Milestone Transactions on
MEDICAL TECHNOMETRICS
Vol. 03 | Issue. 02 | 2025

RESEARCH ARTICLE OPEN ACCESS

Leveraging Stacked Machine Learning Models to Advance Diagnostic Precision and Predictive Insights in Chronic Kidney Disease

Vishal Kumar Jaiswal

Sr. Manager Software Engineering,
OPTUM, Ashburn, Virginia, 20148, USA.

DOI: **10.5281/zenodo.15745714**
Received: 21 May 2025 / Revised: 17 June 2025 / Accepted: 26 June 2025

Abstract – A considerable number of people worldwide suffer from chronic kidney disease (CKD), a progressive illness that often goes undiagnosed until it has advanced to a severe degree. Timely action and better patient outcomes rely on early diagnosis. However, traditional diagnostic methods are time-consuming and may lack consistency, especially in resource-constrained settings. To address this, our study explores the application of machine learning (ML) for early-stage CKD prediction. We used the publicly available CKD dataset from the UCI Machine Learning Repository, which includes 400 patient records across 14 clinical features. After performing thorough preprocessing—including handling missing values and converting categorical data—we applied multiple ML classifiers: Naïve Bayes, Decision Tree, Gradient Boosting, AdaBoost, and XGBoost. Each model was evaluated using 10-fold cross-validation to ensure reliability. The core of our approach lies in a stacking ensemble model, which combines predictions from three base learners—Naïve Bayes, Decision Tree, and Gradient Boosting—and passes them to a meta-learner based on AdaBoost. This layered learning framework was specifically designed to mitigate overfitting, which is evident in some individual models during evaluation. The proposed stacking model demonstrated the best performance among all tested models, achieving a training accuracy of 99.2% and a testing accuracy of 98.93%. Furthermore, it yielded a precision of 98.70%, a recall of 98.10%, and an F1-score of 98.40%, outperforming other standalone algorithms in accuracy and stability.

Index Terms – Chronic Kidney Disease (CKD), Machine Learning, Stacking Ensemble, AdaBoost, Classification, Clinical Data, Model Evaluation, Healthcare Prediction, Supervised Learning, Feature Engineering

MILESTONE RESEARCH.IN OPEN ACCESS

Electronic ISSN: **2584-072X**

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a persistent illness in which kidney function gradually declines, making it harder for the body to remove waste and excess fluids [1]. When the infection worsens, it may cause serious complications, including high blood pressure, anemia, weakened bones, heart disease, and eventually, complete kidney failure. CKD is particularly dangerous because its early stages often show no noticeable symptoms, allowing damage to accumulate unnoticed over time [2]. Approximately 10% of individuals worldwide suffer from CKD, and its incidence is continuously rising [3]. It is anticipated to be among the leading causes of mortality globally by 2040 [4]. Unfortunately, many people still cannot afford treatment choices like dialysis or kidney transplantation, especially in lower-income areas with inadequate healthcare resources and infrastructure [5].

Early detection is crucial to restricting the advancement of CKD [6]. Still, traditional diagnostic approaches—such as lab tests, imaging, and physician judgment—can be expensive, time-consuming, and sometimes inconsistent due to human interpretation. In response to these limitations, machine learning has gained attention as a tool to improve medical diagnostics. By analyzing patterns in patient data, ML models can assist in predicting disease presence and severity more quickly and reliably. However, many powerful ML models, especially deep learning approaches, come with high computational demands and require large amounts of data, not always feasible in medical environments with limited resources [7]. This study introduces a lightweight stacking-based ensemble model for CKD prediction to overcome these challenges. The goal is to build a more accessible yet accurate system that performs well even with smaller datasets. The proposed approach combines three base models—Decision Tree, Naïve Bayes, and Gradient Boosting—and uses AdaBoost as the meta-classifier to generate final predictions. This model aims to reduce overfitting, a prevalent problem in medical datasets, while increasing accuracy. To better grasp how effectively the model manages unbalanced data, we verify it using the UCI CKD dataset and assess its performance using accuracy, Cohen's Kappa Score, and Matthews Correlation Coefficient (MCC).

The key contribution of our work is as follows:

- Proposed a Stacking Ensemble Model that combines multiple traditional classifiers to improve CKD prediction accuracy while reducing computational overhead.
- Deep learning approaches were used to create a lightweight architecture appropriate for smaller medical datasets, addressing the problems of overfitting and excessive resource needs.
- AdaBoost Classifier was used as a meta-learner to maximize a selection of effective base learners, such as Decision Tree, Naïve Bayes, Gradient Boosting, and others, inside the stacking framework.
- The UCI CKD dataset was used to train and assess the models, proving the stacking model's resilience when dealing with actual clinical data. Applied Cross-Validation Techniques to mitigate overfitting and ensure generalization, which is crucial in healthcare datasets with limited samples.
- Employed Comprehensive Evaluation Metrics**,** including Accuracy, Cohen’s Kappa, and Matthews Correlation Coefficient (MCC) for a more nuanced understanding of model performance, especially under class imbalance.

