Chronic Kidney Disease Prediction using Machine Learning: A Comparative Analysis of Ensemble Methods

Md. Sharfuddin

Abstract
Chronic Kidney Disease (CKD) has become a global health crisis, showing rapid increases across all age groups. Frequently termed a "silent killer" , CKD often evades detection in its early stages , necessitating complex diagno--stic assessments.While machine learning (ML) models offer promising pathways for early detection , achieving high predictive accuracy with limited datasets remains a formidable challenge , often complicated by noise and overfitting. This study proposes an optimized framework utilizing advanced feature engineering and data positioning to enhance model reliability. We evaluated the performance of nine distinct machine learning models—including LightGBM, SVC, and Random Forest—using metrics such as accuracy, precision, recall, F1-score, and Mean Squared Error (MSE). Our analysis demonstrates that the LightGBM model achieved an individual accuracy of 90.06% with an MSE of 0.0994. Furthermore, by employing ensemble techniques, we achieved a peak accuracy of 93.07%. These results suggest that with rigorous algorithmic optimization and sophisticated data processing, it is possible to diagnose CKD with high precision even in data-constrained clinical environments. This research provides a scalable, reliable diagnostic tool, marking a significant advancement in medical informatics and early disease management.

Keywords: Chronic Kidney Disease, Machine Learning, LightGBM, Ensemble Learning, Predictive Analytics, Medical Informatics, Early Detection.

1. Introduction

Chronic kidney disease (CKD) is a long-term condition whereby abnormalities of kidney structure or function are present for at least three months and have important clinical consequences [13] . So much so, that due to the slow progression of kidney dysfunction to kidney failure, cardiovascular complications and dialysis dependency, as well as needing a kidney transplantation or early death it has become an important public health problem. CKD (chronic kidney disease) is an enormous worldwide health problem affecting millions of people as reported by the World Health Organization and the burden has been increasing with associated diabetes hyperten--sion, cardiovascular disease and population ageing [14]. As CKD is usually asymptomatic during the early stages, a considerable number of patients remain undiagnosed until significant renal damage has already occurred [14]. It makes it necessary that CKD is detected at an early stage of the disease and classified correctly to ensure timely treatment, prevention of disease progression, and alleviation of healthcare burden in the long term.

The routine diagnosis of CKD is based on clinical assessment and laboratory markers including serum creatinine, estimated glomerular filtration rate (eGFR), albuminuria, blood pressure and other biochemical markers. WHO has stressed others that simple blood and urine testing, especially serum creatinine measurement, can aid in CKD detection within primary healthcare settings [14] . Yet, clinical data interpretation manually may consume time especially when numerous risk factors and biochemical variables influence one another. In real clinical practice, predicting CKD is not as straightforward due to non-linearities in patient data, missingness, class imbalance and the complexity of diagnostic patterns. The above challenges have motivated researchers to employ smart computational techniques in enhancing diseased diagnosis and clinical decision support.

With the ability to model latent patterns found in clinical datasets and classify patients using a combination of health information, machine learning has been identified as a promising tool for CKD prediction. Earlier works have

DOI: https://doi.org/10.5281/zenodo.20705210

indicated the well-known machine-learning algorithms being commonly used for CKD detection and prediction, including Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Naive Bayes, Decision Tree, Random Forests [15] and hybrid models. Furthermore, increasing systematic evidence further shows that predictive performance of artificial intelligence and machine-learning methods is also promising for predicting CKD progression and aiding early diagnosis [16] . These methods are effective because they can evaluate complex nonlinear feature interactions and provide predictive outputs which may help healthcare practitioners in screening and decisions.

The classification performance of individual machine-learning models may be only good to an extent, because their predictive quality and generalizability can vary by data distribution, model structure, feature quality, and class imbalance. Some classifiers might perform good to detect CKD-positive cases but not able to discriminate between various Non-CKD cases. To solve such limitations, one of the most powerful strategies that have been proposed in medical prediction tasks is ensemble learning. Ensemble methods perform better than standalone models by combining their outputs to increase prediction stability and lower approach bias, resulting in improved performance of classification. Ensemble and hybrid methods have been shown to yield robust CKD prediction performance with several single classifiers, especially among adult populations. [15], [16]

In this paper, nine base machine-learning classification models and four voting ensemble models are investigated. The models that were used as base models are Logistic Regression, KNN, SVC, Naive Bayes, Decision tree, Random forest,AdaBoost,XGBoost and LightGBM. The voting ensembles were comprised of combinations created from LightGBM, SVC and Random Forest based on the performance of each individual model. To evaluate the performance of models, confusion matrix analysis, ROC–AUC curves, accuracy/precision/recall/F1-score and mean squared error were performed. The current study investigated the potential of ensemble learning in improving classification performance over base classifiers by comparing both single and ensemble models to find a consistent machine-learning framework for CKD prediction.

The primary contribution of this work is the systematic comparison between a large number of base classifiers and voting ensemble combinations for classification of CKD. These empirically supported findings aim to contribute to the process of strategically constructing smart decision-support systems for earlier detection and management of CKD, ultimately reducing diagnostic delay. Although these medical machine-learning models can potentially demonstrate utility in the real world, they must first be validated with larger balanced independent clinical datasets as their clinical translational properties should include reliability, generalizability, and clinically interpretable.

2. Related Works

Several studies have explored machine learning approaches for chronic kidney disease (CKD) prediction using structured clinical, demographic, lifestyle, and laboratory data. Azizah and Paramitha. [1] used a Kaggle CKD dataset containing 1,659 patient records and applied a Gaussian Naive Bayes classifier after feature selection, scaling, and train-test splitting. Their work showed that even a relatively simple probabilistic classifier can provide acceptable predictive performance for early CKD identification. However, the study mainly focused on a single classifier, leaving room for broader algorithmic comparison and stronger model optimization.

More recently, Anagu et al. [2] used the same 1,659-record, 54-feature Kaggle CKD dataset and compared Random Forest, K-Nearest Neighbors, Support Vector Machine, and Logistic Regression. Their findings indicated that Random Forest achieved the best performance, demonstrating the strength of ensemble learning for high-dimensional and imbalanced medical data. Similarly, Anitha and Rao. [3] investigated CKD prediction using both UCI-CKD and CKD-15 datasets, applying Random Forest with feature ablation and cross-validation. Their work emphasized that numerical clinical variables such as serum creatinine, GFR, blood pressure, and other laboratory indicators are highly valuable for CKD prediction.

Earlier studies using the UCI CKD benchmark established the foundation for applying supervised learning to CKD diagnosis. Polat et

DOI: https://doi.org/10.5281/zenodo.20705210

al. [4] used Support Vector Machine with feature selection and demonstrated that reducing irrelevant features can improve diagnostic performance. Charleonnan et al. [5] compared several machine learning techniques and showed the feasibility of predictive analytics for CKD screening. Al Imran et al. [6] further compared Logistic Regression, Feedforward Neural Network, and Wide & Deep Learning, showing that deep and hybrid learning models can capture more complex relationships among CKD risk factors.

A number of later studies focused on ensemble and explainable methods. Sobrinho et al. [7] conducted a comparative analysis of machine learning techniques for computer-aided CKD diagnosis, especially considering application in developing countries. Qin et al. [8] proposed a machine learning methodology for CKD diagnosis and highlighted the importance of missing-value handling and clinical interpretability. Khan et al. [9] empirically evaluated multiple machine learning techniques for CKD prediction and reported strong performance from ensemble-based classifiers. Chittora et al. [10] provided a broader machine-learning perspective on CKD prediction, reinforcing that preprocessing, feature selection, and model choice strongly influence diagnostic accuracy.

Recent studies have moved toward more robust evaluation and explainable artificial intelligence. Islam et al. [11] compared several machine learning algorithms for CKD prediction and emphasized the importance of performance metrics beyond accuracy. Singamsetty et al. [12] integrated explainable AI into CKD forecasting, using model interpretation to identify influential clinical attributes and improve trust in machine-learning-based decision support. Overall, the literature shows that CKD prediction has progressed from basic classifiers to ensemble, deep learning, and explainable AI models. Nevertheless, research gaps remain in terms of external validation, class imbalance handling, feature explainability, and real-world clinical deployment. Therefore, the present study can contribute by applying a systematic preprocessing pipeline, comparing multiple models, and emphasizing interpretability on the selected CKD dataset.

