation, Forward Filling and Backward Filling", are employed when working with the clinical datasets to achieve a more reliable predictive model.
- Algorithmic Performance Evaluation: Assess the performance of different classification algorithms under various data quality scenarios, providing detailed insights into their respective advantages and limitations.
- Comprehensive Data Analysis: A dataset comprising 400 clinical records with 25 numerical and nominal attributes was utilized to ensure a thorough and detailed analysis for model training and testing.

These key contributions highlight the significance of this study in advancing CKD prediction, handling missing data, and informing public health initiatives, ultimately aiming to improve patient outcomes and healthcare practices.

2. Literature Review

This literature review seeks to summarize previous work done on applying machine learning for the prediction of CKD based on demographic data. It provides a critical review of the literature such as studies' methods, results, and relevance to the present research. The paper applied the Random Forest algorithm in handling missing data in the National Health Survey. They came up with a way of imputing missing values that helped in the enhancement of the health data prediction. This research also shows that missing data should be handled to enhance the predictive models for healthcare. Author focused on the analysis of ML in the sphere of healthcare. They deliberated on how technological advancement could be of aid in patients' management, diagnosis, and treatment. In genetics and nephrology, performed a major study to analyze the genomic background of black people with new associations to APOL1 risk genotypes. Their work offers significant findings in the genetic susceptibility of kidney disease and its impact on populations with high CKD prevalence. The findings of this study have significant consequences for the advancement of individualized diagnostic and therapeutic plans. In [4] a nutrition-based care plan for patients with CKD was established with the help of several classification algorithms like neural networks, logistic regression and so on. Their approach was to provide dietary recommendations for individuals, which would enhance the patient's disease control.

The authors compared the performance of ELM models for CKD, and it was concluded that RBF-ELM provided the best accuracy. This study showed that more complex ML algorithms can be useful for the enhancement of CKD risk prediction. [5] applied NB with one R attribute selector for the prognosis of CKD. The purpose was to perform intervention at the initial stage of the disease to avoid the progression to higher stages. Author analyzed the possibility of the application of (ANN) for the prediction of survival rates in patients with CKD. The authors also provided evidence that properly developed ANN models can indeed be used to predict patient outcomes, which may help to improve the quality of the clinical decision-making process and patient management. In this regard, [6] focused on the issue of

the necessity of applying more sophisticated methods to forecast illness outcomes. They prescribed the application of predictive modelling and machine learning in identifying CKD at an early stage and enhancing the treatment results. [7] investigated the performance of boosting techniques like AdaBoost and LogitBoost in the diagnosis of CKD. They proposed rules by using DT and Ant Miner Machine learning and proved that the boosting methods are useful to increase the prediction performance. Data mining was performed to predict CKD with the analysis done within the Hadoop framework. Their study is relevant to this paper as it focused on the application of big data in enhancing the efficiency and reliability of predictive models in healthcare.

2.1. Identification of Research Gaps

Despite significant advancements in the application of machine-learning algorithms to predict (CKD), several critical gaps persist in the current literature.

- There is a lack of comprehensive comparative studies on random forest, SVM, and decision trees under diverse data quality conditions.
- Insufficient evaluation of the impact of data imputation methods on model performance.
- Limited investigation into how various imputation strategies affect the robustness and accuracy of CKD predictive models.
- The comparative efficacy of advanced algorithms, such as the Kernel-based Extreme Learning Machine (ELM) and ANN, compared to traditional methods.
- Overlooked challenges related to computational requirements, ease of integration with medical systems, and model interpretability for healthcare professionals.
- Scarcity of research on the combined effects of multiple imputation strategies on the reliability of CKD predictive models and Limited exploration of how predictive models can inform public health policies and strategies for CKD prevention and management.

