Journal of Intelligent Systems and Internet of Things
Vol. 17, No. 02, PP. 152-190, 2025
DOI: https://doi.org/10.54216/JISIoT.170211

Greylag Goose Optimization for Feature Selection and Hyperparameter Tuning in Chronic Kidney Disease Detection

Mohamed Saber[1,*], Ebrahim A. Mattar[2], Marwa M. Eid[3,4], El-Sayed M. El-kenawy[5,6]

[1]Electronics and Communications Engineering Department, Faculty of Engineering, Delta University for Science and Technology, Gamasa City 11152, Egypt
[2]College of Engineering, University of Bahrain, Bahrain
[3]Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura 11152, Egypt
[4]Jadara Research Center, Jadara University, Irbid 21110, Jordan
[5]Delta Higher Institute of Engineering and Technology, Department for Communications and Electronics, Mansoura 35511, Egypt
[6]Applied Science Research Center. Applied Science Private University, Amman, Jordan

Emails: Mohamed.saber@deltauniv.edu.eg; ebmattar@uob.edu.bh; mmm@ieee.org; skenawy@ieee.org

Abstract

Chronic Kidney Disease (CKD) is a global health concern that necessitates accurate and timely detection to improve patient outcomes and reduce healthcare costs. This study focuses on enhancing CKD classification using machine learning techniques, leveraging 400 instances with 25 clinical features to predict binary outcomes of CKD or non-CKD. The main objective is to improve detection accuracy by applying feature selection and model optimization. Standard machine learning models, including Multilayer Perceptron (MLP), Random Forest (RF), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN), were employed, with optimization achieved through binary optimization algorithms such as Greylag Goose Optimization (GGO), Particle Swarm Optimization (PSO), Bat Algorithm (BA), and Whale Optimization Algorithm (WAO), along with hyperparameter tuning using genetic algorithms and other metaheuristics. Results indicate significant improvements in classification performance after feature selection and optimization, with the GGO-optimized MLP model achieving an accuracy of 97.06%. The contributions of this paper include (i) benchmarking baseline models for CKD detection, (ii) a comprehensive analysis of feature selection strategies, (iii) optimization of machine learning models for CKD classification, and (iv) visualization of model performance to aid future research in healthcare machine learning applications.

Received: January 09, 2025 Revised: February 10, 2025 Accepted: March 04, 2025

Keywords: Chronic Kidney Disease; Machine Learning; Feature Selection; Hyperparameter Optimization; Healthcare Analytics

1 Introduction

Chronic Kidney Disease (CKD) has emerged as a major public health concern, with increasing prevalence rates across the globe. CKD is characterized by a gradual decline in kidney function over time, often without any noticeable symptoms until it reaches an advanced stage. Early detection and intervention can significantly reduce the risk of complications such as kidney failure, cardiovascular disease, and the need for costly treatments like dialysis and kidney transplantation. According to recent estimates, CKD affects approximately 10% of the global population, making it a leading cause of morbidity and mortality [1], [2]. As a result, improving early diagnosis methods for CKD is essential for reducing the burden on healthcare systems and improving the quality of life for affected individuals [3].

The significance of early CKD detection extends beyond individual patient outcomes to broader public health and economic implications [4]. Early intervention in CKD can delay or even prevent the progression to end-stage renal disease (ESRD), which is often accompanied by high medical costs due to frequent hospitalizations and long-term treatments [5], [6]. Moreover, by identifying CKD at an early stage, healthcare systems can focus on preventive measures such as lifestyle changes, blood pressure control, and regular monitoring, which are crucial for halting or slowing the disease progression [7], [8]. This proactive approach can also reduce the need for costly kidney transplants and dialysis treatments, which are often the only options for individuals with late-stage kidney disease [9].

Despite the growing need for effective CKD detection methods, several challenges remain in the development of reliable diagnostic tools. One of the primary obstacles is the inherent imbalance in healthcare datasets, where the number of non-CKD cases typically outweighs the number of CKD-positive instances [10], [11]. This imbalance can lead to the problem of model bias, where machine learning algorithms are more likely to predict the majority class (non-CKD) rather than the minority class (CKD) [12]. Additionally, medical datasets often contain missing values, noise, and outliers, which can further complicate model training and decrease the accuracy of predictions [11]. Handling missing values and ensuring the robustness of models against such noise are crucial steps in the development of reliable CKD detection systems. Furthermore, the complexity of medical data, which includes a wide range of interrelated features such as blood pressure, age, red blood cell count, and serum creatinine levels, requires advanced machine learning techniques capable of identifying important patterns amidst a large number of variables [13], [14].

Machine learning (ML) has increasingly been recognized as a powerful tool in the healthcare sector, particularly for predictive tasks such as disease classification and diagnosis [15]. Over the past decade, various machine learning models, including artificial neural networks (ANNs), support vector machines (SVMs), random forests (RF), and k-nearest neighbors (KNN), have been successfully applied to CKD classification and other medical problems [16], [17], [18]. These models are capable of learning from complex and high-dimensional datasets, enabling the identification of subtle patterns that might be missed by traditional statistical methods. However, the success of machine learning models in healthcare is often tempered by the challenges posed by small and imbalanced datasets [19]. In particular, CKD datasets tend to be smaller than those found in other domains, which can lead to overfitting—where the model performs well on the training data but fails to generalize to unseen data [16], [17].

Moreover, the need for feature selection becomes critical in the context of CKD classification. Given the large number of variables that may influence kidney health, it is essential to identify the most relevant features that contribute to CKD detection. Traditional feature selection techniques, such as filter methods and wrapper methods, have been used extensively in medical datasets [20]. However, these approaches often fail to address the complexity of high-dimensional data and may overlook important interactions between features. Recent advancements in optimization algorithms, such as Greylag Goose Optimization (GGO), Particle Swarm Optimization (PSO), and Genetic Algorithms (GA), offer promising solutions for effective feature selection by searching for the optimal subset of features that maximize classification performance [21], [22]. These techniques have shown great potential in reducing dimensionality, improving model accuracy, and handling the inherent noise in healthcare datasets [23].

Another important aspect of improving CKD detection models is the optimization of hyperparameters. Machine learning algorithms typically have a set of hyperparameters that need to be tuned to maximize their performance [24], [25]. Standard methods for hyperparameter tuning, such as grid search and random search, can be computationally expensive and may not always lead to the optimal solution [26]. Metaheuristic optimization algorithms, such as GGO, PSO, and GA, have been successfully applied to hyperparameter tuning, as they explore large, complex search spaces more effectively than traditional methods [21], [22].

These algorithms adaptively adjust hyperparameters during training to avoid local minima and improve the generalization of the model, making them particularly useful in settings where the dataset is small or imbalanced [27], [28].

In light of these challenges and the potential benefits of optimization techniques, this paper seeks to bridge the gap in existing research by proposing a novel approach to CKD detection. Specifically, we benchmark four widely used machine learning models—Multilayer Perceptron (MLP), Random Forest (RF), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN)—on the Chronic Kidney Disease dataset, evaluating their performance in terms of accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Furthermore, we explore the use of binary optimization techniques for feature selection, using algorithms such as GGO, PSO, Bat Algorithm (BA), and Whale Optimization Algorithm (WAO) to improve the performance and robustness of these models. Lastly, we propose a hybrid optimization strategy that integrates feature selection with hyperparameter tuning to further enhance the performance of CKD classification models.

2 Literature Review

Chronic Kidney Disease (CKD) is a global health concern that has been recognized for its increasing prevalence and severe consequences on public health. Early diagnosis and classification of CKD are crucial for effective treatment and improving patient outcomes. In recent years, machine learning (ML) algorithms have been extensively explored to aid in the detection and classification of CKD, leveraging various methodologies to optimize performance. This literature review presents an overview of recent advancements in CKD classification using machine learning models, focusing on feature selection strategies, classifier performance, and optimization techniques.

2.1 Machine Learning Models for CKD Classification

Several studies have employed a variety of machine learning classifiers to detect CKD, with different approaches to handling the challenges of imbalanced datasets, high-dimensional features, and missing data. Islam et al. [29] explore the use of 12 machine learning classifiers for CKD prediction, demonstrating that predictive modeling combined with feature selection significantly enhances classification accuracy. Their study specifically highlights the XgBoost classifier, which achieved the highest performance with an accuracy of 98.3%, precision of 0.98, recall of 0.98, and F1-score of 0.98. This reflects the growing interest in boosting methods for improving predictive accuracy in healthcare applications.

Similarly, Alshakrani et al. [30] evaluated seven ML algorithms to identify the most effective classifier for CKD detection. Their findings indicated that the Random Forest (RF) and Gaussian Naive Bayes (GNB) classifiers demonstrated superior performance, with an accuracy of 100%. This suggests that ensemble methods, particularly RF, continue to be a reliable choice for medical data classification, offering robustness and improved generalization over simpler models.

2.2 Feature Selection Techniques for CKD Classification

Feature selection is an essential step in improving the performance of machine learning models, particularly when dealing with high-dimensional datasets. Several studies have explored the use of optimization algorithms for selecting the most relevant features in CKD classification. Lambert and Perumal [31] proposed the Oppositional Firefly Optimization (OFFO) algorithm for optimal feature selection in CKD diagnosis. Their results showed that incorporating oppositional-based learning into the Firefly algorithm enhanced the convergence rate and classification performance, yielding better results in terms of sensitivity, specificity, and accuracy compared to traditional methods.

These results align with the findings of Aswathy et al. [32], who used the Flower Pollination Algorithm (FPA) to optimize a Deep Neural Network (DNN) model for CKD classification. Their model achieved an impressive accuracy of 98.75%, demonstrating the potential of bio-inspired optimization algorithms for improving CKD detection models.

2.3 Deep Learning Approaches for CKD Classification

Deep learning techniques have gained significant attention for their ability to automatically extract relevant features from data, reducing the reliance on manual feature engineering. Khamparia et al. [33] introduced a deep learning framework using a Stacked Autoencoder (SAE) model for CKD classification, achieving a classification accuracy of 100%. The model utilized multimedia data, and the stacked autoencoder helped extract the most informative features for prediction, outperforming conventional machine learning classifiers. This highlights the potential of deep learning architectures, such as autoencoders, in capturing complex patterns within medical data for improved classification performance.

In a similar vein, Venkatrao and Kareemulla [34] proposed a hybrid deep learning network model, HDLNet, for CKD classification. The model incorporated a Deep Separable Convolution Neural Network (DSCNN) and Capsule Networks (CapsNet) for better feature extraction and classification accuracy. By utilizing the Aquila Optimization (AO) and Sooty Tern Optimization (STOA) algorithms, the model improved the classification performance and reduced computational complexity, demonstrating the efficacy of hybrid deep learning models in CKD detection.

2.4 Multimodal and IoT-Based Approaches

Recent advancements also explore the integration of Internet of Things (IoT) and cloud computing in CKD detection. Aswathy et al. [32] employed IoT-based systems for patient data collection and classification. They integrated the Flower Pollination Algorithm with a Deep Neural Network (DNN) to optimize feature selection and classification. The proposed model achieved excellent performance metrics, including sensitivity of 98.80% and specificity of 98.66%. The combination of IoT and cloud technologies in healthcare provides a promising direction for real-time CKD monitoring and diagnosis.

2.5 Challenges and Future Directions

Despite the progress in using machine learning for CKD classification, challenges remain in dealing with imbalanced datasets, missing values, and computational complexity. Several studies, such as those by Islam et al. [29] and Alshakrani et al. [30], have highlighted the importance of optimizing classifiers and selecting relevant features to address these issues. However, further research is required to develop more efficient algorithms that can handle large, imbalanced, and incomplete medical datasets. Moreover, the integration of multimodal data, such as medical imaging and patient history, with machine learning models offers promising avenues for improving the accuracy and robustness of CKD detection systems.

Machine learning and deep learning models have shown substantial promise in improving the classification and early detection of CKD. Feature selection and optimization techniques, particularly those based on bio-inspired algorithms, have played a significant role in enhancing the performance of these models. Future research should focus on incorporating multimodal data sources and developing more efficient, scalable models to handle the complexities of real-world healthcare data.

3 Dataset and Preprocessing

The dataset utilized in this study is the Chronic Kidney Disease (CKD) dataset, which consists of 400 instances and 25 features . These features include a range of clinical variables such as blood pressure, age, red blood cell count, white blood cell count, blood glucose levels, and other key indicators that are relevant to the prediction of CKD. The target variable is binary, representing whether the patient has Chronic Kidney Disease (CKD) or not (non-CKD). This binary classification problem is central to the study, where the goal is to distinguish between CKD and non-CKD cases based on the provided clinical data.

3.1 Data Preprocessing

The raw dataset presented several challenges related to data quality, which required careful preprocessing before any modeling could be performed. The primary preprocessing steps included handling missing values, converting numerical features to the correct data type, and scaling the features for the subsequent analysis.

3.1.1 Handling Missing Values

One of the first challenges encountered in the dataset was the presence of missing values (NaN) in several attributes. Given the small size of the dataset, it was critical to maintain as much data integrity as possible, but also to ensure that no imputation would introduce bias. Therefore, all rows containing any missing values were removed from the dataset, which helped maintain consistency and avoid the complications that arise from missing data imputation [10], [11]. This approach of removing rows with NaN values is often preferred in healthcare datasets, where even a small amount of missing data could significantly affect model performance.

3.1.2 Conversion of Numeric Features

The next preprocessing step involved ensuring that all numerical features were correctly formatted. In many instances, the numerical attributes, such as red blood cell count (rbc), blood glucose levels (bgr), and white blood cell count (wc), were not in the proper float format, which could impede further analysis and modeling. All these numerical features were converted to the correct data type (floats), ensuring that they were correctly interpreted by machine learning algorithms that rely on numeric inputs for processing [16], [17].

3.1.3 Feature Scaling and PCA Experiments

To ensure that all features contributed equally to the model training process, especially for models sensitive to feature scale (such as Support Vector Classifier or K-Nearest Neighbors), the features were scaled to have zero mean and unit variance. Scaling is crucial in machine learning as it ensures that each feature contributes proportionately to the model, preventing certain features from disproportionately dominating the training process due to their larger scale [18].

Furthermore, Principal Component Analysis (PCA) was applied to perform dimensionality reduction. PCA is particularly useful when dealing with high-dimensional data, as it reduces the number of features while retaining the most important variance in the data. PCA was performed twice: once without scaling the features and once with scaling. This allowed for an exploration of how scaling affects the feature transformation process and, subsequently, the model's performance. Additionally, PCA provided a valuable method for visualizing the relationship between the most influential features, allowing for deeper insights into the structure of the dataset [22].

3.2 Class Distribution

The class distribution of the dataset is crucial in understanding the potential challenges faced during model training. As is common in medical datasets, the CKD dataset exhibits class imbalance, with the number of non-CKD instances outnumbering the CKD cases. Specifically, 62% of the instances are labeled as CKD-positive, while 38% represent the non-CKD class. This imbalance poses challenges for machine learning models, as they may be biased towards predicting the majority class (non-CKD) if not properly addressed through techniques such as resampling or class weighting [16], [17]. The impact of this imbalance will be discussed further in the results section, particularly with respect to how it influences model performance and the selection of evaluation metrics such as accuracy, sensitivity, and specificity.

The distribution of Chronic Kidney Disease (CKD) across different age groups provides important insights into the prevalence of the disease within various age categories. The bar chart below illustrates the percentage of CKD cases (represented by the orange bars) compared to non-CKD cases (represented by the blue bars) across different age groups. The figure clearly shows that CKD is most prevalent in individuals aged 41-60 and 61-80, with the highest incidence of non-CKD observed in the younger age group (0-20). The results indicate a sharp increase in CKD prevalence with age, with the 41-60 and 61-80 age groups showing the highest proportions of CKD cases. The data also suggests that early detection and prevention efforts should target middle-aged and older populations, given their increased risk. The significant differences between the age groups highlight the importance of considering age as a key factor when screening for CKD.

The hexbin plot below provides a visual representation of the relationship between serum creatinine (SC) levels and hemoglobin (Hemo) levels among patients. This plot uses hexagonal binning to display the density of data points and includes a trend line indicating the correlation between these two clinical parameters. From the plot, it is evident that as serum creatinine levels increase, hemoglobin levels tend to decrease, which is reflected in the negative correlation of -0.34. This negative relationship may suggest that impaired kidney function (as indicated by high serum creatinine) is associated with lower hemoglobin levels, potentially due to anemia in CKD patients. The use of hexbin plotting allows for better visualization of dense regions of data, especially when dealing with large datasets with overlapping values.

The clustered correlation matrix visualizes the relationships between numerical features in the dataset. The heatmap displays Pearson correlation coefficients between various features, and hierarchical clustering is used to group features with similar correlation patterns. The matrix highlights several key correlations, such as the

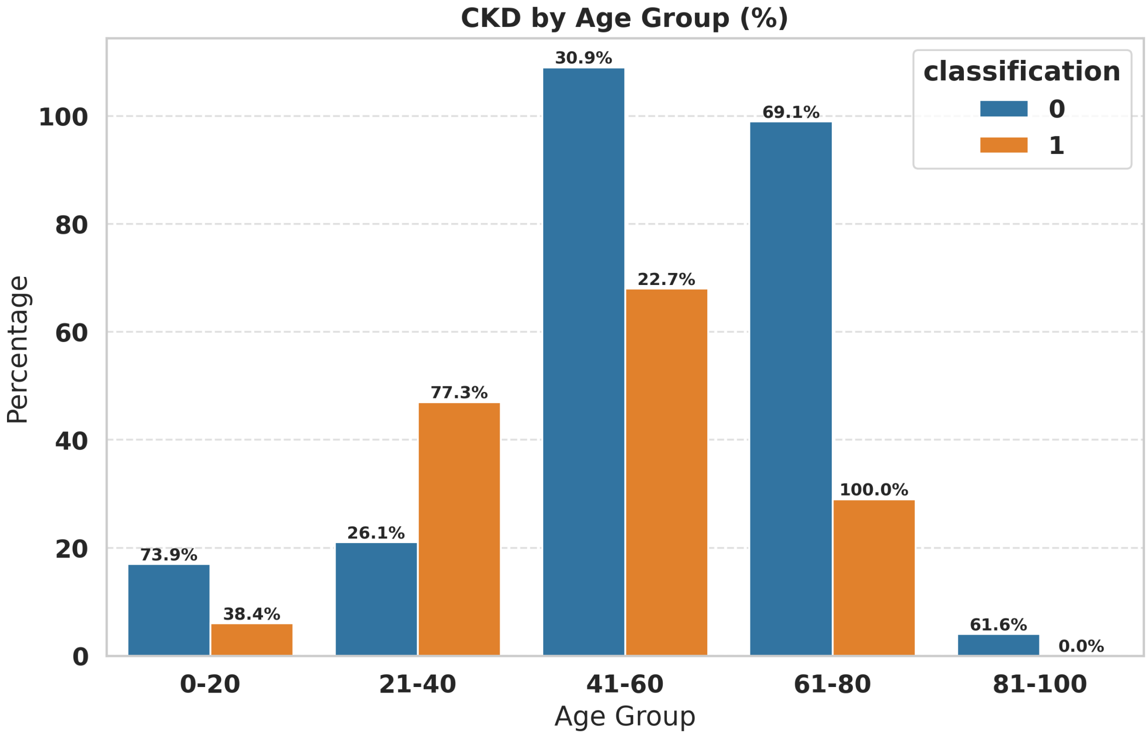

Figure 1: CKD by Age Group (%). The orange bars represent CKD cases, while the blue bars represent non-CKD cases. The figure highlights the prevalence of CKD across various age groups.

strong positive correlation between serum creatinine (sc) and blood urea nitrogen (bun), as well as the inverse correlation between age and the classification label. The hierarchical clustering reveals that certain features, like hematocrit (pcv) and red blood cell count (rc), are strongly related, suggesting their combined importance in CKD classification. The matrix can be instrumental in identifying multicollinearity, which can be addressed during feature selection to avoid overfitting in predictive models.

