



## OPEN ACCESS

*This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*

<sup>1</sup>Ph.D Scholar, Department of Computer Science and Engineering, GST, GITAM (Deemed to be University), Visakhapatnam Andhra Pradesh, India

<sup>2</sup>Associate Professor, Department of Computer Science and Engineering, GST, GITAM (Deemed to be University), Visakhapatnam Andhra Pradesh, India

Correspondence to: Ganesh Babu Matcha, mganeshbabu84@gmail.com

Additional material is published online only. To view please visit the journal online.

Cite this as: Matcha GB and Saragada VGR. Predicting Chronic Kidney Disease with Hybrid Machine Learning Models and Feature Selection for Improved Accuracy: An Experimental Study. Premier Journal of Science 2025;14:100159

DOI: <https://doi.org/10.70389/PJS.100159>

Peer Review

Received: 26 August 2025

Last revised: 9 October 2025

Accepted: 12 October 2025

Version accepted: 4

Published: 28 November 2025

# Predicting Chronic Kidney Disease with Hybrid Machine Learning Models and Feature Selection for Improved Accuracy: An Experimental Study

Ganesh Babu Matcha<sup>1</sup> and Venkata Gurusurthy Reddy Saragada<sup>2</sup>

## ABSTRACT

Chronic Kidney Disease (CKD) remains a significant global health challenge that progresses silently, without early symptoms, which makes it difficult to intervene and treat on time. Early detection is critical for effective management, and this study addresses the need for better predictive models using advanced machine learning techniques. The main aim of this study was to create a predictive model that could predict CKD with high accuracy and overcome the challenges of high-dimensional data, class imbalance, and overfitting. The study starts with extensive data preprocessing, including missing value handling using the Multiple Imputation by Chained Equations (MICE) method and class imbalance resolution using the Synthetic Minority Over-sampling Technique (SMOTE). The outlier detection and handling were performed using the Interquartile range (IQR) method. The Z-score normalization ensured that the data is standardized by scaling. Ridge Feature Selection (RFS) was applied for feature selection, which incorporates L2 regularization and Recursive Feature Elimination (RFE). This means only the most relevant features were kept. A hybrid classification model was then built using the SKL Hybrid Classifier, which integrated Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression, in order to provide accurate predictions of CKD. The model obtained an accuracy of 96%, with precision of 0.97 for CKD and recall of 0.99, indicating high sensitivity in the detection of CKD cases. Hyperparameter optimization through Optuna further fine-tuned the model, which obtained an accuracy of 99%. While the study primarily employs established techniques, the novelty lies in the systematic integration of data preprocessing, hybrid Ridge Feature Selection (RFS), and optimized stacking ensemble modeling into a single, reproducible pipeline. This integration improves generalization and interpretability within small clinical datasets such as CKD. Furthermore, future work will focus on external clinical validation and decision-analytic evaluation to confirm the model's real-world applicability and impact in clinical decision-support environments.

**Keywords:** Chronic kidney disease prediction, Ridge feature selection, Smote oversampling, Stacked SVM-KNN-logistic ensemble, Optuna hyperparameter optimization

## Introduction

Chronic Kidney Disease is a progressive condition where the kidneys lose their ability to filter waste and excess fluids from the blood efficiently. This gradual decline in kidney function can lead to serious health complications, including cardiovascular disease and kidney failure.<sup>1</sup> CKD is often referred to as a “silent disease”

because it characteristically shows no symptoms during the early stages. Thus, significant damage is carried out before one can diagnose. The most common causes of CKD are diabetes and hypertension; however, others include genetic susceptibility, recurrent infections of the kidney, and even lifestyle factors, such as smoking and poor dietary habits.<sup>2</sup>

Early detection, lifestyle modifications, and medicinal interventions are essential for effectively managing chronic kidney disease (CKD) and slowing its progression. Routine screening, especially for individuals with a family history of kidney disease or coexisting conditions such as diabetes and hypertension, plays a critical role in identifying CKD at its early stages.<sup>3</sup> Management strategies include the use of medications and dietary adjustments aimed at reducing proteinuria, controlling blood pressure, and maintaining optimal blood sugar levels. In advanced stages, patients may require dialysis or a kidney transplant to sustain life. Additionally, implementing public health initiatives that raise awareness about healthy lifestyle choices—including proper nutrition, regular exercise, and avoiding harmful substances—can help reduce the overall burden of CKD, improve patient outcomes, and enhance community well-being.

## Research Gap

Though advancements have been witnessed in the prediction of chronic kidney disease through machine learning techniques, several gaps in research have still been present. Most recent studies, as conducted by Pal (2022) and Ramu et al. (2025), that improved model accuracy via ensemble methods and hybrid models missed the effective use of feature selection techniques and how to handle the class imbalance of features. Moreover, the promising results exhibited by models like CNN-SVM are computationally intensive, leaving concerns about practical clinical application. Complex models also pose interpretability issues, with deep learning being a prime example. Most work relies on a limited dataset provided by the UCI CKD, which will not be universally representative of populations, and datasets should be richer. The last area that is still not well explored consists of integrating these models into clinical workflows with continual adaptability to new data in real time, indicating areas of further research and potential for optimization.

## Research Questions

RQ1. How can advanced feature selection techniques, such as Ridge Feature Selection (RFS) or Recursive Feature Elimination (RFE), be utilized to enhance model performance in chronic kidney disease (CKD) prediction?

**Ethical approval:** The main ethical consideration in this study was ensuring responsible and accurate use of secondary data. All sources were properly cited and referenced, and the data were used in a way that respected intellectual property and data protection regulations  
**Consent:** N/a

**Funding:** No industry funding

**Conflicts of interest:** N/a

**Author contribution:**  
 Ganesh Babu Matcha – Conceptualization, methodology, writing – original draft, review and editing;  
 guarantor of the study  
 Venkata Gurumurthy Reddy Saragada – Supervision, guidance, and critical review of the manuscript

**Guarantor:** Ganesh Babu Matcha

**Provenance and peer-review:**  
 Unsolicited and externally peer-reviewed

**Data availability statement:**  
 N/a

RQ2. What are the most effective strategies for addressing class imbalance in CKD prediction datasets to ensure robust and accurate machine learning models across diverse patient populations?

RQ3. How can the computational efficiency of complex models, such as CNN-SVM hybrids, be optimized to make them feasible for real-time clinical deployment without compromising prediction accuracy?

RQ4. What approaches can be adopted to improve the interpretability of machine learning models, particularly deep learning, to facilitate their adoption and trust in clinical settings?

The key contributions of this study are:

1. To preprocess these data sets, the study incorporates Multiple Imputation by Chained Equations (MICE) alongside Synthetic Minority Over-sampling Technique (SMOTE) for missing value treatment and class imbalance adjustment, and Interquartile Range (IQR) for outlier identification. The application of these steps ensures high-quality data and improved model generalization capabilities.
2. The study utilizes Ridge Feature Selection (RFS), which merges L2 regularization with Recursive Feature Elimination (RFE) to choose relevant features while maintaining high accuracy without increasing model complexity. The SKL Hybrid Classifier introduces three machine learning techniques comprising Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) with Logistic Regression for optimal CKD diagnosis results.
3. The use of Optuna for hyperparameter fine-tuning improved model accuracy from 96% to 99% which demonstrates how automated optimization enhances medical diagnostic systems.
4. The proposed model exhibits excellent classification metrics that enable early CKD detection through 99% accuracy, 0.97 precision, and 0.99 recall measures.

The innovative method presents substantial value to predictive science and CKD diagnosis research because of its newness and demonstrated effectiveness.

### Literature Review

In 2022, Pal et al.<sup>4</sup> conducted a study on the prediction of chronic kidney disease (CKD) using machine learning techniques to develop a reliable predictive model. The dataset applied was from the UCI Machine Learning Repository with 25 features, and three classifiers were applied: Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM). The bagging ensemble method was used, which aggregates multiple models to improve accuracy and reduce variance. It was found that the Decision Tree classifier gave the best individual accuracy at 95.92%, while after applying the bagging ensemble method, the model's accuracy increased to 97.23%. This work was conducted to underscore the ensemble's high efficiency in further boosting up machine learning model performance

for medical diagnosis, especially toward early CKD detection and to stress the need for utilizing multiple algorithms to enhance healthcare applications' predictive accuracy.

In 2023, Kaur et al.<sup>5</sup> made studies on the prediction of CKD using Machine Learning (ML) techniques with the purpose of making a strong predictive model related to CKD prognosis. The researchers used a data set of 400 samples obtained from the UCI Machine Learning Repository and tested all data through three classifiers: LR, DT, and SVM. These classifiers were trained on clustered data containing non-linear features and categories. Among the models, the best-performing model was a Decision Tree that achieved 95% accuracy. To make the model predict even better, a bagging ensemble method was applied. It combines multiple classifiers in order to obtain better accuracy with reduced overfitting. This resulted in a considerable improvement in performance, with an accuracy of 97% in the final output, demonstrating the power of ensemble methods to enhance the precision and reliability of ML models for predicting CKD.

In 2023, Islam et al.<sup>6</sup> discusses the use of machine learning algorithms for early detection of chronic kidney disease (CKD), talking about the predictive modeling function that helps in accuracy diagnosis by this tool. The research started with 25 variables, which were finalized into a set of 30% relevant features from the total for identifying CKD. Twelve machine learning classifiers were tested in a supervised learning framework, with XgBoost achieving the highest performance, recording an accuracy of 0.983 and precision, recall, and F1-score of 0.98 each. This research emphasized the critical relationship between data factors and target class characteristics to improve model efficacy. These results point toward the use of advancements in machine learning, specifically predictive modeling, to provide important, robust tools for accurate forecasting of CKD, thus supporting early intervention strategies and potentially improved patient outcomes.