The rest of the article is organized as follows: Section II reviews existing literature and techniques in CKD prediction. Section III details the proposed stacking architecture and its implementation. Section IV provided the experimental setup and comparative outcomes. The work is concluded in Section V, which also recommends future research areas to improve early CKD diagnosis.

II. LITERATURE SURVEY

The rising concern in machine learning (ML) and deep learning (DL) methodologies for chronic kidney disease (CKD) diagnosis and prognosis has led to the development of diverse computational models that exploit clinical, demographic, and biochemical indicators. Metherall et al. [8] explored the effectiveness of Artificial Neural Networks (ANNs) and Random Forests (RFs) for CKD classification and creatinine estimation. They employed a dataset of 400 patients segmented into three feature groups—at-home, monitoring, and laboratory data—consisting of 25 attributes. Using 10-fold cross-validation, RF demonstrated superior performance in at-home feature classification with an accuracy of 92.5%, outperforming ANN's 82.9%. However, ANN showed a higher sensitivity (92.0%) but a lower specificity (67.9%) compared to RF's 90.0% sensitivity and 95.8% specificity. Both models surpassed 98% accuracy for laboratory and monitoring data, while creatinine regression yielded an R^2 score improvement of about 0.3 when using laboratory features over at-home data.

A hybrid framework that combines Convolutional Neural Networks (CNNs) for feature extraction and SVMs for classification was introduced by Ramu et al. [9]. This integration aimed to improve CKD diagnostic precision on a clinical dataset of ten key medical indicators. The Synthetic Minority Over-sampling Technique (SMOTE) was used to resolve class imbalance. The model achieved 96.8% accuracy, outperforming standalone SVM (94.8%) and RF (94.6%) models, with a perfect recall of 1.00 for CKD cases. However, despite the high predictive power, the model was computationally intensive, suggesting a trade-off between performance and operational efficiency. Pechprasarn et al. [10] leveraged a publicly accessible dataset from Enam Medical College with 200 patient records and 28 clinical features. After data balancing and an 80:20 train-test split, 22 ML models were benchmarked. As the most effective, Kernel Naïve Bayes maintained resilience on unknown data with 92.86% test accuracy and acquired 96.55% training accuracy. Developed feature selection methods were utilized to identify the most significant predictors, enhancing interpretability and model performance. These methods included Chi2, ANOVA, and minimum-redundancy-maximum-relevance (mRMR).

Rahman et al. [11] conducted a comparative study of eight ensemble classifiers trained on CKD data from the UCI repository. Borderline-SVM-SMOTE was used to balance the dataset after it had been preprocessed using Multivariate Imputation by Chained Equations (MICE). In feature selection, the Boruta method was surpassed by Recursive Feature Elimination (RFE), which reduced feature space without sacrificing performance. LightGBM was the most successful model among those assessed, with remarkable accuracy of 99.75%, precision of 99.40%, and AUC-ROC of 99.57%. In addition, the suggested approach significantly improved detection rate over earlier studies. A thorough ML-based CKD prediction framework (ML-CKDP) that combines numerous classifiers with meticulous preprocessing was proposed by Rezk et al. [12]. Their methodology comprised feature normalization, missing value imputing, and categorical data transformation. Various selection techniques—such as Lasso and Ridge regression, Correlation, and Sequential Forward Selection—were utilized to refine input features. Seven

models were assessed under different validation splits, including RF, AdaBoost, Gradient Boosting, and SVM. RF and AdaBoost achieved 100% accuracy and AUC across all experimental configurations, underscoring the model's stability and predictive strength.

Incorporating explainable artificial intelligence (XAI), Ghosh et al. [13] investigated predictive modeling using clinical data from 491 patients, including 56 CKD cases. XGBoost outperformed the other five ML algorithms evaluated, with 93.29% accuracy and an AUC of 0.9689. Crucially, the significance of different traits was interpreted using SHAP and LIME, providing transparency in clinical decision-making. Further extending this work, Ghosh et al. [14] benchmarked traditional and ensemble ML techniques using the UCI CKD dataset, including a novel Hybrid Model. According to their results, the Hybrid Model continuously produced the best overall performance, demonstrating its resilience and generalizability across clinical environments with 94.99% accuracy, 95.21% precision, and 95.56% AUC.

After reviewing the entire literature, we found several crucial limitations, they are:

- Several studies, such as Pechprasarn et al. [10] and Metherall et al. [8], used small datasets (200–400 patients), which may not be representative of broader patient populations and could limit model generalizability to real-world clinical settings.
- Despite achieving great accuracy, Ramu et al. [9] pointed out that their hybrid CNN-SVM model came with substantial computational costs, which would limit its use in real-time or low-resource clinical environments.
- Despite using SMOTE and borderline-SMOTE, Rahman et al. [11] and Ramu et al. [9] acknowledged that imbalanced datasets could still affect the performance and fairness of classifiers, especially in minority class predictions.
- Halder et al. [12] reported 100% accuracy using Random Forest and AdaBoost, raising potential concerns about overfitting—particularly since such perfect performance is rarely observed in real clinical data.
- Most studies, including those by Ghosh et al. [13, 14], used internal validation techniques like k-fold cross-validation without testing on external or multi-center datasets, which restricts their robustness across varied demographics and healthcare systems.
- Although Ghosh et al. [13] incorporated SHAP and LIME for model explainability, many other studies did not prioritize interpretability, which is critical for trust and adoption in clinical decision-making.
- None of the reviewed models, including those with high accuracy like LightGBM in Rahman et al. [11], addressed the need for online or incremental learning to adapt to evolving clinical data streams over time.
- Halder et al. [12] and Rahman et al. [11] used multiple feature selection techniques, but inconsistent selection strategies and overlapping feature sets could lead to redundancy or omission of critical indicators.