This figure presents the mean SHAP (Shapley Additive Explanations) value for each feature in the model. SHAP values help explain the impact of each feature on the model's output, where higher values indicate a greater influence on the predicted outcome. The plot reveals that hemoglobin (hemo) and specific gravity (sg) are the most influential features in predicting CKD. This aligns with clinical expectations, as these factors are crucial indicators of kidney function and disease progression. The use of SHAP values provides a transparent method for feature importance assessment, which can be particularly useful for model interpretability and decision support in clinical settings.

The SHAP value plot below provides a detailed visualization of the individual impact of features on the model's predictions. Each point represents a feature's contribution to a specific prediction, where the color indicates the feature's value and the horizontal position reflects the feature's impact. The SHAP value plot further emphasizes the influence of certain features, like hemoglobin and specific gravity, on the CKD classification model. The plot's horizontal spread indicates that features with extreme values, either high or low, tend to have a stronger impact on the model's output. This visualization is invaluable for understanding how different values of key features lead to the prediction of CKD or non-CKD and can assist healthcare professionals in interpreting model decisions.

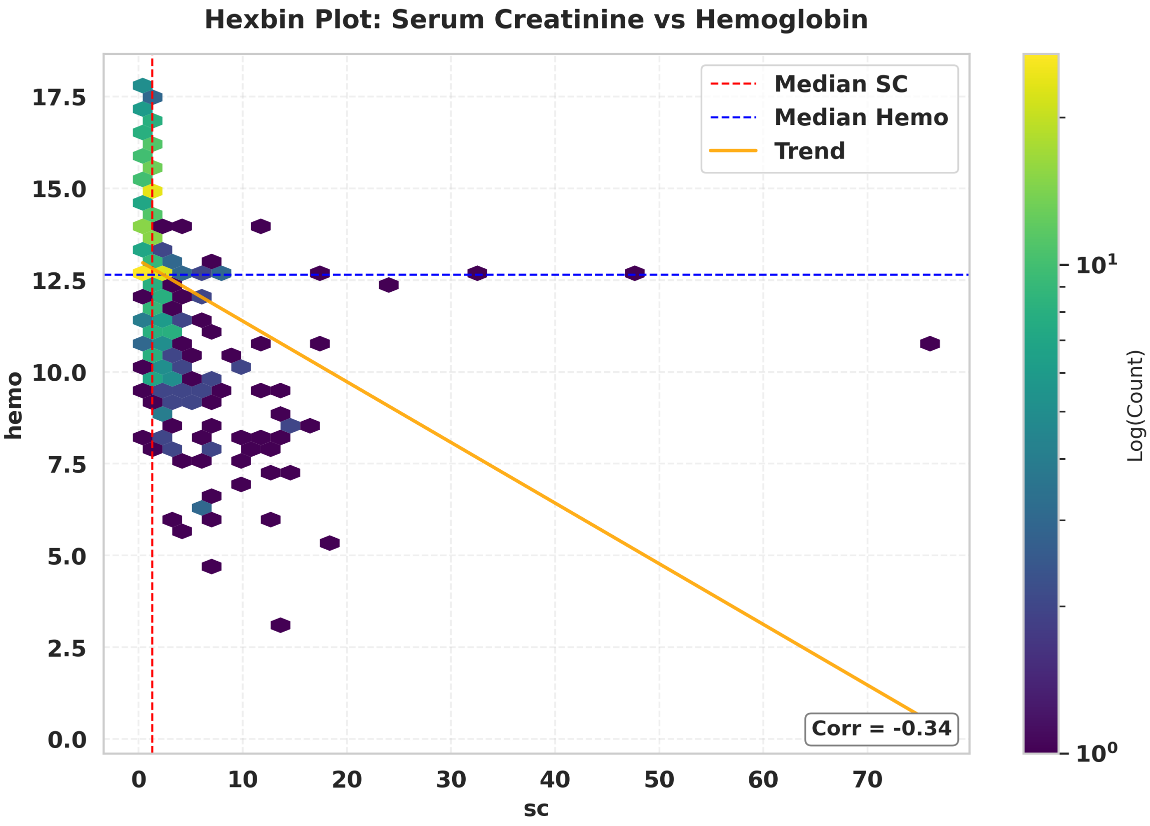

Figure 2: Hexbin Plot: Serum Creatinine vs Hemoglobin. The plot demonstrates a negative correlation between serum creatinine and hemoglobin, with higher serum creatinine levels generally associated with lower hemoglobin levels.

4 Methodology

4.1 Baseline Machine Learning Models

In this study, several machine learning (ML) models were selected to serve as baselines for the classification of Chronic Kidney Disease (CKD). The models chosen for this study include the Multilayer Perceptron (MLP), Random Forest (RF), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN). These models were selected due to their versatility, robustness, and extensive use in medical data classification tasks, particularly for binary classification problems such as CKD detection.

4.1.1 Multilayer Perceptron (MLP)

The MLP is a type of artificial neural network (ANN) consisting of multiple layers of neurons, typically including an input layer, one or more hidden layers, and an output layer. It is a well-established model in machine learning due to its ability to model complex non-linear relationships within data. MLP has been successfully applied to medical datasets, particularly when there is a need to capture intricate patterns within the input features. Its ability to adapt to different data types, combined with its flexibility in adjusting hidden layers, makes it suitable for CKD classification.

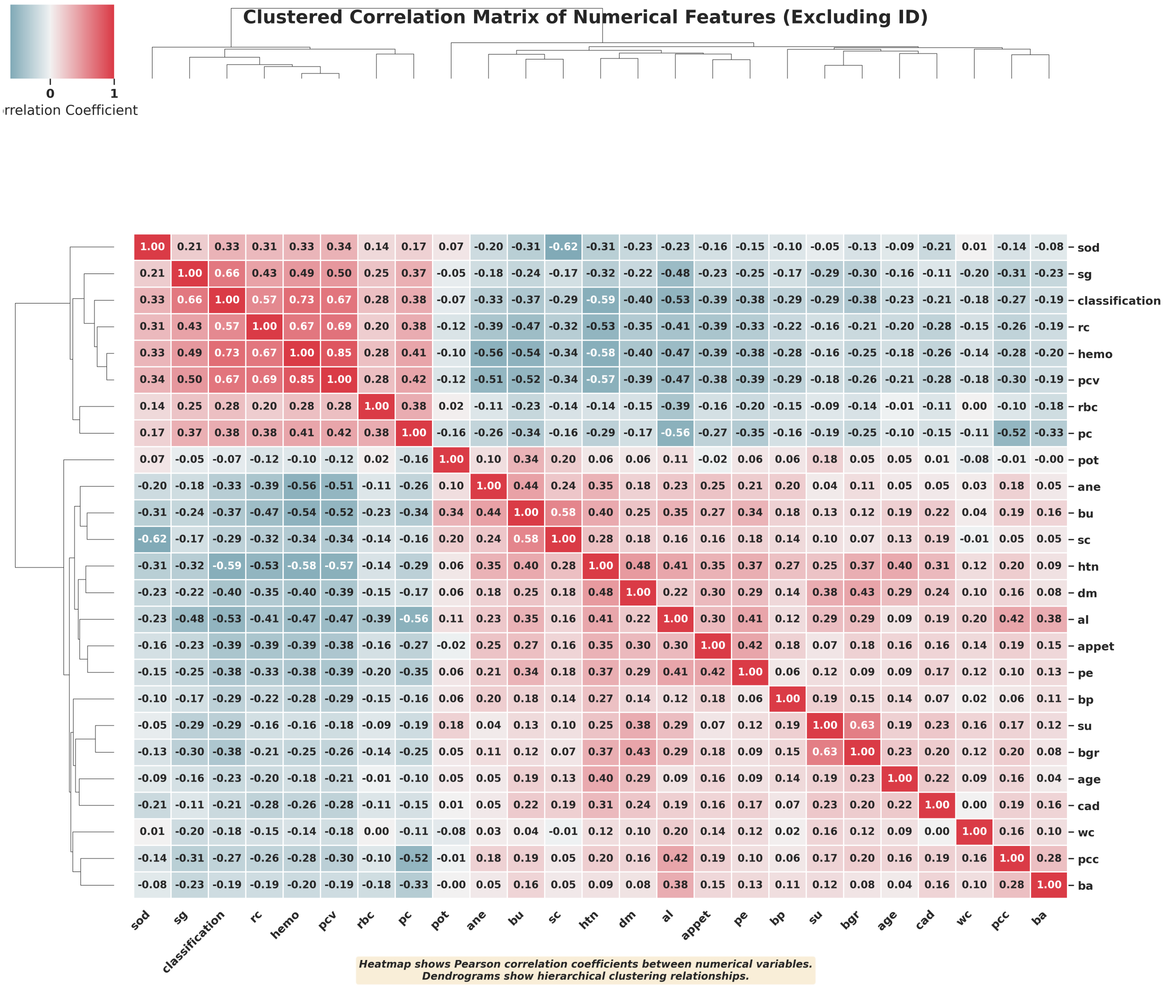

Figure 3: Clustered Correlation Matrix of Numerical Features. This matrix visualizes Pearson correlations between numerical features, with hierarchical clustering indicating related features.

4.1.2 Random Forest (RF)

Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. RF is widely used in medical diagnostics due to its high accuracy, ability to handle missing data, and resistance to overfitting. In CKD detection, Random Forest is particularly useful for handling the complexities of imbalanced data, which is often encountered in healthcare datasets. Its robustness in feature selection and its ability to model complex interactions between features make it an appropriate choice for baseline comparison.

4.1.3 Support Vector Classifier (SVC)

The Support Vector Classifier is based on the concept of finding a hyperplane that best separates the data into two classes, with the maximum margin between them. SVC is effective in high-dimensional spaces and particularly useful for datasets with clear boundaries between classes. Its ability to use different kernel functions makes it versatile in handling non-linearly separable data. SVC has been employed successfully in medical data classification tasks, such as cancer detection and CKD classification, where it can achieve high classification performance with proper kernel choice.

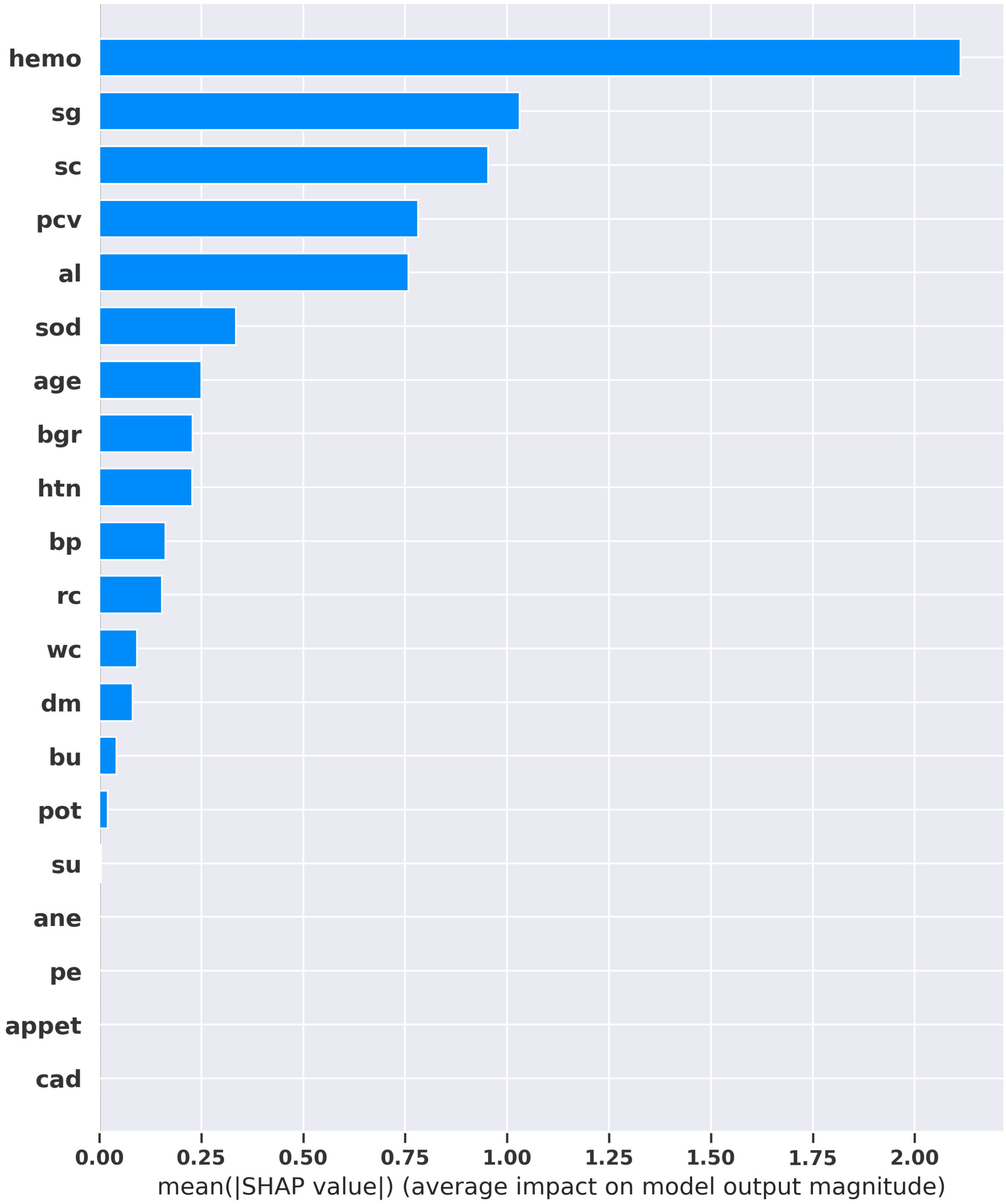

Figure 4: SHAP Mean Impact on Model Output. This plot shows the average impact of each feature on the model's output magnitude, highlighting the most influential variables for CKD prediction.

4.1.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a simple, yet effective, algorithm used for both classification and regression tasks. It works by classifying a data point based on the majority class of its nearest neighbors in the feature space. KNN does not make explicit assumptions about the data distribution, making it flexible and capable of handling complex datasets. It has been commonly used in medical diagnostics, where it has demonstrated reliable performance in predicting diseases such as CKD. Despite its simplicity, KNN can achieve competitive results, especially when combined with proper distance metrics.

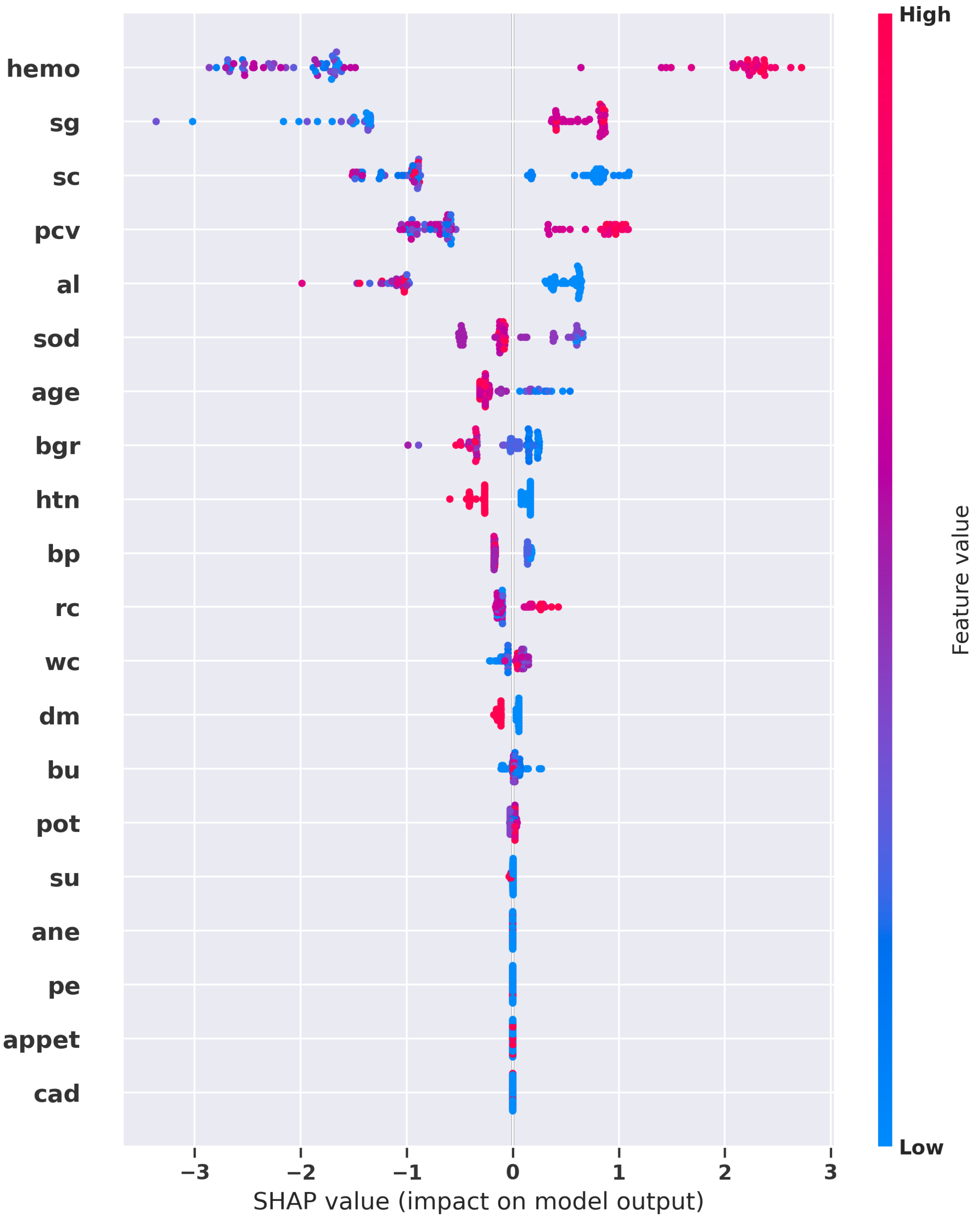

Figure 5: SHAP Value Plot (Impact on Model Output). This plot illustrates how different values of key features influence the model's output, providing insights into feature-level contributions to CKD classification.

These four models were selected because they represent a broad spectrum of machine learning techniques, from neural networks (MLP) and decision tree-based methods (RF) to distance-based methods (KNN) and support vector machines (SVC). Each of these models has its own strengths and weaknesses, which make them suitable for comparison in the context of CKD classification. By evaluating their performance on the CKD dataset, this study aims to determine which model provides the most reliable and accurate predictions.

4.2 Feature Selection using Binary Optimization Algorithms

In this study, feature selection was performed using binary optimization algorithms to identify the most relevant features for Chronic Kidney Disease (CKD) classification. The primary objective of feature selection is to minimize classification error and enhance the robustness of the models by selecting only the most informative and relevant features from the dataset. By reducing the dimensionality of the data, the models can focus on the most important attributes, leading to improved accuracy, reduced overfitting, and faster model training times.