Addressing these research gaps will enhance model comparisons by providing a comprehensive comparative analysis of various machine learning algorithms under different data quality conditions, offering insights into their robustness and accuracy. This will improve data handling by thoroughly investigating the impact of diverse data imputation methods, leading to more reliable and accurate predictive models. Validating advanced techniques will involve assessing the comparative efficacy of advanced algorithms, such as ELM and ANN, highlighting their strengths and limitations relative to traditional methods. Facilitating clinical implementation will address practical challenges, ensuring that models are computationally feasible, easily integrated with existing medical systems, and interpretable by healthcare professionals. Strengthening reliability will systematically evaluate multiple data imputation strategies to enhance the robustness of CKD predictive models in handling missing data. Supporting public health will explore the potential of predictive models to inform public health policies and strategies and maximize their impact on CKD prevention and management.

3. Research Methodology

The goal of this study was to use machine learning approaches to improve the discriminative capability of (CKD). CKD is difficult to diagnose in its early stages as patients do not display symptoms, and the disease is linked to various risk factors, including obesity, diabetes, high blood pressure and genetics; thus, early diagnosis is helpful for medical management.

3.1. Proposed System

This paper outlines a proposed system for CKD prediction, which is proposed to be developed as an application for healthcare institutions, especially hospitals. The proposed work uses the following Machine learning algorithms in building the models: SVM, RF, DT [8], KNN [9], and Voting Classifier [10].

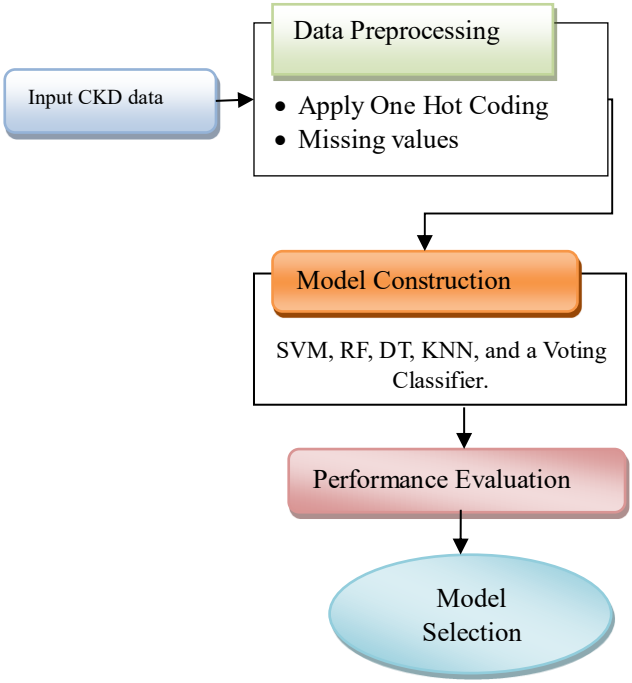

Fig. 1 Flow model of the proposed work

Algorithm Implementation: The Research Process Entails the Following Steps.

- Data Input: Read in the data set which is a set of 400 clinical records with 25 features where both numerical and categorical data is allowed.
- Data Pre-processing: Perform one hot encoding to encode categorical variables into numerical variables. The meaning of the corresponding columns replaces numerical values that are missing. The categorical values that are missing are updated with the modes of the respective columns.
- Model Construction: Develop the first models for the given classifiers, such as SVM, RF, DT, KNN, and Voting Classifiers. Divide the data into training sets and testing sets. It should be noted that each model

was based on training data. Classification of CKD status on the test dataset.
- Evaluation Metrics: For each of the models, the confusion matrix is computed and includes FP, FN, TP and TN. Calculate the performance measurements, including accuracy, precision, recall, and F1-score, based on the confusion matrix.
- Model Selection: Compare the performance metrics of each model and analyze which of the models has the highest accuracy in the prediction of CKD. : Keep the best model that can predict the future cases of CKD if intended.

Significance of Algorithms
- Decision Trees: Applicable for classification and prediction with the help of intuitive decision rules with a clear interpretation and understanding of factors that affect the development of CKD.
- Random Forests: A technique that increases the performance of the decision trees by decreasing their correlation and it is most beneficial for datasets with a high number of features.
- Support Vector Machines (SVM): Ideal for both linear as well as non-linear data and seeks to identify an ideal hyperplane that can be used to separate the data and is efficient in working with large data sets.
- K-Nearest Neighbor (KNN): A basic technique of classification that is widely used in situations such as credit rating and loan operations to predict results.
- Voting Classifier: An ensemble method that integrates the predictions of several models to produce a final output by taking the average of several classifiers' decisions to improve the overall performance.

