Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Leveraging Machine Learning for Chronic Kidney Disease Prognostics

Abdullah Al Mamun[1], Touhid Bhuiyan[2], Aminur Sarker[1], Rakibul hasan[1], Md Maruf Hassan[3]

[1]*Department of CSE, Daffodil International University, Bangladesh*
[2]*School of IT, Washington University of Science and Technology, VA, USA*
[3]*Department of CSE, Southeast University, Bangladesh*

Chronic Kidney Disease (CKD) ranks among the most pressing public health hazards due to the high level of morbidity and mortality correlated with it. Take note therefore, that early and correct forecasting of CKD enhances the timing of the intervention which in turn improves treatment efficacy and also decreases the burden on health care systems. This report surveys the use of machine learning (ML) algorithms for CKD using the CKD dataset that was collected from UCI Machine Learning Repository. Specifically, SVM, DT, RF, and LR model was used for training and analysis of standardizing the dataset using the StandardScaler before applying to K-fold cross-validation (n splits = 5). The dataset was first processed by labelling. Model output was compared using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. In the model, SVM classifier exhibited the top accuracy of 98.48%, while RF achieved 99.24% accuracy after hyperparameter tuning and also 99.24%, and 97.72% for logistic regression, and decision tree respectively. These findings highlight the potential of ML techniques for early detection of CKD, thereby improving treatment strategies and patient prognosis.

Keywords: Chronic Kidney Disease; Machine Learning; support vector machine; decision tree; random forest; logistic regression; Cross-Validation; Hyperparameter Tuning; Kidney Function Prediction; Medical Diagnostics; Early Detection.

1. Introduction

Millions of people worldwide are impacted by chronic kidney disease each year. Over time, this condition causes the kidneys to progressively lose their capacity to function normally. Negative substances can accumulate in the body as kidney function deteriorates, resulting in major complications like heart disease, anemia, and high blood pressure. Symptoms of CKD

typically don't show up until much later, and the disease frequently advances silently. To slow its progression and enhance patient outcomes, early detection is essential. Preliminary detection of "CKD" can be difficult, though, as it necessitates routine kidney function monitoring, which isn't always possible in environments with limited resources. Technological and data science developments have created new avenues for bettering the prognosis and healing of CKD. The goal of this analysis is to predict CKD using clinical and laboratory data by applying machine learning techniques. The goal of this analysis is to create a dependable and effective model for early disease detection by examining the CKD dataset from the UCI ML Repository. This research's findings could enhance patient outcomes and healthcare decision making.

2. Related Work

As you may be aware, machine learning is strongly reliant on classification, and extensive research has been conducted in this field using a range of chronic kidney disease-related medical data sets. The efficacy of these many research efforts varied according to the classifier models used to solve the classification challenge. This is a list of:

The EMPA-KIDNEY Collaborative Group et al.[1][2022], trials assessing medication effects in CKD patients vulnerable to the disease getting worse. When compared to a placebo, the medication decreased the progression of kidney disease or heart disease mortality by 28% in 6,609 participants. (Hazard ratio 0.72; 95% CI 0.64–0.82; P¡0.001) with a median follow-up of 2 years... Benefits were consistent across eGFR ranges and in patients with or without diabetes. Empagliflozin also reduced all-cause hospitalization but showed no significant differences in heart failure hospitalizations, cardiovascular death, or all-cause mortality. Adverse events were similar between groups, highlighting empagliflozin's potential to slow CKD progression and improve outcomes.

Elaine Ku et al.[2][2023]Anemia is a common problem of chronic kidney disease, contributing to significant illness. The 2012 KDIGO suggestions deal with the control of anemia. But since then new facts has emerged approximately the rising treatment. This consists of hypoxia-induced prolyl hydroxylase inhibitors (HIF-PHIs). The 2021 KDIGO Debate Conference reviewed the evidence on HIF-PHIs, discussing their possible position. And the emphasis stays unresolved. Research issues and priorities This report emphasizes the need for similarly studies to refine anemia remedy strategies in continual kidney disease.

