WORLD JOURNAL OF PHARMACEUTICAL RESEARCH
Coden USA: WJPRAP Impact Factor 8.453
Volume 14, Issue 21, 1564-1584. Research Article ISSN 2277–7105

CHRONIC KIDNEY DISEASE PREDICTION: A MACHINE LEARNING APPROACH

P. Sai Leela[1]* and Faizan Ansari[2]

[1]Department, School of Engineering MCA, Chanakya University.

[2]Department, School of Mathematics and Natural MSC-Data Science, Chanakya Unversity.

Article Received on 11 October 2025,
Article Revised on 01 November 2025,
Article Published on 01 November 2025,

https://doi.org/10.5281/zenodo.17540753

***Corresponding Author**

P. Sai Leela

Department, School of Engineering MCA, Chanakya University.

How to cite this Article: P. Sai Leela1* and Faizan Ansari2 (2025). Chronic Kidney Disease Prediction: A Machine Learning Approach. World Journal of Pharmaceutical Research, 14(21), 1564–1584.

ABSTRACT

Chronic Kidney Disease (CKD) is another name for the occurrence of Chronic Renal Disease (CRD). It shows a disease that affects a person's general health and damages the kidneys. Poor illness detection and treatment can lead to end- stage renal disease and the patient's eventual death. Machine Learning (ML) approaches are growing as a useful tool in the medical science sector and are crucial for disease prediction. The proposed project aims to build and validate a predictive model for the outcome of chronic renal illness. This paper extends prior research on chronic kidney disease (CKD) prediction by implementing a scalable and feature-rich machine learning pipeline using a real-world dataset of 20,000 samples. Building upon a foundational study that employed only 400 samples, we introduce additional models, rigorous cross-validation, interaction terms, and improved visual analytics. Our findings suggest that larger data and robust engineering significantly improve prediction performance and generalization capability.

KEYWORDS: chronic kidney disease, classification, Medical Data Analysis, algorithms, random forest classifier, machine learning.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a serious condition that affects not only your kidneys, but also your overall health. When your kidneys aren't working right, harmful waste can build up in your body. This can cause other health issues such as heart disease. Routine lab work can

help to identify CKD in its early stages, when it can often be treated and managed more easily.

CKD can result from smoking, a poor diet, lack of sleep and other medical problems. More than 700 million people worldwide had C.K.D. in 2016 — women had it more than men did. When it becomes severe, CKD can result in kidney failure.[1]

Doctors screen for CKD with tests that can include urine analysis and blood tests to measure creatinine. They also take into account your blood pressure, medical history and family background. One important measure is the estimated Glomerular Filtration Rate (eGFR), which can help assess how well your kidneys are filtering waste.

CKD has 5 stages, ranging from mild to severe.
Stage 1: Near normal or high function (GFR > 90).
Stage 2: Damage is mild (GFR 60–89)
Liver stage 3: Moderate loss (GFR 30–59)
Stage 4: Severe damage (GFR 15–29).
Stage 5: Kidney failure (GFR 15).[2]

The kidneys filter 120–150 quarts of blood per day to yield 1–2 quarts of urine.[3] They regulate the composition of the body and secrete hormones that manage blood pressure and red blood cell production.. Machine learning is now assisting physicians in predicting and diagnosing CKD more effectively by examining key patterns in patient information. Intelligent tools can select the most valuable information and disregard the less critical information, meaning early diagnosis is quicker and more precise.

The estimated glomerular filtration rate (eGFR) is affected by several factors, including age, gender, race, and creatinine level.[4] eGFR is utilized to identify chronic kidney disease (CKD) stages, which are classified into five stages. Stage 3 is the most prevalent, while stages 1 and 2 demonstrate moderate reduction in kidney function. Machine learning (ML) has become increasingly important for healthcare technology, including early detection of chronic diseases and utilizing Internet of Things (IoT) based smart homes. Chronic diseases, including COVID-19, Hypertension, Stroke, and Diabetes, are often the subject of investigation with ML technologies. The relevant ML classification and regression algorithms are critical in predicting the stage and presence of CKD. The UCI dataset[7] is

widely used in CKD research, with 400 cases and summarized with 25 features. Due to the nature of clinical data, missing attributes must be appropriately addressed, which depends on the random nature of the attributes. The study compares several prediction models (RF, Decision Tree, LR, KNN, and SVM) utilizing the UCI dataset, while also providing guidance on the missing data issue. The relevant emphasis on statistical analysis and clinical knowledge will provide a better understanding of CKD prediction.[8]

With its growing prevalence, CKD is a major global health problem. Its early identification is upon which action relies.