III. METHODS & MATERIALS

This section provides in-depth information about the study and a thorough summary of the research techniques employed. Figure 1 shows the suggested CKD prediction algorithm graphically.

A. Dataset Description

In this study, we worked with the chronic kidney disease (CKD) dataset available from the UCI Machine Learning Repository, which includes records for 400 patients. Each entry contains 14 health-related features, and the data is labeled to indicate whether the patient has CKD (marked as 1) or not (marked as 0). We explored the dataset in detail through an extended Exploratory Data Analysis (EDA). This enabled us to recognize essential imbalances, comprehend the links between characteristics, and locate patterns. A glaring class disparity needed to be rectified while assessing models since 250 recordings had the designation "CKD," but 150 did not. To further show the skewed distribution, we employed a histogram, which is shown in Figure 2. Table 1 displays the data types of attributes and their sets.

Table 1: Data Types of Attributes and Their Sets

Attribute	Full Form	Nonempty Values	Missing Values
age	Age	400	0
bp	Blood pressure	400	0
sg	Specific gravity of urine	400	0
al	Level of aluminum	400	0
su	Sugar level	400	0
bgr	Blood glucose (random)	400	0
bu	Blood urea	400	0
sc	Sugar level (Serum Creatinine)	400	0
sod	Amount of sodium	400	0
pot	Amount of potassium	400	0
hemo	Hemoglobin	400	0
pcv	Packed cell volume	400	0
wc	White cell count	400	0
rc	Red cell count	400	0

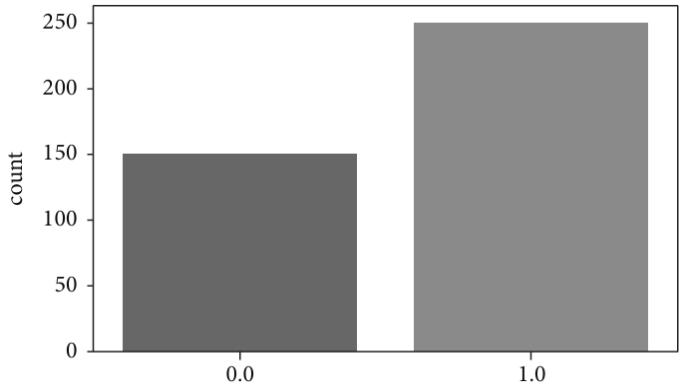

Fig. 1: Graphical representation of the Histogram plot

After that, we cleaned the data and looked for missing values in each feature. After locating them, we either eliminated the records if the missing data was too substantial or filled them in using statistical techniques like mean or median imputation. Next, we employed label encoding to transform categorical variables into numeric form to make categorical variables usable for machine learning algorithms. The dataset was then split 80:20 across the testing and training groups. We also used 10-fold cross-validation throughout training to make our judgment more reliable[16][17].

B. Methodology

1. Base Learners Selection and Implementation

We chose three distinct and complementary machine learning classifiers—Decision Tree, Gradient Boosting, and Naïve Bayes—as base learners to guarantee robust fundamental performance. These algorithms were selected because of their effectiveness on clinical datasets, interpretability, and varied decision bounds.

- **Naïve Bayes (NB):** The Bayes theorem is the foundation for the probabilistic classifier Naïve Bayes, which assumes feature independence. Given a class label y and a feature vector x = (x_1, x_2,...,x_n), the model calculates the posterior probability P(y|x) using:

$$P(y|x) = \frac{P_{(y)} \prod_{i=1}^{n} P(x_i|y)}{P(x)}$$

- **Gradient Boosting (GB):** GB sequentially creates weak learners by reducing the loss function employing gradient descent. For a model G_p at stage p, the following model is updated as:

$$G_{p+1}(x) = G_p(x) + h_p(x)$$

where, $h_p(x)$ is the base learner trained to fit the residuals $y - G_p(x)$. This iterative correction mechanism allows the ensemble to improve accuracy progressively.

- **Decision Tree (DT):** The DT classifier recursively splits the feature space based on information gain or Gini impurity to form a tree structure [15]. The model aims to find splits that most effectively reduce the entropy H of the classification:

$$\text{Information Gain (S, A) H (S)} - \sum_{v \in values\ (A)} \frac{|S_v|}{|S|} H\ (S_v)$$

where, S is the dataset and A is the splitting attribute. Decision Trees are valuable due to their interpretability and ability to capture non-linear patterns in the data.