Michel Burnier et al.[3][2023]High blood pressure is a leading cause of premature death and is closely linked to chronic kidney disease (CKD), with a dual relationship predisposing to both conditions. Effective blood pressure (BP) management in CKD plays an important role in reducing renal and cardiovascular risk. Early diagnosis By using methods such as monitoring blood pressure in outpatients or at home, this can be prevented. It links the spread of masked, drug-resistant, and abnormal blood pressure patterns in CKD. Although traditional blood pressure lowering strategies It will be basic, but newer agents such as SGLT2-hemeric and nonsteroidal mineralocorticoid receptor antagonists are promising in addressing both blood pressure control and the associated risks. New treatments and a focus on blood pressure phenotypes may improve outcomes in hypertension associated with CKD.

Md. Ariful Islam et al.[4][2023] This project investigates the use of ML and predictive modelling for the early detection of CKD, with the goal of delaying or avoiding the progression of end-stage kidney disease, using 25 factors. The research identified a subset of the most predictive factors. This reduces the variables by 70%. Twelve ML classifiers were tested, where XgBoost achieving the highest performance (Accuracy: 98.3%, Precision: 98%, Recall: 98%, F1-score: 98%). These results highlight the promise of sophisticated machine learning methods to improve early detection in predicting chronic kidney disease and pave the way for more effective intervention strategies.

Debabrata Swain et al.[5][2023],This project will investigate the unpredictable fluctuations in chronic renal disease and show how machine learning techniques might improve diagnosis accuracy. It uses SMOTE in conjunction with imputation methods to enable effective data balancing, missing value management, and optimal resource allocation, leveraging publically available datasets. After that, the chi-square test was used to pick resources. A supervised stack of ML models is developed. With SVM and RF providing the best performance where SVM shows the best results. It has a testing accuracy of 99.33% and the lowest false negative rate. Verified through 10- fold cross-validation, this study demonstrates the effectiveness of ML in improving CKD prediction and reducing diagnostic uncertainty.

Hasnain Iftikhar et al.[6][2023], This paper investigates the application of different machine-learning models. To forecast chronic kidney disease (CKD) using data from Buner, Khyber Pakhtunkhwa, Pakistan. Models tested included logistic regression. Probit regression, random trees, decision trees. KNN and SVM with different kernel functions. (Linear basis, Laplacian, Bessel and radial) Outcome criteria such as accuracy, sensitivity, specificity, F1 score and Brier score are used in the evaluation. Together with testing Diebold-Mariano for accuracy in comparative predictions the results show that SVM with Laplacian curves achieves the best performance. Meanwhile, random shooting is also very competitive. These features show the potential of non-invasive ML models for targeted and effective CKD diagnosis.

Vivante, A. et al. [7] [2024] Worldwide, over 800 million people are impacted by from chronic renal disease. It is a significant public health issue. However, the primary causes are excessive blood pressure and diabetes. However, the importance of genetic factors is growing. KDIGO supports genetic testing to improve diagnostic accuracy and support personal choice. Advances in sequencing technology since the discovery of the ADPKD gene locus in 1985 have identified a number of genes involved in CKD. Although genetic causes are more common in children, But it is more well known among adults. This review looks at monogenic CKD and high-risk genetic variations. Polygenic risk scores and secondary risk alleles affecting renal function were excluded.

Sumerah Jabeen et al.[8] [2024] Chronic kidney disease is a long-term disease that affects millions of people around the world. Coronary artery disease is the main reason for death in advanced cases. Traditional determinants such as diabetes, high blood pressure, and abnormal fat levels play an important role while new factors come into play. This chapter discusses oxidative stress, an important new risk factor. It examines the impact of stress on the progression and complications associated with advanced CKD.

Vlado Perkovic et al.[9][2024], This study looked at the effects of semaglutide in people suffering from type 2 diabetes and chronic kidney disease, groups prone to kidney failure.

Cardiovascular disease and death results from 3,533 participants over a median of 3.4 years demonstrated that semaglutide lowered the likelihood of serious kidney injury by 24%, cardiovascular death by 29%, and overall death by 20%. Compared to placebo it helps avoid depressed kidney function and reduces serious side effects. These data indicate that semaglutide has significant benefits in lowering renal and cardiovascular risk in this population.