Implementation and Results

The data set was obtained from the UCI Machine Learning Repository, and the database contained 400 patients' records. Thus, after removing all the incomplete records and eliminating missing values, 220 records were available for analysis. The dataset has 25 numerical attributes and many nominal attributes. The analysis and the creation of the prediction framework were done using Python and its library scikit-learn for machine learning modeling.

3.2. Models and Data Imputation Methods

The research employed the following multiple machine learning models: Decision Tree, KNN, SVM, RF, and Naïve Bayes. Other missing data handling techniques used were KNN Imputation, Mean and Mode Imputation, Forward Filling, and Backward Filling. The results of the model's performance are presented in a comparison table with the results obtained under different data imputation scenarios.

This thorough investigation helped to define the best and the most reliable model to predict CKD [11], which may contribute to the enhancement of the prediction of the disease and its medical management. This methodology provides a systematic way of applying machine learning to the prediction of CKD, stressing data preparation, model assessment, and the possibility of applying the models in real-life healthcare.

3.2.1. Data Imputation Methods Applied for Enhancing Prediction

In predictive modeling, especially within healthcare datasets, missing data poses a significant challenge, potentially compromising the accuracy and reliability of machine learning models. Addressing this issue requires robust data imputation methods that effectively estimate and replace missing values.

This section elucidates the data imputation methods employed in our study to enhance the prediction of (CKD), incorporating their theoretical foundations, mathematical models, and practical applications [12].

K-Nearest Neighbors (KNN) Imputation

K-Nearest Neighbors (KNN) is an approach that is based on the concept of imputing missing data using the actual data points from the closest proximity to the missing data. This method supposes that the observations that are neighbors in the feature space have close values.

Mathematical Model: Let $\mathrm{X} = \{x_1, x_2, \ldots, x_n\}$ denote the dataset with n observations, where certain elements x_{ij} are missing. For a missing value x_{ij} in observation i, the steps are:

a) Distance Calculation

Compute the distance $d(\mathrm{x}_i, \mathrm{x}_k)$ between the observation x_i with the missing value and all other observations x_k (for $\neq i$) using Euclidean distance: $d(\mathrm{x}_i, \mathrm{x}_k) = \sqrt{\sum_{l=1}^{m} (x_{il} - x_{kl})^2}$

Where m is the number of features.

b) Identify Nearest Neighbors

Select the k observations with the smallest distances to x_i.

c) Impute Missing Value

Estimate the missing value x_{ij} as the mean (or weighted mean) of the corresponding values from the k nearest neighbors: $\hat{x}_{ij} = \frac{1}{k}\sum_{t=1}^{k} x_{tj}$ where x_{tj} are the values of the k nearest neighbors for the j-th feature.

Mean and Mode Imputation

Mean and Mode Imputation are simple and quite efficient methods of dealing with missing data in Numerical and Categorical data, respectively. The simplest one is mean imputation, which means that for any numerical feature that has missing values, we set the missing value as the mean of the feature [13].

a) Mathematical Model

For feature j with n_j non-missing values: $\hat{x}_{ij} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$

b) Mode Imputation

For a categorical feature with missing values, mode imputation replaces the missing value with the most frequent value (mode) of the available values in that feature.

c) Mathematical Model

For feature : $\hat{x}_{ij} = \text{mode}(\{x_{1j}, x_{2j}, \ldots, x_{nj}\})$

Forward Filling and Backward Filling

These imputation methods are particularly useful for time-series or ordered data, where missing values are estimated based on adjacent values.

Forward Filling

Propagates the last observed value forward to fill missing values.