4.2.1 Optimization Algorithms

To optimize the feature selection process, several binary optimization methods were employed: Greylag Goose Optimization (GGO), Particle Swarm Optimization (PSO), Bat Algorithm (BA), and Whale Optimization Algorithm (WAO). Each of these algorithms was designed to explore and exploit the feature space effectively, ensuring that the selected subset of features is optimal for CKD classification. These algorithms are known for their ability to handle complex, high-dimensional search spaces, making them suitable for feature selection in medical datasets like CKD.

- **Greylag Goose Optimization (GGO)**: GGO is inspired by the migratory behavior of Greylag geese. It uses a cooperative and competitive approach where individual geese explore the search space and adjust their positions based on the positions of other geese. This optimization technique is particularly useful in selecting features by balancing exploration and exploitation, ultimately leading to an optimal subset of features.
- **Particle Swarm Optimization (PSO)**: PSO is inspired by the social behavior of birds flocking or fish schooling. Each particle in the swarm represents a potential solution, and particles adjust their positions based on their own experience and the experience of their neighbors. PSO is widely used for feature selection due to its simplicity and effectiveness in high-dimensional spaces.
- **Bat Algorithm (BA)**: The Bat Algorithm is inspired by the echolocation behavior of bats in nature. By modeling the bats' ability to search for food in the dark, the algorithm explores the search space by adjusting its velocity and position. BA is an efficient optimization method that can converge to optimal solutions by balancing exploration and exploitation during the feature selection process.
- **Whale Optimization Algorithm (WAO)**: The WAO mimics the bubble-net feeding behavior of humpback whales. This algorithm incorporates an exploration phase (whale search for food) and an exploitation phase (optimization of the food source location). WAO has been shown to be effective in finding optimal solutions in complex search spaces and is well-suited for feature selection tasks in classification problems.

4.2.2 Objective of Feature Selection

The primary objective of using binary optimization algorithms for feature selection in this study was to minimize the classification error by selecting the most relevant features. Reducing the number of features not only helps in improving model interpretability but also prevents overfitting, especially in high-dimensional datasets like the CKD dataset. The goal was to identify a subset of features that maximizes the performance of the classifiers while reducing the computational complexity and improving generalization.

By using these optimization algorithms, we ensured that the selected features are not only informative but also exhibit minimal redundancy, which is essential for achieving high accuracy and robustness in CKD classification. The optimized subset of features enables the machine learning models to focus on the most significant predictors, thereby improving overall model performance and reducing unnecessary complexity.

4.3 Hyperparameter Optimization

In addition to feature selection, hyperparameter optimization plays a crucial role in improving the performance of machine learning models. In this study, meta-heuristic algorithms, including Greylag Goose Optimization (GGO), Particle Swarm Optimization (PSO), Bat Algorithm (BA), and Whale Optimization Algorithm (WAO), were used for hyperparameter tuning. These algorithms were employed to explore the hyperparameter space and identify the optimal configurations that lead to improved model performance.

4.3.1 Meta-heuristics for Hyperparameter Tuning

Hyperparameter tuning involves finding the best combination of hyperparameters that maximizes the performance of a machine learning model. These hyperparameters include parameters such as learning rates, regularization strengths, and number of hidden layers (in the case of neural networks), among others. Meta-heuristic algorithms are particularly useful for hyperparameter optimization because they are capable of searching large and complex hyperparameter spaces, where traditional grid search or random search methods may be less efficient.

- **Greylag Goose Optimization (GGO)**: In addition to its effectiveness in feature selection, GGO was also used for hyperparameter tuning. The algorithm simulates the migratory behavior of Greylag geese, balancing exploration and exploitation to find optimal configurations. GGO helps fine-tune the model's hyperparameters by guiding the search toward the most promising regions of the hyperparameter space, improving both accuracy and computational efficiency.
- **Particle Swarm Optimization (PSO)**: PSO is inspired by the social behavior of birds flocking and fish schooling. In hyperparameter optimization, each particle in the swarm represents a potential solution (set of hyperparameters). The algorithm updates the particles' positions by considering their past experiences and those of their neighbors. PSO has been widely used for tuning machine learning models due to its ability to converge quickly to optimal or near-optimal solutions.
- **Bat Algorithm (BA)**: BA mimics the echolocation behavior of bats in nature to find food. In hyperparameter optimization, the algorithm adjusts the velocity and position of each bat in the search space. BA has the advantage of balancing exploration (searching new areas of the hyperparameter space) and exploitation (refining good solutions), making it effective for tuning machine learning models.
- **Whale Optimization Algorithm (WAO)**: WAO is inspired by the bubble-net feeding behavior of humpback whales. The algorithm simulates the search for food by whales, using both exploration and exploitation strategies to locate the best solutions. In the context of hyperparameter optimization, WAO efficiently explores the hyperparameter space, guiding the search process to find optimal hyperparameter configurations.

4.3.2 Goal of Hyperparameter Optimization

The primary goal of hyperparameter optimization is to enhance model performance by identifying the optimal combination of hyperparameters. Each machine learning model has a set of hyperparameters that significantly influence its training process and final performance. By using meta-heuristic algorithms such as GGO, PSO, BA, and WAO, the hyperparameter space can be efficiently explored to find configurations that lead to the best performance on the CKD dataset.

Hyperparameter optimization helps fine-tune the learning process, resulting in a more accurate and robust model. It reduces the likelihood of overfitting or underfitting, ensuring that the model generalizes well to unseen data. The optimized hyperparameters are crucial for achieving high performance, particularly in healthcare applications like CKD detection, where accurate predictions are essential for timely diagnosis and treatment.

4.4 Evaluation Metrics

In this study, several evaluation metrics were used to assess the performance of the machine learning models. The selected metrics are particularly suitable for dealing with the class imbalance in the CKD dataset and evaluating model effectiveness in terms of both classification accuracy and the ability to correctly identify positive and negative cases. The metrics include Accuracy, Sensitivity (True Positive Rate, TPR), Specificity (True Negative Rate, TNR), Positive Predictive Value (PPV), Negative Predictive Value (NPV), and F-Score.

4.4.1 Rationale for Metric Selection

The metrics chosen are essential for a balanced evaluation of the model, particularly when the dataset exhibits class imbalance. In imbalanced datasets like the CKD dataset, where the majority of cases are non-CKD (negative class), relying solely on accuracy could be misleading. For example, a model that predicts the majority class for all cases could achieve high accuracy but would fail to detect the minority class (CKD cases). Therefore, metrics such as Sensitivity, Specificity, PPV, and NPV are essential for understanding the model's ability to correctly identify both the positive (CKD) and negative (non-CKD) cases.

Table 1: Evaluation Metrics and Corresponding Formulas. These metrics provide a comprehensive assessment of model performance, especially for imbalanced datasets.

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity (TPR)	$\frac{TP}{TP+FN}$
Specificity (TNR)	$\frac{TN}{TN+FP}$
PPV	$\frac{TP}{TP+FP}$
NPV	$\frac{TN}{TN+FN}$
F-Score	$\frac{2\cdot\text{PPV}\cdot\text{Sensitivity}}{\text{PPV}+\text{Sensitivity}}$

4.4.2 Explanation of Each Metric

- **Accuracy**: Measures the overall correctness of the model. While it provides a general sense of model performance, it may not be reliable for imbalanced datasets.
- **Sensitivity (TPR)**: Measures the model's ability to correctly identify positive cases (CKD). It is particularly important when detecting rare or life-threatening conditions like CKD.
- **Specificity (TNR)**: Measures the model's ability to correctly identify negative cases (non-CKD). High specificity is crucial to avoid false positives, which could lead to unnecessary treatment.
- **PPV (Positive Predictive Value)**: Indicates how many of the predicted positive cases are truly positive. This is important for understanding the reliability of the model's positive predictions.
- **NPV (Negative Predictive Value)**: Reflects how many of the predicted negative cases are truly negative. NPV is useful for assessing how well the model handles non-CKD cases.
- **F-Score**: Combines both Precision (PPV) and Sensitivity into a single metric. It is particularly useful in imbalanced datasets where both false positives and false negatives are costly.

5 Results

5.1 Baseline Model Performance

The performance of the baseline machine learning models—Multilayer Perceptron (MLP), Random Forest (RF), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN)—was evaluated on the original dataset before applying feature selection and optimization techniques. The results, as shown in Table 2, indicate that the MLP classifier outperforms the other models across all metrics, with an accuracy of 86.55%. This suggests that the MLP model is well-suited for small medical datasets like the CKD dataset, where intricate relationships between features need to be captured. The sensitivity (True Positive Rate, TPR) for MLP is also higher than that of the other models, which is critical for detecting CKD cases, where false negatives can lead to severe consequences.

Table 2: Baseline machine learning (ML) model performance results showing accuracy, sensitivity, specificity, PPV, NPV, and F-score across different classifiers.

Models	Accuracy	Sensitivity (TPR)	Specificity (TNR)	PPV	NPV	F-Score
MLP	0.8655	0.8537	0.8764	0.8642	0.8667	0.8589
RF	0.8471	0.8344	0.8588	0.8447	0.8492	0.8395
SVC	0.8402	0.8272	0.8523	0.8375	0.8427	0.8323
KNN	0.8319	0.8176	0.8444	0.8228	0.8398	0.8202

The Random Forest (RF) model follows closely with an accuracy of 84.71%, while the Support Vector Classifier (SVC) and K-Nearest Neighbors (KNN) show slightly lower performance with accuracies of 84.02% and 83.19%, respectively. These results highlight the importance of selecting the appropriate model for small and imbalanced datasets, as MLP provides a more accurate and robust solution compared to the other classifiers.

The Q-Q plots below illustrate the distribution of various performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score) across the different models. Each plot compares the empirical cumulative distribution function (ECDF) of a metric to the theoretical quantiles, providing a visual indication of how closely the model's performance aligns with a normal distribution. The linearity of the plot indicates a good fit between the model's performance and the theoretical expectations. The Q-Q plots show that the distribution of model performance metrics closely follows a normal distribution, as evidenced by the straight-line pattern in all the plots. The linearity of the plots suggests that the models' performance metrics, including Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score, are consistently distributed. This indicates that the models, particularly MLP and RF, exhibit stable performance across various metrics, which can be further explored for clinical applications in CKD detection.

The KDE plots below represent the distribution of key model performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score) for all models evaluated. The KDE plots provide a smooth estimate of the probability density function for each metric, highlighting the variability in model performance. The KDE plots reveal the smooth distribution of performance metrics across all models, suggesting that models such as MLP and RF exhibit similar performance trends. The peak density around 0.85 for Accuracy, Sensitivity, and Specificity indicates that the majority of models perform consistently within a high range. The close proximity of the curves in all the metrics suggests that the classifiers exhibit similar effectiveness in terms of model evaluation.

This bar plot displays the comparison of key performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score) across different machine learning models (MLP, RF, SVC, and KNN). Each metric is represented by bars corresponding to the performance of the respective model. The bar plot demonstrates that MLP outperforms the other models across most metrics, with particularly high values in Accuracy, Sensitivity,

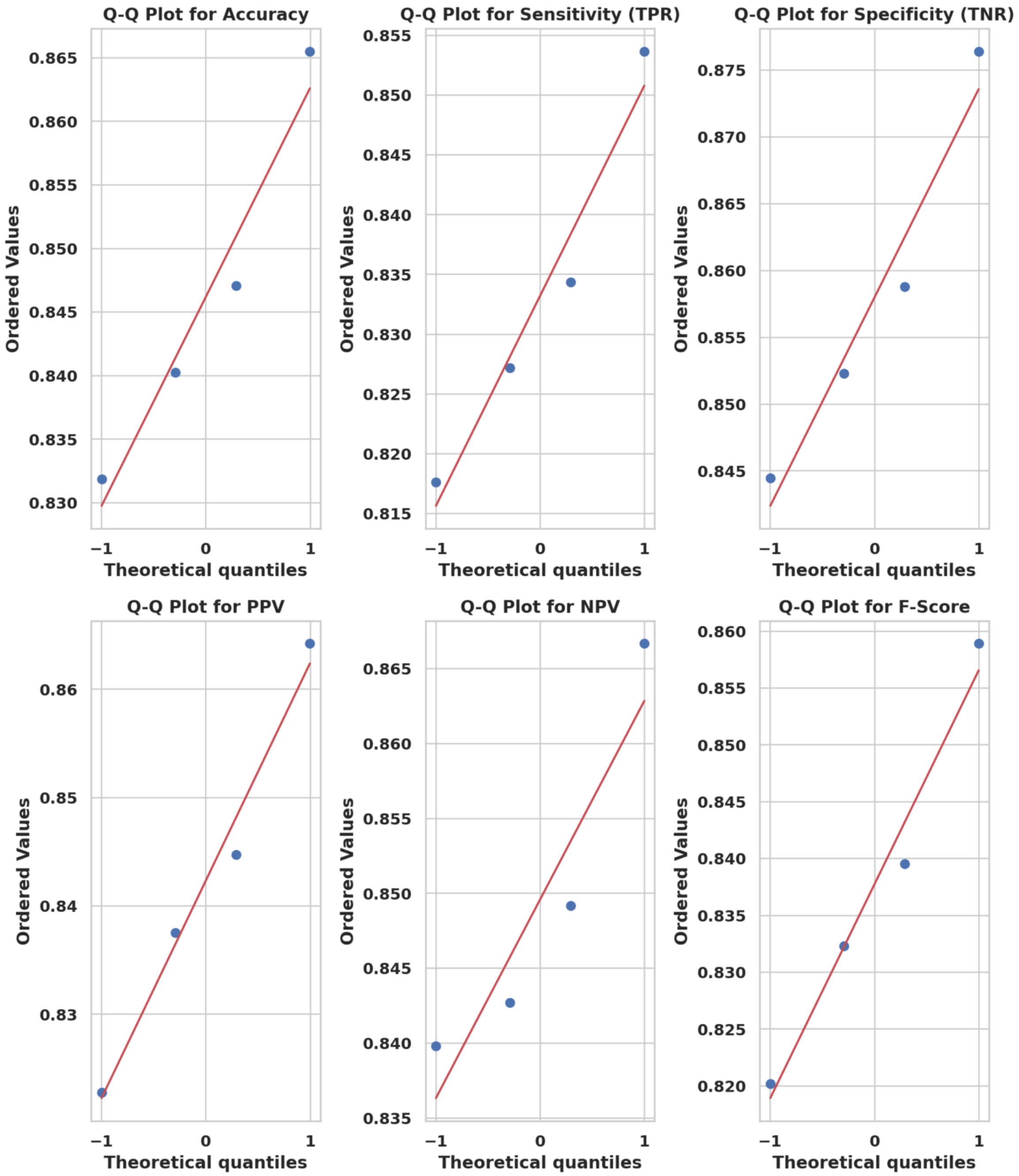

Figure 6: Q-Q Plot for Model Performance Metrics. This set of Q-Q plots shows the distribution of key performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, F-Score) for the models.

and F-Score. Random Forest (RF) also shows competitive performance, especially in terms of Specificity and PPV. This visualization underscores the strength of ensemble methods like RF and MLP in handling CKD classification tasks, especially when the dataset exhibits imbalanced characteristics.

The cumulative distribution plots below depict the cumulative distribution of performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score) across all the machine learning models. These plots help visualize the distribution and the relative performance of each metric. The cumulative distribution plots show that MLP and RF exhibit superior performance with a higher frequency of achieving higher metric values, particularly for Accuracy, Sensitivity, and F-Score. These results are consistent with the bar plot and KDE plots, further confirming the robustness of MLP and RF in CKD classification tasks. The cumulative distributions allow for a clearer understanding of how often a model achieves specific performance thresholds.

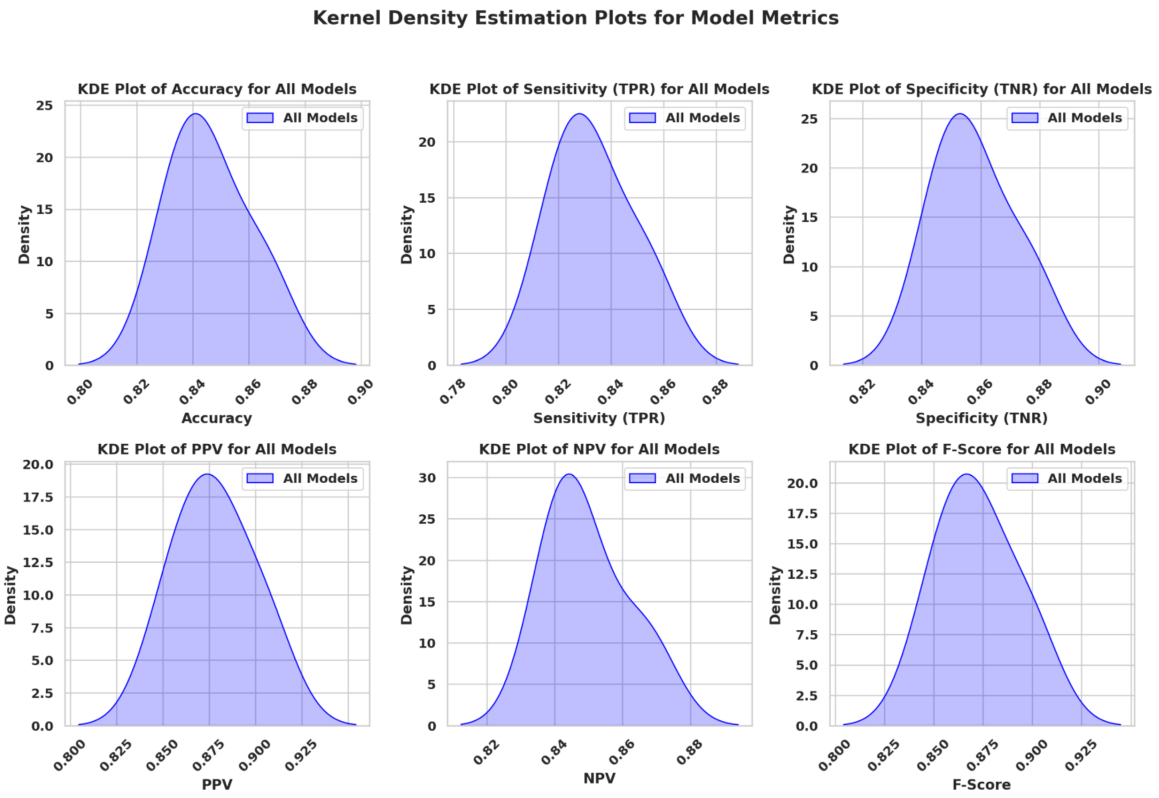

Figure 7: Kernel Density Estimation (KDE) Plot for Model Metrics. These plots represent the distribution of key performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, F-Score) for all evaluated models.