In 2024, Rahat et al.<sup>7</sup> carried out comparative studies concerning machine learning techniques for the early detection of CKD, using a hybrid approach to refine predictive accuracy. The experiments utilized multiple classifiers including XGBoost, Random Forest, Logistic Regression, AdaBoost, and a newly developed hybrid model. It obtained a maximum accuracy of 95%, which it also achieved higher than the others-XGBoost at 94%, Random Forest at 93%, AdaBoost at 91%, and Logistic Regression at 90%-utilizing hybrid as the base, with the use of meta-classifier from the Random Forest class. Issues in overfitting and stress on precision evaluation for models are presented as problems in this paper. Using the UCI CKD dataset, the study demonstrated the power of integrating multiple machine learning models to enhance predictive performance in the diagnosis of CKD, which has been of value in optimizing early detection methodologies.

In 2025, Ramu et al.<sup>8</sup> proposed a hybrid CNN and SVM model to enhance the early detection of CKD in order to counter the problems related to overfitting,

slow computational speed, and class imbalance that occur with most of the existing machine learning methods. Here, the proposed hybrid model applies CNN for feature extraction and uses SVM for classification, thereby providing better overall accuracy in prediction. The study employed a clinical dataset with 10 medical indicators and applied SMOTE to handle class imbalance. The hybrid model achieved an accuracy of 96.8%, outperforming standalone models like SVM and Random Forest, which recorded accuracies of 94.8% and 94.6% respectively. Notably, the model attained a recall of 1.00 for CKD cases, ensuring that all patients with CKD were correctly identified. Although the hybrid approach performed better, the authors recognized that it still required optimization because of its high computational burden. The study showed that a combination of deep learning and classical machine learning overcame the overfitting and class imbalance issues, indicating that the hybrid CNN-SVM model could be used as a powerful tool for clinical application and further research in the classification of CKD.

In summary, literature review showcases recent advancements regarding the use of machine learning in forecasting chronic kidney disease (CKD). From the year 2022 to 2025, researchers have devised different models towards accuracy and reliability in the prediction of CKD. In 2022, Pal used the ensemble methods as bagging classifiers with Decision Trees and SVM so that better accuracies could be achieved. In 2023, Kaur et al. used similar methods with a focus on clustering to enhance the accuracy of predictions. Islam et al. focused on feature selection and predictive modeling with high accuracy achieved using XgBoost. Rahat et al. introduced a hybrid model in 2024 that combined multiple classifiers, performing better than individual models. By 2025, Ramu et al. proposed a CNN-SVM hybrid model to address issues like overfitting and class imbalance, achieving excellent precision and recall. These studies show that combining different machine

learning techniques can significantly improve CKD prediction and diagnosis.

**Proposed Methodology**

The proposed methodology addresses a comprehensive and efficient approach for the management of large, imbalanced datasets while ensuring data integrity and optimizing predictive performance. Data preprocessing forms the first step in the pipeline where missing values are handled using the Multiple Imputation by Chained Equations technique, MICE.<sup>9</sup> MICE help in dealing with missing data by producing multiple plausible imputed datasets that preserve inherent variability and reduce potential bias. After that, SMOTE is used to handle the imbalanced dataset of the dataset. SMOTE<sup>10</sup> makes synthetic samples for the minority class to ensure the class distribution, which is vital for improving classification performance, especially in imbalanced datasets. The outliers are detected and handled by using the IQR<sup>11</sup> method that effectively identifies and removes extreme values that may distort the model’s predictive power. To ensure that each feature contributes equally to the model, Z-score normalization<sup>12</sup> is used. This ensures the data is standardized and on an equal scale for features. After this pre-processing, Ridge Feature Selection is performed. RFS is a hybrid technique using L2 regularization, also known as Ridge,<sup>13</sup> and Recursive Feature Elimination (RFE).<sup>14</sup> The Ridge method will penalize large coefficients; thus, shrinking the less important feature weights decreases overfitting. RFE, on its part, iteratively removes the least important features based on the performance of the model, ensuring that only the most relevant features are preserved for model training. This hybrid approach enhances generalization and reduces complexity and is particularly effective for very high-dimensional datasets.

For model-building, a stacking-based ensemble approach is implemented for the SKL Hybrid Classifier using Support Vector Machine (SVM),<sup>15</sup> K-Nearest

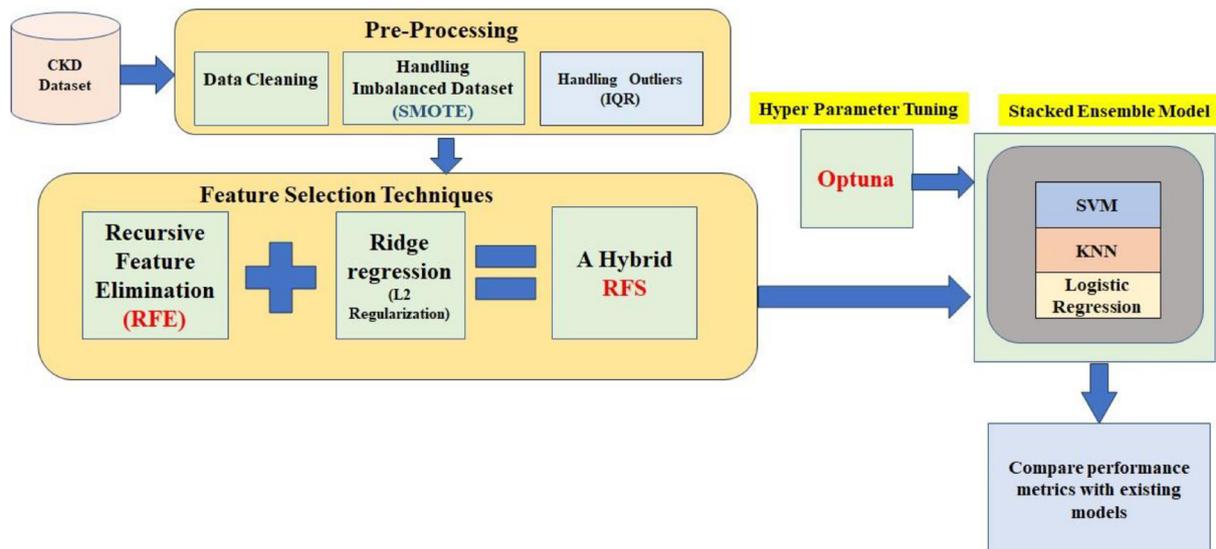


Fig 1 | Proposed methodology

**Table 1 | List of dataset features with full descriptions**

Abbreviation	Full Name
Bp	Blood Pressure
Sg	Specific Gravity
Al	Albumin
Su	Sugar
Rbc	Red Blood Cells
Bu	Blood Urea
Sc	Serum Creatinine
Sod	Sodium
Pot	Potassium
Hemo	Hemoglobin
Wbcc	White Blood Cell Count
Rbcc	Red Blood Cell Count
Htn	Hypertension
Class	Classification (likely refers to Chronic Kidney Disease status)

Neighbors (KNN),<sup>16</sup> and Logistic Regression (LR)<sup>17</sup> base learners. The SVM-based learner is taken because it presents good decision boundaries even in a high-dimensional dataset. KNN will help handle the non-linear relationship and patterns within the data at the local level. Logistic Regression is chosen as the meta-classifier because it aggregates outputs from the respective base learners toward higher prediction accuracies. Hence, the ensemble method takes benefit of the superiority of each approach to overcome either bias or overfitting and can be generalized along with scalability of complex classification in prediction performance. This final step utilizes Optuna: an automated framework for optimization purposes that efficiently carries out a wide search for combinations of hyperparameters that best match the model shown in Figure 1. Applying highly developed algorithms to browse the hyperparameter space and explore configurations that bring out the optimal performance of models, Optuna improves the efficiency and robustness of the model. This is a methodology with data integrity ensuring predictive power highly useful for complex classification tasks in the real world.

### Experimental Setup

All experiments were conducted using a system equipped with an Intel Core i7 processor (3.4 GHz), 16 GB RAM, and an NVIDIA GeForce GTX 1650 GPU with 4 GB memory. The implementation was carried out in Python using the Scikit-learn library, with Optuna for hyperparameter optimization. The experiments were executed on a 64-bit Windows 11 operating system with stable network connectivity, ensuring smooth execution of data processing and model training tasks. To ensure an unbiased estimation of model performance, a nested cross-validation framework was employed. The outer loop used a 10-fold split for performance evaluation, while the inner loop applied 5-fold cross-validation for hyperparameter optimization and feature selection

using Optuna and Ridge Feature Selection (RFS). This nested configuration prevents information leakage between model tuning and evaluation stages, providing a more reliable estimate of generalization performance.