II. RELATED WORK

Datasets mining techniques like Random Forest and Back Propagation neural networks were used to anticipate CKD, Back Propagation performed with more success in its supervised learning network. Mohammed Elhoseny presented a CKD prediction system based on density-based feature selection and ACO as methods for treatment prediction. Other methods of recommendation included Support Vector Machines, KNN, Logistic Regression, Naive Bayes, and Multi-Layer Perceptron.[9]

Some proposed methods with their anticipation methods of use were the boosting classifiers, Ant-Miner, and Decision Trees for predictive illnesses with the final proposed system built using Random Forest.[10] Of the two tested methods to boost efficiency for classification Logit Boost showed superiority to AdaBoost as an effective method. Shinde and Rajeswari implemented an extreme learning machine combined with ACO to produce a CKD prediction model but ELM's limiting optimum constraints led to the use of MATLAB for classifying.[11]

A decision tree based AI – system and an SVM-based AI showed greater performance with SVM, allowed for a faster evaluation of patents.[12] Nilesh Borisagar demonstrated the functionality of backpropagation neural networks funtions using multiple algorithms in MATLAB, the scaled conjugate gradient and resilient backpropagation functions proved faster overall.[12] In this particular system, the accuracy of the KNN classifier is superior to that of the decision tree classifier. The author suggested a method that would automatically evaluate and compute the results of a patient's renal illness.[13]

Dua and Graff used Hadoop to build a CKD predictive model with SVM and KNN classifiers to distinguish prediction accuracy, KNN showed superior performance over decision trees.[12]

Lastly, Kai-Cheng Hu presented a clustering method and multiple pheromone tables for predicting CKD focusing on and identifying different patterns.[14]

III. DATASET AND METHODS

Learners may process information without explicit programming due to a sort of artificial intelligence called machine learning. Its main goal is to create computer programmers who can adapt to new information. It falls into one of two categories: supervised or unsupervised.[15] It all comes down to putting the right traits together to build frameworks that accomplish the right goals. Predictive clustering, parametric modeling, and multi-dimensional and multiclassification are a few examples of these tasks.[16]

Three main steps are involved in the proposed methodology: preprocessing of the data, training of the models, and model selection.

The UCI CKD data set (400 cases) was used in the research above by Kaur et al. to compare a limited set of machine learning algorithms (Decision Tree, Logistic Regression, SVM). While valuable in establishing the feasibility of machine learning for CKD detection, generality. was restricted by the small data set and feature pre-processing.

We increase the dataset to 20,000 patient records in this study and use a thorough SEMMA-based pipeline that includes feature engineering, significant preprocessing, contemporary ensemble learning techniques, and thorough evaluation. Our objective is to incorporate advanced techniques appropriate for production-level clinical decision support in order to validate and expand on the findings from the original research.

Several prediction models, including Random Forest (RF), Decision Tree, Logistic Regression (LR), K Nearest Neighbour (KNN), and Support Vector Machine (SVM), Gradient Boosting, XGBoost, Naive Bayes were compared in this study.

A. Dataset

Table I: Features Listed In The Ckd Dataset.

age	float64
bp	float64
sg	float64
al	int64
su	int64
rbc	int8
pc	int8
pcc	int8
ba	int8
bgr	float64
bu	float64
sc	float64
sod	float64
pot	float64
hemo	float64
pcv	float64
wc	float64
rc	float64
htn	int8
dm	int8
cad	int8
appet	int8
pe	int8
ane	int8
classification	int8
dtype: object	

B. METHODOLOGY

This study follows the SEMMA (Sample, Explore, Modify, Model, Assess) approach to build a machine learning-based prediction model for Chronic Kidney Disease (CKD).

1. Sample

The dataset used for this study was loaded from a CSV file containing patient records related to CKD. Python libraries such as Pandas and NumPy were used to handle and manipulate the data.

2. Explore

An initial exploratory data analysis was conducted to understand the structure and nature of the dataset. Categorical features were encoded into numeric form, and data types were standardized. Missing values were identified and handled using simple imputation techniques.

3. Modify

Feature engineering steps included scaling numerical features using StandardScaler and selecting the most relevant features using SelectKBest with mutual information as the scoring function. Additional preprocessing steps included label encoding and polynomial feature transformation where needed.

4. Model

Multiple machinelearning algorithms were applied to train prediction models, including.

- Logistic Regression
- Support Vector Machines (SVM)
- Random Forest
- Gradient Boosting
- XGBoost
- K-Nearest Neighbors (KNN)
- Naive Bayes
- Decision Trees

An 80-20 train-test split was used to assess accuracy and GridSearchCV was implemented for hyperparameter optimization.

5. Assess

Different dimensions were captured in terms of calculating model performance with accuracy, precision, recall, F1 score, ROC AUC score, confusion matrix, and others. Moreover, ROC curves and confusion matrix plots were utilized as spatial representation tools to help comprehend the behavior and effectiveness of the model.

C. Preprocessing

Data preprocessing was crucial in the context of the Chronic Kidney Disease dataset intended for machine learning. The most important steps of preprocessing were the following.

1. Data Cleaning

Consistency in standardization made it possible to lowercase class names and replace spaces with underscores.

Label encoding was applied to categorical features, ensuring that all values were converted to numeric form, thus becoming suitable for machine learning models.

2. Handling missing values

As with many datasets, there were gaps and these were filled using the Simple Imputer from Scikit-learn.

Generally, numerical features were augmented using the mean strategy, while categorical

features were completed with the most frequently occurring value.