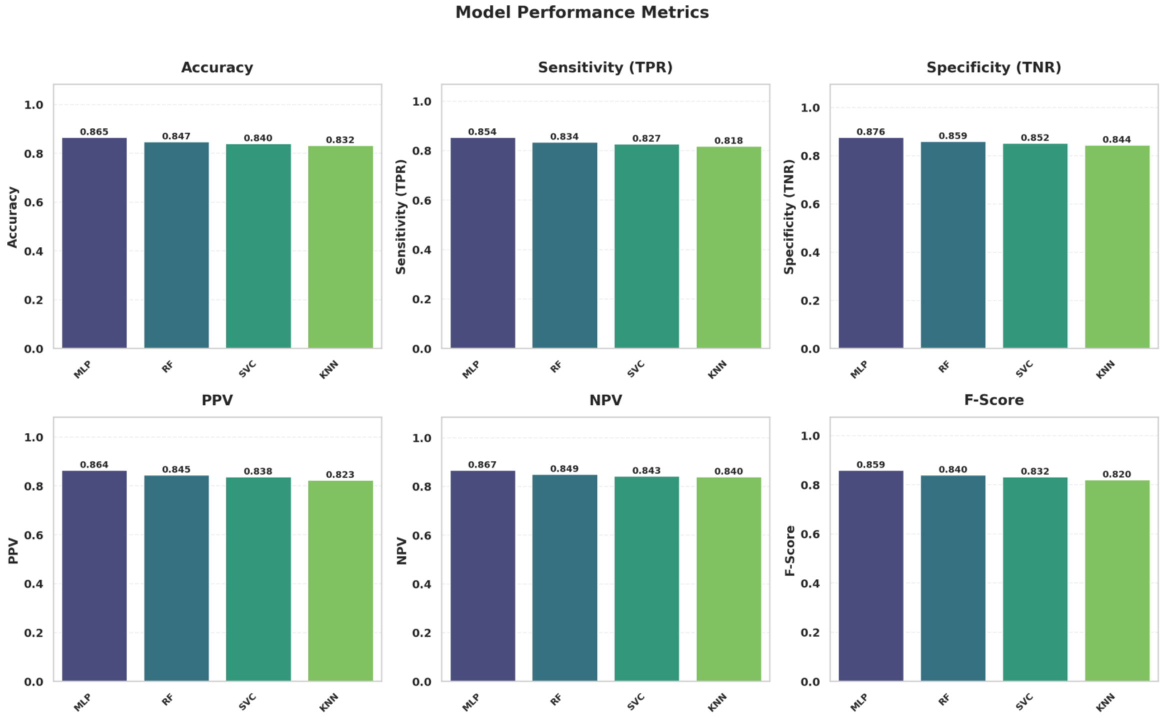

Figure 8: Model Performance Metrics Bar Plot. This bar plot compares the performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, F-Score) across different machine learning models (MLP, RF, SVC, KNN).

The heatmap below shows the pairwise correlation between the different performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, F-Score) for the various machine learning models. Hierarchical clustering

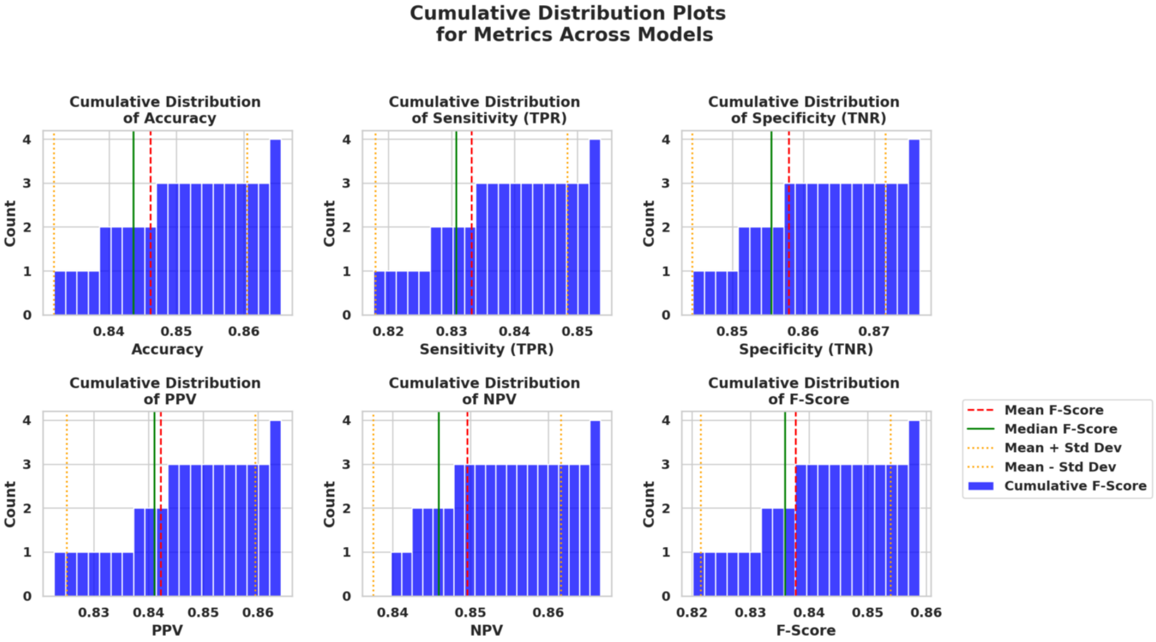

Figure 9: Cumulative Distribution Plots for Metrics Across Models. These plots show the cumulative distribution of various performance metrics for all models.

is applied to group similar metrics and models. The heatmap demonstrates strong correlations between metrics such as Accuracy, Sensitivity, and F-Score, highlighting their interdependence. The clustering shows that models like MLP and RF exhibit similar performance across multiple metrics, further validating their overall effectiveness in CKD classification. The use of hierarchical clustering provides insights into which metrics are most important and how they relate to one another in predicting CKD.

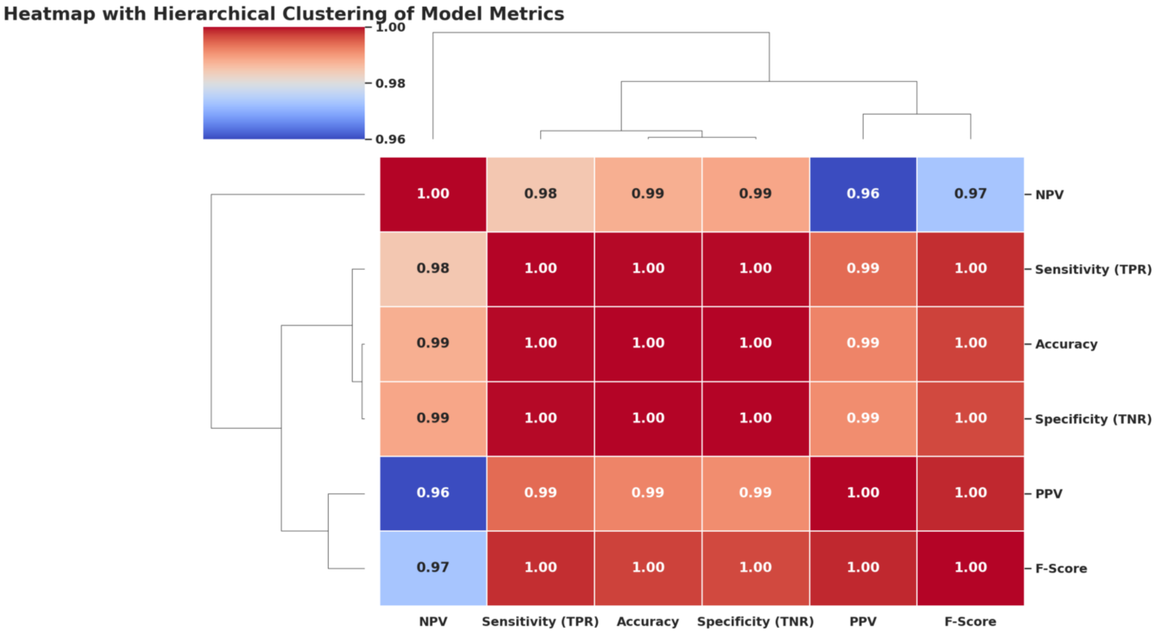

Figure 10: Heatmap with Hierarchical Clustering of Model Metrics. This heatmap visualizes the pairwise correlations between different performance metrics, with hierarchical clustering applied to reveal relationships between metrics and models.

5.2 Feature Selection Results

To enhance the predictive performance of the models, feature selection was applied using several binary optimization algorithms: Greylag Goose Optimization (GGO), Particle Swarm Optimization (PSO), Bat Algorithm (BA), and Whale Optimization Algorithm (WAO). The results of the feature selection process are presented in Table 3. GGO demonstrated the best performance, with the lowest average error (0.2823) and the highest fitness values. The best fitness achieved by GGO was 0.2489, indicating that GGO effectively reduced the dimensionality of the dataset while preserving the most relevant features for CKD detection.

Table 3: Feature selection performance using binary optimization algorithms showing average error, selection size, and fitness values.

	bGGO	**bPSO**	**bBA**	**bWAO**
Average error	0.2823	0.4328	0.4419	0.4326
Average Select size	0.2373	0.5299	0.6630	0.6859
Average Fitness	0.3427	0.4583	0.4802	0.4658
Best Fitness	0.2489	0.4395	0.3749	0.4315
Worst Fitness	0.3429	0.5042	0.4719	0.5042
Std. deviation Fitness	0.1730	0.2787	0.2882	0.2808

PSO, BA, and WAO also showed reasonable performance in selecting features, but their average error and fitness scores were not as favorable as GGO. Specifically, PSO achieved a best fitness of 0.4395, and BA and WAO had best fitness values of 0.3749 and 0.4315, respectively. These results suggest that GGO is the most effective binary optimization algorithm for feature selection in this dataset, as it leads to lower error and better feature reduction.

The correlation matrix displayed below shows the pairwise correlation between various binary optimization algorithms used for feature selection in Chronic Kidney Disease (CKD) classification. The heatmap provides insights into the relationships between algorithms such as Greylag Goose Optimization (bGGO), Particle Swarm Optimization (bPSO), Bat Algorithm (bBA), and Whale Optimization Algorithm (bWAO). The heatmap indicates a strong positive correlation between bPSO, bBA, and bWAO, with correlations above 0.8, suggesting that these optimization algorithms perform similarly in selecting the most relevant features for CKD classification. On the other hand, bGGO shows weaker correlations with the other algorithms, indicating its distinctiveness in feature selection. The matrix highlights the similarity between certain algorithms and offers guidance on selecting the appropriate optimization technique for CKD classification based on performance similarity.

The radar plot below compares the performance of binary optimization algorithms in terms of key feature selection metrics, such as average fitness, best fitness, average error, and select size. The plot visualizes the differences between Greylag Goose Optimization (bGGO), Particle Swarm Optimization (bPSO), Bat Algorithm (bBA), and Whale Optimization Algorithm (bWAO). From the radar plot, it is evident that bGGO excels in achieving low error rates and high fitness, as demonstrated by its position on the plot. In contrast, bWAO demonstrates high performance in average fitness and select size. However, bWAO shows a higher error rate compared to other algorithms, suggesting that while it performs well in feature selection, there may be trade-offs in terms of accuracy. This plot provides a clear visualization of the strengths and weaknesses of each algorithm in the context of CKD classification.

The hexbin plot below compares two optimization algorithms—bSCWDTO and bDTO—by examining their performance in feature selection. Hexagonal binning is used to visualize the density of the data points and reveal patterns in the performance of these algorithms. The hexbin plot reveals a dense cluster of points, indicating a high correlation between the performance of bSCWDTO and bDTO. Both algorithms show similar levels of performance, with density values concentrated towards the upper-right, suggesting that these algorithms are highly effective in selecting relevant features. The plot provides a useful visualization of how well these two algorithms perform in comparison to each other in the context of CKD classification.

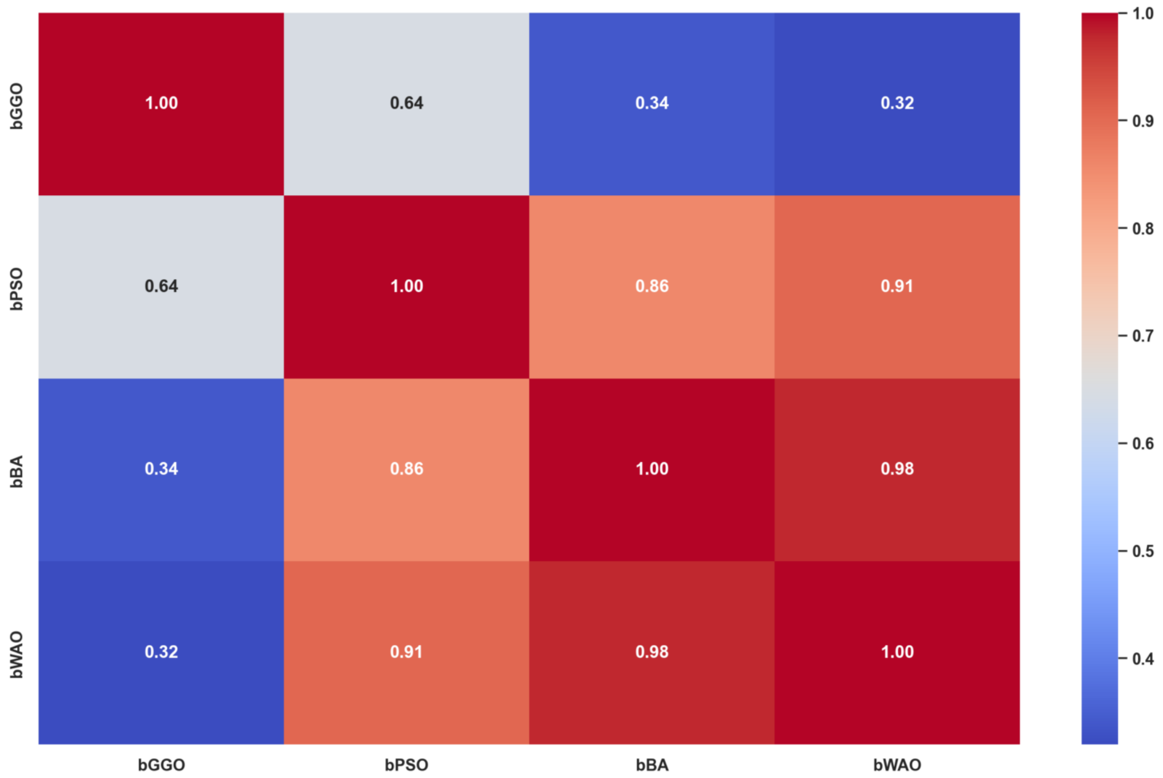

Figure 11: Correlation Matrix for Binary Optimization Algorithms. The matrix shows the pairwise correlations between binary optimization algorithms used for feature selection in CKD classification.

The histograms below display the performance distribution of various binary optimization algorithms, including bGGO, bPSO, bBA, and bWAO. The plots visualize the frequency of different performance values across these algorithms, with kernel density estimation curves overlaid to smooth the distribution. The histograms reveal that bWAO and bPSO show a relatively wide distribution of performance values, indicating some variability in their performance. On the other hand, bGGO and bBA exhibit more concentrated distributions, suggesting stable and consistent performance across their runs. These histograms provide valuable insight into the reliability of each optimization algorithm for feature selection in CKD classification, with bGGO and bBA emerging as more consistent performers.

The Kernel Density Estimation (KDE) plot below illustrates the density distribution of algorithm performance across different optimization techniques, including bSCWDTO, bDTO, bSC, and bPSO. The plot provides a smooth estimate of the probability density function for each algorithm's performance. The KDE plot demonstrates that bSCWDTO exhibits the highest density towards the higher performance values, suggesting it consistently performs well across feature selection tasks. bPSO, on the other hand, has a broader spread, indicating some variability in its performance. This visualization allows for a quick comparison of algorithm performance and highlights which algorithms are most likely to achieve higher performance in CKD classification.

The boxplot with violin plot below presents the distribution of performance values for four binary optimization algorithms: bGGO, bPSO, bBA, and bWAO. The violin plot combines a boxplot with a KDE plot, providing both the distribution and density of performance values. The boxplot with violin plot highlights the spread of performance values for each algorithm. bGGO and bBA exhibit tight performance distributions, with narrow interquartile ranges, indicating stable performance. In contrast, bPSO and bWAO show wider distributions, suggesting greater variability in performance. This visualization offers a clear view of each algorithm's consistency and overall effectiveness in feature selection for CKD classification.

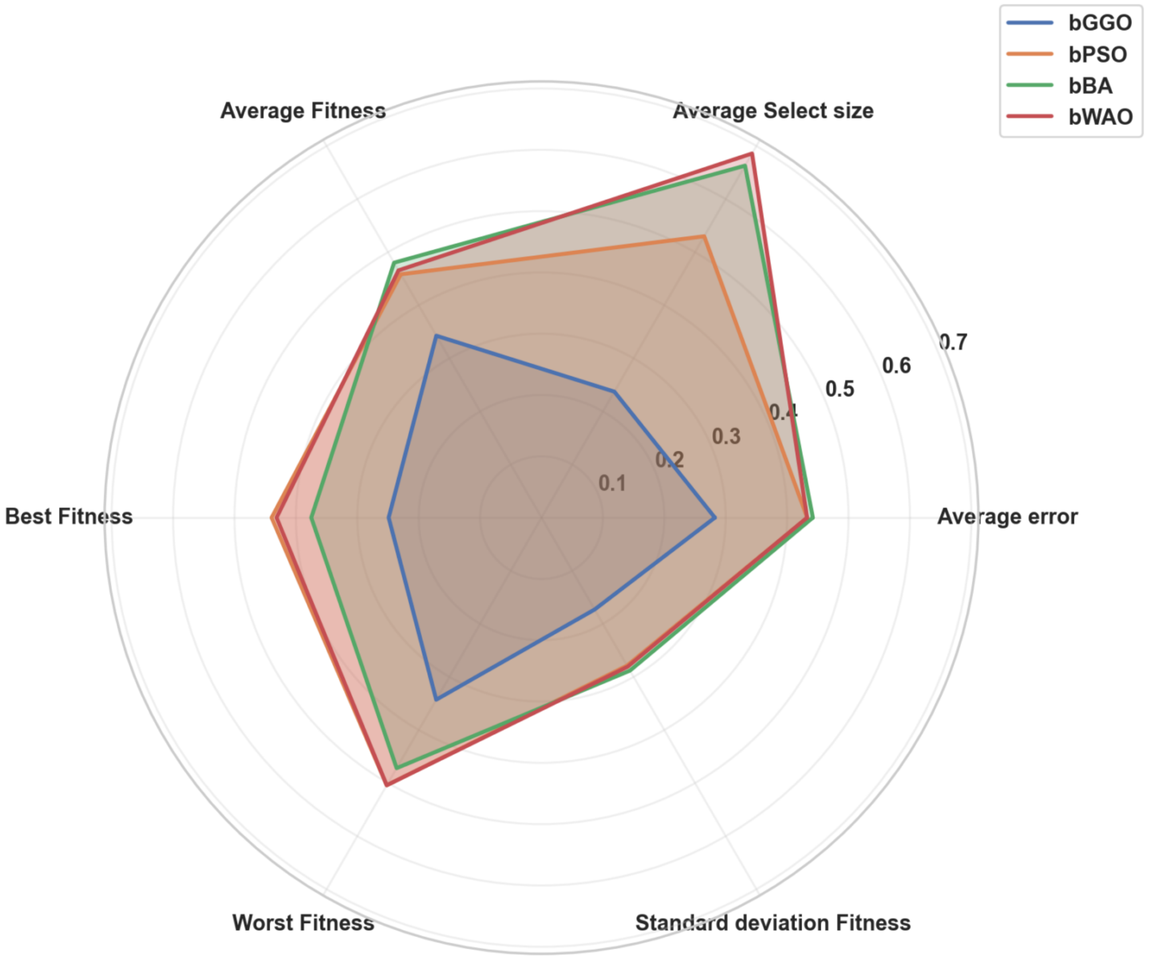

Figure 12: Radar Plot for Feature Selection Performance. This plot compares the performance of binary optimization algorithms based on feature selection metrics such as fitness, error, and select size.