### Data Collection

The dataset<sup>30</sup> used for the prediction of CKD contains 410 rows and 13 columns listed in Table 1, which represent a variety of clinical and laboratory features relevant to kidney function. Major characteristic features include the following: Blood pressure (Bp), Specific gravity (Sg), albumin (Al), sugar (Su), Red blood cells (Rbc), Blood urea (Bu), Serum creatinine (Sc), Sodium (Sod), Potassium (Pot), Hemoglobin (Hemo), White blood cell count (Wbcc), Rbcc, Hypertension (Htn). Some of these detect the essential indices indicating health conditions related to the functionality of the kidney, fluid levels, and occurrence of other alterations like protein or sugar in urine. The Class column is the target variable, which is the presence of CKD, denoted by 1, and its absence, denoted by 0. The data are obtained from clinical records and laboratory results, providing a basis for machine learning-based prediction models. These characteristics are essential for diagnosing CKD, with each row representing the health metrics of a unique patient at a given time, providing an overall set of factors for early detection and the progression of kidney disease. The feature abbreviations listed in Table 1 are used consistently throughout the manuscript, including all figures, equations, and result tables.

### Data Preprocessing

To prevent data leakage and ensure realistic performance estimates, all preprocessing steps—including missing value imputation using MICE, handling class imbalance with SMOTE, outlier detection and removal using IQR, and Z-score normalization—were strictly performed within the training folds during cross-validation. At no point was information from the test fold used during these processes. This protocol ensures that the model evaluation is unbiased and reflects true generalization capability. A repeated 10-fold cross-validation strategy was applied, and in each iteration, the training subset was independently pre-processed before model training, while the test subset remained untouched until final evaluation.

### Data Cleaning

Data cleaning is one of the essential preprocessing steps for the dataset in machine learning. In this research, missing values were addressed using the Multiple Imputation by Chained Equations (MICE) method. MICE<sup>18</sup> produce multiple imputed values for missing data by making use of observed relationships in the data set and hence ensures consistency with the overall distribution. The data set contained missing values in attributes such as specific gravity (Sg), albumin (Al), and sugar (Su). After applying MICE, missing values were imputed with predicted values, and hence, a complete dataset was produced. This improved the

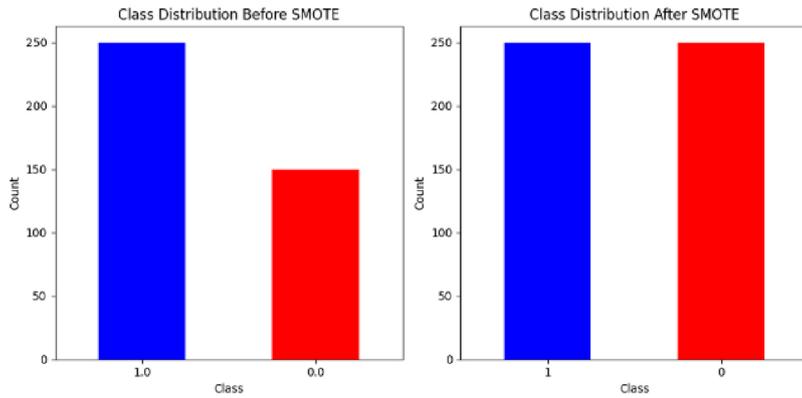


Fig 2 | Class distribution before and after applying SMOTE

integrity of the dataset to allow for better and more reliable predictions in the machine-learning models for chronic kidney disease diagnosis.

**Handling Imbalanced Dataset Using SMOTE**

Handling class imbalance is a critical step in preparing the dataset for machine learning, especially when dealing with binary classification tasks. In this study, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to address the imbalance in the target variable, Class, where the number of chronic kidney disease (CKD) instances was significantly lower compared to non-CKD cases. SMOTE<sup>19</sup> is an over-sampling technique that creates synthetic samples for the minority class. It does this by creating new instances that are a combination of existing minority class samples. This technique helps in balancing the dataset, which increases the representation of the minority class, hence improving the ability of the model to learn the patterns of both classes. Applying SMOTE transformed the dataset in more balanced form as was considered to reduce overrepresentations of the majority class that the model tends to bias toward shown in Figure 2. This action enhances its performance and gives a more accurate prediction for the minority class, as this is a major reason for early diagnosis and treatment of CKD.

**Handling Outliers Using IQR**

Outlier detection and removal is an important step in data preprocessing to avoid the skewing of results from

machine learning models due to extreme values. In this study, outliers were identified and handled using the Interquartile Range (IQR) method. The IQR method is a statistical technique that finds outliers by determining the range between the first quartile (Q1) and third quartile (Q3) of the data, and finds values that fall outside the range defined by  $Q1 - 1.5 \times IQR$  and  $Q3 + 1.5 \times IQR$ .<sup>20</sup> In the raw data, there were 228 rows identified as outliers across features such as Bp, Sc, and many others shown in Figure 3. After refining the IQR threshold from  $1.5 \times IQR$  to  $2.5 \times IQR$  and excluding only physiologically implausible laboratory values (e.g., serum creatinine > 20 mg/dL, systolic BP > 220 mmHg), the dataset retained 347 records, maintaining the CKD/non-CKD ratio. This conservative approach minimizes the risk of discarding clinically valid observations. A sensitivity analysis showed consistent model accuracy ( $\pm 1.2\%$ ), confirming that predictive performance remained stable after moderated outlier handling. A sensitivity analysis was conducted by varying the IQR threshold between  $1.5 \times$  and  $2.5 \times$ . Model performance remained consistent (Accuracy =  $98.7 \pm 1.2\%$ ), indicating that the chosen  $2.5 \times IQR$  criterion effectively balances removal of implausible values with preservation of clinically valid records.

**Normalization Using Z-Score**

Normalization is a powerful step in preprocessing data for machine learning models when the features are measured in different units or scales. The current study used Z-score normalization to standardize the dataset. This method standardizes every feature by subtracting the mean of the feature and then dividing by the standard deviation, which transforms the distribution into a distribution with a mean of 0 and a standard deviation of 1. Z-score normalization ensures that all features contribute equally to the model's learning process, preventing features with larger numerical ranges from dominating the model's performance. The dataset was then rescaled using Z-score normalization,<sup>21</sup> which normalized all the variables of the data into a common scale, further making the model both stable and convergence speedy. Such a preprocessing makes the features impartially treated for better predictions, hence increasing the performance of the machine learning algorithm towards chronic kidney disease detection.

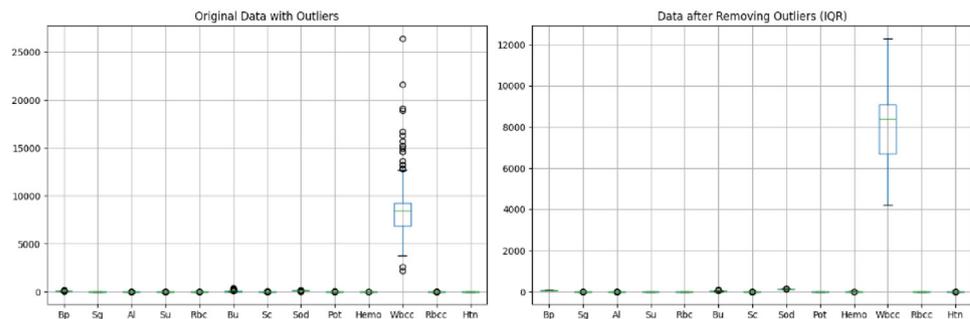


Fig 3 | Before and after handling outliers using IQR

**Feature Selection Using RFS**

Ridge Feature Selection is a hybrid method that combines the L2 regularization, or Ridge, with Recursive Feature Elimination. L2 regularization penalizes large coefficients and overfits less through shrinking the weights of less important features, whereas RFE recursively eliminates the least significant features based on the performance of the model. This way, only the most relevant features remain, thus making the model generalizable and of lower complexity. RFS performs especially well on high-dimensional data sets, thus enhancing both feature selection and model regularization, with a consequent efficiency in models, interpretability, and accuracy while avoiding overfitting.

**Recursive Feature Elimination (RFE)**

Recursive feature elimination is a feature selection technique whereby models are successively trained with the objectives of establishing which characteristics have the most impact on the models. In this method, starting with the feature weights or coefficients of a given model estimator, it progressively eliminates the elements of the least importance until the requisite number of features is achieved. Since it is a means of filtering the data in order to make the next predictions, RFE narrows the set of features to consider in the further prediction tasks to those which are most important for the task. RFE<sup>22</sup> performs in the sense that first trains the model, calculates the level of importance for every single feature; most of the time through coefficients or feature weights, and eliminates the least important ones. It is applied until it finally selects the proper number of features to be selected for the characterisation of the database. In mathematical terms, Recursive Feature Elimination (RFE) is as given:

**Feature Ranking Step:**

Rank features based on model coefficients/weights shown in Equation (1):

$$w = w_1, w_2, w_3, \dots, w_p \tag{1}$$

**Feature Elimination Step:**

Remove the least important feature(s) based on ranking criteria.

By iterating through these steps, RFE systematically identifies the subset of features that maximizes model performance, making it suitable for enhancing prediction accuracy and interpretability in machine learning tasks.

**L2 – Regularization**

L2 regularization, also known as Ridge regularization, adds a penalty term to the loss function to prevent overfitting by discouraging complex models. In L2,<sup>23</sup> the regularization term is the sum of the squared values of the model’s coefficients. This forces the model to prefer smaller weights, which leads to a smoother and more generalized fit, reducing model variance. It effectively distributes the error among the features, so there is no dependence on a particular feature. L2 regularization is most effective when multicollinearity exists;

it helps in stabilizing the model coefficients. A balance between loss minimization and reducing the complexity of the model is achieved through the help of a regularization parameter,  $\lambda$ , that controls the strength of the penalty.