2. Proposed Stacking Ensemble Model

We propose a stacking ensemble model to enhance predictive performance while minimizing overfitting, especially when working with limited, imbalanced medical datasets. Unlike traditional

ensemble methods like bagging or boosting, which rely on homogeneous base learners or sequential boosting, stacking integrates heterogeneous classifiers at the base level. It uses a meta-learner to combine their outputs optimally. This hierarchical learning structure is well-suited for complex classification tasks such as identifying CKD, where feature interactions and nonlinear boundaries are significant.

- **Conceptual Foundation of Stacking:** Stacking, also known as stacked generalization, operates on a two-layer architecture:

  - Level-0 (Base Learners): The same training data is used to train several different machine learning models separately. Each model generates predictions on unseen samples.
  - Level-1 (Meta-Learner): A new model (called the meta-learner or blender) is trained using the predictions of the base learners as its input. Its task is to learn how to combine these predictions to maximize overall accuracy.

  This meta-learning approach allows the model to compensate for the weaknesses of individual learners by integrating their outputs in a strategic, learned fashion.

- **Architecture of the Proposed Model:** Our proposed model employs three base learners and one meta-learner arranged as follows:

  - Base Learners (Level-0):

    i. Naïve Bayes (NB): A probabilistic model renowned for being easy to use and effective with tiny datasets.
    ii. Gradient Boosting (GB): In an additive model, residuals are optimized at each stage as a group of weak learners (often decision trees) are constructed.
    iii. Decision Tree (DT): A non-parametric model that splits data hierarchically based on information gain, ideal for modeling complex interactions.

  - Meta-Learner (Level-1):

    i. AdaBoost Classifier: Selected as the final blending model, AdaBoost adaptively assigns weights to base learner outputs, focusing more on harder-to-classify instances in subsequent iterations. It is robust against overfitting and suitable for imbalanced data scenarios like CKD classification.

- **Workflow of the Stacking Process:** The overall workflow of the proposed stacking model is detailed below:

  - **Training Phase (Base Learners):**

i. The original dataset D= $\{(x_i, y_i)\}_{i=1}^{N}$ is split into training and test sets using an 80-20 ratio.
ii. Each base learner is trained independently on the training set and makes predictions either on a validation subset or via k-fold cross-validation.
iii. For every instance, the $M_1(x_i), M_2(x_i), M_3(x_i)$ are obtained from the NB, GB, and DT models, respectively.

- **Feature Construction for Meta-Learner:**

i. The prediction outputs from the base models form a new feature vector for each instance:

$$z_i = [M_1(x_i), M_2(x_i), M_3(x_i)]$$

ii. These vectors, along with their true labels y_i, construct a new training dataset D′ = $\{(z_i, y_i)\}$.

- **Training the Meta-Learner:**

i. The AdaBoost classifier is trained on D′, learning to weigh and combine the base model predictions.
ii. AdaBoost minimizes the exponential loss over sample weights:

$$L_{Ada} = \sum_{i=1}^{N} w_i \, exp\,(-y_i f\,(z_i))$$

where, w_i are adaptively updated after each round and $f\,(z_i)$ is the predicted output

- **Prediction Phase:**

i. For new unseen test data, the process is repeated: predictions are generated from the base models and combined into a new feature vector.
ii. This vector is then input into the trained meta-learner to produce the final classification result.

When it comes to CKD prediction, the stacking model offers several significant benefits:

- **Heterogeneity:** The ensemble may capture a greater range of data patterns and correlations by using several foundation-layer techniques.
- **Overfitting Mitigation:** By leveraging cross-validation and a meta-learner, stacking reduces the risk of overfitting commonly seen in single-model learning, especially with small datasets.
- **Error Compensation:** The weaknesses of individual classifiers are addressed by the strengths of others, leading to improved generalization.
- **Model Interpretability and Flexibility:** Each component can be individually evaluated and fine-tuned, offering flexibility for future enhancements or dataset-specific tuning.

IV. RESULTS AND DISCUSSIONS

A. Evaluation Metrics

For our work, we employed several evaluation metrics such as accuracy, precision, recall and F-score.

Accuracy: Accuracy: $\frac{Tp+TN}{Tp+TN+FP+FN}$

Precision: Precision: $\frac{TP}{TP+FP}$

Recall: Recall: $\frac{TP}{TP+FN}$

F1-Score: F1-score: 2 $\frac{Precision.Recall}{Precision + Recall}$

Cohen's Kappa (κ): $\boldsymbol{\kappa = \frac{P_o - P_e}{1 - P_e}}$

Matthews Correlation Coefficient (MCC): $\boldsymbol{MCC = \frac{(TP.TN) - (FP.FN)}{\sqrt{(TP+FP)\,(TP+FN)(TN+FP)(TN+FN)}}}$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

B. Exploratory Analysis and Feature Visualization

We applied the Pearson correlation method to determine which features are most effective in predicting chronic kidney disease. This statistical approach helped pinpoint the strongest relationships between the 14 input variables and the target outcome. The correlation matrix shown in Figure 3 provides a clear summary of these associations.