Johannes F. E. Mann et al.[10][2024],This study examined individuals diagnosed with type 2 diabetes and chronic kidney disease, categorising them according to whether they had previously used SGLT2 inhibitors (SGLT2i). The purpose was to examine the advantages of semaglutide. The study revealed that semaglutide lowered the risk of major kidney and heartrelated events by 24%. Notably, those who did not use SGLT2i showed significant improvements. Secondary outcomes were similar across all groups, including a slower decline in kidney function (eGFR), fewer heart-related events, and lower overall mortality. These data suggest that semaglutide, either alone or combined with SGLT2 inhibitors, can enhance health outcomes for individuals with chronic kidney disease.

3. PERLIMINARY SECTION

The data information and evaluation matrices for this study are explained in this section

A. Data Description: The UCI machine learning dataset is the source of the chronic kidney disease dataset. Information gathered from people who might or might not have CKD is included in this dataset. It is intended to make research on CKD detection predictive modeling and classification tasks easier. This collection includes 400 instances and 25 features where features are divided into categorical and numerical values. Categorical features representing qualitative data (yes or no). Target Variable is labeled with ”class” attribute, indicating whether the individual has CKD (ckd) or not (notckd). The dataset contains missing values which are effectively handled by imputing. The median is applied to numerical columns, and the most frequent value is used for categorical columns to guarantee that the dataset is appropriate for machine learning analysis

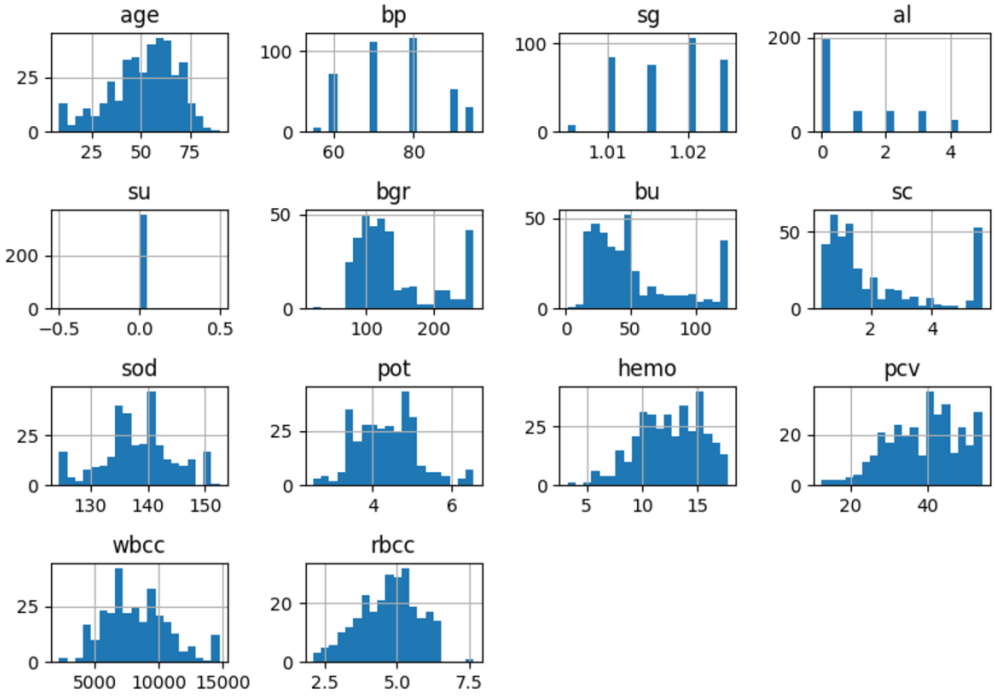

Figure 1. Histograms for Numerical Columns.