3. Methodology

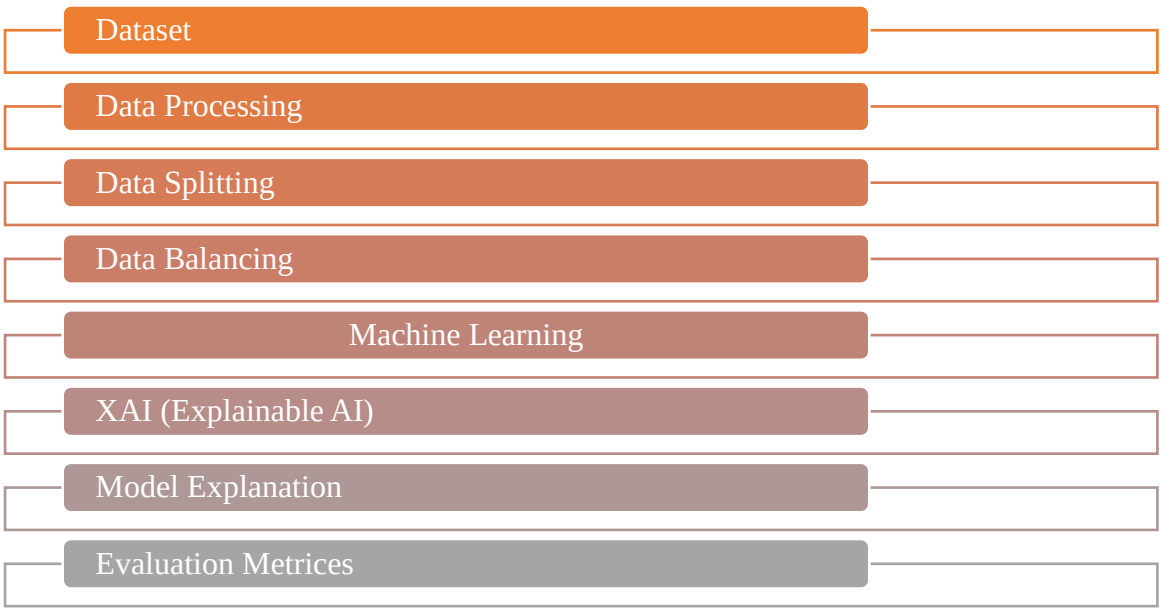


Fig 3.0: Methodology Process

This study contributes a novel systematic machine learning framework for CKD prediction, from the preparation of data to interpretable clinical diagnostics. It starts with Dataset Collection where patient clinical attributes are extracted and then undergoes a

DOI: https://doi.org/10.5281/zenodo.20705210

tedious Data Processing, missing value imputation, categorical encoding followed by feature engineering. For avoiding biased evaluation, the dataset is divided into separate sections through Data Splitting (70% training data and 20% testing data, along with 10% validation) method where a sample of each class builds validation set to correct any class imbalance in Data Balancing. We then train and compare diverse ML classifiers, closely integrated with Explainable AI (XAI) methodologies. This facilitates whole Model Explanation using SHAP and LIME to provide both global and local feature transparency with the overall performance of the framework measured through a comprehensive suite of clinical Evaluation Metrics.

3.1 Dataset

Using the Chronic Kidney Disease dataset available on kaggle which has 54 features. The dataset includes clinical and demographic information from 1,659 patients, which makes it a robust basis for performing analysis on risk factors associated with kidney impairment. The target variable is chronic kidney disease either present or absent which makes it a good binary classification benchmark dataset.

3.2 Data Processing

3.2.1 NaN*:* The missing values were handled to keep data quality high. High missingness percentage features were assessed utilizing domain-specific knowledge and statistical techniques (mean, median or mode imputation) to carry out data integrity with no sample size drops.

3.2.2 ***Data Encoding:*** Categorical features were encoded for ML use cases Nominal features underwent One-Hot Encoding, whilst ordinal features were represented through Label Encoding to convert qualitative clinical characteristics into quantitative values without creating any false hierarchies.

3.2.3 ***Feature:*** Feature engineering was performed to improve model input quality. We took best 24 feature for evaluate machine learning model performance.

3.3 Data Splitting

The data were separated into training (70%), testing (20%) and validation (10%) sets for strict evaluation of the model. A training set was utilized to fit the machine learning classifiers, a validation set supported hyperparameter tuning to avoid overfitting and the final test set was kept separate solely to evaluate generalizability and performance on out-of-sample data.

3.4 Data Balancing

All classes are weighted equal using data balancing methods. In order to deal with the natural imbalance between healthy and CKD groups, SMOTE (Synthetic Minority Over-sampling Technique) was used. This allows us to calculate evaluation metrics (F1-score, Sensitivity etc) that perfectly represent the true diagnosing power of the models.

3.5 Machine Learning:

We tested the machine learning models in two ways: once with the base model and once with the voting ensemble model. The Accuracy, Precision, Recall, F1 score, and MSE of each model were observed.

3.5.1 Machine Learning Model

We tested a total of 9 machine learning models and observed their performance.
These are:

- K Nearest Neighbor
- Support Vector Classifier
- Naive bayes
- Logistics Regression
- Decision Tree
- Random Forest
- Adaptive Boosting
- Extreme Gradient Boosting
- Light Gradient Boosting Machine

3.5.2 Ensemble Model

We selected the best 3 models based on accuracy and declared them as a,b,c to build an ensemble model. Where:

a = Best Accuracy
b = 2nd Best Accuracy
c = 3rd Best Accuracy
The 4 ensemble models are:
a+b = Best Accuracy + 2nd Best Accuracy
b+c = 2nd Best Accuracy + 3rd Best Accuracy
a+c = Best Accuracy + 3rd Best Accuracy
a+b+c = Best Accuracy + 2nd Best Accuracy + 3rd Best Accuracy

DOI: https://doi.org/10.5281/zenodo.20705210

Ensemble model Build for achieve highest value of Accuracy, Precision, Recall & F1-Score.

3.6 XAI (Explainable AI)

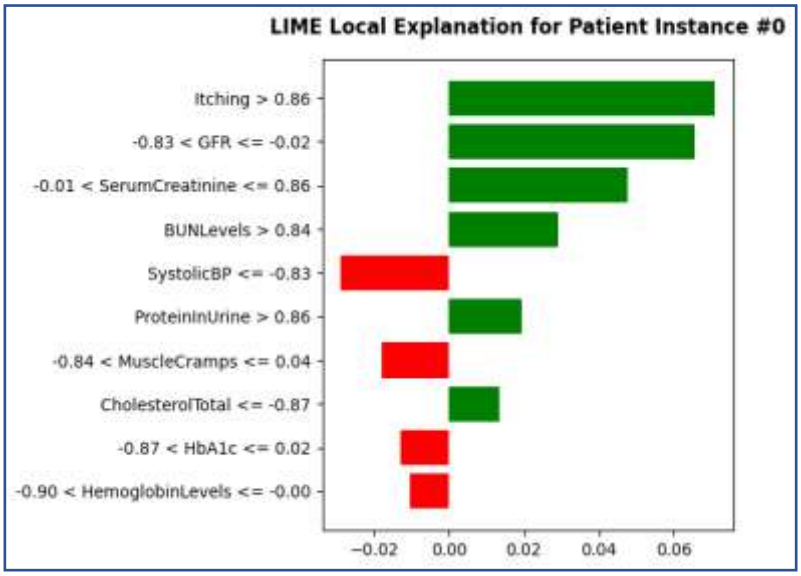
Future iterations will incorporate XAI to provide transparency into the diagnostic reasoning of the ensemble models.

Here, figure 3.1 shows the global importance of 24 clinical predictors that were implemented into the model to identify a case for CKD A long bar indicates that the feature was more important to the model prediction. Serum Creatinine is the feature that contributes most to the prediction, followed closely by GFR, Protein in Urine, Itching and BUN Levels. As expected, they are very well correlated to kidney function and hence the model relies heavily on them. The other features that are significant include Muscle Cramps, HbA1c, Fasting Blood Sugar, Systolic BP, Cholesterol HDL and BMI. There is very little feature scoring under Family History of Kidney Disease and Smoking in this chart. That is, they played a smaller role in prediction relative to the direct kidney-function markers.