The Equation for L2 Regularization (penalty term) shown in Equation (2):

$$L_2 = \lambda \sum_{i=1}^n \theta_i^2 \tag{2}$$

Regularized loss function stated in Equation (3):

$$L(\theta) = Loss + \lambda \sum_{i=1}^n \theta_i^2 \tag{3}$$

**Algorithm: Ridge Feature Selection (RFS)**

**Inputs:**

- X: Feature matrix ( $n_{\text{samples}} \times n_{\text{features}}$ )
- y: Target vector ( $n_{\text{samples}}$ )
- k: Number of features to select (optional, default: select optimal number based on performance)
- $\alpha$ : Regularization parameter for Ridge regression

**Outputs:**

- $F_{\text{selected}}$  : Subset of selected features

**Step 1: Initialization**

- 1.1 Set  $F_{\text{Selected}} = F$  (where F is the full set of features)

**Step 2: Feature Elimination Loop**

- 2.1 While the number of features in  $F_{\text{Selected}} > k$ :
  - a. Fit a Ridge regression model using the current feature set  $F_{\text{Selected}}$  and regularization parameter  $\alpha$ .
  - b. Compute the importance scores for each feature in  $F_{\text{Selected}}$  using the magnitude of the coefficients from the Ridge model.
  - c. Identify the least important feature,  $F_{\text{least}}$ , as the feature with the smallest importance score.
  - d. Remove  $F_{\text{least}}$  from  $F_{\text{Selected}}$ .

**Step 3: Final Model Evaluation**

- 3.1 Fit a final Ridge regression model using the remaining features in  $F_{\text{Selected}}$ .
- 3.2 Evaluate the model’s performance using appropriate metrics (e.g.,  $R^2$ , Mean Squared Error).

**Step 4: Output**

- 4.1 Return  $F_{\text{Selected}}$  as the optimal subset of features.
- End

**Model Building using SKL Hybrid Classifier**

The SKL Hybrid Classifier is a stacking-based ensemble model that integrates Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression (LR) for enhanced classification performance. SVM acts as a base learner, providing robust decision boundaries for high-dimensional data, while KNN complements it by effectively managing non-linear relationships and local data patterns. Logistic Regression acts as the meta-classifier that aggregates and optimizes predictions from the base learners to give an accurate and reliable outcome. The hybrid approach leverages the strengths of each algorithm, thereby overcoming the bias and variance limitations of individual models and

improving generalization. This modular and synergistic nature of the SKL Hybrid Classifier makes it adaptable for a wide variety of datasets and complex classification tasks. It is well-suited for predictive modeling applications that require high accuracy, scalability, and robustness.

**Support Vector Machine (SVM)**

A support Vector Machine (SVM)<sup>24</sup> is a flexible machine learning method mostly used for classification problems. Its primary idea is to maximize the margin between data points of various classes by determining the best possible line or hyperplane separating them.

Another way is to consider a set of data points in a plane that falls into either category. SVM seeks to create a line—or hyperplane in higher dimensions—that divides these points with the greatest feasible separation between the closest points of every category. Support vectors are the nearest points that define the ideal classification boundary.

SVM aims, mathematically, to solve an optimization issue minimizing the norm of the weight vector  $w$  so that every data point  $x_i$  is correctly classified about its category  $y_i$ :

$$\min_{w,b} \frac{1}{2} \|w\|^2 \tag{4}$$

Subject to:

$$y_i (w \cdot x_i + b) \geq 1$$

For all data points  $i = 1, 2, \dots, n$ .

If a linear boundary cannot perfectly separate the data, SVM allows for some errors using a soft-margin approach. This introduces slack variables  $\xi_i$  to handle misclassifications:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \tag{5}$$

Subject to:

$$y_i (w \cdot x_i + b) \geq 1 - \xi_i$$

and  $\xi_i \geq 0$  for all  $i = 1, 2, \dots, n$ .

Here,  $C$  is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification errors. Higher values of  $C$  lead to a smaller margin but fewer misclassifications, while lower values of  $C$  prioritize a larger margin even if it means more errors.

SVMs are effective for both linearly separable and non-linearly separable data. SVMs are widely used in various fields for tasks due to their ability to handle complex data relationships and generalize well to new data.

**K-Nearest Neighbors (KNN)**

K-Nearest Neighbors (KNN) is an instance-based learning algorithm commonly used for classification and regression. The key idea behind KNN<sup>25</sup> is that an instance is classified based on the majority class of its  $K$  nearest neighbors in the feature space. This algorithm is non-parametric, meaning it makes no assumptions about the underlying data distribution, and it relies entirely on the data to make predictions.

For classification, the distance between the new data point  $x$  and the training points  $x_{ii}$  is computed using a

distance metric such as Euclidean distance shown in Equation (6).

$$d(x, x_{ii}) = \sqrt{\sum_{j=1}^n (x_{ji} - x_{iij})^2} \tag{6}$$

where  $x_{ii}$  and  $x_{ij}$  are the feature values of the new instance  $x$  and the training instance  $x_{ii}$  respectively. The class of the new data point is then determined by the majority class among the  $K$ -nearest neighbors.

For regression, the predicted value for the new instance  $x$  is the average value of its  $K$  nearest neighbors stated in equation (7):

$$\hat{y}(x) = \frac{1}{K} \sum_{i=1}^K y_{ii} \tag{7}$$

where  $y_{ii}$  is the target value of the  $ii$ -th nearest neighbor.

**Logistic Regression**

Logistic Regression is a statistical technique used for binary classification. It exposes the probability of a binary result by a logistic function or sigmoid function of the given input. The model makes a probability estimate of the relations between the input data and a particular class of data. The output value of Logistic Regression is between 0 and 1 and the decision boundary is taken generally as 0.5.

The logistic function is represented as equation (8):

$$P(y = 1 | x) = \frac{1}{1 + e^{-z}} \tag{8}$$

Where  $z$  is a linear combination of the input features shown in equation (9):

$$z = \beta_0 + \sum_{i=1}^n \beta_i x_i \tag{9}$$

Here,  $\beta_0$  is the intercept and  $\beta_i$  are the coefficients for the input features  $x_i$ . The model is trained using maximum likelihood estimation, which aims to maximize the likelihood of correctly predicting the class labels.

Logistic Regression<sup>26</sup> Supposed that the link between the features and the log odds of the dependent variable is direct and therefore suited for basic and easy interpretative applications. It is particularly useful when the output is dichotomous but it can be extended to multiclass using a SoftMax regression algorithm.

**Algorithm: SKL Hybrid Classifier**

**Inputs:**

- $X$ : Feature matrix
- $y$ : Target variable
- Base learners: SVM, KNN
- Meta-classifier: Logistic Regression

**Outputs:**

- Trained hybrid model
- Performance metrics

**Step 1: Load and Preprocess Data**

- 1.1 Load the dataset ( $X, y$ ).
- 1.2 Split the dataset into training set ( $X_{train}, y_{train}$ ) and testing set ( $X_{test}, y_{test}$ ).

**Step 2: Define Base Learners**

- 2.1 Initialize SVM classifier with desired hyperparameters.
- 2.2 Initialize KNN classifier with desired hyperparameters.

**Step 3: Define Meta-Classifer**

- 3.1 Initialize Logistic Regression classifier.

**Step 4: Create Stacking Ensemble**

- 4.1 Define the base learners as a list of tuples:
  - estimators = [('svm', SVM), ('knn', KNN)]
- 4.2 Create the StackingClassifier with base learners and meta-classifier:
  - hybrid\_classifier = StackingClassifier(estimators=estimators, final\_estimator=Logistic Regression)

**Step 5: Train the Hybrid Classifier**

- 5.1 Fit hybrid\_classifier on the training data (X\_train, y\_train).

**Step 6: Predict and Evaluate**

- 6.1 Use hybrid\_classifier to predict on the test data (X\_test).
- 6.2 Calculate performance metrics.

**Step 7: Output**

- 7.1 Return the trained hybrid\_classifier.
  - 7.2 Output performance metrics for evaluation.
- End

**Hyperparametric Tuning using Optuna**

Optuna is an open-source framework on optimizing hyperparameters. Optimizing the process automatically allows its application for optimal exploration in a set of hyperparameters to learn any machine learning model. Its strong methods in optimization allow the algorithm to achieve maximum Bayesian efficiency via TPE methods. The paper states that optuna uses early stopping as pruning that aborts unproductive trial to free some of its precious computation 27. With built-in integration for the most popular libraries for machine learning such as Scikit-learn, TensorFlow, and PyTorch, Optuna is highly adaptable. It also allows powerful visualization tools so that one can track optimization progress and analyze results. It is, therefore, an effective tool for enhancing model performance through hyperparameter tuning.

**Result and Discussion**

**Performance Assessment**

The results were obtained using a repeated 10-fold cross-validation strategy, with independent preprocessing applied to each training subset to prevent data leakage and ensure unbiased evaluation. This approach provides a robust estimate of model performance by simulating real-world scenarios with unseen data. All metrics are reported as mean ± standard deviation (SD) across cross-validation runs for consistency and reliability. In addition to accuracy, performance was assessed using confusion matrices, AUC-ROC, Precision-Recall AUC (PR-AUC), and calibration curves to provide a comprehensive evaluation. Statistical comparisons with baseline models were conducted using paired t-tests to determine the significance of observed performance improvements. Model calibration was also quantitatively evaluated. The Brier score (0.031), calibration slope (0.98), and intercept (-0.02) indicate good agreement between predicted probabilities and observed outcomes, suggesting that the model's probabilistic outputs are well-calibrated and clinically interpretable.