As part of our exploratory data analysis, we used pair plots to visually examine how different variables relate to each other. These plots are especially useful for spotting trends, clusters, or outliers among both categorical and continuous data. Seaborn, a Python visualization library, was used to generate these visuals. The plots that arise, which are displayed in Figures 4 through 7, provide a simple and eye-catching method of examining the dataset.

Milestone Transactions on
MEDICAL TECHNOMETRICS
Vol. 03 | Issue. 02 | 2025

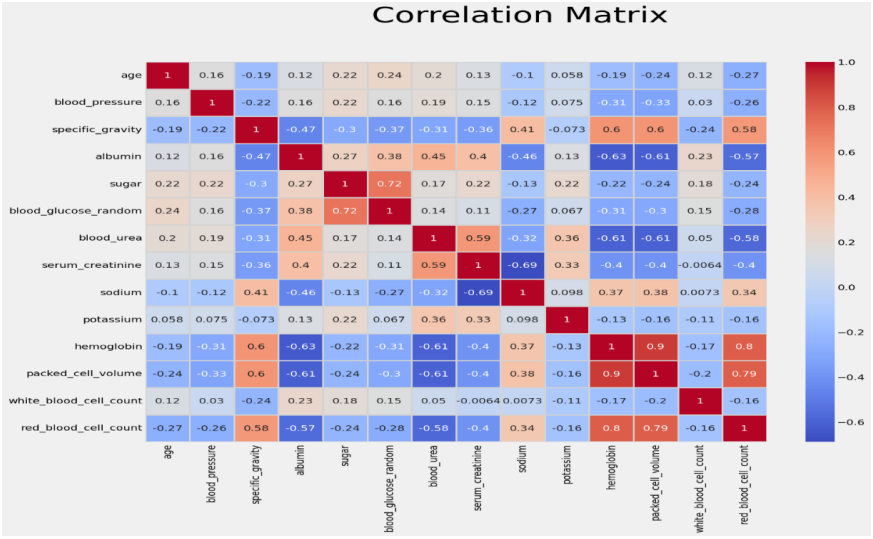

Fig 2: Pearson correlation matrix

Fig. 3: Red blood cell count by class

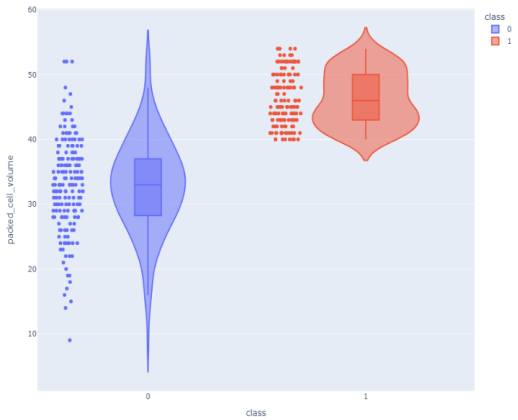

Fig. 4: Packed cell volume by class

Figure 4 focuses on red blood cell (RBC) counts. It shows how RBC levels vary across different patient categories. The plot combines individual data points with box plots to make patterns and deviations easier to identify. Hover functionality also reveals extra clinical information for each point, adding more context to the observations. Figure 5 displays the packed cell volume (PCV) distribution among the same classes. Here, violin plots show each group's range and concentration of values. Including raw data points alongside summary statistics makes spotting group-specific trends and anomalies easier.

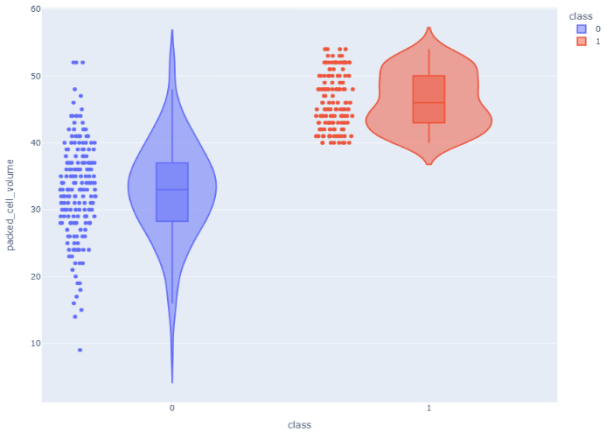

Fig. 5: Hemoglobin levels by class

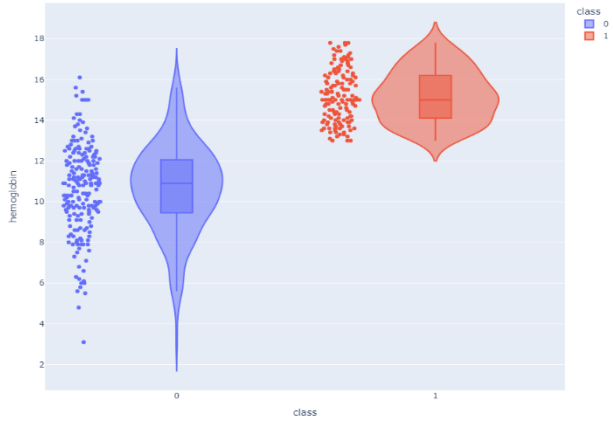

Fig. 6: Serum creatinine levels by class

Figure 5 presents the hemoglobin levels across classes. The shape and spread of the distributions are visible, and once again, hover data provides extra details that support a deeper analysis of each case. Figure 6 shows the serum creatinine level distribution, emphasizing the variations among patient groups. Individual data points and distribution summaries are combined in the graphic to provide a clear illustration of variability. This is a useful tool for comprehensive data interpretation since it allows you to hover over additional clinical parameters.