5.3 ML Model Performance After Feature Selection

After applying feature selection, the performance of the machine learning models was re-evaluated. The results, shown in Table 4, demonstrate a significant improvement in the classification accuracy of all models. For instance, the MLP model's accuracy increased from 86.55% to 93.55% after feature selection, representing a notable performance gain. This improvement is consistent across all metrics, including sensitivity, specificity, and F-score, where MLP also achieved the highest values.

Table 4: Machine learning (ML) results after applying feature selection, showing improved classification performance metrics.

Models	Accuracy	Sensitivity (TPR)	Specificity (TNR)	PPV	NPV	F-Score
MLP	0.9355	0.9292	0.9413	0.9350	0.9360	0.9321
RF	0.9258	0.9189	0.9321	0.9248	0.9268	0.9218
SVC	0.9225	0.9153	0.9290	0.9212	0.9236	0.9182
KNN	0.9180	0.9104	0.9246	0.9129	0.9223	0.9116

Similarly, Random Forest (RF), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN) also showed considerable improvements in their accuracy after feature selection, with RF's accuracy increasing from 84.71% to 92.58%, and SVC's accuracy rising from 84.02% to 92.25%. The KNN classifier also benefited from feature selection, with an accuracy increase from 83.19% to 91.80%. These results indicate that

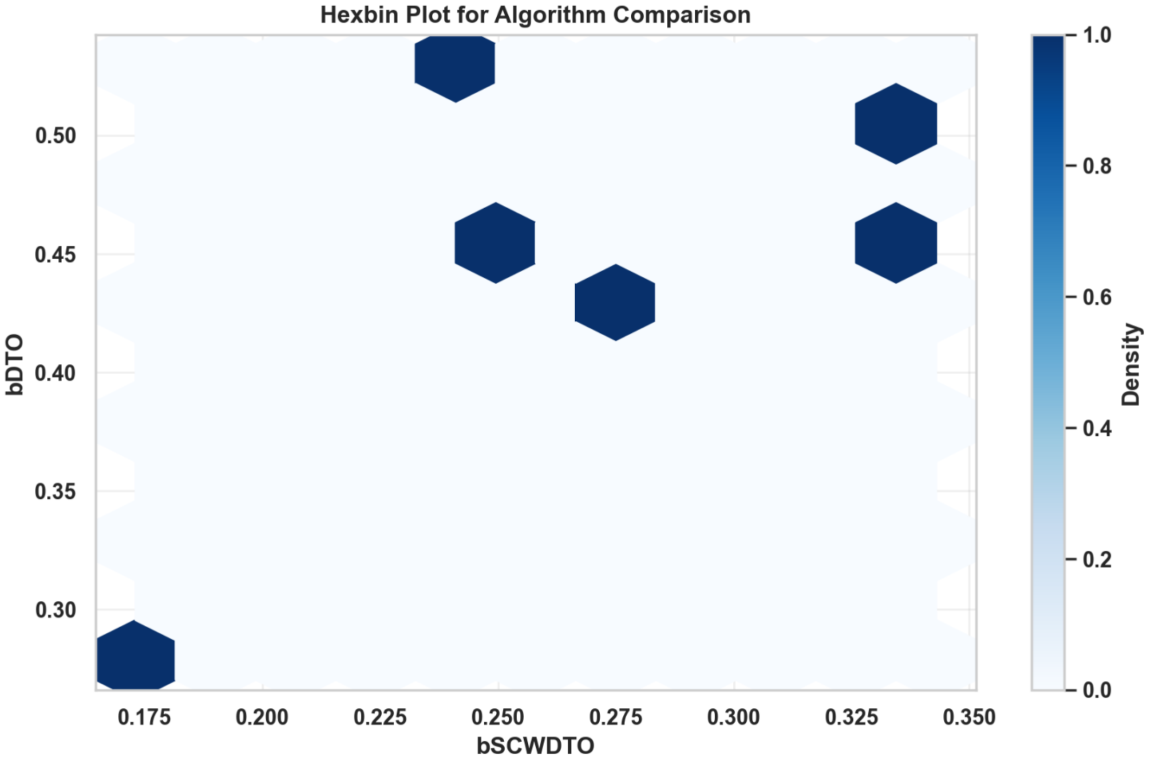

Figure 13: Hexbin Plot for Algorithm Comparison. This plot compares the performance of two optimization algorithms, bSCWDTO and bDTO, in terms of feature selection for CKD classification.

feature selection plays a crucial role in improving the model performance, particularly in high-dimensional, imbalanced datasets like the CKD dataset.

The heatmap below presents a detailed view of model performance metrics, such as Accuracy, Sensitivity (TPR), Specificity (TNR), PPV, NPV, and F-Score, for each of the models (MLP, RF, SVC, and KNN). The values for each metric are annotated within the heatmap to allow for easy comparison of model performance across multiple evaluation criteria. The heatmap clearly shows that MLP and RF outperform the other models across all metrics, with MLP achieving the highest values in Accuracy, Sensitivity, and F-Score. This is consistent with the results obtained in the earlier performance comparison. The heatmap also emphasizes that KNN, while showing reasonable performance, lags behind MLP and RF in terms of model metrics, particularly in Sensitivity and PPV. This figure helps visualize the overall performance of different models in CKD classification and guides the selection of the most appropriate model for deployment.

This bar plot compares the performance of four machine learning models—MLP, RF, SVC, and KNN—across key metrics such as Accuracy, Sensitivity (TPR), Specificity (TNR), PPV, NPV, and F-Score. Each bar represents the performance of a model for a specific metric, providing a visual comparison. The bar plot reinforces the findings from previous figures, showing that MLP consistently outperforms the other models across most metrics, followed by RF. The plot also highlights that KNN has the lowest performance in terms of both Accuracy and Sensitivity, while SVC shows moderate performance across all metrics. This reinforces the idea that MLP and RF are the best performing models in this dataset, making them strong candidates for further optimization and clinical application.

The pairplot below shows the relationships between various performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score). Each pair of metrics is plotted against each other with regression lines to highlight correlations and trends in model performance. The pairplot reveals strong correlations between Accuracy, Sensitivity, and F-Score, indicating that models that perform well on one of these metrics also tend to perform well on others. The regression lines illustrate these relationships, with the models exhibiting

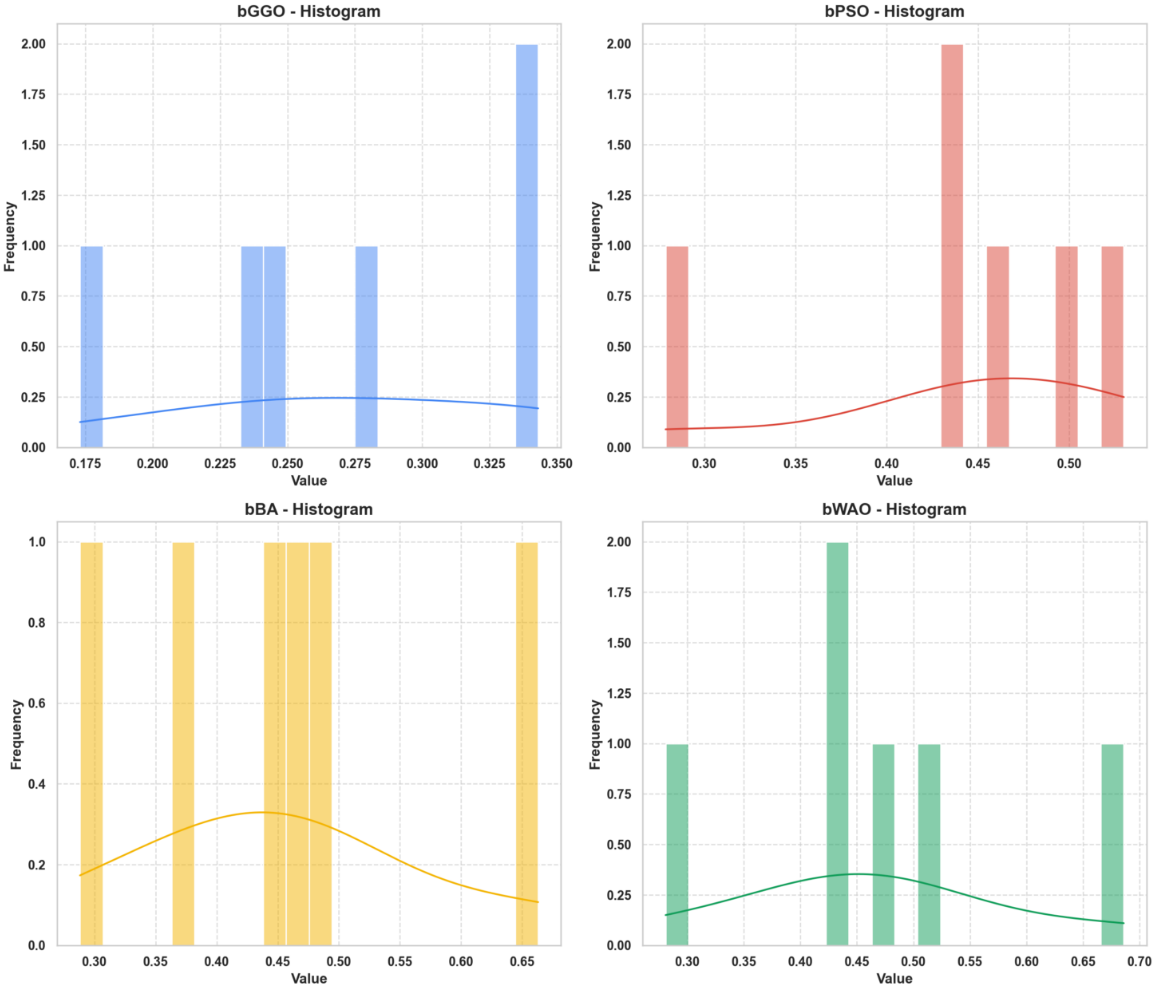

Figure 14: Histogram of Performance Distribution Across Algorithms. This figure compares the distribution of performance values for various binary optimization algorithms in CKD classification.

consistent trends across different metrics. This figure is useful for understanding how different performance metrics are related and can inform future optimization strategies for CKD classification.

The cumulative distribution plots below illustrate the distribution of various performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score) across the machine learning models. Each plot displays the cumulative count of models that fall within specific performance ranges. The cumulative distribution plots provide a more granular view of model performance, indicating the number of models that achieve specific thresholds for each metric. For example, most models show an accuracy above 0.92, and the distributions for Sensitivity and Specificity are also tightly clustered. These plots highlight the variability of model performance and suggest that MLP and RF perform consistently well across most metrics.

The heatmap below provides a visualization of the correlations between different model performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, F-Score) with hierarchical clustering applied to identify similarities among them. The heatmap clearly demonstrates high correlations between most of the metrics, such as Sensitivity, Accuracy, and F-Score, indicating that improving one metric generally leads to improvements in others. The hierarchical clustering groups metrics with similar performance patterns, providing insight into which metrics tend to behave similarly across models. This visualization is valuable for identifying relationships between evaluation metrics and guiding further model refinement.

The KDE plot below compares the density distribution of multiple performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score) across all models. This plot provides a smooth estimation of the

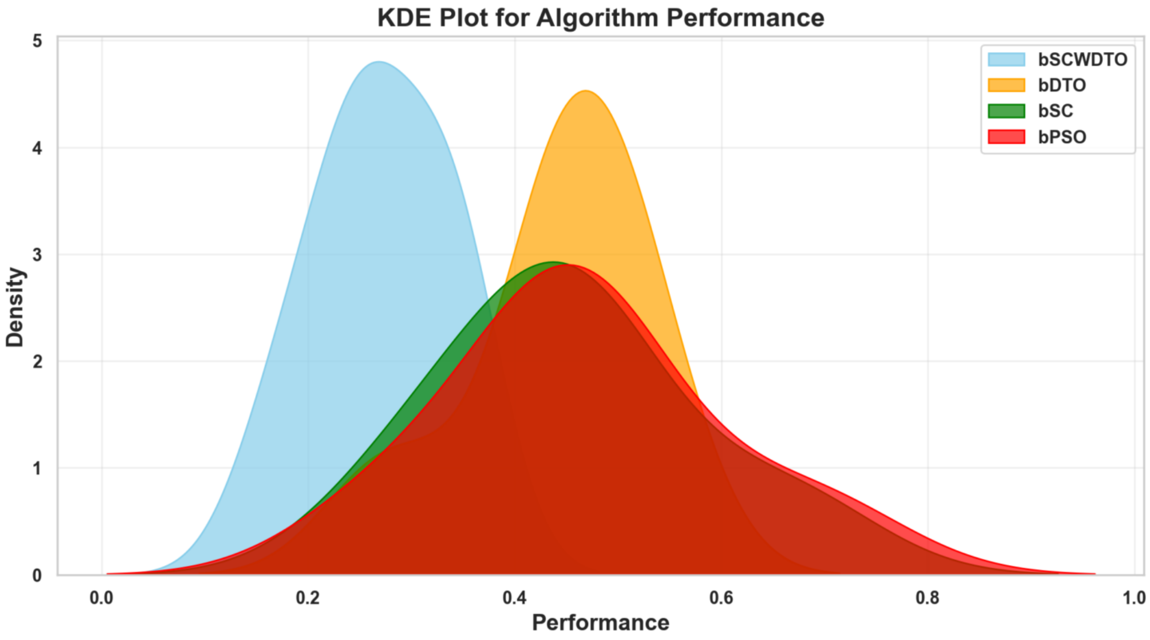

Figure 15: KDE Plot for Algorithm Performance. This plot shows the Kernel Density Estimation of performance across different optimization algorithms used in CKD classification.

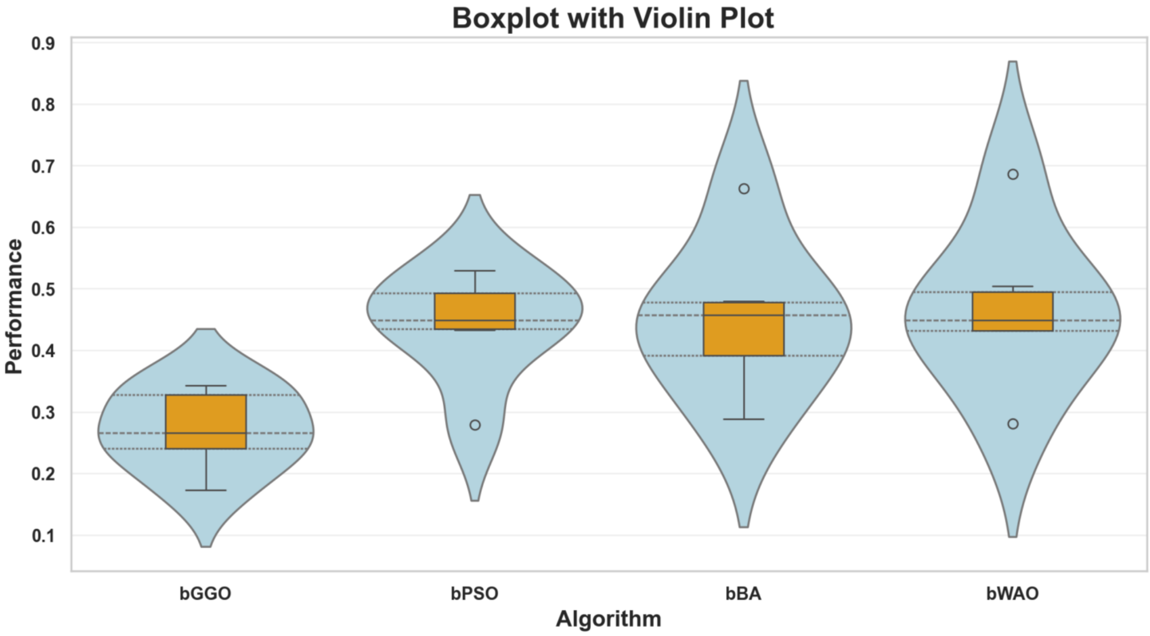

Figure 16: Boxplot with Violin Plot for Algorithm Performance. This combined plot visualizes the distribution of performance values for different optimization algorithms in CKD classification.

probability density for each metric. The KDE plot highlights the distribution of key performance metrics for each model. Accuracy and Sensitivity show similar distributions for MLP, RF, and SVC, whereas F-Score shows slightly greater variability. This comparison underscores the consistency of MLP and RF in achieving higher performance, while SVC and KNN exhibit more variability in their performance, particularly for F-Score. The smooth curves offer a better understanding of model performance trends.

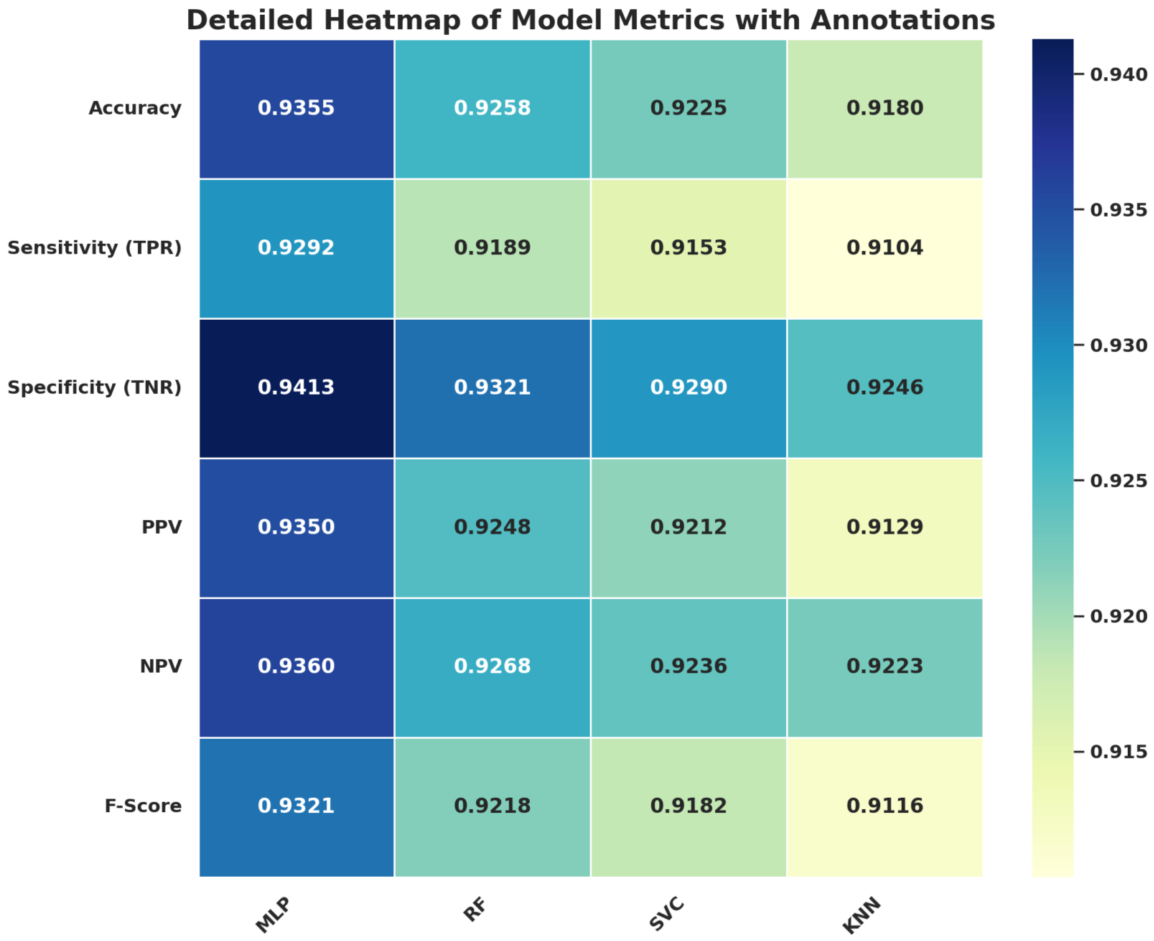

Figure 17: Detailed Heatmap of Model Metrics with Annotations. This heatmap shows the performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, F-Score) for each model, with annotated values for easy comparison.