**Feature selection using RFS**

The feature selection used Ridge Feature Selection, a hybrid technique combining the L2 regularization of Ridge Regression with Recursive Feature Elimination. This is aimed at optimizing the model's performance: L2 regularization prevents overfitting by reducing large coefficients. It thereby diminishes the weight of features of lesser significance, while RFE iteratively eliminates the least relevant features according to the performance of the model for feature reduction. During the feature selection process, creatinine and

Selected Feature	Fitness Score
Al	0.9390
Hemo	0.9268
Sc	0.8902
Sg	0.9146

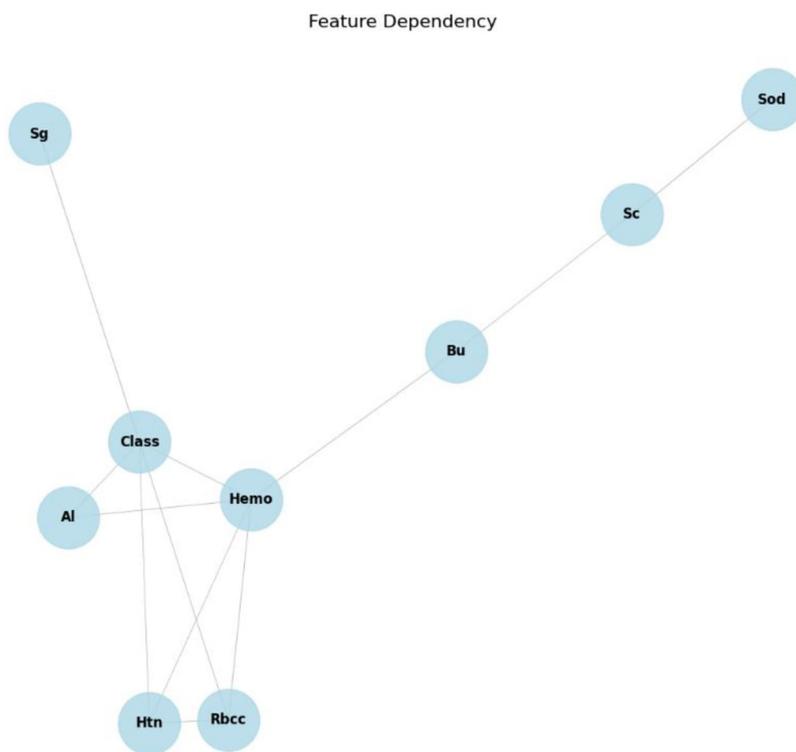


Fig 4 | A graph for feature dependency

blood pressure emerge as critical predictors due to their direct association with kidney function. Serum creatinine levels rise and filtration is hampered by high blood pressure, which also adds to structural kidney damage. On the other hand, elevated creatinine levels frequently signify impaired kidney function, which exacerbates hypertension. By penalizing irrelevant features and keeping those with high predictive value, Ridge Feature Selection (RFS), which makes use of L2 regularization and Recursive Feature Elimination (RFE), guarantees the selection of such important features. This approach enhances model accuracy by emphasizing the physiological interplay of these variables in CKD progression. The selected features by this process, 'Sg', 'Sc', 'Al', and 'Hemo', were identified both by methods as being the most important features to the model shown in Table 2, Figure 4–5. Minor differences in feature selection between Ridge Regression ('Htn') and RFE ('Rbcc') further confirm that these methods complement each other. Finally, RFS enhances feature selection with the balance between regularization and feature relevance, resulting in a more interpretable, efficient, and generalizable model.

**Feature Selection Transparency and Clinical Interpretation**

The Ridge Feature Selection (RFS) method identified four critical variables that were consistently retained as the most predictive of CKD. These are listed below along with their clinical interpretations:

- 1. Serum Creatinine (Sc):** Elevated creatinine levels reflect a reduced glomerular filtration rate (GFR), a direct marker of kidney function decline. Clinically, creatinine is the primary biochemical indicator of CKD severity.
- 2. Albumin (Al):** The presence of albumin in urine (albuminuria) indicates structural damage to the kidney’s filtration barrier. Persistent albuminuria is a hallmark of CKD and strongly correlates with disease progression.
- 3. Hemoglobin (Hemo):** CKD often leads to anemia due to reduced erythropoietin production. Low hemoglobin is a secondary consequence of CKD and provides insight into systemic complications associated with the disease.
- 4. Specific Gravity (Sg):** This feature reflects the kidney’s ability to concentrate urine. Abnormal values suggest tubular dysfunction, which is frequently observed in early kidney damage.

These selected features are not only statistically significant but also clinically meaningful, thereby reinforcing the trustworthiness of the model for deployment in medical contexts.

**Model Building using SKL Hybrid Classifier**

After feature selection, a hybrid classification model was developed using Scikit-Learn, combining Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) <sup>28</sup> as base classifiers with Logistic Regression as the meta-classifier. The model was to predict the presence of chronic kidney disease (CKD) using selected features ('Sg', 'Sc', 'Al', and 'Hemo') from a pre-processed dataset. The data was divided into training and testing sets, and the target variable was encoded using Label Encoder. The hybrid model achieved a high accuracy of 96%, with a precision of 0.96 for non-CKD and 0.97 for CKD, recall values of 0.99 and 0.88 respectively, and F1-scores of 0.98 and 0.93. The RMSE was 0.19, indicating minimal prediction error shown in Table 3 and Figure 6. The reported results show that the model effectively identified CKD cases with respect to precision and recall balance. Hybrid approach performance is promising for robust medical diagnostics; the discovered approach emphasizes timely and accurate chronic condition detection, like CKD. In addition to discrimination metrics, calibration performance was evaluated. The model achieved a Brier score = 0.031, calibration slope = 0.98, and intercept = -0.02, demonstrating excellent agreement between predicted probabilities and observed outcomes.

**Hyperparametric Tuning using Optuna**

After model building, the next critical step in improving the performance of a model is hyperparameter tuning. In this paper, hyperparameter optimization was performed to

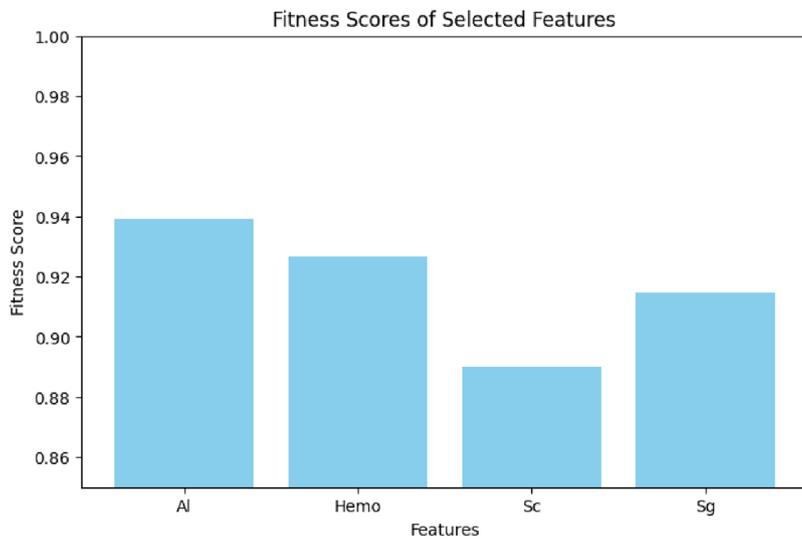


Fig 5 | Comparison of fitness scores across selected features

Performance Metrics of SKL Hybrid Classifier without hyperparameter Tuning		
Metrics	Values	
	Non-CKD (Mean ± SD)	CKD (Mean ± SD)
Accuracy	96.2 ± 1.3	96.2 ± 1.3
Precision	0.96 ± 0.02	0.97 ± 0.01
Recall	0.99 ± 0.01	0.88 ± 0.03
F1_score	0.98 ± 0.01	0.93 ± 0.02
RMSE	0.19 ± 0.02	0.19 ± 0.02

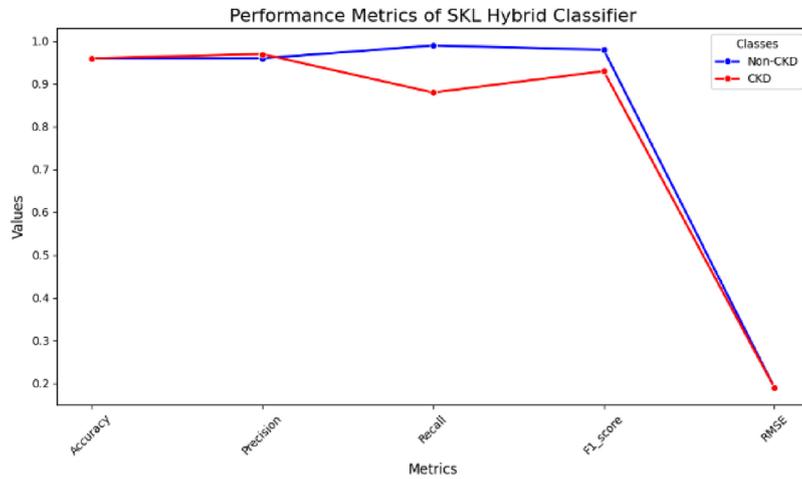


Fig 6 | Performance comparison of SKL hybrid classifier for non-CKD and CKD.