So another feature importance figure, but average absolute SHAP values. It shows the average impact of top 20 most influential features on model prediction In this figure, GFR is the most crucial feature, Serum Creatinine next followed by Itching, Muscle Cramps, HbA1c and BUN Levels. It also takes into Protein.in.Urine, Systolic.BP, Family History of Kidney Disease, Fasting Blood Sugar, Age and Fatigue levels. This chart differs a little from the 24-feature importance plot. In the first graph, Serum Creatinine The first is and (in the SHAP chart) GFR expressed number two. This difference is expected, since SHAP depicts the real contribution of each feature to a model output.

This is a SHAP summary plot that shows the aggregation of how features affect the model prediction across many samples. Each dot represents one patient/sample. The x-axis is the SHAP value. A positive SHAP value has an increased chance of CKD prediction, while a negative SHAP value reduced the chances of CKD prediction. Such graph mainly focuses on Serum Creatinine and GFR. Serum Creatinine has a significant predictive power, as higher values of creatinine usually correlate with reduced kidney function. GFR comes up with a lot of importance too, as obviously lower GFR is identified commonly CKD. Figure a confirms this point: showing that the model is primarily leveraging kidney function biomarkers to predict CKD.

This is an effect of Serum Creatinine on the model prediction. In the figure, X-axis shows Serum Creatinine values and Y-Axis shows the SHAP value for Serum Creatinine. The SHAP value mostly negative when Serum Creatinine is low. So it decreases the chance of CKD prediction. This means that the SHAP value is positive when there is an increase in Serum Creatinine. That is, it enhances the prediction of CKD. GFR is represented as color, so this figure illustrates the interaction between Serum Creatinine and GFR. Overall, the combination of high Serum Creatinine with low GFR provides stronger evidence for CKD prediction.



DOI: https://doi.org/10.5281/zenodo.20705210

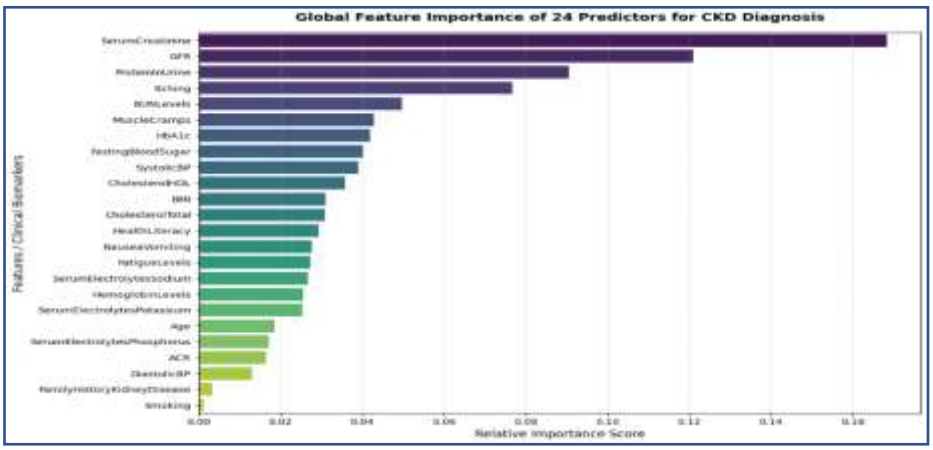


Fig 3.1: Feature Importance (24 Features)

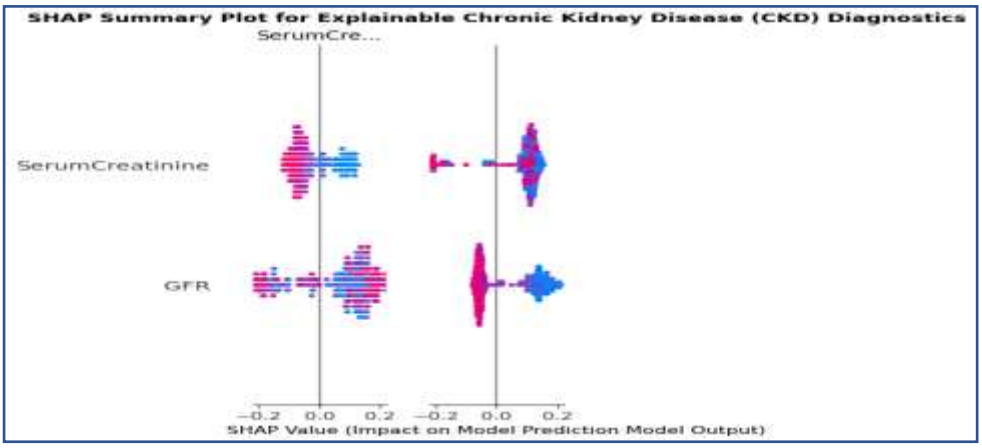


Fig 3.2: Feature Importance (Predicted By SHAP)

