International Journal of Informatics and Communication Technology (IJ-ICT)
Vol. 14, No. 2, August 2025, pp. 684~707
ISSN: 2252-8776, DOI: 10.11591/ijict.v14i2.pp684-707

An innovative approach for predictive modeling and staging of chronic kidney disease

Safa Boughougal[1], Mohamed Ridda Laouar[1], Abderrahim Siam[2], Sean Eom[3]
[1]Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria
[2]Knowledge Engineering and Computer Security Laboratory (ICOSI), University of Abbes Laghrour, Khenchela, Algeria
[3]Department of Management, Southeast Missouri State University, Cape Girardeau, USA

Article Info

Article history:

Received May 24, 2024
Revised Dec 16, 2024
Accepted Jan 19, 2025

Keywords:

Chronic kidney disease
Classification models
Estimated glomerular
Filtration rate
Modification of diet in renal disease

ABSTRACT

Diagnosing silent diseases such as chronic kidney disease (CKD) at an early stage is challenging due to the absence of symptoms, making early detection crucial to slowing disease progression. This study addresses this challenge by introducing a novel feature, the estimated glomerular filtration rate (eGFR), calculated using the modification of diet in renal disease (MDRD) formula. We enriched our dataset by incorporating this feature, effectively increasing the volume of data at our disposal. eGFR serves as a critical indicator for diagnosing CKD and assessing its progression, thereby guiding clinical management. Our focus is on developing machine learning and deep learning models for the efficient and precise prediction of CKD. To ensure the reliability of our approach, we employed robust data collection and preprocessing techniques, resulting in refined information for model training. Our methodology integrates various machine learning and deep learning models, including four machine learning algorithms: adaptive boosting (AdaBoost), random forest (RF), Bagging, and artificial neural network (ANN), as well as a hybrid model. Our proposed ANN_AdaBoost model not only introduces a novel perspective by addressing an identified gap but significantly enhances CKD prediction.

Corresponding Author:

Safa Boughougal
Laboratory of Mathematics, Informatics and Systems (LAMIS)
Echahid Cheikh Larbi Tebessi University
Tebessa, Algeria
Email: safa.boughougal@univ-tebessa.dz

1. INTRODUCTION

Chronic kidney disease (CKD) represents a significant global health challenge, characterized by a gradual loss of kidney function over time. One of the leading contributors to CKD is diabetes mellitus, a condition affecting approximately 422 million people worldwide, according to the World Health Organization (WHO) [1]. Prolonged hyperglycemia associated with diabetes can damage vital organs, including the kidneys, eyes, heart, nerves, and blood vessels [2].

CKD frequently manifests as a complication in individuals with diabetes, with a majority experiencing this condition during their lifetime [3]. The metabolic consequences of diabetes play a significant role in the development of CKD, making it one of the primary causes of kidney damage. Approximately 40% of diabetic patients develop CKD, a silent yet potentially fatal condition that commonly afflicts adults, particularly those with comorbidities such as diabetes or hypertension [4].

Diagnosis and assessment of CKD severity often rely on parameters such as estimated glomerular filtration rate (eGFR), the presence of albuminuria, age, dietary habits, and underlying health conditions like

diabetes and hypertension [5]. Notably, eGFR is a critical indicator of kidney function and is calculated based on various factors, including blood creatinine levels, age, and sex [6]. CKD diagnosis typically involves identifying consistent increases in urinary albumin excretion (albuminuria), diminished eGFR, or other signs indicating kidney impairment [7]. Laboratory tests assessing kidney function, such as elevated creatinine and blood urea levels as well as anemia due to reduced red blood cell count, contribute to diagnostic evaluation [8].

The silent progression of CKD poses challenges in early detection, often leading to advanced stages with debilitating consequences. Patients with CKD, especially those with concomitant diabetes, face numerous complications that necessitating meticulous management and potentially regular dialysis [9]. The economic burden associated with managing both diabetes and CKD escalates due to expenses incurred for medications and treatments, significantly impacting patients' quality of life.

Early detection of CKD among diabetic individuals enables timely interventions to mitigate disease progression, thereby minimizing complications and sustaining a better quality of life [10]. Furthermore, early diagnosis aids in averting complications such as end-stage renal disease and cardiovascular disorders. To address the imperative need for early detection, sophisticated approaches such as artificial intelligence offer promising solutions [11].