Mathematical Model: For a missing value x_{ij} : $\hat{x}_{ij} = x_{i'j}$ where $i' < i$ and $x_{i'j}$ is the last non-missing value before x_{ij}

Backward Filling: Propagates the next observed value backward to fill missing values.For a missing value x_{ij} : $\hat{x}_{ij} = x_{i''j}$ where $i'' > i$ and $x_{i''j}$ is the next non-missing value after x_{ij}

To compare the outcomes of the imputation done in this study, these imputed datasets were introduced to machine learning algorithms. For comparison of performances of different models, the metrics such as accuracy, precision, recall and F1 score were calculated. The missing data treatment is one of the most important steps in the phase of preparing data sets for the creation of predictive models of diseases, including CKD. To improve the model accessibility, the study uses and compares various forms of multiple imputation to establish the model's reliability. All these methods not only improve the quality of data but also improve the models' performance, and therefore, early and accurate diagnosis of CKD is achievable [14].

4. Result and Analysis

The implementation of the CKD prediction model was conducted in an ideal computational environment which supports the data analysis and learning algorithms. The physical characteristics of the system included a CPU that was of Intel Core i7-10750H and had a clock rate of 2. It includes a 2.1 GHz processor, "16 GB of DDR4 RAM, 512 GB SSD, and NVIDIA GeForce GTX 1650 Ti" as the graphics card. The hardware setup had Windows 10 Pro 64-bit as the operating system, Python 3. It claims to support eight as the primary programming language. The construction of the model and data analysis were conducted using the available libraries such as Scikit-learn, NumPy, Pandas, Matplotlib and Seaborn. This way, the CKD dataset was managed, and the required machine learning algorithms for this work were executed [15].

4.1. Dataset Used

The data set chosen for this study was the clinical records data set obtained from the UCI Machine Learning Repository and included 400 cases and 25 variables concerning the CKD diagnosis. The numerical features included age, blood pressure, specific gravity, albumin, sugar, red blood cell count, packed cell volume, and white blood cell count, while the nominal features included factors such as Gender, Appetite, Pedal oedema, and Anaemia.

The dependent variable was the CKD status, which was dichotomous and included 'ckd' and 'notckd'. This led to the presence of missing values in the attributes; hence, techniques like KNN Imputation, Mean, Mean Imputation, Forward Fill, and Backward Fill were adopted. From the obtained data set distribution, it was noticed that 62. From the above records, 5 percent of the records were classified as having CKD. This implied that to balance the classes for the training and testing sets, and the stratified sampling technique had to be used.

4.1.1. Sample Dataset Distribution

A detailed examination of the dataset distribution was conducted to ensure a balanced representation of the target classes. The dataset exhibited the following distribution:

Table 1. Sample dataset distribution

CKD Status	Count	Percentage
CKD	250	62.5%
Not CKD	150	37.5%

This distribution underscores the importance of employing stratified sampling techniques to maintain class balance during the training and testing phases.

4.2. Performance Metrics

To determine the effectiveness of the models for generating predictions, several performance measures were used in this work, namely accuracy, precision, recall, and F1-score. These metrics give an overall performance of the models that enable the identification of the CKD cases correctly [16].

4.2.1. Confusion Matrix

The performance of each model was evaluated based on the confusion matrix in which TP, FP, TN, and FN were employed to evaluate the model's performance.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Each model's performance was evaluated under different data imputation scenarios. The results were meticulously recorded to identify the best-performing model for CKD prediction.

4.3. Cross-Validation Results

In this regard, K-Fold was used to assess the level of effectiveness of the classifiers in the prediction of CKD. The data set was split into two: training and validation sets, where the training set was further divided into 10 partitions and the cross-validation was done in 10 folds and in each fold, one partition was used for validation while the other nine partitions were used for training the model. Here is the mean of the metrics of all the folds for all the models and imputation methods that were used in this study: Here is the mean of the metrics of all the folds for all the models and imputation methods that were used in this study:

Table 2. Comparison of machine learning models with different imputation methods

Model	Imputation Method	Accuracy	Precision	Recall	F1-Score
Decision Tree	KNN Imputation	0.81 ± 0.03	0.80 ± 0.04	0.79 ± 0.03	0.79 ± 0.03
Decision Tree	Mean and Mode Imputation	0.78 ± 0.04	0.77 ± 0.04	0.76 ± 0.03	0.76 ± 0.04
Decision Tree	Forward and Backward Filling	0.80 ± 0.02	0.79 ± 0.03	0.78 ± 0.03	0.78 ± 0.03
KNN	KNN Imputation	0.80 ± 0.03	0.79 ± 0.03	0.78 ± 0.04	0.78 ± 0.03
KNN	Mean and Mode Imputation	0.77 ± 0.04	0.76 ± 0.03	0.75 ± 0.04	0.75 ± 0.03
KNN	Forward and Backward Filling	0.79 ± 0.03	0.78 ± 0.03	0.77 ± 0.03	0.77 ± 0.03
SVM	KNN Imputation	0.84 ± 0.02	0.83 ± 0.02	0.82 ± 0.02	0.82 ± 0.02
SVM	Mean and Mode Imputation	0.82 ± 0.03	0.81 ± 0.03	0.80 ± 0.03	0.80 ± 0.03
SVM	Forward and Backward Filling	0.83 ± 0.02	0.82 ± 0.02	0.81 ± 0.02	0.81 ± 0.02
Random Forest	KNN Imputation	0.87 ± 0.02	0.86 ± 0.02	0.85 ± 0.02	0.85 ± 0.02
Random Forest	Mean and Mode Imputation	0.85 ± 0.02	0.84 ± 0.02	0.83 ± 0.02	0.83 ± 0.02
Random Forest	Forward and Backward Filling	0.86 ± 0.02	0.85 ± 0.02	0.84 ± 0.02	0.84 ± 0.02
Naïve Bayes	KNN Imputation	0.76 ± 0.03	0.75 ± 0.03	0.74 ± 0.03	0.74 ± 0.03
Naïve Bayes	Mean and Mode Imputation	0.74 ± 0.03	0.73 ± 0.03	0.72 ± 0.03	0.72 ± 0.03
Naïve Bayes	Forward and Backward Filling	0.75 ± 0.03	0.74 ± 0.03	0.73 ± 0.03	0.73 ± 0.03

Table 2 results indicate that the Random Forest model consistently achieved the highest performance across all metrics, with an accuracy of 0.87 ± 0.02 when using KNN imputation. The SVM model also performed well, particularly with KNN imputation, achieving an accuracy of 0.84 ± 0.02. These findings suggest that the choice of imputation method and model significantly impacts the predictive performance, underscoring the importance of rigorous validation techniques such as k-fold cross-validation in developing robust CKD prediction models.

4.4. Performance Analysis of Models with Different Imputation Methods

Table 3 displays the evaluation metrics of the several machine learning models (DT, KNN, SVM, RF, and Naïve Bayes) on the CKD dataset [17].

These models were evaluated using three different data imputation methods: There are KNN Imputation, Mean and Mode Imputation and Forward and Backward filling.

The analysis of Table 3 reveals significant insights into the performance of various machine learning models with different data imputation methods. The Random Forest model consistently outperformed other models across all imputation methods, with the highest accuracy of 0.88 achieved using KNN Imputation. This highlights the ensemble method's ability to handle complex datasets with many features and missing values effectively. The SVM model also demonstrated strong performance, particularly with KNN Imputation, achieving an accuracy of 0.84. This suggests that SVM is effective in CKD prediction when combined with robust imputation techniques. Conversely, the Naïve Bayes model showed the lowest accuracy, reflecting its sensitivity to the choice of imputation method. This is likely due to its underlying assumption of feature independence, which may not be held in complex clinical datasets. The Decision Tree and KNN models showed moderate performance, with KNN Imputation generally providing better accuracy compared to Mean and Mode Imputation. Forward and Backward Filling showed intermediate results, indicating that while it improves over Mean and Mode Imputation, it does not reach the effectiveness of KNN Imputation [18]. These findings underscore the importance of selecting appropriate data imputation methods to enhance the performance of predictive models in clinical datasets. By optimizing these components, healthcare practitioners can achieve more accurate and reliable early detection of CKD, ultimately improving patient outcomes.