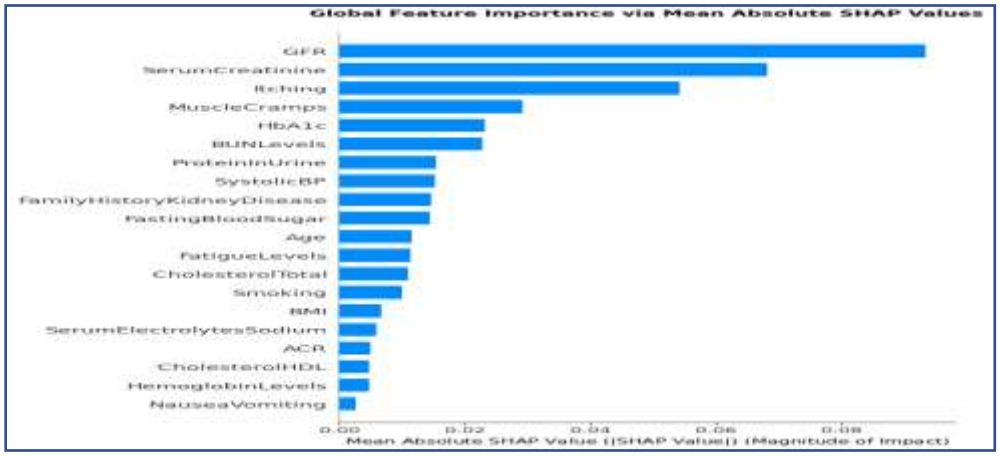


Fig 3.3: SHAP Summary Plot

DOI: https://doi.org/10.5281/zenodo.20705210

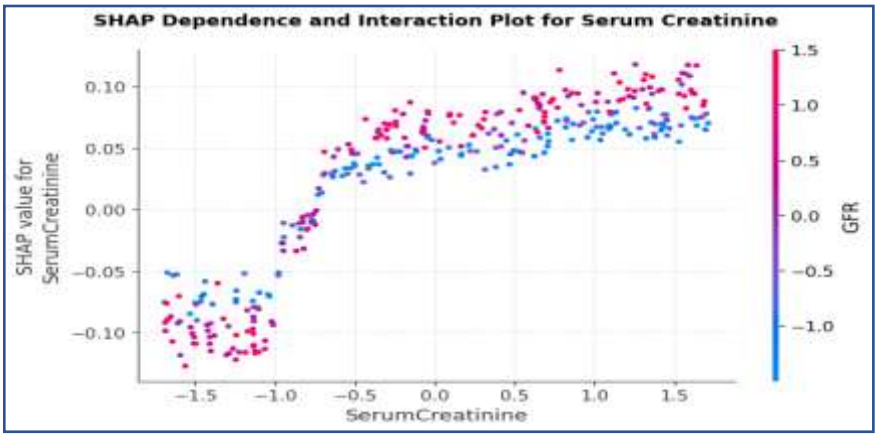


Fig 3.4: SHAP Dependence and Interaction plot for Seum Creatinine

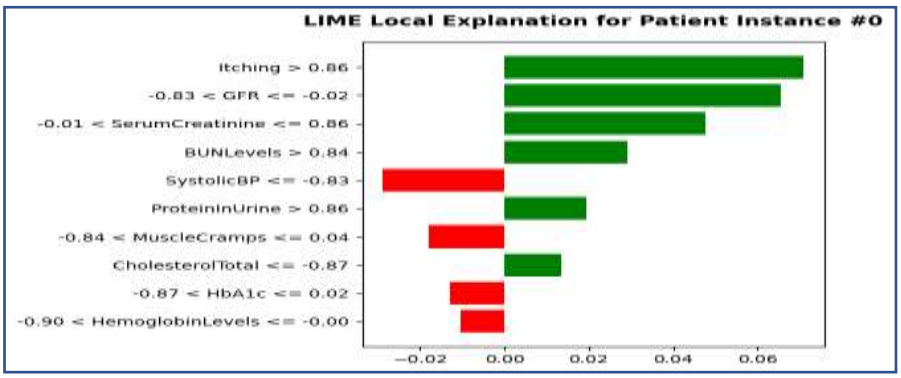


Fig 3.5: LIME Explainer

In this figure we can explain the model decision for this patient specifically. In contrast, LIME explains why a model predicted the label for a given instance — not global feature importance. The green bars supporting and the red bars against the CKD prediction by our model. For this patient, Itch>0.86, GFR range and Serum Creatinine range, BUN Levels >0.84 and Protein in Urine >0.86 all confirmed the CKD diagnosis. Conversely, Systolic BP and Muscle Cramps improved the prediction score each another 0.8 points by default as did Hemoglobin and HbA1c resulting in a total of 52 confidence scores In summary, this figure indicates that itching, GFR, creatinine, BUN and urine protein were the potential factors influencing model decision for this patient.

Overall Brief Summary

In general, the results indicate that the model is primarily driven by kidney-function-specific features. Most Influential Predictors are GFR, Serum Creatinine, Protein in Urine, BUN Levels, Itching, HbA1c,Blood Sugar, Blood Pressure,Musscle Cramps and Hemoglobin levels.

The 24-feature importance plot orders all predictors on a high level, while the SHAP top-20 figure dives deeper into understanding the features that matter most. Collectively, these numbers indicate that the model CKD prediction is mainly attributable to clinical biomarkers, urine-associated properties, metabolic components, blood influence and patient symptoms.

3.7 Model Explanation

Logistic Regression

Logistic Regression− Logistic regression is the only proper classification algorithm in this list

DOI: https://doi.org/10.5281/zenodo.20705210

of simple yet effective algorithms for binary prediction. It was used in this study to classify patients into CKD and Non-CKD groups based on clinical features.

K-Nearest Neighbors
Compared with the distance data points of the training dataset, K-Nearest Neighbors classifies a sample. The class is predicted based on the majority class of its neighbours.

Support Vector Classifier
A Support Vector Classifier looks for the optimal line that separates CKD from Non-CKD classes. It can be beneficial for coping with complex classification patterns.

Naive Bayes
Naive bayes being a probabilistic classifier first calculates the probability that which category an entry would belong to It's quick, it works for medical classification tasks.

Decision Tree
Decision Tree — It uses tree based structure of rules to classify data. It takes decisions iteratively on a value of the features and gives a human understandable classification output.

Random Forest
Random Forest is an ensemble model that combines a number of decision trees to give a more accurate and robust prediction. It compensates for the weakness of a single decision tree.

AdaBoost
AdaBoost combines the classification of multiple weak learner to achieve better predictions. It allows more emphasis on samples that are misclassified and reduces the prediction performance incrementally.

XGBoost
XGBoost is a sophisticated boosting algorithm that constructs models one after the other to fix the mistakes of earlier models. It is fast and works well on structured medical datasets.

LightGBM
LightGBM LightGBM is a fast gradient boosting model. It also accepts a more complicated relationship among features and is ideal for increasing CKD classification accuracy.

3.8 Evaluation Metrices

True Positive (TP) : model predicts a positive case correctly. This study defined TP (True positive) if a CKD patient can correctly gets classified as having CKD.

True Negative (TN): The model predicts a negative case correctly. For this study, TN stands for a Non-CKD patient is correctly predicted as Non-CKD.

False Positive (FP): Represents an incorrect prediction of a negative example being predicted as positive rather than correct. FP : Non-CKD is predicted as CKD in this study.

False Negative (FN) : where the model predicts a positive case as negative. In this study, FN shows that a CKD patient being labelled as Non-CKD.

Accuracy: Accuracy is a measure of the overall correctness of a model. It shows how many samples were correctly classified.

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

Precision: Precision is the fraction of relevant instances among the retrieved instances. It indicates how correct positive predictions are.

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

Recall: Recall indicates to what extent the actual positive cases were correctly detected by the model. It is also called sensitivity.

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

F1-Score: It is a harmonic mean of precision and recall. This provides a balanced measure where both false positives and false negatives are concerned.

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

Mean Squared Error (MSE): This metric is used to compute the average squared deviation of a value from the predicted variable. The lower the MSE, the better the model.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The relationship between the True Positive Rate and False Positive Rate enables us to analyse performance of binary classifiers using ROC curve and AUC metric.

True Positive Rate (TPR / Sensitivity): This estimates the ratio of positive cases that

DOI: https://doi.org/10.5281/zenodo.20705210

actually were positive to actual positives This is represented on the Y-axis of an ROC curve.

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

False Positive Rate (FPR): The fraction of actual negative cases that the model wrongly flagged as positive This is plotted on the X axis.

$$\text{FPR} = \frac{FP}{FP + TN}$$

Specificity: The ability of the model to correctly identify actual negative cases. And it is actually inseparable from the FPR.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The AUC — Area Under ROC Curve

AUC is a single scalar summarizing the overall performance of the whole classifier at discriminating between classes. It is the level of disjointedness where a greater value means an improved model.

Mathematical Foundation

Mathematically, the AUC is defined as the area under the receiver operating characteristic curve, which can be calculated by integrating the true positive rate with respect to false positive rate from 0 to 1:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}) \cdot d(\text{FPR})$$

Practical Implementation

However, in practice machine learning we usually do not have a smooth function defining the curve. Instead, we consider a finite number of discrete threshold points to test the model on an example test dataset. In these situations the area under the curve is usually calculated by using the trapezoidal rule for approximation:

$$\text{AUC} = \frac{1}{2} \sum_{i=1}^{m-1} (\text{FPR}_{i+1} - \text{FPR}_i) \times (\text{TPR}_{i+1} + \text{TPR}_i)$$

This gives data scientists a heuristic for classifier quality that can be performed on discrete, empirical samples, rather than smooth curves.

4.2 Confusion Matrix

4. Results

The experimental results of the proposed machine learning model have been presented in this chapter. Firstly, the system specification provides information about the hardware and software environment for model training and testing. Next, the confusion matrix displays total correctly as well as incorrectly classified outputs. The results are featured in AUC curve to see the efficiency of this model for different classes. Last but not least, we evaluate the performance of our machine learning model using common evaluation metrics (i.e., accuracy, precision, recall and F1-score) to measure how well the model perform as a whole.

4.1 System Specification

The experimental work was performed in a Python-based machine-learning environment. The dataset was processed, trained, tested, and evaluated using common machine-learning libraries. The study was conducted using a laptop or desktop computer with an Intel Core i5 processor, 8 GB RAM, 512 GB SSD storage, integrated graphics, and a 64-bit operating system.

For software implementation, Windows 10 or Windows 11 was used as the operating system. Python was used as the main programming language, and Jupyter Notebook was used for model development and experiment execution. NumPy was used for numerical calculation, while Pandas was used for dataset handling and preprocessing. Scikit-learn was used for developing and evaluating machine-learning models. XGBoost and LightGBM libraries were used to implement the XGBoost and LightGBM classifiers. Matplotlib and Seaborn were used for visualizing graphs, confusion matrices, ROC–AUC curves, and other performance results.

The proposed system used different machine-learning algorithms to classify CKD and Non-CKD cases. After preprocessing the dataset, the models were trained and tested. Finally, performance was evaluated using accuracy, precision, recall, F1-score, confusion matrix, ROC–AUC curve, and mean squared error.