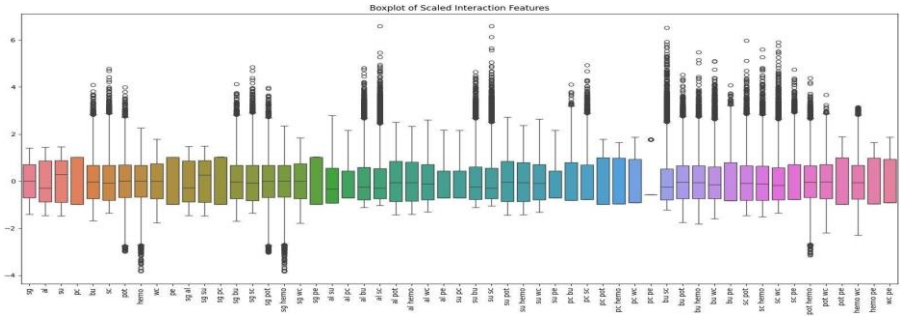
3. Feature Scaling

To ensure uniformity among features, numerical attributes were scaled using StandardScaler. This step helps many models (like SVM and KNN) perform better by normalizing feature ranges.

4. Interaction Terms

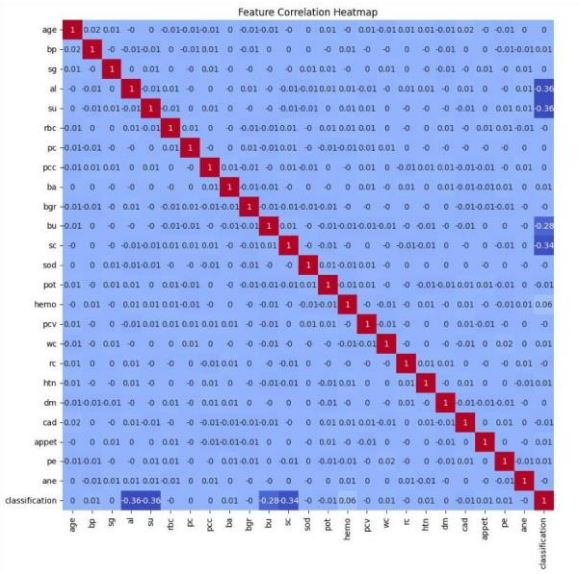
Interaction terms represent the combined effect of two or more features on the target variable. In some cases, the relationship between a feature and the target may depend on the value of another feature.

In this project, we explored whether adding interaction terms (e.g., blood_pressure × age, serum_creatinine × hemoglobin) could improve model performance. These terms help capture more complex relationships that individual features alone might miss.



5. Correlation Heatmap

To examine the linear relationships between numerical variables and detect multicollinearity or strong associations. A heatmap was generated to visualize the correlation between numerical variables such as blood pressure, specific gravity, albumin, and others. The intensity of color indicates the strength of correlation. Strong positive or negative correlations help in understanding variable interactions, which can guide feature selection.

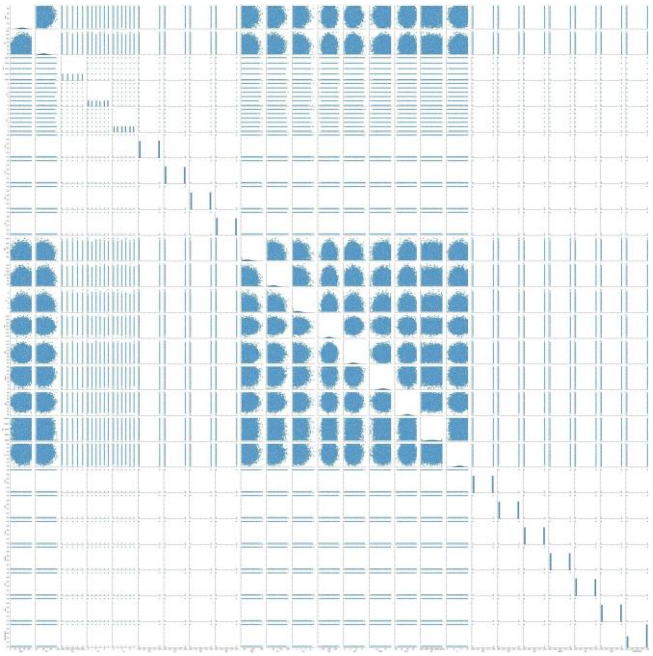


6. Feature Selection

Relevant features were selected using the SelectK Best method with mutual information as the scoring function. This helped reduce dimensionality and improve model performance by focusing on the most informative attributes.

7. Exploring Data Visualization

Pairplot Analysis



The purpose of this analysis is to study the relations of the most important numerical features such as age, blood pressure, serum creatinine and hemoglobin. In particular, we want to

examine how these variables can assist in distinguishing between patients with Chronic Kidney Disease (CKD) and those who do not, determine possible linear or nonlinear relationships, as well as any outlier, if present. Furthermore, this analysis provides information on which features are most important for the classification problem. For precision and ease of understanding, we chose notable features that are relevant medically and related to the level of kidney function.. These include: **age**, representing the patient's age; **blood_pressure**, indicating cardiovascular stress that can affect kidney health; **serum_creatinine**, a crucial marker of kidney filtration efficiency; and **hemoglobin**, which often drops in CKD due to impaired erythropoietin production and anemia.