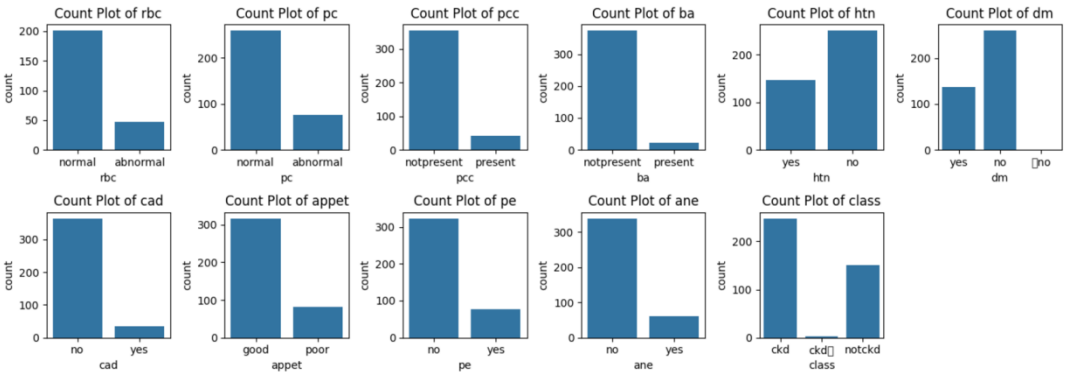

Figure 2. Count-plot for Categorical Columns

B. Preprocessing:

Table 1. CATEGORIZATION OF FEATURES IN THE CKD DATASET.

NO.	Feature	NO.	Feature	NO.	Feature
1	Age	10	bgr	19	htn
2	bp	11	bu	20	dm
3	sg	12	sc	21	cad
4	al	13	sod	22	appet

5	su	14	pot	23	pe
6	rbc	15	hemo	24	ane
7	pc	16	pcv	25	Class(target)
8	pcc	17	wbcc		
9	ba	18	rbcc		

C. Performance Evalatuion Metrices: Several assessment metrics have been used on this have a look at to gauge how well the device studying fashions achieved. These metrics offer facts approximately how well the version predicts Chronic Kidney Disease (CKD). The metrics indexed underneath were employed:

Precision (P): Tells us the percentage of fine predictions that have been really accurate. The model has fewer false alarms (fake positives) whilst its precision is high.

$$P = \frac{TP}{TP + FP} \qquad (1)$$

Recall (R): The model's take into account gauges how nicely it recognizes real CKD instances. It shows the share of CKD patients that the model efficaciously identified. In scientific studies, in which failing to stumble on a wonderful case will have foremost repercussions, that is mainly crucial.

$$R = \frac{TP}{TP + FN} \qquad (2)$$

F1-Score: Precision and Recall are blended into a single price referred to as the F1-Score. It gives a truthful angle, particularly in cases wherein the distribution of CKD and nonCKD cases within the dataset is uneven.

$$\text{F1-score} = \frac{2 \times P \times R}{P+R} \qquad (3)$$

Accuracy (A): The version's typical accuracy is determined by how many predictions it efficaciously made. It is determined via dividing the overall range of predictions by the percentage of correct predictions (CKD and non-CKD).

$$A = \frac{TP + TN}{TP + TN + FN + FP} \qquad (4)$$

D. Methodology:

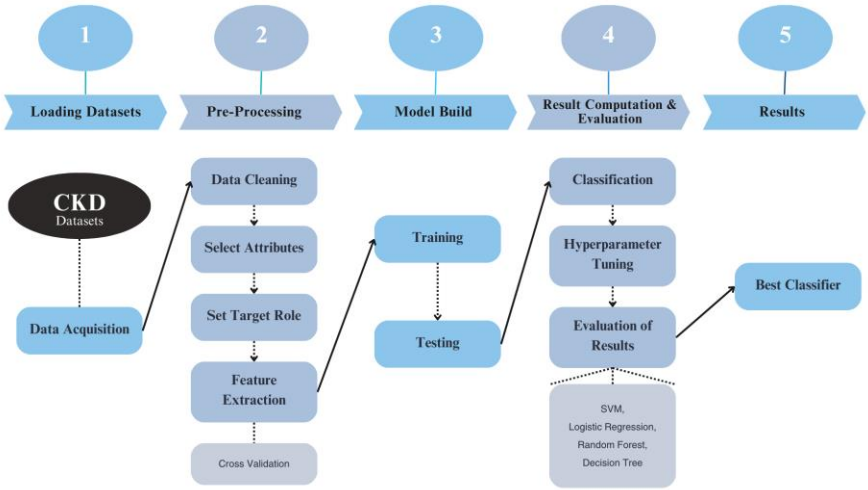

Figure 3. Process Flow Diagram.