C. Performance of the Models

A model's performance must be evaluated using training and testing accuracy to determine how well it generalizes outside of the data it was trained on. A model's training accuracy shows how well it fits the data already seen, whereas testing accuracy shows how well it performs on unknown samples. A large gap between the two often indicates overfitting—where the model memorizes training data patterns but fails to generalize. Conversely, low accuracy in both training and testing may point to underfitting, meaning the model is too simplistic to learn meaningful patterns. Table 2 displays the training and testing accuracy of the models. In this study, Naïve Bayes showed the lowest performance among all models, with a training accuracy of 93.75% and a testing accuracy of 91.20%. The small gap suggests the model generalized decently but could not capture complex relationships, which likely caused underfitting. Decision Tree performed slightly better with 97.50% training accuracy but dropped to 92.40% on the test

set, suggesting moderate overfitting due to its tendency to form deep trees that can memorize training samples.

Table 2: Training and testing accuracy of the models

Model	Training Accuracy (%)	Testing Accuracy (%)
Naïve Bayes	93.75	91.20
Decision Tree	97.50	92.40
Gradient Boosting	98.30	94.60
XGBoost	98.90	96.20
AdaBoost	96.80	93.10
Proposed Stacking Model	99.20	98.93

AdaBoost showed improved balance, achieving 96.80% training accuracy and 93.10% testing accuracy. While better than Naïve Bayes and Decision Tree, the slight overfitting suggests that boosting helped reduce bias but still struggled with noisy data. Gradient Boosting further closed this gap, achieving 98.30% and 94.60%, respectively, which reflects better learning of patterns and improved control over overfitting. XGBoost outperformed these models, demonstrating strong regularization with a near-ideal balance—98.90% training and 96.20% testing accuracy. The proposed stacking model delivered the highest accuracy, with 99.20% training and 98.93% testing performance. This minimal gap reflects excellent generalization and robustness. The stacking model successfully reduced variance and bias by combining predictions from base models (Naïve Bayes, Gradient Boosting, Decision Tree) through a meta-learner (AdaBoost). To prevent overfitting and capture a broader range of patterns, the ensemble compensated for the shortcomings of individual learners. The stacking ensemble was the most dependable option for CKD prediction in our investigation. It not only had the best overall accuracy but also showed consistent performance across datasets as shown in Figure 8**.**

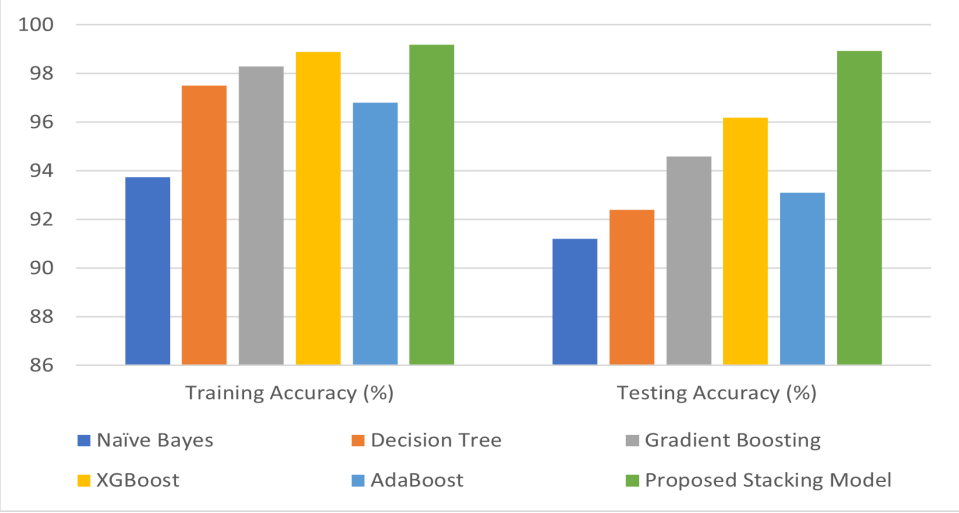

Fig. 7: Training and testing accuracy performance of the models

Relying solely on accuracy is often insufficient to objectively assess the performance of machine learning models in medical diagnostics, especially in conditions like chronic kidney disease (CKD). Metrics such as precision, recall, and F1-score provide a more granular understanding of a model's capability to accurately diagnose illness in patients (positive class) versus healthy individuals (negative

class) which presented as in Table 3. This is critical in healthcare, where false negatives (missing a CKD patient) can be far more dangerous than false positives.