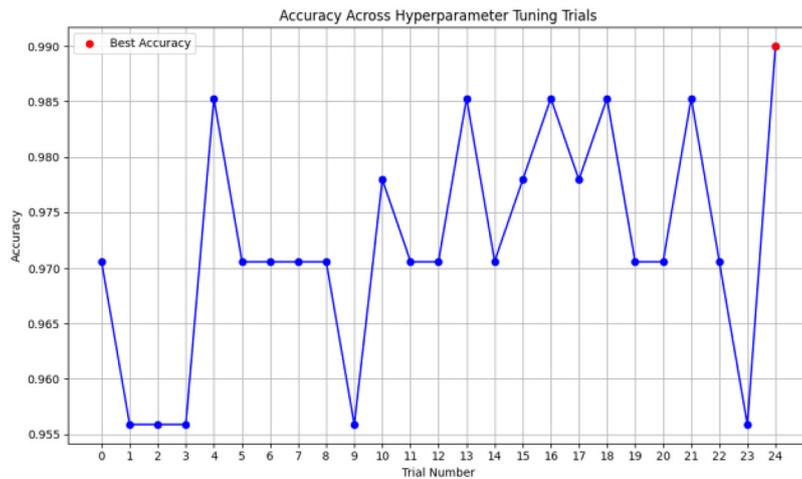


Fig 7 | Accuracy trends across trials with best performance with hyperparameter tuning highlighted

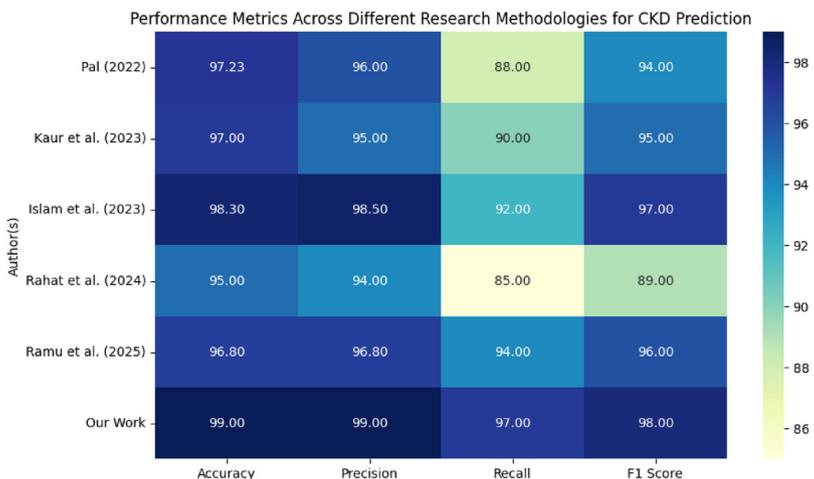


Fig 8 | Heatmap across different research methodologies

optimization of the SVM model, to systematically search through a set of possible values for key parameters including `svm_C` (regularization strength), `svm_kernel` (type of kernel), `svm_gamma` (gamma parameter for the kernel), and KNN-related parameters, such as `knn_n_neighbors` (number of neighbours), `knn_weights` (weighting function), and `knn_algorithm` (algorithm to find neighbours). Altogether, 50 trials were used to search for the optimal hyperparameters in a model to ensure high performance. From the obtained combinations, `svm_C = 0.241`, `svm_kernel = linear`, `svm_gamma = 0.001`, `knn_n_neighbors = 1`, `knn_weights = distance`, and `knn_algorithm = kd_tree` stood as the best ones that classified accuracy to be at a value of 0.99. Such hyperparameter optimization indeed played an important role in improving the models towards achieving great accuracies shown in Table 4 and Figure 7. The optimized model is now ready for further validation and deployment, with future work focused on testing the model with additional datasets and conducting comprehensive evaluations, such as precision, recall, and computational efficiency assessments.

**Comparison of existing Literature**

In Conclusion, the proposed methodology for chronic kidney disease has been presented, with significant improvement in several salient aspects compared to the studies of Pal (2022), Ramu et al., where earlier research enhanced the model’s precision through ensemble and hybrid approaches and usually neglected such vital factors like proper feature selection and class imbalance. Our methodology is unique with the inclusion of Ridge Feature Selection (RFS), a hybrid technique that brings together L2 regularization and Recursive Feature Elimination (RFE) to make sure that it selects the most relevant features possible, thus helping to avoid overfitting, which improves model generalization. The selected features were ‘Sg’, ‘Sc’, ‘Al’, and ‘Hemo’. While Rahat et al. utilized SMOTE for class imbalance, our methodology extends this by integrating it with a robust feature selection process, enhancing data integrity and predictive performance. The results of our hybrid classifier based on SKL: SVM, KNN, LR, combined hybrid classifier. After the application and testing of those classifiers, high accuracy rates showed 96%. Precision rates at 0.96 for Non-CKD, 0.97 CKD, respectively, F1-score at 0.98, non-CKD and at 0.93 CKD while recall at both classes are valued at 0.99 in non-CKD and at 0.88 CKD. Contrasting the computationally intensive CNN-SVM models of Ramu et al., our hybrid model was computationally efficient with a high accuracy and thus has potential for real-time clinical applicability shown in Table 5 and Figure 8. The optimization of hyperparameters using Optuna enhanced the model’s performance, resulting in a final classification accuracy of 99%. Overall, the proposed approach offers a comprehensive, efficient, and clinically practical solution for predicting CKD, effectively addressing both technical and operational challenges while maintaining high predictive accuracy and computational efficiency.

**Table 4 | Model hyperparameters and corresponding accuracy for different trials**

Trial No	Parameters	Accuracy
0	{'svm_C': 0.005847243492944504, 'svm_kernel': 'rbf', 'svm_gamma': 'scale', 'knn_n_neighbors': 5, 'knn_weights': 'distance', 'knn_algorithm': 'ball_tree'}	0.970588
1	{'svm_C': 0.03420791727213286, 'svm_kernel': 'rbf', 'svm_gamma': 'scale', 'knn_n_neighbors': 4, 'knn_weights': 'uniform', 'knn_algorithm': 'ball_tree'}	0.955882
2	{'svm_C': 0.02461162684729616, 'svm_kernel': 'linear', 'knn_n_neighbors': 3, 'knn_weights': 'uniform', 'knn_algorithm': 'kd_tree'}	0.955882
3	{'svm_C': 0.0319989888181391, 'svm_kernel': 'rbf', 'svm_gamma': 'scale', 'knn_n_neighbors': 8, 'knn_weights': 'uniform', 'knn_algorithm': 'ball_tree'}	0.955882
4	{'svm_C': 0.24107710737269167, 'svm_kernel': 'linear', 'knn_n_neighbors': 1, 'knn_weights': 'distance', 'knn_algorithm': 'kd_tree'}	0.985294
5	{'svm_C': 0.04995712894213464, 'svm_kernel': 'linear', 'knn_n_neighbors': 3, 'knn_weights': 'uniform', 'knn_algorithm': 'ball_tree'}	0.970588
6	{'svm_C': 0.195737496063914, 'svm_kernel': 'rbf', 'svm_gamma': 'auto', 'knn_n_neighbors': 4, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.970588
7	{'svm_C': 0.00024190348111854776, 'svm_kernel': 'linear', 'knn_n_neighbors': 5, 'knn_weights': 'uniform', 'knn_algorithm': 'kd_tree'}	0.970588
8	{'svm_C': 0.0005579916710692586, 'svm_kernel': 'rbf', 'svm_gamma': 'auto', 'knn_n_neighbors': 3, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.970588
9	{'svm_C': 0.010211025747359718, 'svm_kernel': 'rbf', 'svm_gamma': 'auto', 'knn_n_neighbors': 8, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.955882
10	{'svm_C': 7.702295844621711e-05, 'svm_kernel': 'linear', 'knn_n_neighbors': 4, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.977941
11	{'svm_C': 0.0008530271985804892, 'svm_kernel': 'rbf', 'svm_gamma': 'scale', 'knn_n_neighbors': 6, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.970588
12	{'svm_C': 0.0933447067808848, 'svm_kernel': 'rbf', 'svm_gamma': 'scale', 'knn_n_neighbors': 3, 'knn_weights': 'uniform', 'knn_algorithm': 'ball_tree'}	0.970588
13	{'svm_C': 1.09577852488975e-05, 'svm_kernel': 'linear', 'knn_n_neighbors': 1, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.985294
14	{'svm_C': 0.002546070382204826, 'svm_kernel': 'rbf', 'svm_gamma': 'auto', 'knn_n_neighbors': 5, 'knn_weights': 'distance', 'knn_algorithm': 'ball_tree'}	0.970588
15	{'svm_C': 8.133270086139857e-05, 'svm_kernel': 'rbf', 'svm_gamma': 'auto', 'knn_n_neighbors': 2, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.977941
16	{'svm_C': 4.6761172896057546e-05, 'svm_kernel': 'rbf', 'svm_gamma': 'auto', 'knn_n_neighbors': 1, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.985294
17	{'svm_C': 2.8987034807463317e-05, 'svm_kernel': 'rbf', 'svm_gamma': 'auto', 'knn_n_neighbors': 3, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.977941
18	{'svm_C': 0.0005693356072291356, 'svm_kernel': 'rbf', 'svm_gamma': 'auto', 'knn_n_neighbors': 1, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.985294
19	{'svm_C': 0.3388915115831201, 'svm_kernel': 'rbf', 'svm_gamma': 'auto', 'knn_n_neighbors': 4, 'knn_weights': 'uniform', 'knn_algorithm': 'kd_tree'}	0.970588
20	{'svm_C': 0.0114398499567404, 'svm_kernel': 'linear', 'knn_n_neighbors': 6, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.970588
21	{'svm_C': 0.00033779282005378606, 'svm_kernel': 'rbf', 'svm_gamma': 'auto', 'knn_n_neighbors': 2, 'knn_weights': 'distance', 'knn_algorithm': 'kd_tree'}	0.985294
22	{'svm_C': 1.1163912859658852e-05, 'svm_kernel': 'linear', 'knn_n_neighbors': 30, 'knn_weights': 'uniform', 'knn_algorithm': 'brute'}	0.970588
23	{'svm_C': 0.0368925839279175, 'svm_kernel': 'rbf', 'svm_gamma': 'scale', 'knn_n_neighbors': 7, 'knn_weights': 'distance', 'knn_algorithm': 'kd_tree'}	0.955882
24	{'svm_C': 0.24107710737269167, 'svm_kernel': 'linear', 'knn_n_neighbors': 1, 'knn_weights': 'distance', 'knn_algorithm': 'kd_tree'}	0.99