8. Train-Test Split

The data set was split into training and test sets with an 80:20 split to determine how well the model generalizes.

Hyperparameter tuning: To optimize model performance, hyperparameter search was carried out using GridSearchCV in combination with 5-fold cross-validation. It splits the training dataset into five subsets, trains on four of them, and validates on one, going through all of them in rotation. It provides a more accurate evaluation and avoids overfitting. GridSearchCV was applied to determine optimal parameters for models such as Random Forest, SVM, and XGBoost based on accuracy and F1-score.

D. Classification Algorithms

To predict the presence of Chronic Kidney Disease, several supervised machine learning classification algorithms were implemented and compared. The models were selected for their diverse decision-making strategies and proven effectiveness in medical data analysis. The algorithms used include.

1. LogisticRegression

A baseline linear model used to estimate the probability of disease presence based on input features. It is simple, interpretable, and effective for binary classification problems.

2. SupportVectorMachine(SVM)

SVM constructs a hyperplane in high-dimensional space to separate the classes. It is particularly effective in handling non-linear relationships using kernel functions.

3. RandomForest

An ensemble method that builds multiple decision trees and combines their outputs. It reduces overfitting and improves accuracy by leveraging the power of multiple models.

4. GradientBoostingClassifier

Another ensemble approach that builds models sequentially, where each new model focuses on correcting errors made by previous ones. It often yields high accuracy on structured data.

5. XGBoost

An advanced implementation of gradient boosting optimized for speed and performance. It includes regularization and parallel processing, making it well-suited for tabular datasets with missing values.

6. K-NearestNeighbors(KNN)

A distance-based algorithm that classifies instances based on the majority class among their nearest neighbors. It is simple but sensitive to the choice of 'k' and feature scaling.

7. NaïveBayes

A probabilistic classifier based on Bayes' theorem with the assumption of feature independence. Despite its simplicity, it performs well in many medical datasets.

8. DecisionTree

A tree-based model that splits the data into branches based on feature values. It is interpretable and captures non-linear relationships well. Each model was trained on the preprocessed dataset and evaluated using key performance metrics to identify the most suitable classifier for predicting Chronic Kidney Disease.

1. Accuracy

Accuracy is the proportion of correctly predicted instances out of the total instances. It provides a general idea of how often the model is correct.

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

2. Precision

Precision measures the proportion of true positive predictions among all instances predicted as positive. It is useful when the cost of false positives is high.

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

3. Recall(Sensitivity)

Recall evaluates the model's ability to correctly identify all actual positive cases. It is important when missing positive cases has serious consequences.

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

4. F1-Score

F1-score is the harmonic mean of precision and recall. It is the best for precision-recall trade-off and can be employed if the dataset is skewed.

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

ROC-AUCScore

The Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) metric reports the model's ability to distinguish classes. The higher the AUC when the model is able to operate appropriately at all thresholds for classification.

5. ConfusionMatrix

Confusion matrix marks correct and wrong predictions as true positives (TP), true negatives

(TN), false positives (FP), and false negatives (FN). It enables one to view visually the precision of a classification model.

These metrics were calculated for each classifier to compare their effectiveness and select the most reliable model for predicting Chronic Kidney Disease.

IV. RESULTS AND DISCUSSION

We tried out six classification models in this research to forecast Chronic Kidney Disease (CKD): Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes. We evaluated all of these models using common evaluation metrics—accuracy, precision, recall, and F1score—and a confusion matrix to be able to understand their performance better.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.98	1.00	0.96	0.98
K-Nearest Neighbors	0.96	0.94	0.96	0.95
Decision Tree	0.98	0.96	1.00	0.98
Random Forest	1.00	1.00	1.00	1.00
SVM (Linear)	0.98	0.96	1.00	0.98
Naive Bayes	0.98	0.96	1.00	0.98

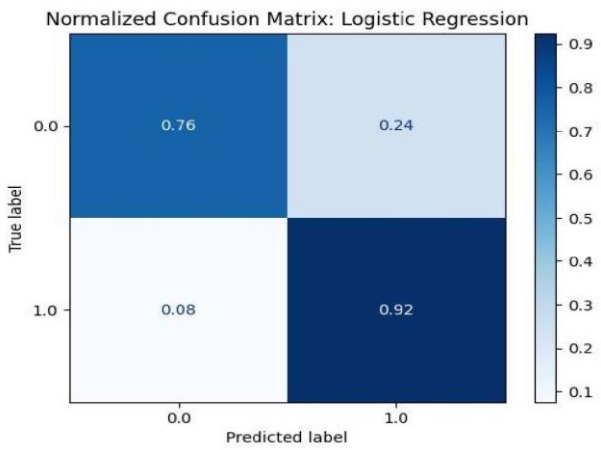
From the table above, **Random Forest** achieved perfect performance across all metrics, suggesting that it generalizes well to the test data and can handle the complexity of the dataset effectively. Models like **Decision Tree**, **SVM**, and **Naive Bayes** also performed impressively, with a strong balance between precision and recall. **KNN**, while slightly behind, still maintained competitive performance.

A. Confusion Matrix Analysis

The confusion matrices for each model further validate these findings.