DOI: https://doi.org/10.5281/zenodo.20705210

ML Model:

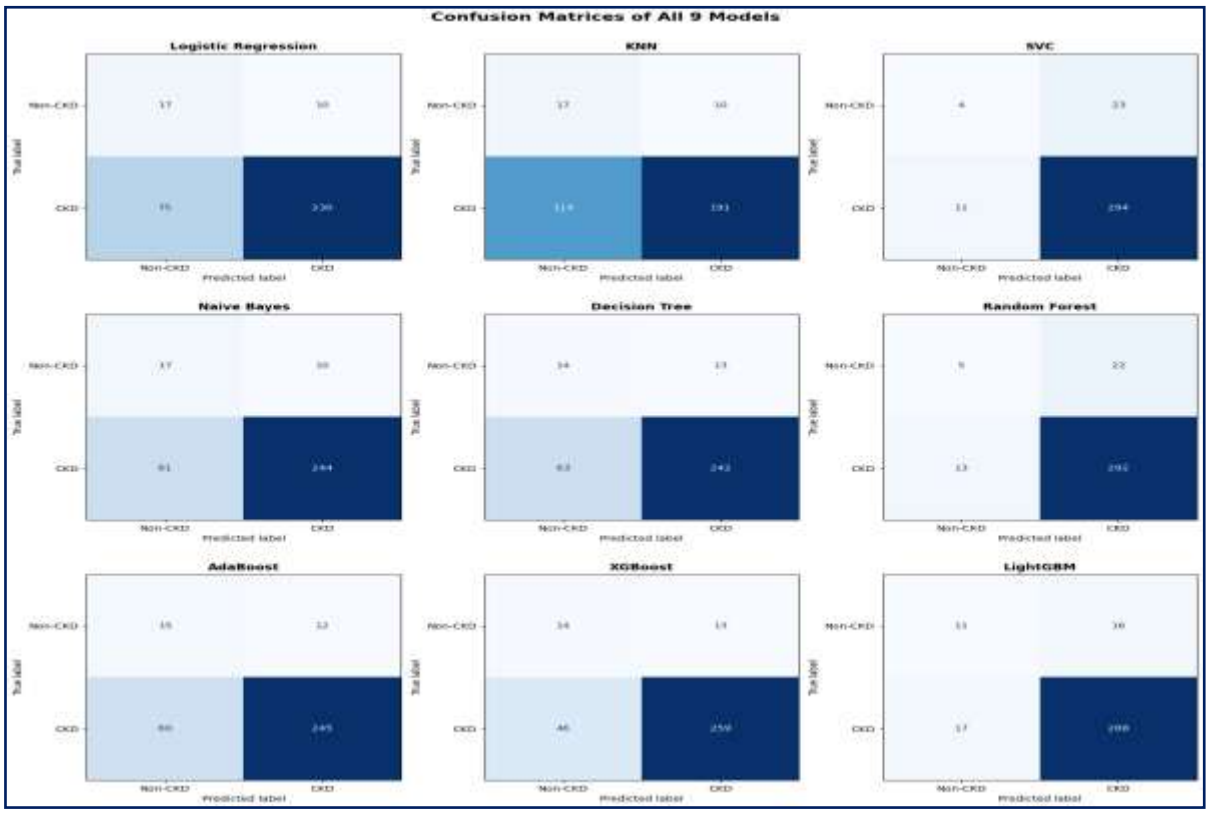


Fig. 4.1: Confusion Matrix (9 base Models)

The confusion matrices of the nine single machine-learning classifiers for CKD classification are indicated in 4.1. As seen from the figure, more CKD cases were classified correctly than Non-CKD cases by all models implying that our models had a higher sensitivity to the non-minority CKD class. Among these algorithms, LightGBM, SVC and Random Forest performed relatively better in terms of correctly classified CKD patients numbers and low total misclassification than other algorithms. LightGBM, SVC and Random Forest accurately classified 288 (99.3%) and 11 (1.7%),294 as CKD cases,302 non-CKD,292 as CKD in turn. However, many of the predictive models generated a large number of false positives incorrectly categorized as Non-CKD suggesting that whilst some of the models performed well for CKD detection they did not discriminate between healthy or Non-CKD cases.

Summary Table:

Model	TN	FP	FN	TP
Logistic Regression	17	10	75	230
KNN	17	10	114	191
SVC	4	23	11	294
Naive Bayes	17	10	61	244
Decision Tree	14	13	63	242
Random Forest	5	22	13	292
AdaBoost	15	12	60	245
XGBoost	14	13	46	259

DOI: https://doi.org/10.5281/zenodo.20705210

LightGBM	11	16	17	288

Table 1 : Confusion Matrix Summary Table(9 Base Model)

Ensemble Model:

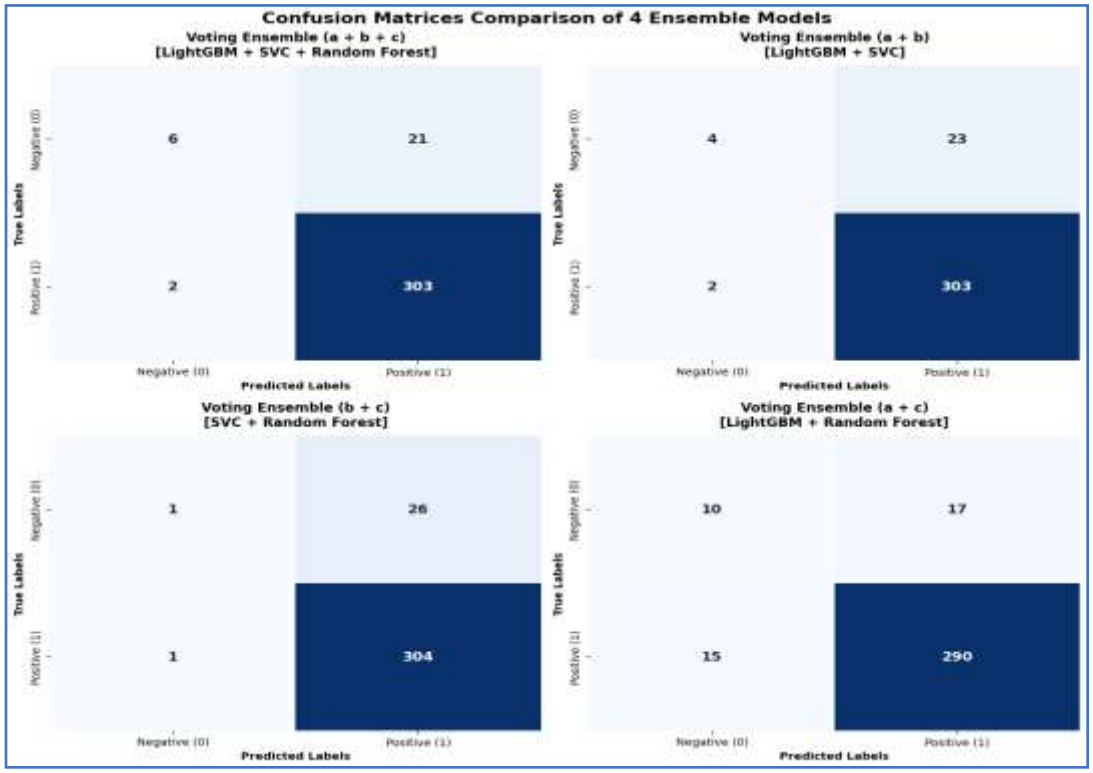


Fig. 4.2: Confusion Matrix (4 Ensemble Models)

The confusion matrices of four voting ensemble models combining LightGBM, SVC and Random Forest (RF) in various combinations are shown in Fig. 4.2. The use of ensemble models resulted in better performance on classification when compared with the majority of individual models and especially lower false-negatives for CKD cases. Voting ensemble of three models, LightGBM-SVC-Random Forest with best overall performance (Some CKD identified and only 2 CKD cases misclassified into Non-CKD) The LightGBM + SVC ensemble had similar CKD detection performance with 303 true-positives as well. The results of this study suggest that the prediction stability can be enhanced by combining multiple classifiers and therefore ensemble learning improves performance due to its ability of combining the advantages of different classifiers while false positive characterization as Non-CKD was still a major limitation.