Chronic kidney disease (CKD) is predicted in this study using machine learning algorithms. The ckd dataset, sourced from the UCI ML Repository, contains 400 instances with 25 features, including both numerical and categorical data. The target variable"class" indicates whether CKD may be present or not. Data preprocessing involved handling missing values through imputation (using the median for numerical and the most frequent value for categorical features) and standardizing the dataset using StandardScaler. The dataset was split into training (67%) and testing (33%) subsets. Four machine learning models (SVM, DT, RF, and LR) were trained using the processed dataset. K-fold cross validation (n splits = 5) was applied to ensure reliable evaluation. Model performance was assessed using accuracy, precision, recall, F1-score, and confusion matrix.

GridSearchCV from the scikit-learn library was used for hyperparameter optimization, leading to a significant improvement in model accuracy. The best accuracy of 99.24% was obtained by optimized Random Forest and Logistic Regression models, followed by SVM with 98.48% and Decision Tree with 97.72%. The ability of machine learning techniques for the quick detection of CKD is demonstrated by these findings, providing a base for future research and clinical application.

4. Machine Learning Algorithms

We successfully applied machine learning techniques for predictive analysis in our study. The following methods were applied:

A. Support Vector Machine (SVM)

SVM, a classifier that determines categorically different data by an optimal hyper plane margin between them.

$$f(x) = \sum_{i=1}^{n} \alpha i \, yi \, K(xi, x) + b$$

Where:

• x: Input data (features) for the sample that we want to classify.

• xi : Support vectors (training samples that closest to decision boundary).

• yi : True labels of support vectors (+1 for one class, −1 for other class).

• αi : Weights (Lagrange multipliers) assigned to support vectors.

• K(xi , x): Kernel function that measures how similar xi is to x.

• b: Bias term (helps shift decision boundary).

The predicted class (yˆ) is determined by the sign of f(x):

$$\hat{y} = \begin{cases} +1 \text{ if } f(x) > 0 \\ -1 \text{ if } f(x) < 0 \end{cases}$$

When probability estimation is enabled, the model applies a sigmoid function to f(x) to estimate class probabilities:

$$P(y = 1|x) = \frac{1}{1 + \exp(-Af(x) - B)}$$

Here:

• P(y = 1|x): The likelihood that sample belongs to class +1.

• A and B: These are values learned during training to fit sigmoid function.

• f(x): The result of decision function.

B. Random Forests (RF)

Random forests, or random decision forests, are a type of ensemble method and specifically a type of decision tree used for classification, regression, and other tasks. They aggregate many trees that have been trained on the same dataset, with the final output representing either the majority vote or average prediction. Random Forests are very robust; they manage large datasets efficiently and also constrain over fitting. Equations are given below:

Classification

The prediction for single decision tree is denoted as ht(x), where t represents t-th tree in forest and x is input. The final predicted class (yˆ) is attained by predominant voting over T trees:

$$\hat{y} = \arg\max c \sum_{t=1}^{T} I(ht(x) = c)$$

Where:

• c: Label or category of class.

• I: Function that gives 1 if ht(x) = c, and 0 if it is not.

Regression

Each decision tree gives a numerical prediction(x). The final prediction ($y^$) obtained by averaging the predictions of all the trees:

$$y^ = \frac{1}{T} \sum_{t=1}^{T} ht(x)$$

Where:

• T: Total trees in forest

• ht(x): Prediction made by t-th tree.

C. Logistic Regression:

Logistic Regression predicts the probability of a condition or disease based on risk factors. IT connects a dependent variable (outcome) to one or more independent variables (predictors). It applies mostly to binary or multiclass classification problems.

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta n Xn)}}$$

Here:

• β0 : This is intercept or bias term, which is starting point for model's prediction.

• βi (for i = 1, 2, . . . , n) : These are coefficients linked with each input feature, and they are determined during model training. They show how much each feature affects output.

• Xi : These are the values of the input features, which are the data used to make predictions.

A threshold is set, and the logistic regression formula is used to calculate the predicted probability P(y = 1 | X). The predicted outcome is decided based on this probability.

$$y^ = \begin{cases} 1 \text{ if } P(y = 1 \mid X) \geq \text{threshold (usually 0.5)} \\ 0 \text{ otherwise.} \end{cases}$$

These rules can be modified depending on the specific application, such as prioritizing sensitivity or specificity in a health-related study.

D. Decision Tree (DT)

Decision Tree, such as the C4.5 model, is an analytical tool that classifies data into different categories based on a set of conditions. It has found applications in many disciplines due to its explicit and reasonable approach to making predictions.