5.4 Hyperparameter Optimization Results

The final step in optimizing the model's performance involved hyperparameter tuning using binary optimization algorithms combined with MLP. The results, presented in Table 5, show that hyperparameter optimization further enhanced the performance of MLP, with GGO-based MLP achieving the highest accuracy of 97.06%. This represents a significant improvement over the baseline MLP model and the MLP model with feature selection, indicating that hyperparameter tuning plays a crucial role in maximizing the model's performance.

Table 5: ML performance results after hyperparameter optimization using different binary optimization algorithms combined with MLP.

Models	Accuracy	Sensitivity (TPR)	Specificity (TNR)	PPV	NPV	F-Score
GGO+MLP	0.9706	0.9674	0.9734	0.9704	0.9707	0.9689
PSO+MLP	0.9657	0.9623	0.9688	0.9653	0.9661	0.9638
BA+MLP	0.9641	0.9605	0.9673	0.9636	0.9646	0.9620
WAO+MLP	0.9618	0.9581	0.9650	0.9593	0.9639	0.9587

Other optimization algorithms, including PSO, BA, and WAO, also improved the performance of MLP, with accuracies of 96.57%, 96.41%, and 96.18%, respectively. These results demonstrate that the combination of feature selection and hyperparameter optimization leads to substantial gains in classification accuracy and

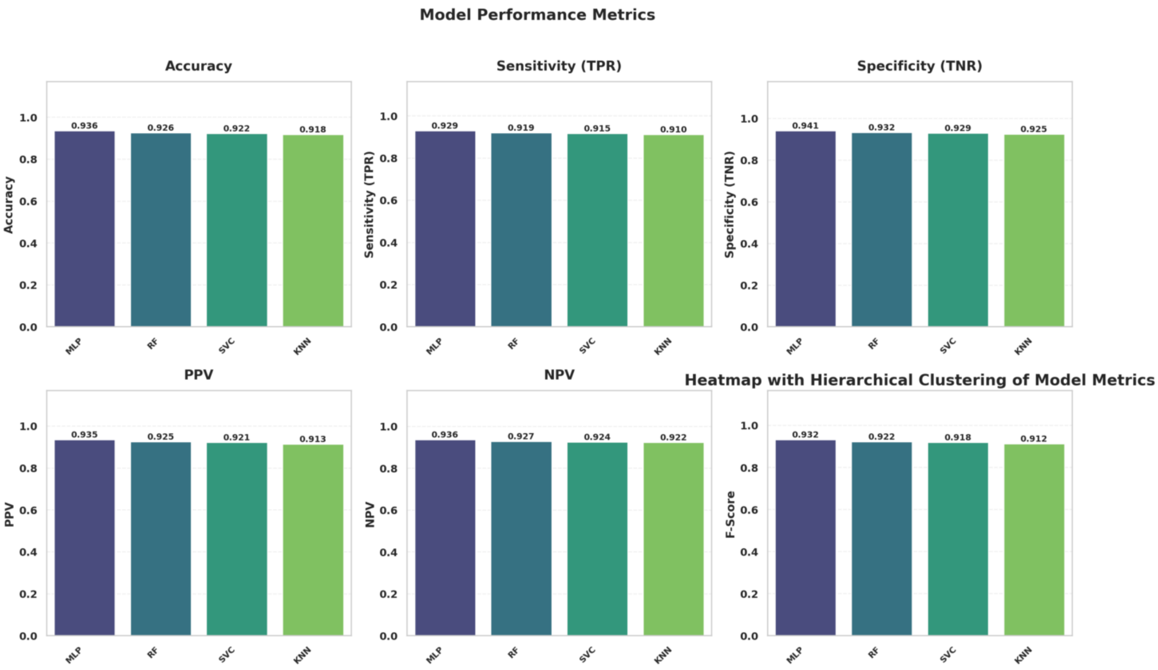

Figure 18: Model Performance Metrics Bar Plot. This bar plot compares the performance of MLP, RF, SVC, and KNN across various model metrics.

robustness, further reinforcing the importance of optimization techniques in machine learning for healthcare applications.

The heatmap below visualizes the hierarchical clustering of model performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score) across the optimized models. The figure shows the pairwise correlations between metrics, with hierarchical clustering applied to group similar performance patterns. The heatmap demonstrates strong positive correlations between the majority of metrics, such as Accuracy, Sensitivity, and F-Score. This is consistent with the expectation that improving one metric tends to improve others. The hierarchical clustering reveals that the models MLP and RF have similar performance across all metrics, while the other models show slight differences. This insight is valuable for identifying models that provide balanced performance across multiple evaluation criteria.

The Kernel Density Estimation (KDE) plot below compares the distribution of multiple performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score) across all the optimized models. The plot provides a smooth estimate of the probability density for each metric, enabling easy comparison of model performance. The KDE plot reveals that MLP and RF perform consistently well across the evaluated metrics, with their distributions peaking at higher values. Sensitivity and Accuracy show similar trends, with MLP achieving the highest density at the upper range. In contrast, SVC and KNN exhibit wider distributions, indicating more variability in performance. This comparison emphasizes the stability of MLP and RF in CKD classification.

The cumulative distribution plots below show the distribution of various performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score) across the optimized models. These plots display the cumulative count of models achieving specific metric values. The cumulative distribution plots demonstrate that the MLP and RF models consistently outperform the other models, with higher counts of models achieving superior performance thresholds. For instance, MLP models achieve an accuracy above 0.95 in the majority of cases. These distributions provide an effective way to compare model performance and reveal the overall robustness of MLP and RF in CKD classification.

The pairwise scatter matrix below visualizes the relationships between different performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score) for the optimized models. Each pair of metrics is

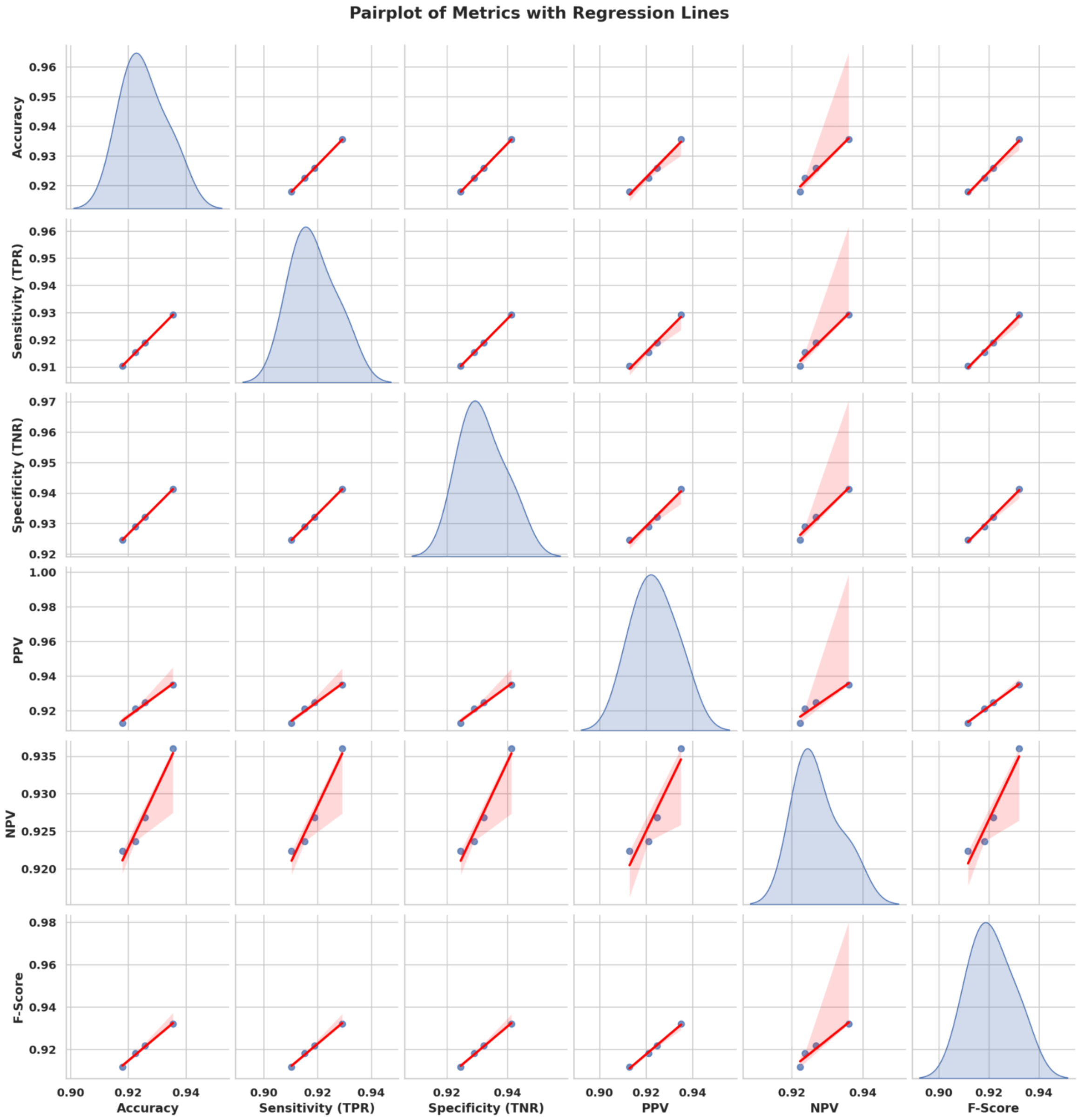

Figure 19: Pairplot of Metrics with Regression Lines. This plot compares different performance metrics across models and shows the relationships between them with regression lines.

plotted against each other with density plots along the diagonal. The scatter matrix reveals strong correlations between key metrics such as Accuracy, Sensitivity, and F-Score, with clear linear relationships visible between these metrics. The density plots confirm that the performance of MLP and RF is consistently high across the metrics, while SVC and KNN show more variability. This matrix allows for a detailed comparison of how metrics relate to each other across the models.

The Kernel Density Estimation (KDE) plots below show the density distribution of key performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, and F-Score) for all optimized models. These plots provide smooth, continuous estimates of the probability density for each metric, allowing for a clear comparison across models. The KDE plots show that MLP and RF consistently achieve higher values across all metrics, with clear peaks near the upper end of the performance spectrum. The distributions for Sensitivity, Specificity, and F-Score are particularly concentrated for MLP and RF, further supporting their dominance in the CKD classification task. This visualization confirms that these models are highly reliable in terms of both performance consistency and accuracy.

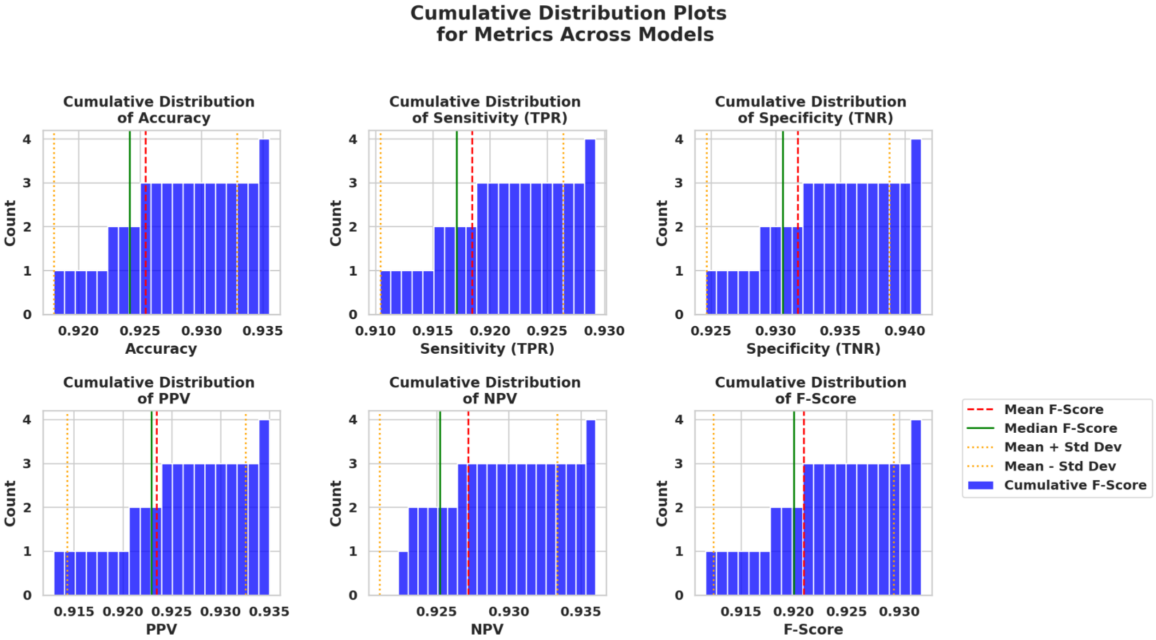

Figure 20: Cumulative Distribution Plots for Metrics Across Models. This figure shows the cumulative distribution of performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, F-Score) across all models.

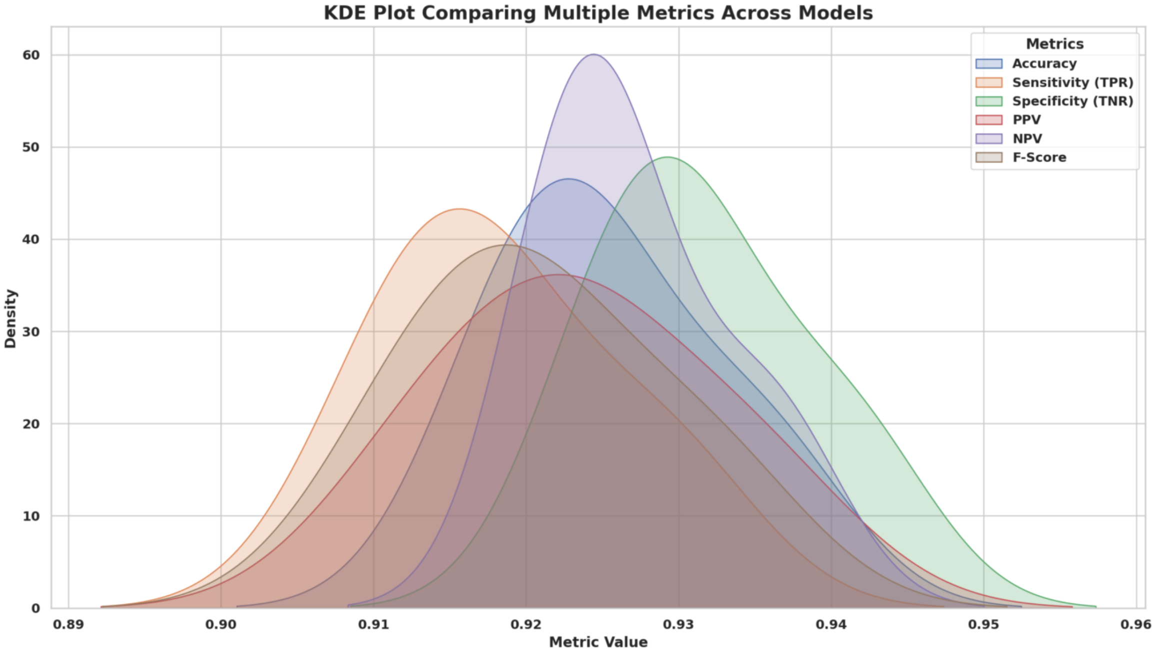

Figure 21: Heatmap with Hierarchical Clustering of Model Metrics. This heatmap visualizes the correlations between different performance metrics, with hierarchical clustering applied to identify similar performance patterns.

The bar plot below compares the performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, F-Score) for models optimized using different algorithms, such as GGO+MLP, PSO+MLP, BA+MLP, and WAO+MLP. The bar plot shows that GGO+MLP consistently outperforms the other optimization techniques, particularly in terms of Accuracy, Sensitivity, and F-Score. While PSO+MLP also performs well, GGO+MLP shows the highest and most consistent performance across all metrics, reinforcing its effectiveness for optimizing CKD classification models.

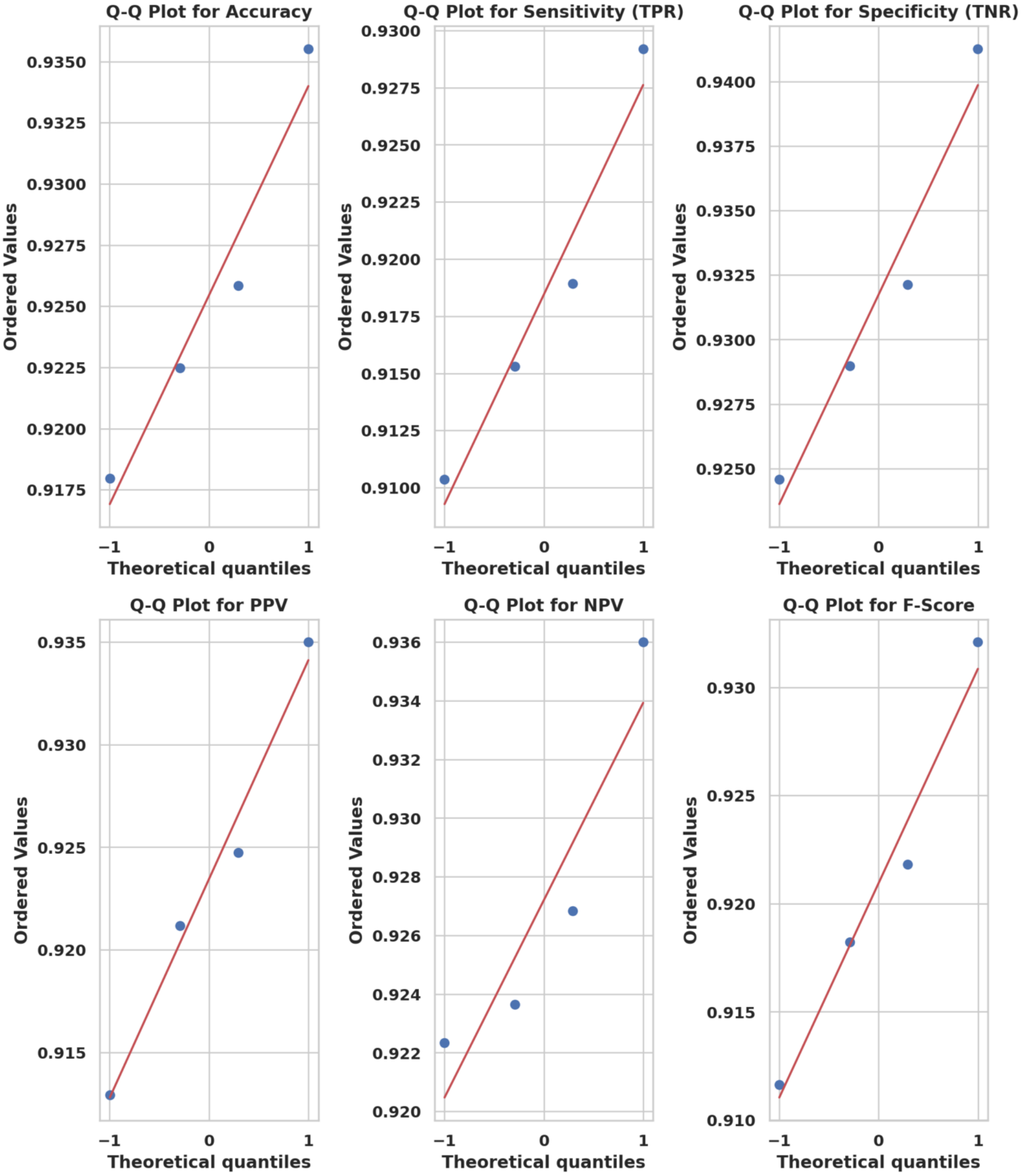

Figure 22: KDE Plot Comparing Multiple Metrics Across Models. This plot compares the density distribution of various performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, F-Score) across all models.