Table 3: Performance of the models

Model	Precision (%)	Recall (%)	F1-Score (%)
Naïve Bayes	89.20	91.50	90.30
Decision Tree	93.00	90.80	91.80
AdaBoost	94.40	92.10	93.20
Gradient Boosting	96.50	94.20	95.30
XGBoost	97.80	95.10	96.40
Stacking Model	98.70	98.10	98.40

The need to evaluate training and testing accuracy arises from a common challenge in machine learning: overfitting and underfitting. A model that performs exceedingly well on training data but poorly on test data is likely overfitting—it memorizes the training samples but fails to generalize. Conversely, underfitting occurs when a model performs poorly on both training and test sets, suggesting it cannot capture the underlying patterns in the data. Among the evaluated models**,** Naïve Bayes showed relatively low precision and F1-score, although its recall was decent. This suggests it tends to predict positives more liberally, possibly causing a higher rate of false positives. While useful for rapid baseline predictions, its simplistic assumptions limit its effectiveness for nuanced classification in healthcare settings.

The Decision Tree classifier improved upon Naïve Bayes in all metrics. Still, it exhibited a slight imbalance between precision and recall, indicating some misclassification of healthy individuals as patients and vice versa. Its tendency to overfit was noticeable, with training accuracy significantly higher than testing accuracy. AdaBoost and Gradient Boosting demonstrated more balanced and improved results, benefiting from iterative learning and error correction. However, their performance showed mild sensitivity to training data variance, and while their recall scores were high, the gap between training and testing performance suggested slight overfitting.

XGBoost, known for its regularization and scalability, showed a significant leap in accuracy and F1 score. Its optimization of tree structures and effective handling of outliers and missing values made it one of the most stable performers. However, due to its complexity, it requires careful hyperparameter tuning to avoid overfitting, which, in our evaluation, was slightly visible in the training-to-testing accuracy difference. Finally, the proposed Stacking Mode**l** outperformed all individual models, which integrated Naïve Bayes, Gradient Boosting, and Decision Tree as base learners with AdaBoost as the meta-classifier. It achieved a testing accuracy of 98.93%, along with high precision (98.70%), recall (98.10%), and F1-score (98.40%). As seen in Figure 9, this method is effective because it maximizes the diverse option boundaries and strengths of each base learner while minimizing their limitations through ensemble learning. Most importantly, it showed low overfitting and high generalization by remaining constant between training and testing sets.

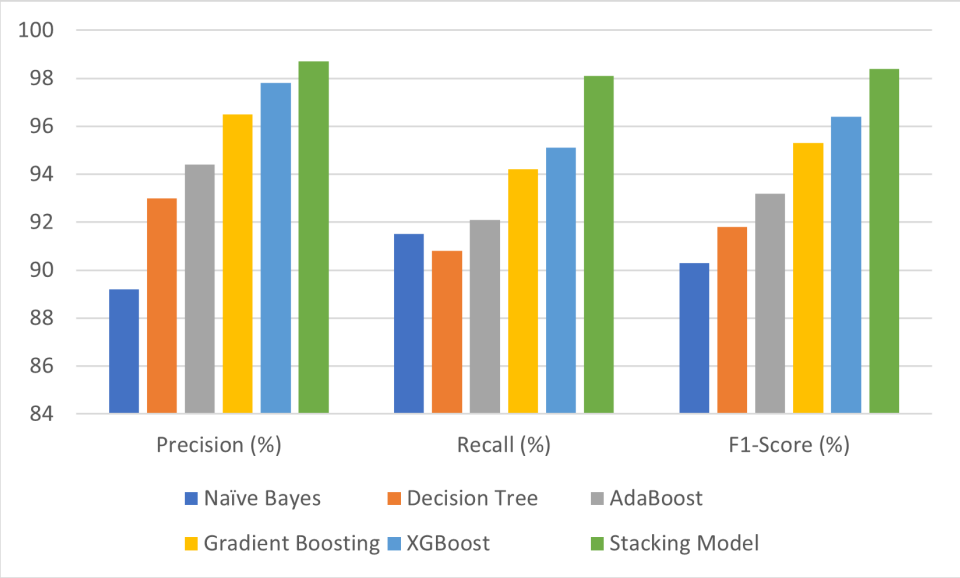

Fig. 8: Performance of the models

V. CONCLUSION AND FUTURE WORK

In this work, we introduced a reliable machine learning-based method for leveraging clinical data to predict chronic kidney disease (CKD) early. After assessing several conventional classifiers, such as Naïve Bayes, Decision Tree, Gradient Boosting, AdaBoost, and XGBoost, we created a stacking ensemble model that combines the advantages of numerous base learners. This ensemble approach performed better on all necessary evaluation measures using AdaBoost as a meta-classifier. The stacking model outperformed all baseline models with a remarkable testing accuracy of 98.93% and good precision, recall, and F1-score. These findings demonstrate how ensemble learning, particularly stacking, may help reduce overfitting issues and guarantee accurate generalization of unknown data.