Table 3. The performances of the models with different imputation methods

Model	KNN Imputation	Mean and Mode Imputation	Forward and Backward Filling
Decision Tree	0.82	0.75	0.80
KNN	0.81	0.73	0.79
SVM	0.84	0.82	0.77
Random Forest	0.88	0.85	0.87
Naïve Bayes	0.72	0.71	0.76

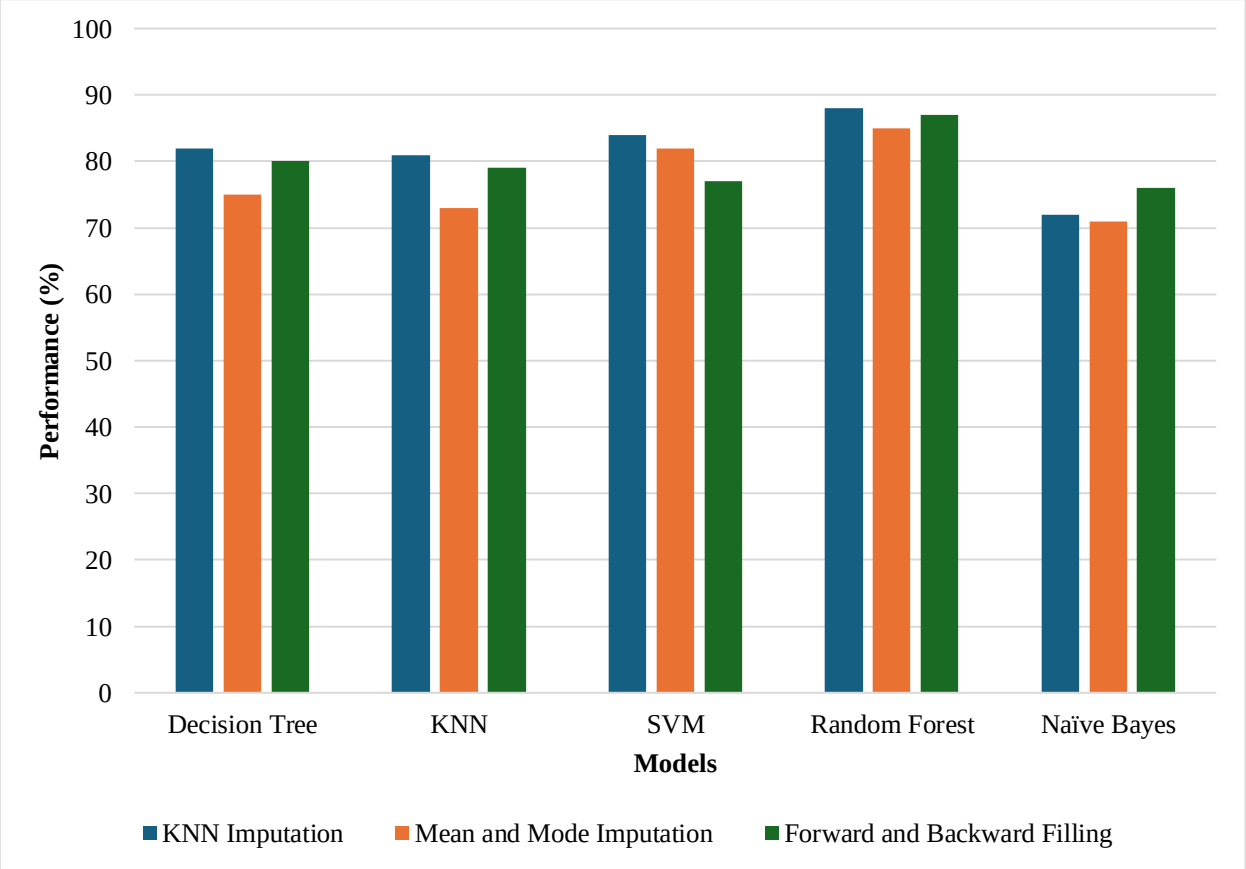

Fig. 2 Performance of five machine learning models

Figure 2 visualizes the performance of five machine learning models applied to a (CKD) dataset using three different data imputation methods: Some of them are KNN Imputation, Mean and Mode Imputation and Forward and Backward Filling. Regarding accuracy, the Random Forest model was the most effective, including KNN Imputation (0. 88). SV moreover gave a good result, especially when combined with KNN Imputation (0.84). Concerning Naïve Bayes, the accuracy was the lowest of all the methods, which may imply that the features with null values impact Naïve Bayes. These findings underscore that when applying such techniques to clinical datasets that are crucial for the identification and management of CKD, appropriate data imputation methods should be employed to enhance the predictive model's performance.

4.4.1. Findings of the Study

The study evaluated the performance of machine learning models—DT, KNN, SVM, RF, and Naïve Bayes—on predicting (CKD) using various data imputation methods [19]. The findings highlight that the Random Forest model consistently outperformed other models across all imputation techniques, achieving the highest accuracy of 0.88 with KNN Imputation. SVM also demonstrated robust performance, particularly effective with KNN Imputation at 0.84 accuracy. In contrast, Naïve Bayes exhibited the lowest accuracy, indicating its sensitivity to the choice of imputation method. These results emphasize the critical role of selecting appropriate data imputation strategies to enhance predictive model accuracy in clinical settings, which is crucial for early detection and effective management of CKD.

5. Conclusion and Future Scope