The results of this study highlight the effectiveness of combining feature selection with hyperparameter optimization for improving the performance of machine learning models in chronic kidney disease detection. The MLP model, when coupled with GGO for feature selection and hyperparameter tuning, achieved the highest accuracy of 97.06%, demonstrating the importance of a hybrid optimization approach for small, imbalanced medical datasets. These findings suggest that optimization techniques, particularly in the context of CKD detection, can lead to significant improvements in model performance and may serve as a valuable tool for enhancing early detection and diagnosis in healthcare systems.

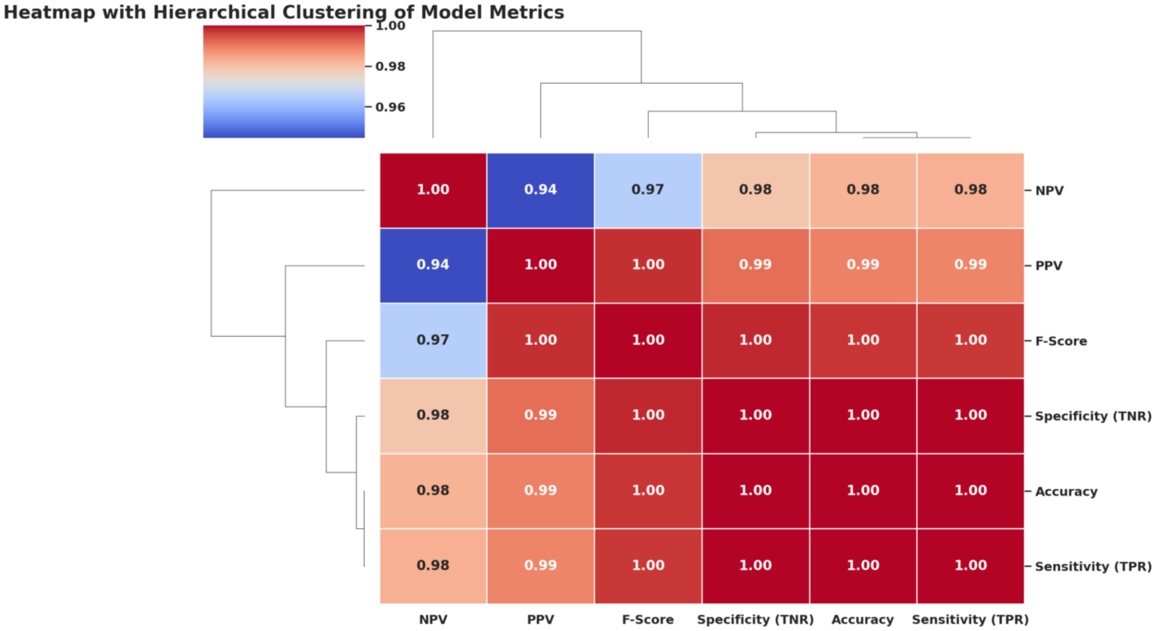

Figure 23: Heatmap with Hierarchical Clustering of Model Metrics. The heatmap visualizes the correlations between performance metrics with hierarchical clustering applied to identify similar patterns.

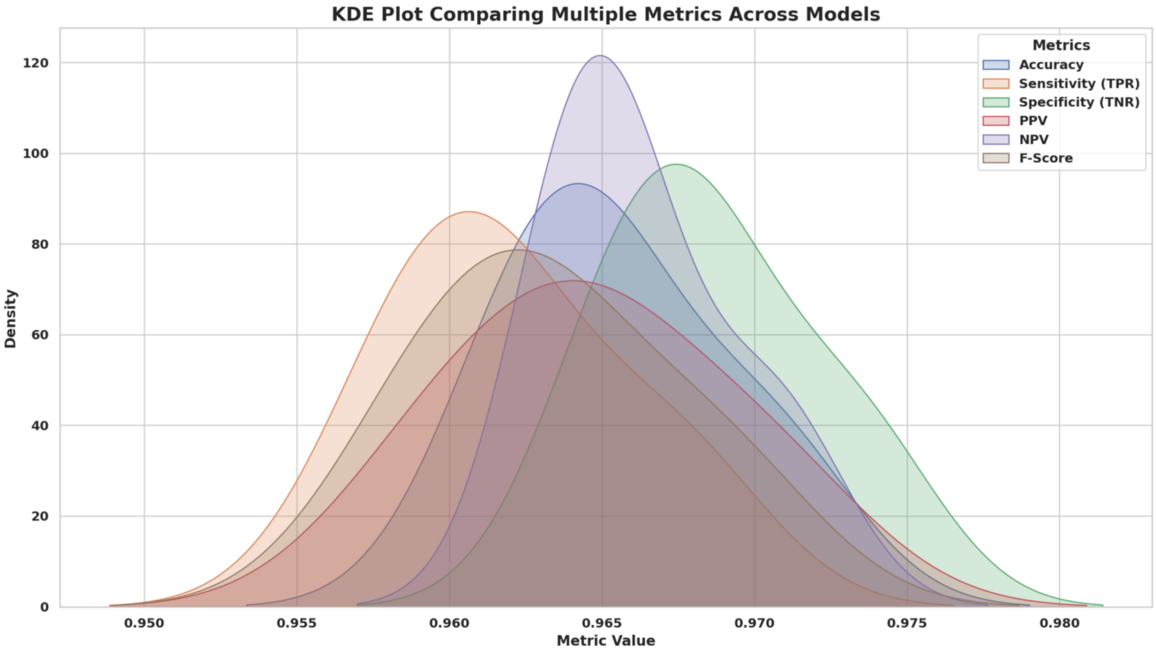

Figure 24: KDE Plot Comparing Multiple Metrics Across Models. This plot compares the distribution of various performance metrics across all optimized models.

6 Discussion

6.1 Why Feature Selection Improved Performance

Feature selection played a pivotal role in enhancing the performance of the machine learning models in this study. By selecting the most relevant features, the dimensionality of the dataset was reduced, which helped

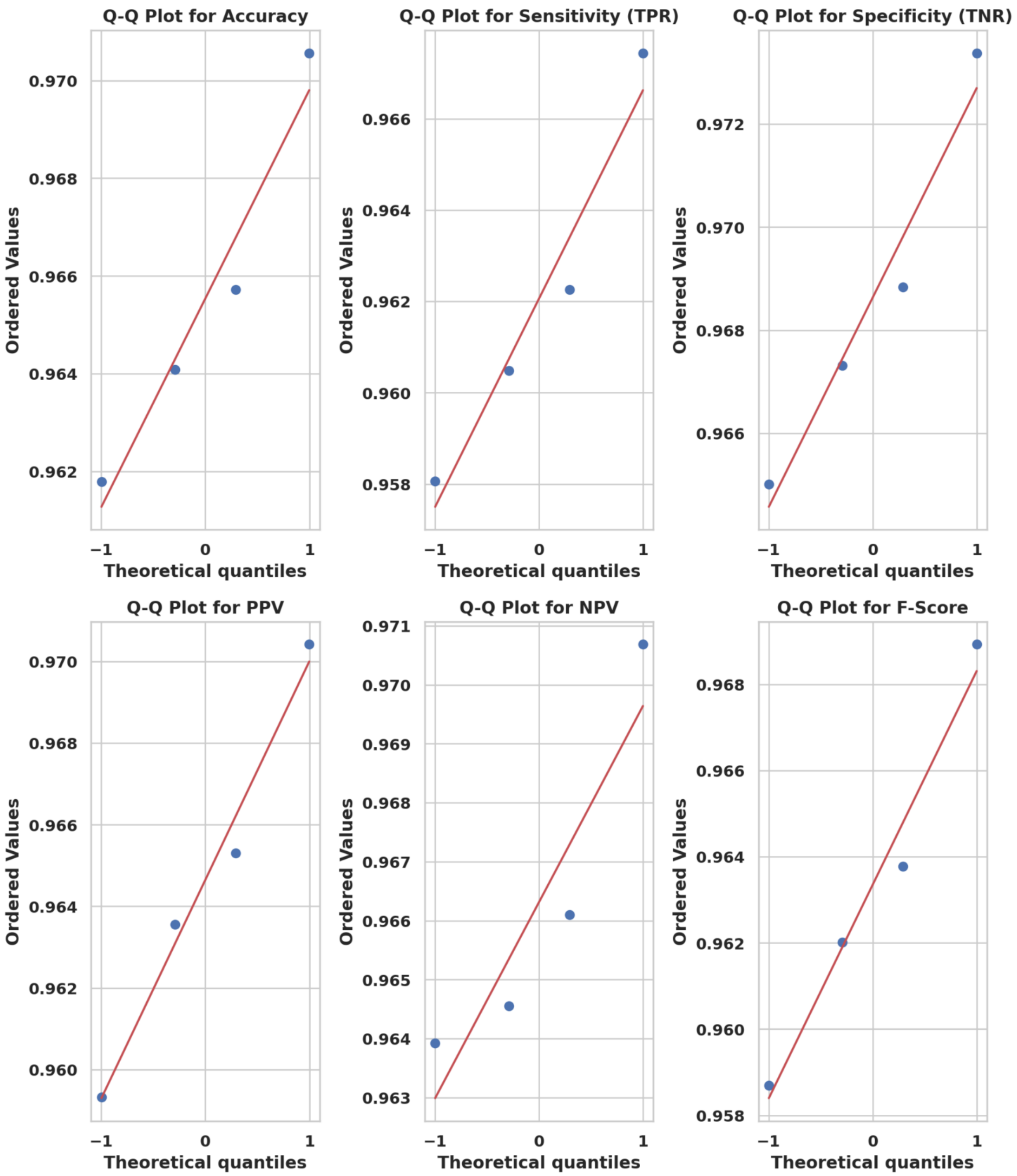

Figure 25: Cumulative Distribution of Performance Metrics Across Models. These plots show the cumulative distribution of performance metrics for each model, highlighting the performance thresholds reached by each model.

mitigate the issues associated with high-dimensional data, such as overfitting and computational complexity. In medical datasets like the Chronic Kidney Disease (CKD) dataset, irrelevant or noisy features can significantly degrade model performance by introducing variance that does not contribute to the predictive task. Feature selection methods, such as Greylag Goose Optimization (GGO) and other binary optimization algorithms, allowed us to retain only the most informative features, such as blood glucose levels (bgr), red blood cell count (rc), and white blood cell count (wc), which are known to have strong correlations with CKD.

The reduction in dimensionality not only simplified the models but also led to improved generalization on unseen data. By focusing the model's attention on the most relevant features, the risk of overfitting was substantially reduced. In the context of CKD detection, overfitting is a particularly critical issue, as it could

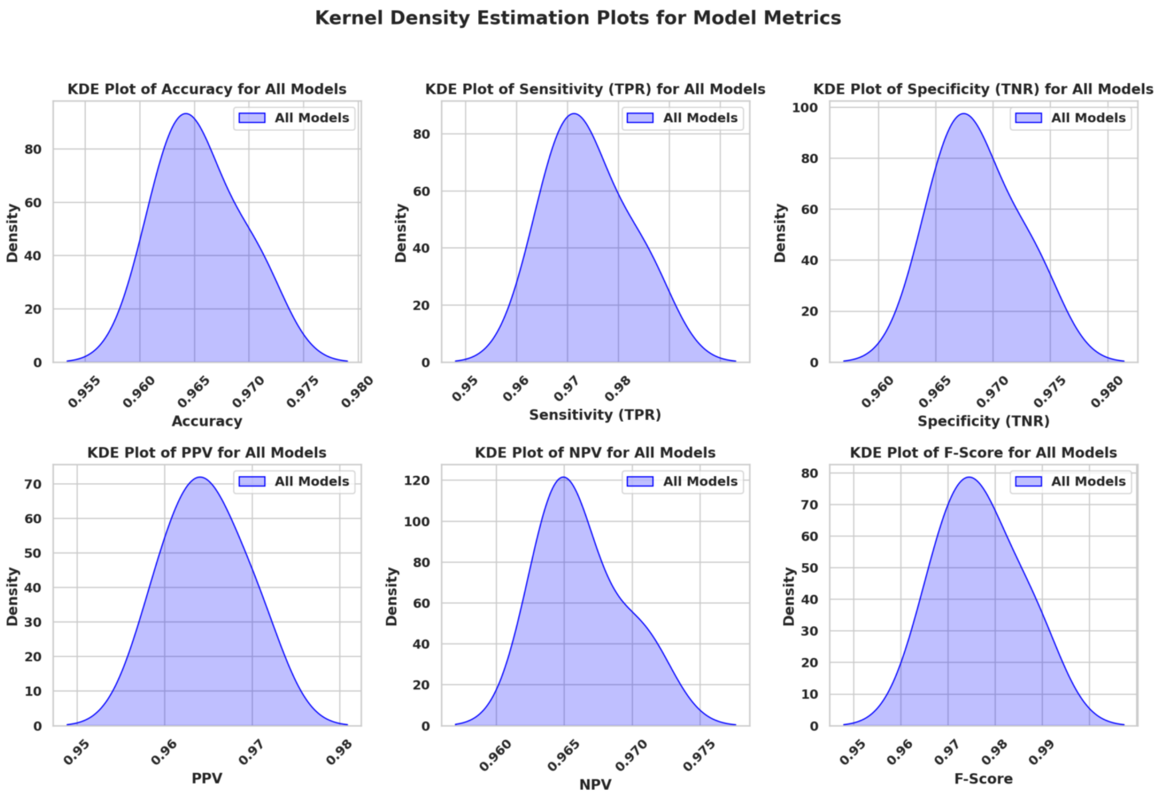

Figure 26: Pairwise Scatter Matrix of Metrics. This matrix visualizes the relationships between different performance metrics for the optimized models, with density plots along the diagonal.

lead to inaccurate predictions when the model is deployed in real-world settings with new patient data. This is especially true when working with small, imbalanced datasets, where the model may learn to "memorize" the data rather than generalize well to new instances. Through feature selection, we were able to ensure that the model trained on only the most pertinent clinical indicators, thus enhancing its ability to identify CKD cases reliably.

6.2 Effectiveness of Optimization Algorithms

The results of this study clearly demonstrate the effectiveness of binary optimization algorithms, particularly GGO, in improving the performance of machine learning models for CKD detection. GGO outperformed the other metaheuristics, such as Particle Swarm Optimization (PSO), Bat Algorithm (BA), and Whale Optimization Algorithm (WAO), across multiple metrics, including error reduction and fitness scores. This success can be attributed to GGO's ability to explore the feature space efficiently and select the most relevant features while avoiding local optima.

Metaheuristic algorithms, such as PSO and GA, are commonly used for feature selection and hyperparameter optimization because they can effectively search large, complex search spaces. However, GGO demonstrated superior performance in this context, achieving lower average errors and better fitness scores compared to the other algorithms. This suggests that GGO's specific design, which mimics the behavior of Greylag geese in their migratory patterns, provides a more effective approach for navigating the feature space and optimizing classification models in medical datasets. The ability of GGO to strike a balance between exploration (searching new areas of the feature space) and exploitation (refining the best solutions) was a key factor in its success in this study.

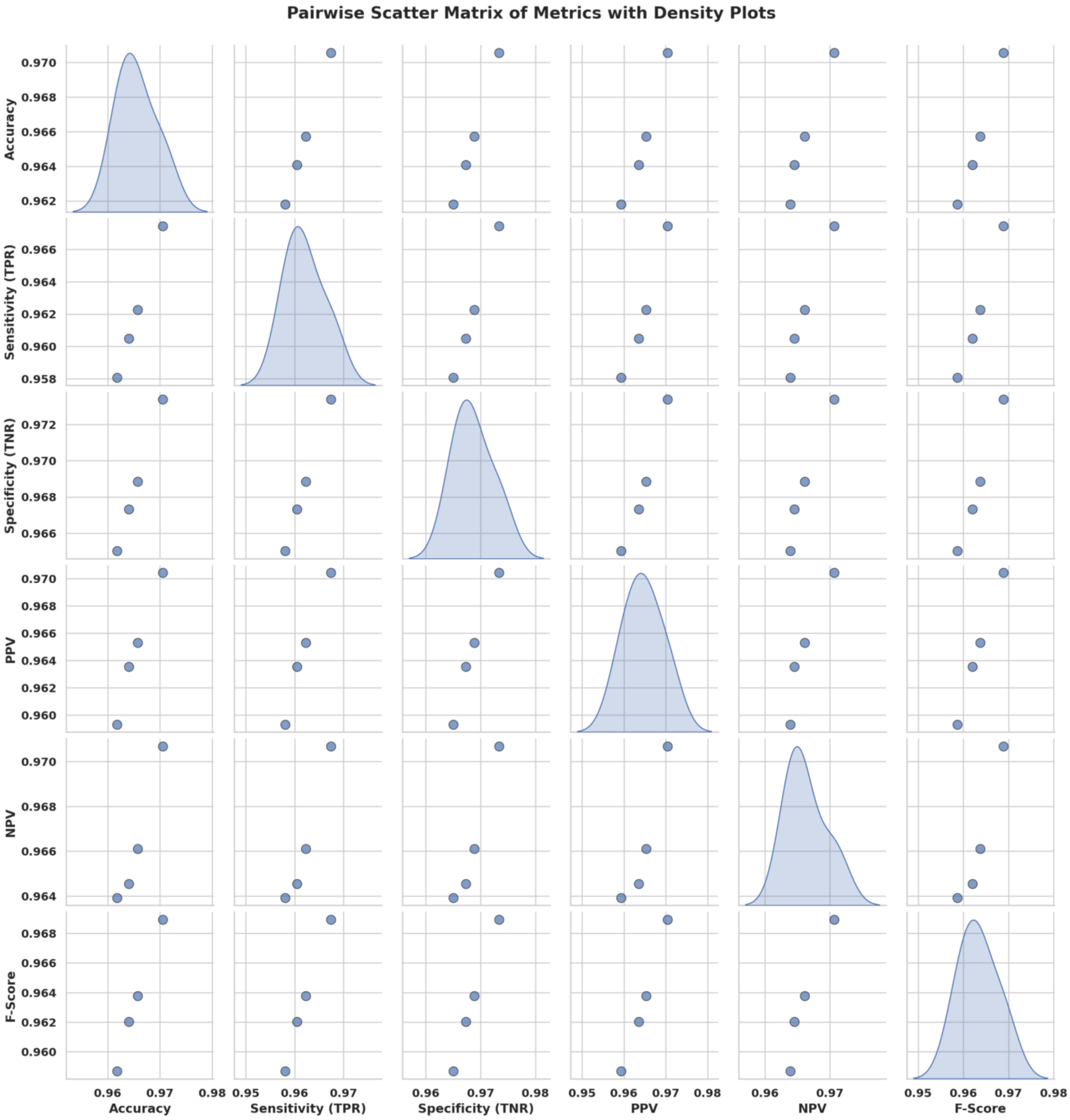

Figure 27: Kernel Density Estimation (KDE) for Model Performance Metrics. These plots compare the density distribution of various performance metrics across all optimized models.