Summary Table:

Ensemble Model	TN	FP	FN	TP
a + b + c	6	21	2	303
a + b	4	23	2	303
b + c	1	26	1	304
a + c	10	17	15	290

Table 2: Summary Table of Ensemble Model (TN, FP, FN, TP)

DOI: https://doi.org/10.5281/zenodo.20705210

4.3 ROC – AUC Curve

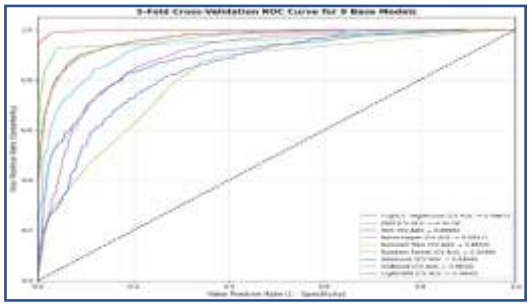


Fig. 4.3: 5 Fold Cross Validation ROC curve for 9 Base Model

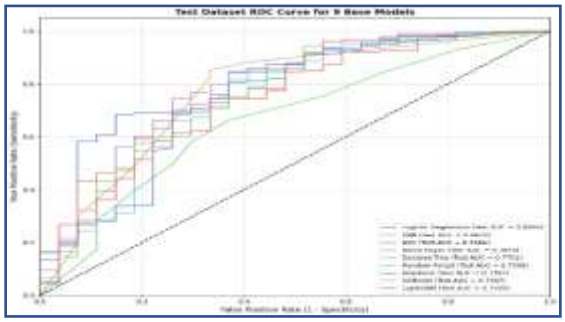


Fig.4.4: Testing ROC curve for 9 Base Model

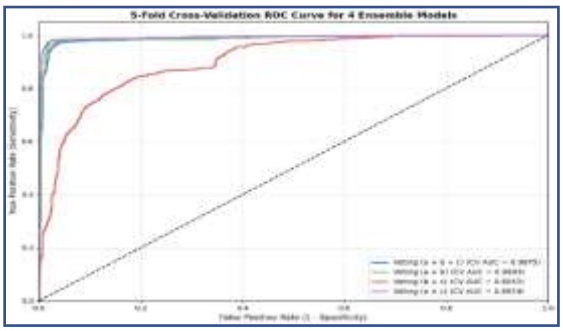


Fig. 4.5: 5 Fold Cross Validation ROC curve for 4

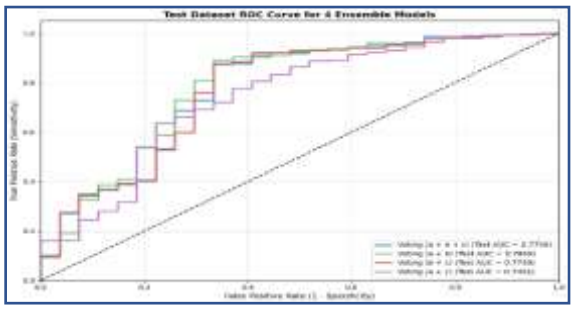


Fig. 4.6: Testing ROC curve for 4 ensemble Model

Here, widely used hold-out testing set presented in Figs. 4.5 and 4.6 show ROC–AUC performance for trained models based on training data used during cross-validation (CP) for nine single classifiers and four ensemble methods. As indicated in the cross-validation ROC curves, a number of high AUC values suggest the strong discriminative ability for CKD and Non-CKD classes; SVC,XGBoost, LightGBM and Random Forest exhibited particularly good separation patterns. The ensemble ROC curves also demonstrated good cross-validation performance, and almost all the ensemble combinations resulted in predictive performance that exceeded those of individual classifiers, as evidenced by their proximity to the upper-left corner of the plot. The testing ROC curves were, however, lower than the cross-validation curves (i.e. moderate generalisation ability on unseen data). This discrepancy indicates that the models learned strong prediction patterns on seen data during validation but would need to be tested externally or on larger datasets in order to confirm their robustness.

4.4 Machine Learning Performance (Individual)

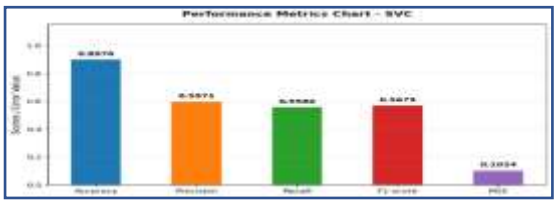


Fig. 4.7: SVC Performance

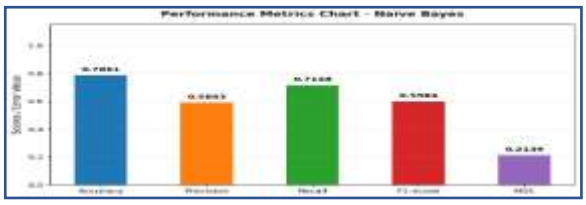


Fig. 4.8: NB Performance

DOI: https://doi.org/10.5281/zenodo.20705210

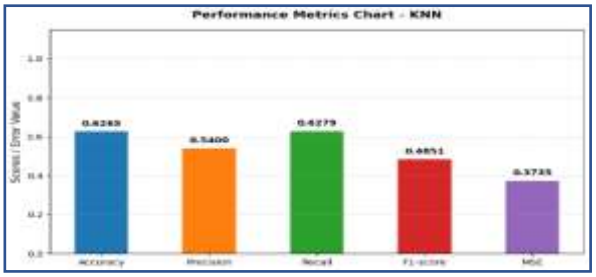


Fig. 4.9: KNN Performance

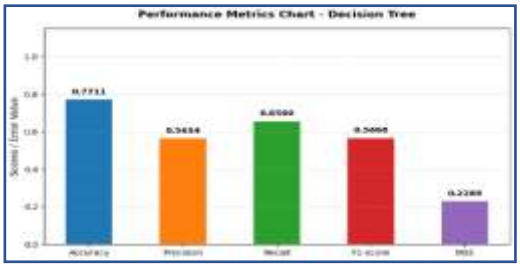


Fig. 4.10: DT Performance

Fig. 4.11: LR Performance

Fig. 4.12: RF Performance

Fig. 4.13: XGBoost Performance

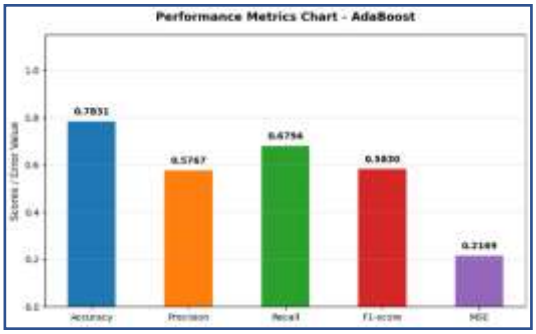


Fig. 4.14: Adaboost Performance

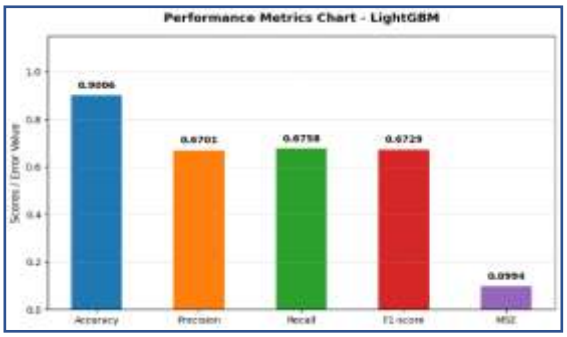


Fig. 4.15: LightGBM Performance

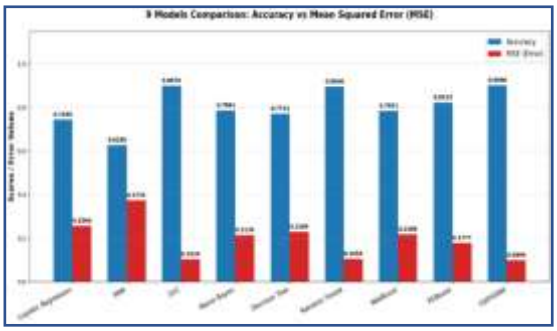


Fig. 4.16: All Base Model Comparison (Accuracy & MSE)

DOI: https://doi.org/10.5281/zenodo.20705210

Summary Table (Base Model)

Model Name	Accuracy	Precision	Recall	F1-score	MSE
LightGBM	**0.9006**	**0.6701**	**0.6758**	**0.6729**	**0.0994**
SVC	**0.8976**	**0.5971**	**0.556**	**0.5679**	**0.1024**
Random Forest	**0.8946**	**0.6039**	**0.5713**	**0.5828**	**0.1054**
XGBoost	**0.8223**	**0.5928**	**0.6838**	**0.6098**	**0.1777**
Naive Bayes	**0.7861**	**0.5893**	**0.7148**	**0.5984**	**0.2139**
AdaBoost	**0.7831**	**0.5767**	**0.6794**	**0.583**	**0.2169**
Decision Tree	**0.7711**	**0.5654**	**0.656**	**0.5668**	**0.2289**
Logistic Regression	**0.744**	**0.5716**	**0.6919**	**0.5649**	**0.256**
KNN	**0.6265**	**0.54**	**0.6279**	**0.4851**	**0.3735**

Table 3: Summary Table 9 Machine Learning Model

Individual Performance of 9 Base Machine-Learning Models Using Accuracy, Precision, Recall, F1-Score and Mean Squared Error are shown in Table4.16 As a result, we see that LightGBM has the best base-model performance with an accuracy of (0.9006) and MSE (0.0994), closely followed by SVC ((0.8976)) and Random Forest ≈(0.8946). They achieved better accuracy and operational means than Logistic Regression, Naive Bayes, Decision Tree, AdaBoost, XGBoost and KNN models. KNN performed the worse with lowest accuracy score and a very high MSE. Nevertheless, the macro precision, recall and F1-score values obtained were lower than those for accuracy, which highlights that balance classification performance was influenced by class imbalance in this study with a minority Non-CKD class.