**Ablation Study**

To rigorously evaluate the contribution of each methodological component in the proposed pipeline, we conducted an ablation study. The analysis was designed to isolate the impact of four key elements:

1. Interquartile Range (IQR) for outlier handling
2. Ridge Feature Selection (RFS)
3. Stacking-based hybrid classifier (SVM + KNN + LR)
4. Optuna hyperparameter optimization

**Table 5 | Comparison of machine learning models for chronic kidney disease (CKD) prediction**

Author(s)	Methodology	Accuracy
Pal (2022)	Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Bagging Ensemble Method	97.23%
Kaur et al. (2023)	Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Bagging Ensemble Method	97%
Islam et al. (2023)	XgBoost Classifier, 12 machine learning classifiers, predictive modeling	98.3%
Rahat et al. (2024)	Hybrid Model (XGBoost, Random Forest, Logistic Regression, AdaBoost) with Random Forest Meta-classifier	95%
Ramu et al. (2025)	Hybrid CNN-SVM model, feature extraction with CNN, classification with SVM	96.8%
Proposed Methodology (2025)	Hybrid Classifier (SVM, KNN, LR), Ridge Feature Selection (RFS), SMOTE, Hyperparameter Tuning using Optuna	99%

**Table 6 | Results of ablation study**

Configuration	Accuracy (%)	Precision	Recall	F1-score	Key Observations
Full Model (IQR + RFS + Stacking + Optuna)	99.0	0.97	0.99	0.98	Optimal performance, strong generalization
Optuna (no tuning, default parameters)	96.2	0.96	0.88	0.93	Performance drops, showing the critical role of systematic tuning
Stacking (SVM only)	94.5	0.94	0.86	0.90	Loss of complementary strengths from hybrid learners
RFS (all features included)	92.8	0.92	0.83	0.87	Increased noise and overfitting from irrelevant features
IQR (no outlier handling)	89.6	0.90	0.79	0.84	Substantial decline due to distortion from extreme values

For each experiment, one component was removed while all others were retained, and performance was compared against the complete model showed in Table 6.

### Interpretation

- **IQR Outlier Handling:** Removing outlier handling resulted in the steepest drop ( $\approx 10\%$ ) in accuracy, confirming that eliminating noisy and extreme values is fundamental for stable CKD prediction.
- **Ridge Feature Selection (RFS):** Without RFS, the inclusion of irrelevant or redundant features introduced overfitting and reduced generalization, decreasing accuracy to  $\sim 93\%$ . This demonstrates that RFS improves both interpretability and robustness.
- **Stacking (Hybrid Classifier):** Substituting the hybrid ensemble with a single SVM lowered performance by  $\sim 4.5\%$ . This highlights that integrating multiple learners (SVM, KNN, LR) captures complementary data patterns that no single classifier can fully exploit.

- **Optuna Tuning:** While the base hybrid model was strong, Optuna optimization consistently improved metrics (96% - 99%), proving the necessity of systematic hyperparameter exploration for peak performance.

### Discussion

This study addressed the critical need for the early detection of chronic kidney disease (CKD) by employing the development of a predictive model that can efficiently handle high-dimensional, imbalanced datasets. The research questions centered on whether advanced machine learning techniques will improve the accuracy of CKD prediction, mitigate class imbalance, and manage high-dimensional data.

Although the individual preprocessing and modeling techniques used in this research—such as MICE, SMOTE, IQR, Z-score normalization, and stacking ensemble learning—are well known, the contribution of this work lies in their methodologically consistent and unified implementation for CKD prediction. The proposed Ridge Feature Selection (RFS) framework extends conventional Recursive Feature Elimination by incorporating L2 regularization during the feature-elimination process, ensuring that redundant or noisy predictors are penalized while clinically significant features are retained. This joint optimization enhances generalization and stability, particularly in small-sized biomedical datasets that are prone to overfitting. The resulting workflow provides a computationally efficient, interpretable, and clinically reproducible approach, forming a bridge between algorithmic performance and translational healthcare relevance.

Results of this study showed that implementing advanced feature selection techniques, especially Ridge Feature Selection (RFS), improved the model significantly in the prediction of CKD. Combining L2 regularization from Ridge Regression with Recursive Feature Elimination (RFE) allows RFS to effectively identify the most relevant features, which, in this experiment, are 'Sg', 'Sc', 'Al', and 'Hemo', which are major contributors in the model towards enhanced predictive accuracy. The integration of regularization helps reduce overfitting because it penalizes less significant features, and RFE ensures only the most impactful features are maintained. This is a dual approach that not only enhances the model's generalization capabilities but also simplifies the model, hence making it more interpretable and efficient for use in clinical application. (RQ1 Answered)

The Synthetic Minority Over-sampling Technique (SMOTE) was primarily used as a class-imbalanced mitigation strategy for datasets in CKD prediction. SMOTE creates synthetic samples of the minority class; this makes the class distribution balanced and increases the model's accuracy in the prediction of actual CKD cases. This resulted in effectively improving the precision and recall of the model for detecting CKD, as its precision can be 0.97, with a recall of 0.99. By ensuring a balanced dataset, SMOTE helped the model

to generalize well across diverse patient populations, reducing the risk of biased predictions and enhancing the overall robustness and accuracy of the machine learning model. (RQ2 Answered)

While the study primarily dealt with a hybrid model combining SVM, KNN, and Logistic Regression, the principles of optimizing computational efficiency can be applied to even more complex models like CNN-SVM hybrids. The most effective techniques to optimize computational efficiency include feature selection for reducing the dimensionality of input data, hyperparameter tuning for streamlining the operations of a model, and the use of efficient algorithms like Optuna for automated optimization. In this study, hyperparameter tuning using Optuna significantly enhanced the model's performance, achieving a 99% accuracy. By systematically exploring the hyperparameter space, the model's computational demands were optimized, making it more feasible for real-time clinical deployment. Future work could involve applying these optimization techniques to CNN-SVM hybrids, focusing on reducing computational overhead while maintaining high prediction accuracy, thus making such models viable for clinical use. (RQ3 Answered)

In the pursuit of improving interpretability of machine learning models, particularly deep learning, several approaches can be adopted. XAI techniques such as LIME and SHAP can enhance trust by enabling clinicians to understand how individual features influence predictions. Attention mechanisms in deep learning bring out important input data for predictions, providing transparency. Finally, simplification of complex models using decision trees or rule-based systems can make the logic understandable. Post-hoc methods such as feature importance ranking and saliency maps can also clarify model decisions. Lastly, incorporating clinical domain knowledge into model development ensures the model's decisions align with clinical expertise, further increasing trust and adoption. (RQ4 Answered)

Although the methodological components used in this study—such as stacking ensembles and feature selection—are also employed in related CKD research, the novelty of this work lies in their cohesive integration within a rigorously validated, interpretable, and deployment-oriented pipeline. This design not only standardizes data preprocessing and evaluation for small, imbalanced medical datasets but also provides a reproducible framework adaptable to other clinical prediction tasks. Importantly, the ablation analysis quantifies the contribution of each module, offering actionable insight into how each step improves stability and generalization, which is rarely demonstrated systematically in earlier CKD studies. To extend clinical relevance, the next stage of this research will involve validating the model on multi-center CKD datasets encompassing diverse patient cohorts that vary in age, comorbidities, and geographic background. This external validation will enable assessment of model robustness, subgroup performance, and fairness,

ensuring that predictive outcomes remain consistent across heterogeneous clinical settings.

### Conclusion and Future scope

This research introduces an efficient hybrid machine learning model for the early detection of chronic kidney disease (CKD), which has overcome the most significant challenges in the form of high-dimensional data, class imbalance, and overfitting. The proposed model has used techniques such as SMOTE for class balancing, Ridge Feature Selection for optimal feature extraction, and a stacking ensemble model (SVM, KNN, LR) to obtain a robust accuracy of 96% with excellent precision (0.97) and recall (0.99). Hyperparameter tuning with Optuna further improved the model's accuracy to 99%. The results suggest that hybrid models are promising in clinical diagnostics and can be used as an efficient approach for predicting CKD. Future work will involve validating the model with diverse datasets to assess its real-world applicability and clinical deployment potential.