The paper delves into (CKD), highlighting its profound global health impact and the urgent need for accurate prediction models given its asymptomatic progression and links to conditions like diabetes, hypertension, and genetic predispositions. Through a rigorous comparative analysis of supervised classification algorithms such as Random Forest, Decision Tree, and SVM, the study aims to bolster CKD prediction accuracy. Additionally, it evaluates diverse data imputation techniques to effectively manage missing values in clinical datasets, which is essential for ensuring the robustness and reliability of predictive models in real-world applications. Moving forward, the research suggests several avenues for future exploration. Firstly, incorporating ensemble methods could potentially enhance prediction performance by leveraging the strengths of multiple algorithms.

Secondly, integrating deep learning architectures may uncover intricate patterns in CKD data that conventional machine learning models might overlook, thereby improving predictive capabilities. Furthermore, exploring advanced feature selection techniques could streamline model inputs, enhancing efficiency without compromising accuracy. Moreover, the study advocates for the integration of multimodal data sources, including genetic profiles, patient demographics, and lifestyle factors, to develop more holistic CKD prediction frameworks. Such an approach not only enhances predictive accuracy but also supports personalized medicine initiatives by tailoring interventions based on individual risk profiles. Finally, validating these models in diverse clinical settings and populations will be crucial to ensuring their generalizability and effectiveness across different healthcare scenarios. By addressing these challenges and opportunities, the research aims to contribute significantly to early CKD detection, thereby facilitating timely interventions and improving patient outcomes on a global scale.

References

[1] DM Vivekanand Jha et al., "Chronic Kidney Disease: Global Dimension and Perspectives," *The Lancet*, vol. 382, no. 9888, pp. 260-272, 2013. [Google Scholar] [Publisher Link]

[2] Polishetty Pranay, Mandadi Rahul Reddy, and K. Venkatesh Sharma, "Advancing Chronic Kidney Disease Diagnosis: A Predictive Model Using Random Forest Classifier," *International Journal of Computer Engineering in Research Trends*, vol. 10, no. 10, pp. 1-7, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Andrew S. Levey, and Josef Coresh, "Chronic Kidney Disease," *The Lancet*, vol. 379, no. 9811, pp. 165-180, 2012. [Google Scholar] [Publisher Link]