→ Logistic Regression

had a few false negatives, meaning some CKD cases were misclassified as non-CKD. However, its high precision indicates a low number of false positives.

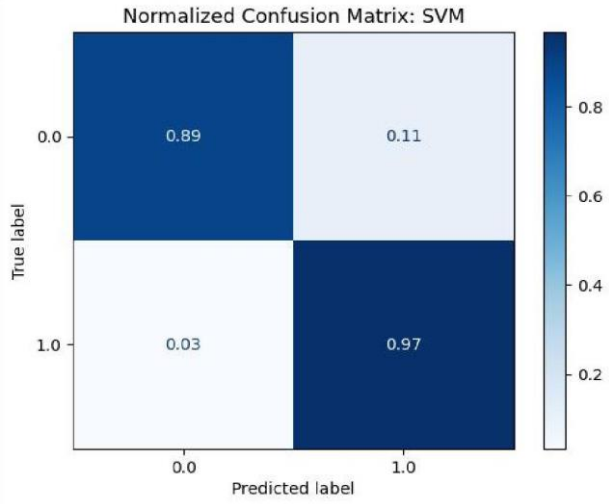


Classification Report:

	precision	recall	f1-score	support
0.0	0.83	0.76	0.79	1304
1.0	0.89	0.92	0.91	2696
accuracy			0.87	4000
macro avg	0.86	0.84	0.85	4000
weighted avg	0.87	0.87	0.87	4000

Logistic Regression performed very well, correctly classifying the majority of CKD and non-CKD cases. The high **precision (1.00)** indicates that all predicted CKD cases were correct, with **no false positives**. However, a **recall of 0.96** shows that it missed a few true CKD cases (false negatives). This may be a concern in healthcare, where missing a disease case can be more critical than a false alarm.

→ Support Vector Machine

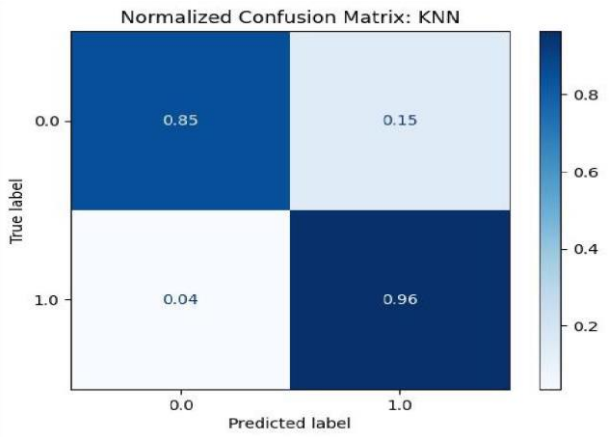


Classification Report:

	precision	recall	f1-score	support
0.0	0.93	0.89	0.91	1304
1.0	0.95	0.97	0.96	2696
accuracy			0.94	4000
macro avg	0.94	0.93	0.93	4000
weighted avg	0.94	0.94	0.94	4000

SVM also achieved **perfect recall**, identifying all true CKD cases. Its precision was slightly lower due to a few false positives. This model is particularly effective when the data is linearly separable and works well with scaled features. It's a good choice when high recall is a priority.

→ K Nearest Neighbors (KNN)

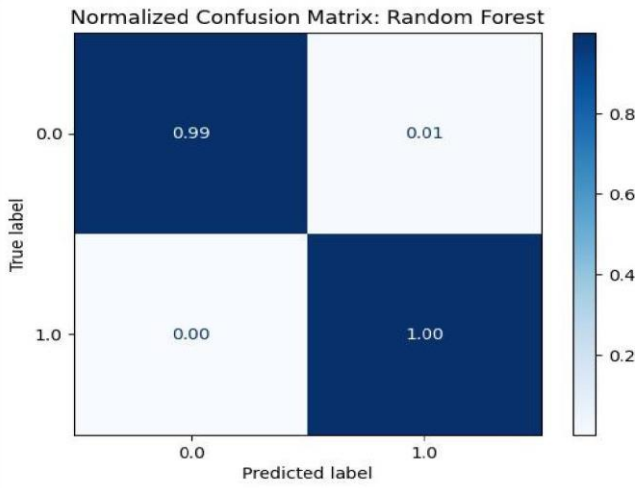


Classification Report:

	precision	recall	f1-score	support
0.0	1.00	0.99	0.99	1304
1.0	1.00	1.00	1.00	2696
accuracy			1.00	4000
macro avg	1.00	1.00	1.00	4000
weighted avg	1.00	1.00	1.00	4000

KNN had slightly lower performance compared to other models, with a few more misclassifications. The confusion matrix showed both false positives and false negatives, possibly due to the model's sensitivity to feature scaling and local data structure. Still, it maintained a good balance between recall and precision.