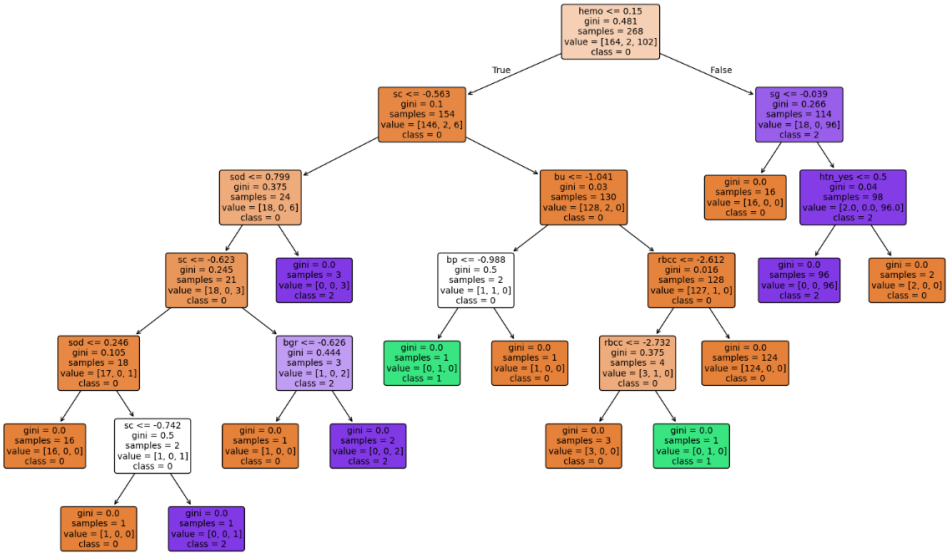

Fig. 4. Decision Tree Visualization (Train Data)

E. Parameter Optimization

Parameter optimization is an important process for improving a model's performance by determining the best values for its settings. Hyperparameter optimization can be done using a variety of techniques, such as Grid Search, Random Search, and Bayesian Optimization. In our research, we used GridSearchCV from the scikit-learn library to fine-tune the hyperparameters of all of our predictive models. This approach investigates various parameter combinations to determine which one produces the best results. As a result, these optimization techniques significantly improved the accuracy of our predictive analysis.

5. Experimental Results & Discussions

A. Experimental Result: The dataset for Chronic Kidney Disease utilized in this research consists of 400 samples and 25 features, which encompass both clinical and laboratory information. Missing values for numeric column and categorical column were addressed through SimpleImputer(strategy=median ´ ´) and SimpleImputer(strategy=most ´ frequent´), and standardizing was applied to all features to ensure consistency. In order to assess the effectiveness of the machine learning models, the dataset was split into subsets for training and testing. Specifically, 33% of data was designated for testing purposes, whereas the remaining 67% was utilized to train the models. Figure 4 display the accuracy from different machine learning models.