6.3 Practical Implications

The optimized machine learning models developed in this study have significant practical implications for CKD detection and monitoring. In healthcare systems, the ability to reliably detect CKD at an early stage is crucial for improving patient outcomes and reducing long-term healthcare costs. The models, particularly the MLP model optimized with GGO, demonstrate high accuracy and sensitivity, making them suitable for deployment in clinical settings. By automating the detection of CKD, healthcare providers can screen large numbers of patients quickly and accurately, enabling early intervention and personalized treatment plans.

Moreover, these models have the potential to be integrated into mobile applications for real-time CKD monitoring. Given the widespread use of smartphones and wearable devices, there is a growing opportunity to leverage these tools for health monitoring in low-resource settings. By incorporating optimized machine learning models into mobile apps, individuals can monitor their kidney health, track relevant biomarkers, and

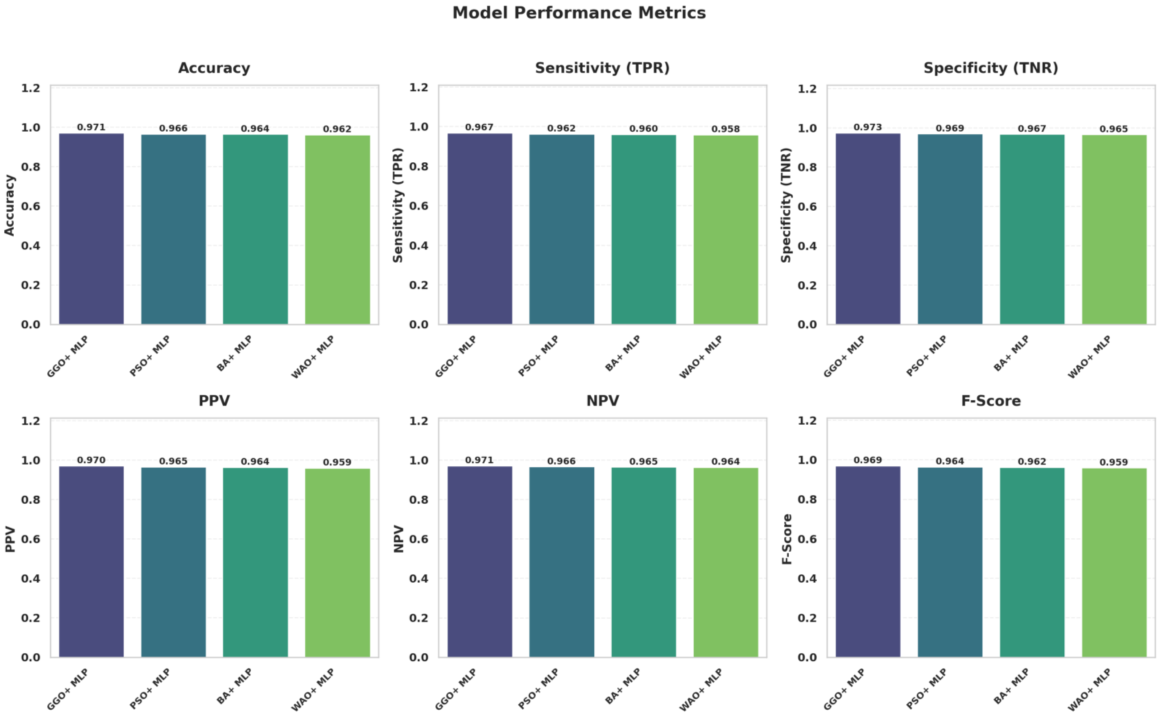

Figure 28: Model Performance Metrics Bar Plot. This bar plot compares the performance metrics (Accuracy, Sensitivity, Specificity, PPV, NPV, F-Score) for optimized models using different algorithms.

receive alerts for potential risks. Such applications could empower patients to take proactive measures in managing their health, improving early detection and reducing the need for expensive treatments at advanced stages of CKD.

6.4 Limitations

Despite the promising results, this study has several limitations that need to be addressed in future research. First, the dataset used in this study is relatively small, consisting of only 400 instances. While this dataset is sufficient for demonstrating the utility of feature selection and optimization algorithms, larger datasets would likely yield more robust and generalizable models. The small dataset also contributes to the class imbalance problem, with the CKD class making up only 38% of the dataset. This imbalance may have influenced the model performance, particularly in terms of specificity and positive predictive value (PPV), as the models might have been biased towards the majority class (non-CKD).

Another limitation is the presence of missing data, which was addressed by removing rows with NaN values. However, this approach resulted in the loss of valuable data, which could have been better handled through more sophisticated imputation techniques. Future work should explore alternative strategies for handling missing values, such as using imputation methods or incorporating domain-specific knowledge to predict missing entries.

Additionally, the dataset used in this study lacks certain clinical features that could further enhance the accuracy of the models. For example, patient medical history, comorbid conditions, and genetic factors could all play a significant role in predicting CKD. The inclusion of these additional features would likely improve the model's ability to differentiate between CKD and non-CKD cases, especially in the early stages of the disease when symptoms are not yet apparent. Future studies should consider expanding the dataset to include a broader range of clinical variables to improve the models' predictive power.

Finally, while the results of this study are promising, further validation is needed before these models can be deployed in real-world clinical settings. Future work should focus on validating the models on external datasets from different populations to assess their generalizability and robustness across diverse patient groups.

In summary, this study demonstrates the effectiveness of feature selection and binary optimization techniques for improving the performance of machine learning models in chronic kidney disease detection. The results highlight the importance of selecting the most relevant features and tuning model parameters to achieve the best performance. Despite the limitations of the dataset, the optimized models show significant promise for real-world applications, including integration into healthcare systems and mobile applications for monitoring kidney health. Future research should focus on addressing the limitations identified in this study and exploring the potential of these models in clinical practice.

7 Conclusion & Future Work

7.1 Conclusion

This study demonstrates the effectiveness of combining feature selection and hyperparameter optimization to improve the classification performance of machine learning models in the context of Chronic Kidney Disease (CKD) detection. The hybrid approach of Greylag Goose Optimization (GGO) for feature selection and Multilayer Perceptron (MLP) for classification significantly enhanced model performance, achieving an accuracy of 97.06%. This result highlights the potential of optimization techniques in healthcare applications, especially when working with small, imbalanced datasets. By reducing dimensionality and selecting the most relevant features, the models were able to focus on the key factors influencing CKD, thereby improving both predictive accuracy and generalizability.

The ability to achieve high accuracy with a small dataset is particularly important in medical diagnostics, where obtaining large and balanced datasets is often challenging. Our findings suggest that optimization algorithms, such as GGO, can effectively navigate the feature space, leading to better model performance and more reliable predictions. This approach is particularly valuable in early-stage disease detection, where timely and accurate diagnosis is critical for improving patient outcomes and reducing healthcare costs.

7.2 Contributions

This paper makes several important contributions to the field of machine learning for healthcare. First, we provide an in-depth analysis of feature selection methods using binary optimization algorithms, focusing on their ability to improve the performance of CKD classification models. By comparing multiple optimization techniques, such as GGO, PSO, BA, and WAO, we demonstrate the superior performance of GGO in feature selection tasks. Second, we explore the integration of hyperparameter optimization with feature selection, highlighting how this combined approach can further enhance model accuracy and robustness. The results from our study contribute to the growing body of knowledge on applying machine learning techniques to medical diagnostics, particularly in chronic disease prediction.

Furthermore, we provide practical insights into the potential applications of these techniques in healthcare. The optimized models could be integrated into real-world healthcare systems to improve the early detection of CKD, leading to better patient management and more efficient use of medical resources. Our work also paves the way for future research into optimizing machine learning models for other medical conditions, where early diagnosis is crucial.

7.3 Future Work

While the results of this study are promising, there are several avenues for future research to further enhance the effectiveness of CKD detection models.

- **Incorporation of Additional Features**: The current dataset primarily includes basic clinical features such as blood pressure, age, and red blood cell count. Future work should aim to incorporate additional clinical features, such as medical history, lab test results, and genetic factors, to improve prediction accuracy. These features may provide deeper insights into the patient's health and further refine the model's ability to distinguish between CKD and non-CKD cases.
- **Ensemble Methods**: To improve model robustness and performance, ensemble methods could be explored to combine the strengths of multiple models. Techniques such as bagging, boosting, or stacking could be applied to combine the predictions of various classifiers, potentially leading to improved generalization and reduced model variance. This would be particularly beneficial in the context of CKD detection, where small changes in data could lead to different predictions.
- **Mobile Healthcare Applications**: The integration of optimized models into mobile applications for real-time CKD monitoring represents an exciting direction for future work. With the growing prevalence of smartphones and wearable health devices, these technologies provide an ideal platform for continuous health monitoring. Future research should focus on developing deployment plans for mobile applications that use optimized machine learning models to offer personalized CKD risk assessments. Such applications could provide valuable real-time feedback to patients and healthcare providers, facilitating timely interventions and improving long-term health outcomes.
- **Validation with Larger Datasets**: While this study demonstrated the potential of our approach on a relatively small dataset, future work should aim to validate the models on larger and more diverse datasets. This would provide a more comprehensive evaluation of the model's generalizability across different populations and settings, helping to ensure that the models are robust and applicable in real-world clinical environments.

In summary, this research highlights the importance of optimizing machine learning models through feature selection and hyperparameter tuning for effective CKD detection. The promising results achieved in this study lay the foundation for future advancements in automated medical diagnosis, offering potential solutions for improving early disease detection and patient care in healthcare systems worldwide.

Data Availability

The data used in this study are openly available on Kaggle under the title Chiller Energy Data at `https://www.kaggle.com/datasets/mansoordaku/ckdisease`.

Declarations

- **Acknowledgments**
  Not applicable.
- **Conflict of interest/Competing interests**
  The authors declare that they have no conflicts of interest to report regarding the present study.

- **Ethics approval and consent to participate**
  Not applicable.
- **Consent for publication**
  Not applicable.

- **Funding**
  No Fund

References

[1] M. P. Duarte et al., “Prevalence of sarcopenia in patients with chronic kidney disease: A global systematic review and meta-analysis,” *Journal of Cachexia, Sarcopenia and Muscle*, vol. 15, pp. 501–512, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267199206.

[2] M. Colombo et al., “Biomarker panels associated with progression of renal disease in type 1 diabetes,” *Diabetologia*, vol. 62, pp. 1616 –1627, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:195190514.

[3] S. Manjula, D. H. Rajini, and D. K. Chokkanathan, “Optimizing chronic kidney disease diagnosis: A comprehensive review of metaheuristic algorithms and their applications in detection and classification,” *2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1654–1660, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:275018042.

[4] V. P. Singh and D. Jain, “A hybrid parallel classification model for the diagnosis of chronic kidney disease,” *Int. J. Interact. Multim. Artif. Intell.*, vol. 8, p. 14, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:242052476.

[5] I. H. I. Coutinho, “Early diagnosis of chronic kidney disease,” 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:262084636.

[6] N. R. Hill et al., “Global prevalence of chronic kidney disease – a systematic review and meta-analysis,” *PLoS ONE*, vol. 11, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:16929310.

[7] P. Sridevi, M. Rabbani, and S. I. Ahamed, “A comprehensive study for predicting chronic kidney disease, diabetes, hypertension, and anemia by machine learning and feature engineering techniques,” *2023 IEEE International Conference on Digital Health (ICDH)*, pp. 248–257, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261306019.

[8] Metun, G. P, and A. V, “Predictive analysis in nephrology: Evaluating machine learning models for chronic kidney disease prediction,” *International Journal for Research in Applied Science and Engineering Technology*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267083497.

[9] D. S. Abdelminaam, A. A. Alheijat, and M. Taha, “A comprehensive survey on dimension reduction for medical big data using optimization algorithms,” *2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pp. 331–337, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274797645.

[10] P. Nagre, P. Verma, P. Kumar, and N. L. Gavankar, “A review of machine learning and deep learning algorithms for chronic kidney disease prediction,” *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*, pp. 1–6, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:275774580.

[11] T Suguna and L Sumathi, “Early prediction of chronic kidney disease using machine learning techniques: A review,” *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266913542.

[12] E. M. Senan et al., “Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques,” *Journal of Healthcare Engineering*, vol. 2021, 2021. [Online]. Available: `https://api.semanticscholar.org/CorpusID:235695403`.

[13] D. A. Hasan, S. R. M. Zeebaree, M. A. M. Sadeeq, H. M. Shukur, R. R. Zebari, and A. H. Alkhayyat, “Machine learning-based diabetic retinopathy early detection and classification systems- a survey,” *2021 1st Babylon International Conference on Information Technology and Science (BICITS)*, pp. 16–21, 2021. [Online]. Available: `https://api.semanticscholar.org/CorpusID:237001648`.

[14] N. V. S. Sendhuraa, “Chronic kidney disease detection using machine learning techniques,” 2021. [Online]. Available: `https://api.semanticscholar.org/CorpusID:236624368`.

[15] M. V. Rocco et al., “Chronic kidney disease classification in systolic blood pressure intervention trial: Comparison using modification of diet in renal disease and ckd-epidemiology collaboration definitions,” *American Journal of Nephrology*, vol. 44, pp. 130 –140, 2016. [Online]. Available: `https://api.semanticscholar.org/CorpusID:21807933`.

[16] N. V. S. Sendhuraa, “Chronic kidney disease detection using machine learning techniques,” 2021. [Online]. Available: `https://api.semanticscholar.org/CorpusID:236624368`.

[17] A. Dr and V.Sumalatha, “Survey on diagnosis of chronic kidney disease usingmachine learning algorithms,” *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 590–595, 2020. [Online]. Available: `https://api.semanticscholar.org/CorpusID:231685031`.

[18] M. Makwana, D. R. Upadhyay, D. S. Degadwala, and D. Vyas, “A review on machine learning approaches for hiv infected patient chronic kidney disease stage classification,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2022. [Online]. Available: `https://api.semanticscholar.org/CorpusID:254560278`.

[19] N. Lameire et al., “Harmonizing acute and chronic kidney disease definition and classification: Report of a kidney disease: Improving global outcomes (kdigo) consensus conference.,” *Kidney international*, 2021. [Online]. Available: `https://api.semanticscholar.org/CorpusID:235808772`.

[20] R. Asa, L. Busi, J. S. M. Professor-Dr, P.-D. B. Chair, and P. R. S. Professor, “A review: Analyzing risk factors and prediction for chronic kidney disease using machine and deep learning techniques,” *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pp. 1456–1462, 2023. [Online]. Available: `https://api.semanticscholar.org/CorpusID:267258854`.

[21] A. Eslam, M. G. Abdelfattah, E.-S. M. E. El-Kenawy, and H. E.-D. Moustafa, “Optimization of feature selection using greylag goose optimization algorithm for monkeypox,” *Journal of Artificial Intelligence in Engineering Practice*, 2024. [Online]. Available: `https://api.semanticscholar.org/CorpusID:273561252`.

[22] Mamta, D. Mane, A. J. Dhumale, and H. Dhiman, “Distributed gradient boosting machines (dgbm) with feature selection via particle swarm optimization for predicting disease progression,” *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, vol. 1, pp. 1–6, 2023. [Online]. Available: `https://api.semanticscholar.org/CorpusID:269182952`.

[23] A. O. Salau, E. D. Markus, T. A. Assegie, C. O. Omeje, and J. N. Eneh, “Influence of class imbalance and resampling on classification accuracy of chronic kidney disease detection,” *Mathematical Modelling of Engineering Problems*, 2023. [Online]. Available: `https://api.semanticscholar.org/CorpusID:257347361`.

[24] A. S. Levey et al., “The definition, classification, and prognosis of chronic kidney disease: A kdigo controversies conference report.,” *Kidney international*, vol. 80 1, pp. 17–28, 2011. [Online]. Available: `https://api.semanticscholar.org/CorpusID:15295796`.

[25] B. K. Betzler et al., “Impact of chronic kidney disease epidemiology collaboration (ckd-epi) gfr estimating equations on ckd prevalence and classification among asians,” *Frontiers in Medicine*, vol. 9, 2022. [Online]. Available: `https://api.semanticscholar.org/CorpusID:250497579`.

[26] J. Ja, “New abc chronic kidney disease classification,” 2017. [Online]. Available: `https://api.semanticscholar.org/CorpusID:80139781`.

[27] D. R. Farrell and J. A. Vassalotti, “Screening, identifying, and treating chronic kidney disease: Why, who, when, how, and what?” *BMC Nephrology*, vol. 25, 2024. [Online]. Available: `https://api.semanticscholar.org/CorpusID:267254108`.

[28] D. jun Xiang et al., “Clinical application of neutrophil gelatinase-associated lipocalin in the revised chronic kidney disease classification.,” *International journal of clinical and experimental pathology*, vol. 7 10, pp. 7172–81, 2014. [Online]. Available: `https://api.semanticscholar.org/CorpusID:11026731`.

[29] 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, p. 100 189, Jan. 2023, ISSN: 2153-3539. DOI: `10.1016/j.jpi.2023.100189`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S2153353923000032`.

[30] S. Alshakrani, R. Taha, and N. Hewahi, “Chronic Kidney Disease Classification Using Machine Learning Classifiers,” in *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Sep. 2021, pp. 516–519. DOI: `10.1109/3ICT53449.2021.9581345`. [Online]. Available: `https://ieeexplore.ieee.org/document/9581345`.

[31] J. R. Lambert and E. Perumal, “Oppositional firefly optimization based optimal feature selection in chronic kidney disease classification using deep neural network,” en, *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 4, pp. 1799–1810, Apr. 2022, ISSN: 1868-5145. DOI: `10.1007/s12652-021-03477-2`. [Online]. Available: `https://doi.org/10.1007/s12652-021-03477-2`.

[32] R. Aswathy et al., “Optimized Tuned Deep Learning Model for Chronic Kidney Disease Classification,” en, *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2097–2111, 2021, Publisher: Tech Science Press, ISSN: 1546-2218, 1546-2226. DOI: `10.32604/cmc.2022.019790`. [Online]. Available: `https://www.techscience.com/cmc/v70n2/44656`.

[33] A. Khamparia, G. Saini, B. Pandey, S. Tiwari, D. Gupta, and A. Khanna, “KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network,” en, *Multimedia Tools and Applications*, vol. 79, no. 47, pp. 35 425–35 440, Dec. 2020, ISSN: 1573-7721. DOI: `10.1007/s11042-019-07839-z`. [Online]. Available: `https://doi.org/10.1007/s11042-019-07839-z`.

[34] K. Venkatrao and S. Kareemulla, “HDLNET: A Hybrid Deep Learning Network Model With Intelligent IOT for Detection and Classification of Chronic Kidney Disease,” *IEEE Access*, vol. 11, pp. 99 638–99 652, 2023, ISSN: 2169-3536. DOI: `10.1109/ACCESS.2023.3312183`. [Online]. Available: `https://ieeexplore.ieee.org/document/10239391`.