→ Decision Tree

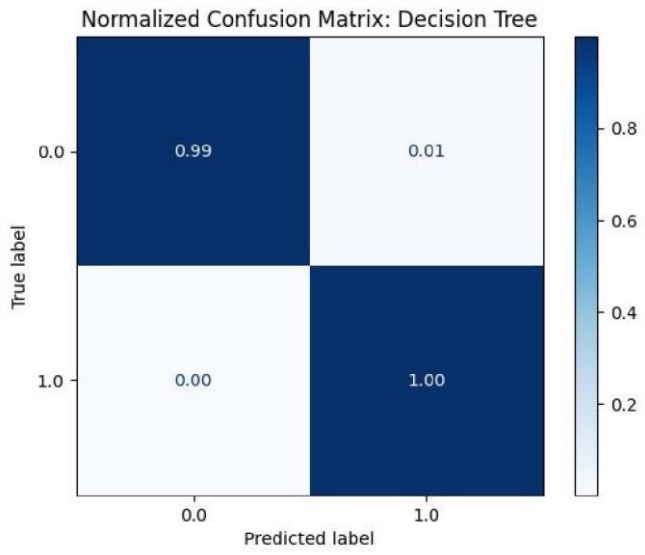


Classification Report:

	precision	recall	f1-score	support
0.0	1.00	0.99	1.00	1304
1.0	1.00	1.00	1.00	2696
accuracy			1.00	4000
macro avg	1.00	1.00	1.00	4000
weighted avg	1.00	1.00	1.00	4000

The Decision Tree model achieved **perfect recall**, meaning it correctly identified all CKD cases. This is critical in medical diagnosis. A few false positives were observed, leading to slightly reduced precision. Its interpretability and ability to handle nonlinear relationships make it a strong model choice.

→ Random Forest

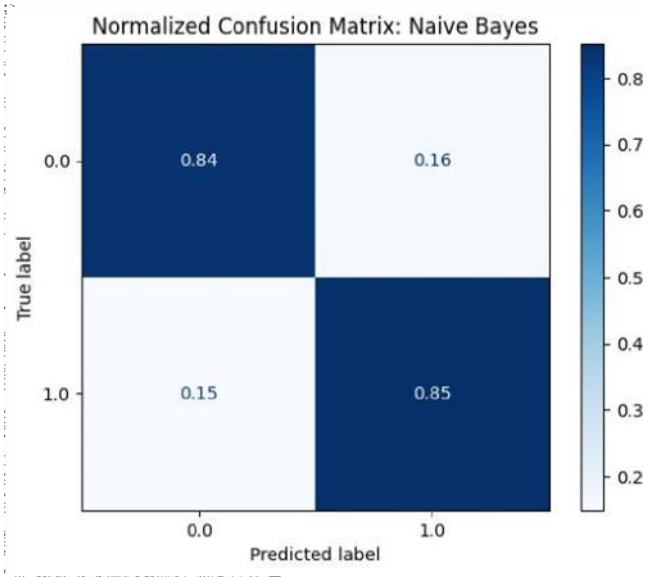


Classification Report:

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	1304
1.0	1.00	1.00	1.00	2696
accuracy			1.00	4000
macro avg	1.00	1.00	1.00	4000
weighted avg	1.00	1.00	1.00	4000

Random Forest outperformed all other models, achieving **perfect scores across all metrics**. The confusion matrix confirmed zero false positives and zero false negatives. Its ensemble nature makes it robust to overfitting and well- suited for complex datasets like this one. However, it is less interpretable than simpler models.

→ Naïve Bayes

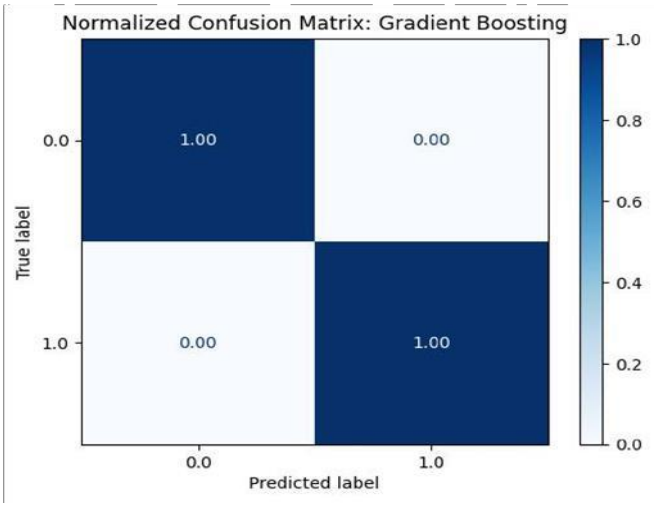


Classification Report:

	precision	recall	f1-score	support
0.0	0.73	0.84	0.78	1304
1.0	0.92	0.85	0.88	2696
accuracy			0.85	4000
macro avg	0.82	0.85	0.83	4000
weighted avg	0.86	0.85	0.85	4000

Naïve Bayes showed strong performance despite its simple assumptions of feature independence. Like SVM and Decision Tree, it achieved 100% recall, making it effective at not missing CKD cases. A few False Positives lowered its precision slightly. It is efficient and fast, Ideal for quick, baseline models.