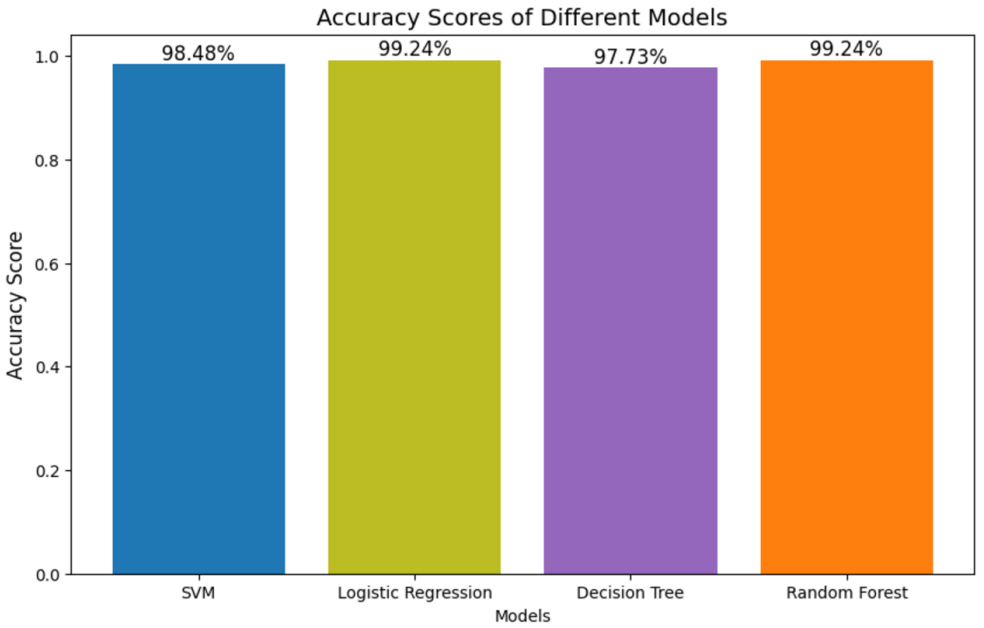

Fig. 4. Accuracy score of Different Models

B. Discussions: Using chronic kidney data, we demonstrated several machine learning (ML) algorithms for disease detection. Hyperparameter optimization, conducted using GridSearchCV, significantly enhanced the performance of all models. After calibration, Random Forest and logistic regression achieved an accuracy of 99.24%, thus demonstrating their performance on high-dimensional classification problems. The sound in reducing redundancy showed his lower-performing Decision Tree model improved to 97.72% after optimization as well. The focus was on hyper-parameter optimization in improving model performance; SVM models attained an accuracy of 98.48%. These results thus exhibit the potential of ML approaches post-meta-parameter adjustment for earlier identification of renal disease. Hence needed larger and more heterogeneous datasets to validate these findings for better general clinical applicability in future studies.

6. Conclusion

The dataset on Chronic Kidney Disease employed various machine learning techniques, including 'support vector machine (SVM)', 'logistic regression (LR)', 'decision tree (DT)', and 'random forest (RF)'. Through hyperparameter optimization, the accuracy for each model significantly improved. Among these models, both LR and RF achieved impressive accuracies of 99.24%. SVM also demonstrated strong performance with an accuracy of 98.48%, while the Decision Tree model reached 97.72%. These findings suggest that machine learning methods, particularly when paired with metaparameter adjustments, can serve as accurate and effective tools for addressing CKD. However, a limitation of this study is the reliance on numerical data for machine learning. In the future, we intend to explore the effective use of images through various extraction techniques.

Acknowledgment

This research work is supported by School of IT, Washington University of Science and Technology, VA, USA

References

1. EMPA-Kidney Collaborative Group. (2023). Empagliflozin in patients with chronic kidney disease. New England Journal of Medicine, 388(2), 117-127.
2. Ku, E., Del Vecchioet. al. (2023). Novel anemia therapies in chronic kidney disease: Conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney International, 104(4), 655-680.
3. Burnier, M., & Damianaki, A. (2023). Hypertension as a cardiovascular risk factor in chronic kidney disease. Circulation Research, 132(8), 1050-1063.
4. Islam, M. A., et. al. (2023). Chronic kidney disease prediction based on machine learning algorithms. Journal of Pathology Informatics, 14, 100189. https://doi.org/10.4103/jpi.jpi_67_23
5. Swain, D., Mehta, et. al (2023). A robust chronic kidney disease classifier using machine learning. Electronics, 12(1), 212.
6. Iftikhar, H., et. al (2023). A comparative analysis of machine learning models: A case study in predicting chronic kidney disease. Sustainability, 15(3), 2754. https://doi.org/10.3390/su15032754
7. Vivante, A. (2024). Genetics of chronic kidney disease. New England Journal of Medicine, 391(7), 627-639. https://doi.org/10.1056/NEJMra2208253
8. Jabeen, S., & Noor, S. (2024). Oxidative stress and chronic kidney disease. In Fundamental Principles of Oxidative Stress in Metabolism and Reproduction (pp. 151-165). Springer. https://doi.org/10.1007/978-3-030-78699-2_9
9. Perkovic, V., Tuttle, et. al(2024). Effects of semaglutide on chronic kidney disease in patients with type 2 diabetes. New England Journal of Medicine, 391(2), 109-121.
10. Mann, J. F., Rossing, et. al. (2024). Effects of semaglutide with and without concomitant SGLT2 inhibitor use in participants with type 2 diabetes and chronic kidney disease in the FLOW trial. Nature Medicine, 1-8.