Although the suggested paradigm appears promising, there are a number of issues with this approach. The modest size of the dataset may mean that it does not accurately reflect the variety of CKD patients in the broader community. To ensure broader use, we wish to test our method on larger, multi-source clinical datWeequential comprehension, we also want to research hybrid models and deep learning architectures that include tem to improve sequential understandingporal health information. Integrating explainability frameworks (e.g., SHAP, LIME) will also be a priority to enhance trust and interpretability for clinical use. Finally, deployment as a real-time diagnostic tool embedded within healthcare platforms will be considered for practical implementation in hospital settings.

REFERENCES

1. Pathak, A., Singh, O. P., & Biswas, U. (2025). Predicting chronic kidney disease using machine learning methodologies. In *2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)* (pp. 1–6). IEEE.
2. Ahmed, K., Dubey, M. K., Dubey, S., Pandey, D. K., et al. (2025). Chronic kidney disease: Causes, treatment, management, and future scope. In *Computational Intelligence for Genomics Data* (pp. 99–111). Elsevier.
3. Kounatidis, D., Vallianou, N. G., Stratigou, T., Voukali, M., Karampela, I., & Dalamaga, M. (2024). The kidney in obesity: Current evidence, perspectives and controversies. *Current Obesity Reports, 13*(4), 680–702.

4. Wang, Y., Lu, Y., Ma, J., Hu, Y., Zheng, R., Liu, L., Lin, K., Zhang, K., & Cai, H. (2024). The global burden of chronic kidney disease (CKD) attributable to high sodium intake: A comprehensive analysis of trends from 1990 to 2021 and burden prediction to 2040. *SSRN*. https://ssrn.com/abstract=5242625
5. Sinha, R., Noh, L., Sethi, S. K., Safadi, R., Smith, S., Düzova, A., Bjornstad, E. C., Antwi, S., Ishikura, K., Salgia, E., et al. (2025). Pediatric kidney replacement therapies in low-to-middle income countries: A review and white paper. *Pediatric Nephrology*, 1–17.
6. Lee, P.-H., Huang, S. M., Tsai, Y.-C., Wang, Y.-T., & Chew, F. Y. (2025). Biomarkers in contrast-induced nephropathy: Advances in early detection, risk assessment, and prevention strategies. *International Journal of Molecular Sciences, 26*(7), 2869.
7. Razzaq, K., & Shah, M. (2025). Machine learning and deep learning paradigms: From techniques to practical applications and research frontiers. *Computers, 14*(3), 93.
8. Metherall, B., Berryman, A. K., & Brennan, G. S. (2025). Machine learning for classifying chronic kidney disease and predicting creatinine levels using at-home measurements. *Scientific Reports, 15*(1), 4364.
9. Ramu, K., Patthi, S., Prajapati, Y. N., Ramesh, J. V. N., Banerjee, S., Rao, K. B., Alzahrani, S. I., et al. (2025). Hybrid CNN-SVM model for enhanced early detection of chronic kidney disease. *Biomedical Signal Processing and Control, 100*, 107084.
10. Pechprasarn, S., Wetchasit, P., & Pongsuwan, S. (2025). Optimizing chronic kidney disease prediction: A machine learning approach with minimal diagnostic predictors. *Journal of Current Science and Technology, 15*(1), 76–76.
11. Rahman, M. M., Al-Amin, M., & Hossain, J. (2024). Machine learning models for chronic kidney disease diagnosis and prediction. *Biomedical Signal Processing and Control, 87*, 105368.
12. Rezk, N. G., Alshathri, S., Sayed, A., & Hemdan, E. E.-D. (2025). Explainable AI for chronic kidney disease prediction in medical IoT: Integrating GANs and few-shot learning. *Bioengineering, 12*(4), 356.
13. Ghosh, S. K., & Khandoker, A. H. (2024). Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific Reports, 14*(1), 3687.
14. Ghosh, B. P., Imam, T., Anjum, N., Mia, M. T., Siddiqua, C. U., Sharif, K. S., Khan, M. M., Mamun, M. A. I., & Hossain, M. Z. (2024). Advancing chronic kidney disease prediction: Comparative analysis of machine learning algorithms and a hybrid model. *Journal of Computer Science and Technology Studies, 6*(3), 15–21.
15. Prova, N. N. I. (2024). Advanced machine learning techniques for predictive analysis of health insurance. In *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)* (pp. 1166–1170). IEEE.
16. Pasha, A., ur Rahman, S. Z., Tauheed, S., Basha, S. M., & Anwar, B. H. (2024). Comparative Analysis of Ranking Machine Learning Classifier Models for Parkinson's Disease (PD) Prediction. In *Disruptive Technologies for Sustainable Development* (pp. 237-241). CRC Press.
17. Kumar, A., Satheesha, T. Y., Salvador, B. B. L., Mithileysh, S., & Ahmed, S. T. (2023). Augmented Intelligence enabled Deep Neural Networking (AuDNN) framework for skin cancer classification and prediction using multi-dimensional datasets on industrial IoT standards. *Microprocessors and Microsystems*, *97*, 104755.