4.3 Ensemble Performance

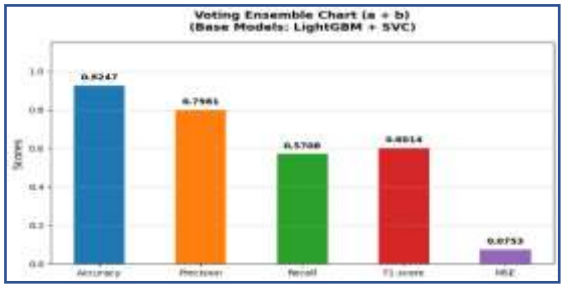


Fig 4.17: Ensemble Performance (a+b)

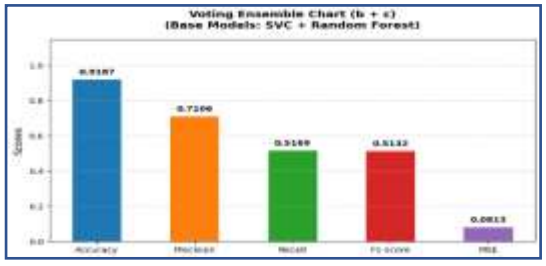
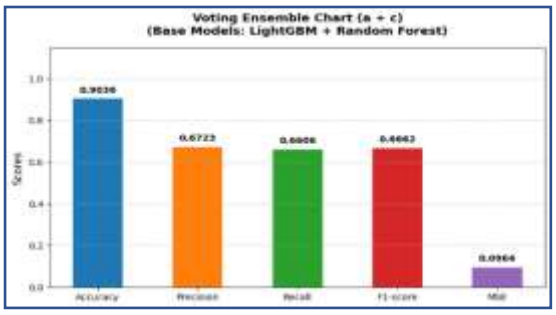
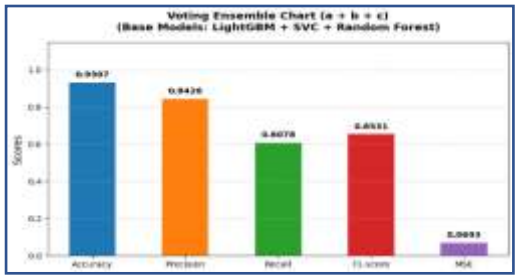


Fig 4.18: Ensemble Performance (b+c)





DOI: https://doi.org/10.5281/zenodo.20705210

Fig 4.19: Ensemble Performance (a+c)

Fig 4.20: Ensemble Performance (a+b+c)

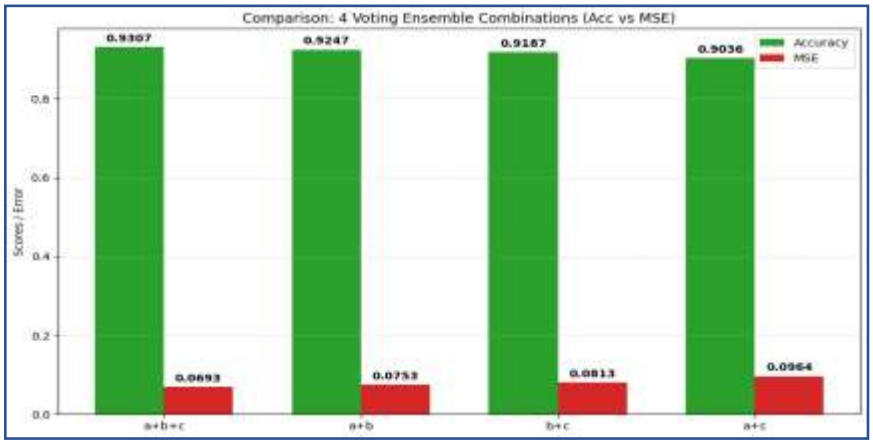


Fig 4.21: 4 Ensemble Model Performance Comparison

Summary Table (Base Model)

Ensemble Combination	Accuracy	Precision	Recall	F1-score	MSE
a+b+c (LGBM+SVC+RF)	**0.9307**	**0.8426**	**0.6078**	**0.6531**	**0.0693**
a+b (LGBM+SVC)	**0.9247**	**0.7981**	**0.6014**	**0.5708**	**0.0753**
b+c (SVC+RF)	**0.9187**	**0.7106**	**0.5169**	**0.5132**	**0.0813**
a+c (LGBM+RF)	**0.9036**	**0.6723**	**0.6606**	**0.6662**	**0.0964**

Table 4: Summary Table 4 Ensemble Model

Performance of the four voting ensemble combinations in 4.21 The ensemble of three model between LightGBM, SVC and Random Forest achieves the best overall accuracy at 0.9307 and the lowerst MSE at 0.0693 also so it indicates that has the best overall predictive performance among others. The second is the LightGBM + SVC ensemble, with 0.9247 accuracy, while at third place, we find SVC + Random Forest (0.9187) and then LightGBM + Random Forest (0.9036). Fig. 4.21 shows that the complete ensemble model outperformed all other combinations on prediction error reduction. However, the macro-level recall and F1-score demonstrate that both CKD and Non-CKD classes need to see better levels of recognition.

5. Discussion

In conclusion, the results of this study show that machine-learning models are adequate for distinguishing CKD cases, but that their performance can vary extensively with algorithmic structure and ensemble strategy. The best overall classification capability was seen with LightGBM, SVC and Random Forest among the nine base models. The single best accuracy (0.9006) and base-model MSE (0.0994), which is the lowest among all base models, were achieved with LightGBM indicating that gradient boosting was also very effective at capturing nonlinear relationships needed to predict CKD status. SVC and Random Forest also performed closely, with accuracies at 0.8976 and 0.8946 respectively. Nonlinear and tree-based learning methods work much better than simple or weak performing models, such as KNN for CKD classification.

Moreover, confusion-matrix results also show that the models had a strong prediction performance for CKD-positive cases. From a clinical screening perspective, this is important since the use of false-negative CKD

DOI: https://doi.org/10.5281/zenodo.20705210

predictions can defer onset and treatment. The SVC model was trained to predict a total of 294 CKDs, while the Random Forest and the LightGBM identified 292 and 288 CKD cases respectively. However, all the models performed reasonably poor with Non-CKD cases being misclassified as CKD which indicates that generally specificity is weaker than sensitivity in detection. This behavior might be due to the class imbalance, where CKD samples were significantly greater than other Class Non-CKD samples. While the average accuracy was high, macro-level precision, recall and f1 score give a more conservative assessment of model performance.

The ROC–AUC curves offer further perspective on the models' power of discrimination. In 5-fold cross-validation, most base and ensemble models had very strong AUC values to distinguish CKD from Non-CKD in the validation stage. Nevertheless, the testing ROC curves demonstrated a lower AUC compared to cross-validation. This indicates that the models might have fitted dataset specific signals and further validations of their generalisation performance on unseen external datasets is warranted. Such a gap between validation and testing performance is not uncommon in medical machine-learning ranging studies, especially with small or imbalanced datasets.