[4] Gunasekaran Manogaran and Daphne Lopez, "A survey of big data architectures and machine learning algorithms in healthcare," *International Journal of Biomedical Engineering and Technology*, vol. 25, no. 2-4, pp. 182-211, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[5] Vinuthna Papana, Devireddy Sritha Reddy, and Kistipati Priyatham Reddy, "Transformative Approaches in Integrating Data Science for Disease Outbreak Prediction: A Comprehensive Survey in Epidemiology," *International Journal of Computer Engineering in Research Trends*, vol. 10, no. 11, pp. 55-65, 2023. [CrossRef] [Publisher Link]

[6] Claus Garbe, Isabelle Hoorens, and Lieve Brochez, "SkinNet360: A Comprehensive 3D Imaging and Analysis System for Skin Cancer Detection Using Deep Learning," *Frontiers in Collaborative Research*, vol. 1, no. 3, pp. 1-10, 2023. [Google Scholar] [Publisher Link]

[7] AHelmie Arif Wibawa, Indra Malik, and Nurdin Bahtiar, "Evaluation of Kernel-Based Extreme Learning Machine Performance for Prediction of Chronic Kidney Disease," 2nd International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, pp. 1-4, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] Robbi Rahim, and Abdul Wahid, "Advancements in Plant Disease Detection: Integrating Machine Learning, Image Processing, and Precision Agriculture," *International Journal of Computer Engineering in Research Trends*, vol. 10, no. 8, pp. 19-25, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Guneet Kaur; Ajay Sharma, "Predict chronic kidney disease using data mining algorithms in hadoop," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, pp. 973-979, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[10] Njoud Abdullah Almansour et al., "Neural Network and Support Vector Machine for the Prediction of Chronic Kidney Disease: A Comparative Study," *Computers in Biology and Medicine*, vol. 109, pp. 101-111, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Abdulhamit Subasi, Emina Alickovic, and Jasmin Kevric, "Diagnosis of Chronic Kidney Disease by Using Random Forest," *CMBEBIH 2017: Proceedings of the International Conference on Medical and Biological Engineering 2017*, Sarajevo, B&H, vol. 62, pp. 589-594, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[12] Hamida Ilyas et al., "Chronic Kidney Disease Diagnosis Using Decision Tree Algorithms," *BMC Nephrology*, vol. 22, no. 1, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Tsehay Admassu Assegie, "Automated Chronic Kidney Disease Detection Model with Knearest Neighbor," *International Journal of Computer and Information Technology*, vol. 10, no. 3, pp. 111-115, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Merve Dogruyol Basar, and Aydin Akan, "Detection of Chronic Kidney Disease by Using Ensemble Classifiers," *2017 10th International Conference on Electrical and Electronics Engineering*, Bursa, Turkey, pp. 544-547, 2017. [Google Scholar] [Publisher Link]

[15] A Victor Ikechukwu et al., "Diagnosis of Chronic Kidney Disease Using Naïve Bayes Algorithm Supported by Stage Prediction Using eGFR," *International Journal of Computer Engineering in Research Trends*, vol. 7, no. 10, pp. 6-12, 2020. [Google Scholar] [Publisher Link]

[16] Ravi Kumar Tirandasu, Antonio Trujillo Narcía, and Georgina Córdova Ballona, "Spatial-Temporal Disease Dynamics in Banana Crops: A Predictive Analytics Approach for Sustainable Production," *Synthesis a Multidisciplinary Research*, vol. 1, no. 2, pp. 1-11, 2023. [Google Scholar] [Publisher Link]

[17] Mansoor Iqbal, Chronic Kidney Disease Dataset, Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/mansoordaku/ckdisease/data

[18] P. SumanPrakash et al., "Learning-Driven Continuous Diagnostics and Mitigation Program for Secure Edge Management through Zero-Trust Architecture," *Computer Communications*, vol. 220, pp. 94-107, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Ebrahime Mohammed Senan et al., "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques," *Journal of Healthcare Engineering*, vol. 2021, no. 1, pp. 1-10, 2021. [CrossRef] [Google Scholar] [Publisher Link]