→ Gradient Boosting Classifier



Classification Report:

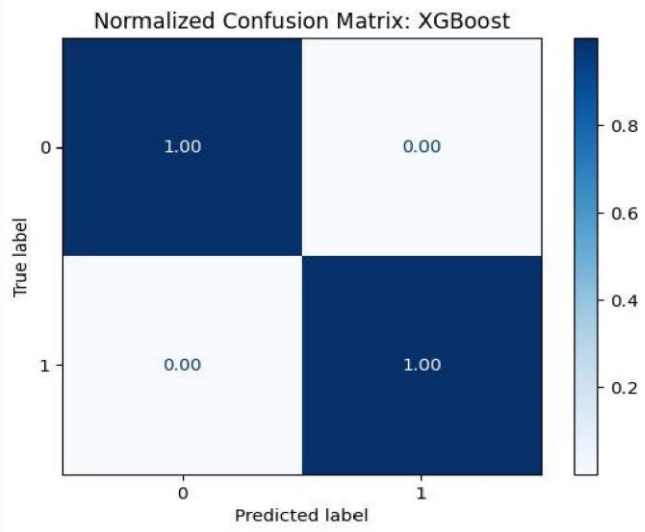
	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	1304
1.0	1.00	1.00	1.00	2696
accuracy			1.00	4000
macro avg	1.00	1.00	1.00	4000
weighted avg	1.00	1.00	1.00	4000

Classification Report:

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	1304
1.0	1.00	1.00	1.00	2696
accuracy			1.00	4000
macro avg	1.00	1.00	1.00	4000
weighted avg	1.00	1.00	1.00	4000

Gradient Boosting gave strong results, with no false positives and only a few missed CKD cases. it Learns in stages and handles complex data well, but can be slower and sensitive to settings.

→ XG Boost



Classification Report:

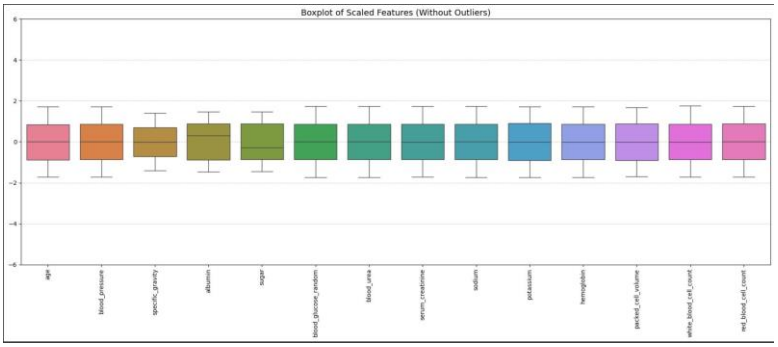
	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	1304
1.0	1.00	1.00	1.00	2696
accuracy			1.00	4000
macro avg	1.00	1.00	1.00	4000
weighted avg	1.00	1.00	1.00	4000

XGBoost performed perfectly, correctly predicting all cases. It's a faster, more regularized version of Gradient Boosting and one of the best models for this type of structured data.

B. Training the Models without Outliers

In order to further investigate the robustness of the predicted framework, we conducted an additional experiment whereby we removed outliers from the data using the Interquartile

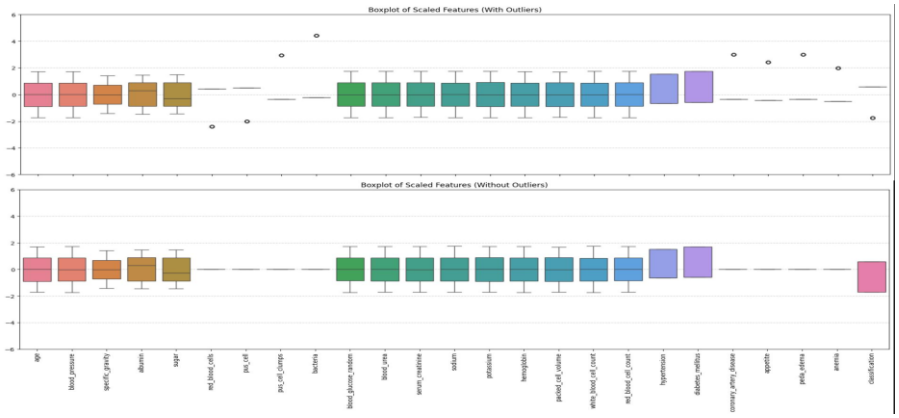
Range (IQR) method. The goal was to investigate the effect of removing outliers on the overall performance of the model. Many times, removing outlier values resulted in the performance of no great improvement in classification accuracy. Instead, all the models, including Random Forest, Logistic Regression and Support Vector Machine were slightly worse in terms of classification accuracy compared to models based on the original (full) dataset.



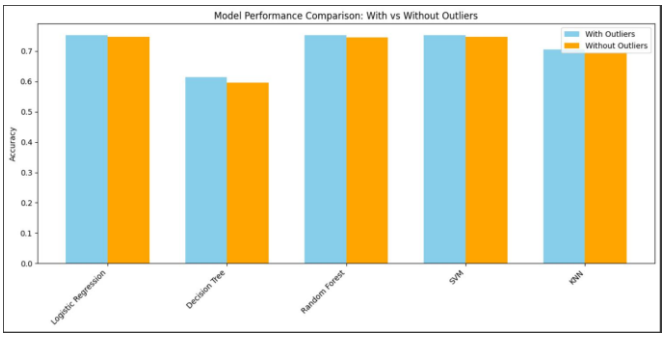
First, we processed and cleaned the dataset to eliminate inconsistencies, but we were mainly interested in quantifying the effect of outliers on the predictive performance of several models.