The ensemble-learning results yielded the most encouraging overall performance. The best accuracy of 0.9307 and MSE of 0.0693 was achieved using the voting ensemble LightGBM + SVC + Random Forest. This enhancement shows promise in overcoming individual model weaknesses and stabilizing predictions, especially with numerous strong classifiers. The three-model ensemble yielded only 2 additional false-negative CKD positives — an important aspect for CKD screening evaluations because the reduction of missed CKD is clinically significant. However, some of the false-positive Non-CKD cases were still generated from this ensemble which suggests that the model is likely to refer several healthy individuals for further clinical assessment. This trade-off may be acceptable in a medical screening context, since the clinical diagnosis can ultimately validate true disease status — while missing CKD cases could have more serious consequences.

The results sug- gest that ensemble learning generally gives a more robust and accurate CKD prediction framework than individual classifiers. The complete voting ensemble reaches the best trade-off between high votes+accuracy and minor forecasting error. However, the more modest macro recall and F1-score indicate that further work on handling class imbalance (with techniques like stratified-sampling, SMOTE, cost-sensitive learning, threshold optimization or larger balanced datasets) is warranted before drawing definitive conclusions from this data. Practical use also warrants external validation with multicenter clinical data.

6. Conclusion

This study assesses CKD classification based on nine base machine-learning models and four voting ensemble models. The results indicates that LightGBM, SVC and Random Forest were strongest single classifiers performance-wise; out of those, the base-model accuracies is 0.9006 (mean) its MSE(0.0994) was lowest as well in these three based models - among others no strong performance against LightGBM-. We found that LightGBM, SVC and Random Forest gave the lowest mean squared error (MSE) on their own, but it was the combination of all three in a voting ensemble model that outshone all other combinations with an accuracy metric of 0.9307 and an MSE score of 0.0693 Additionally, the ensemble model minimized overestimation of CKD prediction, which may cause mislabeling with a false-negative result in an early-stage CKD detection and screening.

The study shows that ensemble learning could markedly enhanced CKD prediction performance in terms of predicting CKD compared with machine-learning models on their own. The suggested voting ensemble shows high potential as a decision-support tool for classification of CKD. So the models exhibited poorer balanced accuracy for the minority Non-CKD class and lower testing ROC–AUC values compared to cross-validation results, necessitating larger, better-balanced independent clinical dataset validation. In the future, identify which type of specifiction can reduce false positives & build experimentally to prove this concept in clinical practice.

DOI: https://doi.org/10.5281/zenodo.20705210

Reference:

[1] M. F. Azizah and A. T. Paramitha, "Predictive Modelling of Chronic Kidney Disease Using Gaussian Naive Bayes Algorithm," *International Journal of Artificial Intelligence in Medical Issues*, vol.2,no.2,2024.
DOI:https://doi.org/10.56705/ijaimi.v2i2.160
Paper link: https://jurnal.yoctobrain.org/index.php/ijaimi/article/view/160

[2] E. J. Anagu, G. T. Abe, V. Z. Sabo, and S. J. Lamiri, "Optimizing Chronic Kidney Disease Prediction Via Ensemble Learning On Imbalanced Multi-Feature Clinical Data," *Brilliance:Research ofArtificialIntelligence*,vol.6,no.2,2026.
DOI:https://doi.org/10.47709/brilliance.v6i2.8210
Paperlink: https://jurnal.itscience.org/index.php/brilliance/article/view/8210

[3] K. Anitha and B. B. Rao, "Early Prediction of Chronic Kidney Disease Using an Ensemble Machine Learning-Based Clinical Decision Support System," *ShodhKosh: JournalofVisualand PerformingArts*,vol.7,no.13s,pp.109–125,2026.
DOI:https://doi.org/10.29121/shodhkosh.v7.i13s.2026.8433
Paperlink: https://www.granthaalayahpublication.org/Arts-Journal/ShodhKosh/article/view/8433

[4] H. Polat, H. Danaei Mehr, and A. Cetin, "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods," *Journal of Medical Systems*, vol. 41, no. 4, Art. no. 55,2017.
DOI:https://doi.org/10.1007/s10916-017-0703-x
Paperlink: https://pubmed.ncbi.nlm.nih.gov/28243816/

[5] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques," in *Proc. 2016 Management and Innovation Technology International Conference (MITicon)*, 2016, pp.MIT-80–MIT-83.
DOI:https://doi.org/10.1109/MITICON.2016.8025242
Paperlink: https://www.semanticscholar.org/paper/Predictive-analytics-for-chronic-kidney-disease-Charleonnan-Fufaung/462663b9075c11af4c4b24c85b013af91cb08ab7

[6] A. A. Imran, M. N. Amin, and F. T. Johora, "Classification of Chronic Kidney Disease Using Logistic Regression, Feedforward Neural Network and Wide & Deep Learning," in *Proc. 2018 International Conference on Innovation in Engineering and Technology (ICIET)*, 2018, pp. 1–6.
DOI:https://doi.org/10.1109/CIET.2018.8660844
Paperlink: https://www.semanticscholar.org/paper/Classification-of-Chronic-Kidney-Disease-using-and-Imran-Amin/ce5ce12d324517a033d1990d55781329034ecf82

[7] A. Sobrinho, A. C. M. D. S. Queiroz, L. D. da Silva, E. D. B. Costa, M. E. Pinheiro, and A. Perkusich, "Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries: A Comparative Analysis of Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 25407–25419, 2020.
DOI:https://doi.org/10.1109/ACCESS.2020.2971208
Paperlink:https://www.researchgate.net/publication/339014686_Computer-Aided_Diagnosis_of_Chronic_Kidney_Disease_in_Developing_Countries_A_Comparative_Analysis_of_Machine_Learning_Techniques

[8] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease," *IEEE Access*, vol. 8, pp. 20991–21002,2020.
DOI:https://doi.org/10.1109/ACCESS.2019.2963053
Paperlink: https://www.researchgate.net/publication/338239369_A_Machine_Learning_Methodology_for_Diagnosing_Chronic_Kidney_Disease

[9] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy," *IEEE Access*, vol. 8, pp. 55012–55022, 2020.
DOI:https://doi.org/10.1109/ACCESS.2020.2981689
Paperlink: https://pure.kfupm.edu.sa/en/publications/an-empirical-evaluation-of-machine-learning-techniques-for-chroni/

DOI: https://doi.org/10.5281/zenodo.20705210

[10] P. Chittora *et al*., "Prediction of Chronic Kidney Disease—A Machine Learning Perspective," *IEEEAccess*,vol.9,pp.17312–17334,2021. DOI:https://doi.org/10.1109/ACCESS.2021.3053763
Paperlink: https://ieeexplore.ieee.org/document/9336034
[11] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic Kidney Disease Prediction Based on Machine Learning Algorithms," *Journal of Pathology Informatics*, vol. 14, Art. no. 100189, 2023. DOI:https://doi.org/10.1016/j.jpi.2023.100189
Paperlink: https://pmc.ncbi.nlm.nih.gov/articles/PMC9874070/
[12] S. Singamsetty, S. Ghanta, S. Biswas, and A. Pradhan, "Enhancing Machine Learning-Based Forecasting of Chronic Renal Disease with Explainable AI," *PeerJ Computer Science*, vol. 10, Art. no.e2291,2024. DOI:https://doi.org/10.7717/peerj-cs.2291
Paper link: https://peerj.com/articles/cs-2291/
[13] Kidney Disease: Improving Global Outcomes CKD Work Group, "KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease," *Kidney International*, vol. 105, no. 4S, pp. S117–S314, 2024. Available: https://pubmed.ncbi.nlm.nih.gov/38490803/
[14] World Health Organization, "Kidney disease," WHO Fact Sheet, Apr. 20, 2026. Available: https://www.who.int/news-room/fact-sheets/detail/kidney-disease
[15] H. Khalid, A. Khan, M. Z. Khan, S. Ahmed, and K. K. Singh, "Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease," *Computational Intelligence and Neuroscience*, vol. 2023, Article ID 9266889,2023.Available: https://pubmed.ncbi.nlm.nih.gov/36959840/
[16] F. Khalid et al., "Predicting the Progression of Chronic Kidney Disease: A Systematic Review of Artificial Intelligence and Machine Learning Approaches," 2024. Available: https://pubmed.ncbi.nlm.nih.gov/38864072/
[17]DatasetLink : https://www.kaggle.com/datasets/usmanshafeeqdit/chronic-kidney-dsease-data

DOI: https://doi.org/10.5281/zenodo.20705210