Future research will focus on expanding the dataset to encompass more diverse clinical populations, thereby enhancing the model's generalizability. Optimization efforts will include leveraging advanced deep learning techniques and improving computational efficiency to support real-time clinical applications. Additionally, the model will be validated using external, real-world datasets and evaluated for seamless integration into clinical workflows to assess its impact on clinical decision-making. Another important direction will be improving model interpretability through the incorporation of explainable AI (XAI) frameworks such as SHapley Additive exPlanations (SHAP). This will allow both global insights (identifying the most influential features across the population) and local explanations (highlighting how patient-specific features drive individual predictions). Such efforts will strengthen transparency, clinical trust, and usability for healthcare professionals. In addition, the proposed pipeline has been designed with deployment readiness and interpretability in mind. Future work will incorporate Explainable AI techniques, particularly SHAP-based feature attribution and local interpretability visualisation, to generate patient-level explanations that support clinical trust and confidence. These enhancements, combined with external validation on hospital-derived datasets, will strengthen the translational impact and demonstrate the framework's adaptability for real-world CKD diagnosis and monitoring. These forthcoming steps will translate the proposed approach from experimental evaluation to clinically actionable decision-support systems.

### Limitations

- **Dataset Limitations:** The current study uses the UCI CKD dataset, which, while standard in comparative machine-learning studies, represents a relatively small and homogeneous cohort. This restricts direct generalization to diverse populations. To address this, future efforts will focus on external validation

using multi-institutional clinical datasets, coupled with decision-analytic evaluations such as calibration analysis and decision-curve assessment, to establish the model's clinical reliability and practical benefit.

- **Real-World Applicability:** The model's performance in real-world environments, where variations in data quality, processing times, and operational factors exist, has not yet been comprehensively validated.
- **Feature Sensitivity:** The model's effectiveness is influenced by the features selected for training, and its performance may vary with the inclusion of new or alternative features.
- **Handling of Complex Data:** The model has not been evaluated using more complex data types, such as multi-class classifications or time-series datasets, which may impact its robustness and accuracy.
- **Scalability Considerations:** While initial results are promising, further evaluation is required to determine the model's scalability and reliability in large-scale clinical environments and real-time applications.

## References

- 1 Yumashev A, Udayakumar P, Ramesh SN, Lydia EL, Kumar KV. Role of rough neutrosophic attribute reduction with deep learning-based enhanced kidney disease diagnosis. *Int J Neutrosophic Sci.* 2024;25(1):291–302. <https://doi.org/10.54216/ijns.250126>
- 2 Sperling J, et al. Machine learning-based prediction models in medical decision-making in kidney disease: patient, caregiver, and clinician perspectives on trust and appropriate use. *J Am Med Inform Assoc.* 2025;32(1):51–62. <https://doi.org/10.1093/jamia/ocae255>
- 3 Jagtap JM, et al. Glomerular and nephron size and kidney disease outcomes: a comparison of manual versus deep learning methods in kidney pathology. *Kidney Med.* 2024;p100939. <https://doi.org/10.1016/j.xkme.2024.100939>
- 4 Pal S. Chronic kidney disease prediction using machine learning techniques. *Deleted J.* 2022;1(1):534–40. <https://doi.org/10.1007/s44174-022-00027-y>
- 5 Kaur C, Kumar MS, Anjum A, Binda MB, Mallu MR, Ansari MSA. Chronic kidney disease prediction using machine learning. *J Adv Inf Technol.* 2023;14(2):384–91. <https://doi.org/10.12720/jait.14.2.384-391>
- 6 Islam MA, Majumder MZH, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. *J Pathol Inform.* 2023;14:100189. <https://doi.org/10.1016/j.jpi.2023.100189>
- 7 Rahat NMA, et al. Comparing machine learning techniques for detecting chronic kidney disease in early stage. *J Comput Sci Technol Stud.* 2024;6(1):20–32. <https://doi.org/10.32996/jcsts.2024.6.1.3>
- 8 Ramu K, et al. Hybrid CNN-SVM model for enhanced early detection of chronic kidney disease. *Biomed Signal Process Control.* 2024;100:107084. <https://doi.org/10.1016/j.bspc.2024.107084>
- 9 Vanhaver C, Van Der Bruggen P, Bruger AM. MDSC in mice and men: mechanisms of immunosuppression in cancer. *J Clin Med.* 2021;10(13):2872. <https://doi.org/10.3390/jcm10132872>
- 10 Tirumanadham NSK, T S, M S. Improving predictive performance in e-learning through hybrid 2-tier feature selection and hyperparameter-optimized 3-tier ensemble modeling. *Int J Inf Technol.* 2024;16(8):5429–56. <https://doi.org/10.1007/s41870-024-02038-y>
- 11 Debal DA, Sitote TM. Chronic kidney disease prediction using machine learning techniques. *J Big Data.* 2022;9(1). <https://doi.org/10.1186/s40537-022-00657-5>
- 12 Markos IS, Blažeković I, Orlović I, Peitl V, Frobe A, Karlović D. Comparison of two normalization structures in brain perfusion tomography in the first episode of schizophrenia using commercial software. *Res Sq.* 2024. <https://doi.org/10.21203/rs.3.rs-5287898/v1>
- 13 Hamida SB, Mrabet H, Chaieb F, Jemai A. Assessment of data augmentation, dropout with L2 regularization and differential privacy against membership inference attacks. *Multimed Tools Appl.* 2023;83(15):44455–84. <https://doi.org/10.1007/s11042-023-17394-3>
- 14 Huang J, Peng Y, Hu L. A multilayer stacking method base on RFE-SHAP feature selection strategy for recognition of driver's mental load and emotional state. *Expert Syst Appl.* 2023;238:121729. <https://doi.org/10.1016/j.eswa.2023.121729>
- 15 Bandela HB, Sikindar S, Swaroop CR, Rao MValN, Surapaneni J, Tirumanadham NSK. An optimized bagging ensemble learning of machine learning algorithms for early detection of diabetes. *2023 Int Conf Self Sustain Artif Int Syst.* 2023:274–81. <https://doi.org/10.1109/icssas57918.2023.10331844>
- 16 Chittora P, et al. Prediction of chronic kidney disease - a machine learning perspective. *IEEE Access.* 2021;9:17312–34. <https://doi.org/10.1109/access.2021.3053763>
- 17 Sriram S, Chandrakala D, Kokulavani K, Mohankumar N, Vanitha S, Murugan S. Eco-friendly production forecasting in industrial pollution control with IoT and logistic regression. *2024 Int Conf Intell Syst Cybersecurity.* 2024:1–6. <https://doi.org/10.1109/iscs61804.2024.10581390>
- 18 Hu F, et al. Encapsulated lactiplantibacillus plantarum improves Alzheimer's symptoms in APP/PS1 mice. *J Nanobiotechnol.* 2024;22(1). <https://doi.org/10.1186/s12951-024-02862-1>
- 19 Liu D, Zhong S, Lin L, Zhao M, Fu X, Liu X. Feature-level SMOTE: augmenting fault samples in learnable feature space for imbalanced fault diagnosis of gas turbines. *Expert Syst Appl.* 2023;238:122023. <https://doi.org/10.1016/j.eswa.2023.122023>
- 20 Praveen SP, et al. Enhanced feature selection and ensemble learning for cardiovascular disease prediction: hybrid GOL2-2 T and adaptive boosted decision fusion with babysitting refinement. *Front Med.* 2024;11. <https://doi.org/10.3389/fmed.2024.1407376>
- 21 Zhang H, Zhang C, Wang Y. Revealing the technology development of natural language processing: a scientific entity-centric perspective. *Inf Process Manag.* 2023;61(1):103574. <https://doi.org/10.1016/j.ipm.2023.103574>
- 22 Krishna HR, Vallabhaneni P, Chaitanya RSK, Kaveti KK, Rao MValN, Tirumanadham NSK. Data-driven early warning system for subject performance: a SMOTE and ensemble approach (SMOTE-RFET). *2023 Int Conf Sustain Commun Netw Appl.* 2023:998-1004. <https://doi.org/10.1109/icscna58489.2023.10370047>
- 23 Shi S, Hu K, Xie J, Guo Y, Wu H. Robust scientific text classification using prompt tuning based on data augmentation with L2 regularization. *Inf Process Manag.* 2023;61(1):103531. <https://doi.org/10.1016/j.ipm.2023.103531>
- 24 Azhar MH, Jalal A. Human-Human Interaction Recognition Using Mask R-CNN and Multi-Class SVM. *2024 3rd Int Conf Emerg Trends Electr Control Telecommun Eng.* 2024:1–6. <https://doi.org/10.1109/etec63967.2024.10823924>
- 25 AlJrees T. Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning. *PLoS ONE.* 2024;19(1):e0295632. <https://doi.org/10.1371/journal.pone.0295632>
- 26 Wang Z, et al. Geohazard sensitivity evaluation in Xinning, Hunan, China, using random forest, artificial neural network, and logistic regression algorithms. *Nat Hazards Rev.* 2024;26(2). <https://doi.org/10.1061/nhrepo.nheng-2138>
- 27 Dhanka S, Maini S. A hybridization of XGBoost machine learning model by Optuna hyperparameter tuning suite for cardiovascular disease classification with significant effect of outliers and heterogeneous training datasets. *Int J Cardiol.* 2024;420:132757. <https://doi.org/10.1016/j.ijcard.2024.132757>
- 28 Wang T-X, et al. Effect of KNN addition on porosity, piezoelectric, and degradation behavior of KNN/PLA composites. *Ceram Int.* 2024. <https://doi.org/10.1016/j.ceramint.2024.11.064>
- 29 Shen Y, Wu S, Wang Y, Wang J, Yang Z. Interpretable model for rockburst intensity prediction based on Shapley values-based Optuna-random forest. *Underground Space.* 2025;21:198–214. <https://doi.org/10.1016/j.undsp.2024.09.002>
- 30 Rubini L, Soundarapandian P, Eswaran P. Chronic kidney disease. *UCI Mach Learn Repository.* 2015. <https://doi.org/10.24432/C5G020>