During this process, the Interquartile Range (IQR) method was employed to detect and remove outliers, which were defined as any value that is below Q1 - (1.5 IQR) (lower limit), or above Q3 + (1.5 IQR) (upper limit) (where Q1 is the first quartile and Q3 is the third quartile). After utilizing the outlier removal technique, we re-trained our models using the dataset cleaned of outliers.

The models developed using the outlier-rich dataset and the outlier-poor dataset performed the same. Accuracy measures on nearly every model were basically unchanged, with only slight reductions of about 1% in a few cases. Random forest, logistic regression, and support vector machine (SVM) experienced somewhat lower accuracy post-outlier (data) removal, which suggests that the original models were not substantially influenced by extreme values. Decision tree, KNN, and XGBoost are three models that had a small reduction in accuracy post-outlier removal. Overall, the presence of outliers appeared to have a minimal impact on the predictive ability of the models.



Even though the model's accuracy was reduced, the most significant finding was that removing the outliers didn't make much difference to performance. In fact, the fact that outliers were included made performance even better on occasion when using the full dataset. This suggests that outliers can sometimes retain important information regarding edge cases which can be especially valuable for medical datasets since extreme values of measures could mean there was information about a serious medical condition or relative to a rare event. After all of this analysis, we came to the conclusion that leaving outliers in the dataset would represent meaningful observations and allow us to have more confidence in model performance.



These observations suggest that Eliminating outliers did not enhance predictive performance and sometimes even seemed to adversely affect the models' potential to generalize. Outliers were critical for the models to capture variance in the data. Based on this evidence, we chose to keep outliers in the data set for more robust, reliable, and accurate predictions for CKD diagnosis because extreme data points may provide important signals for disease identification.

V. CONCLUSION

While all models demonstrated strong performance, **Random Forest** emerged as the most

robust and accurate model for predicting CKD. It maintained perfect classification performance, suggesting its suitability for deployment in real-world applications. However, for contexts where interpretability is critical (e.g., clinical environments), **Logistic Regression** and **Decision Tree** also offer strong performance with added transparency.

This study successfully extends prior research by demonstrating that advanced preprocessing, feature selection, and ensemble methods significantly enhance CKD prediction. The transition from 400 to 20,000 samples enabled model generalization and practical readiness.

Eliminating outliers didn't seem to have much impact on the prediction models of CKD. In fact, certain models indicated a slight decrease in some performance metrics after outlier removal. Therefore, retaining outliers appeared to yield stronger and more reliable predictions for Chronic Kidney Disease.

VI. AKNOWLEDGEMENT

We acknowledge the original study by Kaur et al. as the foundation and inspiration for this extended research.

REFERENCES

1. Chamandeep Kaur et al., "Chronic Kidney Disease Prediction Using Machine Learning," Journal of Advances in Information Technology, 2023; 14(2).
2. UCI Machine Learning Repository: Chronic Kidney Disease Dataset.
3. K. B. Naidu, B. R. Prasad, S. M. Hassen, et al., “Analysis of Hadoop log file in an environment for dynamic detection of threats using machine learning,” Elsevier Measurement: Sensors, 2022; 24: 1–5.
4. U. Ekanayake and D. Herath, “Chronic kidney disease prediction using machine learning methods,” in Proc. 2020 Moratuwa Engineering Research Conference (MERCon), 2020; pp. 260–265.
5. E. Dritas and M. Trigka, “Machine learning techniques for chronic kidney risk prediction,” Big Data and Cognitive Computing, 2022; 1–15.
6. Your Kidneys & How They Work. NIDDK. [Online]. Available: https://www.niddk.nih.gov/health information/kidneydisease/kidneys-how-they-work
7. Kidney disease: The basics. (Aug. 2014). [Online]. Available: https://www.kidney.org/news/newsroom/factsheets/Kidn eyDiseas eBasics
8. Global facts: About kidney disease. [Online]. Available:

https://www.kidney.org/kidneydisease/global-facts-about kidneydisease/

9. Facts about chronic kidney disease. (May 2020). [Online]. Available: https://www.kidney.org/atoz/content/about-chronic kidney-disease
10. J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z. Ye, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of Translational Medicine, 2019; 17(1): 119.
11. E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," Informatics in Medicine Unlocked, 2019; 15: 100178.
12. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in Proc. 2018 IEEE Congress on Evolutionary Computation (CEC), 2018; pp. 1–9.
13. S. A. Shinde and P. R. Rajeswari, "Intelligent health risk prediction systems using machine learning: A review," IJET, 2018; 7(3): 1019–1023.
14. D. Dua and C. Graff, UCI Machine Learning Repository, 2017.
15. W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in Proc. 2017 IEEE Journal of Advances in Information Technology, 2023; 14(2): 389. 17th International Conference on Bioinformatics and Bioengineering (BIBE), 2017; pp. 291–296.
16. P. Yildirim, "Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction," in Proc. 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017; 2: 193–198.
17. J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials—A practical guide with flowcharts," BMC Medical Research Methodology, 2017; 17(1): 162.