In this study, our primary objective is to introduce an intelligent approach for predicting and diagnosing CKD. Instead of limiting ourselves to an innovative machine learning and deep learning model, we deliberately opted for an approach that intelligently leverages data for more efficient and accurate disease prediction. Our methodology begins with a thorough exploration various analytical techniques, culminating in enriching our dataset. Specifically, we enriched our dataset by incorporating the eGFR attribute, a previously identified aspect for future work in previous studies, which we accomplished. This addition enhances our dataset's comprehensiveness and improves aspect for future work by incorporating a crucial biomarker for CKD diagnosis. We calculated the eGFR using the modification of diet in renal disease (MDRD) formula, which is renowned for its applicability in approximating the glomerular filtration (GFR) rate of the kidney. We then conducted a comprehensive analysis of specific risk factors for CKD within this particular group. The inclusion of the eGFR indicator in our dataset constitutes a significant breakthrough, recognized as a crucial yet neglected factor in previous research on predicting CKD. This indicator is essential for diagnosing the presence of CKD and contributes to determining its stage of progression, assisting clinicians in planning appropriate treatment. To enhance the accuracy of our predictions, we adopted an intelligent approach by using rigorous techniques such as least absolute shrinkage and selection operator (LASSO) for feature selection and principal component analysis (PCA) for dimensionality reduction.

While previous studies have explored machine learning and deep learning models for CKD prediction, many have not explicitly addressed the integration of crucial biomarkers such as the eGFR, despite its recognized importance in CKD diagnosis and staging. Additionally, prior research often focused on using the same dataset for training, testing, and validation, which limits the generalizability of the results Unlike many previous studies, which relied on using the same dataset for training, testing, and validation, our approach evaluates the models' ability to generalize to new data, confirming the validity of our conclusions and contributing to more reliable CKD prediction.

In this paper, we extend our investigation by validating model results using a second distinct dataset to ensure the reliability and robustness of our results. This decision is motivated by previous findings indicating that many prior studies have relied solely on a single dataset for model training, testing, and validation, which may potentially limit the generalizability of the results. By utilizing a separate dataset for testing, we aim to assess the ability of our models to generalize to new data and confirm the validity of our conclusions.

2. RELATED WORK

The use of machine learning algorithms for chronic disease prediction has become increasingly important in recent years because of this disease's effective prediction and improved accuracy [12]. Ani *et al.* [13] developed a decision support system for clinicians based on machine learning models: LDA classifier, neural network (NN)-based back propagation (BPN), probability-based Naive Bayes, random subspace classification algorithms, decision tree (DT), and K-nearest neighbor (KNN) to predict CKD. After evaluating the accuracy results, the random forest (RF) model achieved the highest accuracy of 94%, while the LDA model had an accuracy of 76%. Predictive models are developed and compared for early prediction of CKD using biochemical analyses. Xioa *et al.* [14] employed nine machine learning algorithms: logistic regression (LR), Elastic Net, LASSO regression, ridge regression, support vector machine (SVM), RF, XGBoost, NN, and KNN on a dataset of 551 patients. Elastic Net, LASSO regression, ridge regression, and LR yielded the best results with accuracies of 0.82 and 0.81, respectively. Hassan *et al.* [15] proposed a model that can predict patients with early-stage CKD by utilizing the seven most relevant features: hemo, sg, pcv, al, pe, se, and pb, using the XGBoost feature selection technique based on XGBoost. Hassan *et al.* [15] applied five

machine learning models, NN, RF, SVM, random tree, and Bagging tree model, to their XGBoost-based dataset to find the best predictive model. The SVM model achieved the highest accuracy of 100% for the seven relevant features, while the NN model achieved 100% accuracy on the entire UCI machine learning repository dataset.

To impede or arrest the progression of CKD, Baidya *et al.* [16] employed a set of twelve machine learning classifiers, encompassing adaptive boosting (AdaBoost), DT, XGBoost, CatBoost, KNN, RF, Gradient boosting, Stochastic gradient boosting, light gradient boosting machine (LGBM), Extra tree, SVM, and artificial neural network (ANN). These classifiers underwent training on ten crucial features, namely hemoglobin, specific gravity, albumin, sugar, random blood glucose, serum creatinine, potassium, packed cell volume, white and red blood cell count, and diabetes mellitus. The feature selection method used was PCA. Notably, the XGBoost classifier demonstrated an impressive accuracy of 98% in the early detection of CKD. The approach of Baidya *et al.* [16] highlights the effectiveness of machine learning models in the early detection of CKD, carrying substantial implications for the timely management and treatment of this disease. The outcomes obtained emphasize the significance of feature selection and the utilization of advanced machine learning techniques to enhance the precision of CKD-related predictions.

To predict whether a person has CKD or not, Chittora *et al.* [17] applied seven machine learning classifiers, including ANN, C5.0, Chi-square Automatic interaction detector, LR, linear SVM with penalty L1 & L2, and random tree, on the CKD dataset from the UCI machine learning repository. They applied three feature selection techniques for each classifier: correlation-based feature selection, wrapper method feature selection, and LASSO regression. After comparing the results, Chittora *et al.* [17] found that the LSVM classifier with L2 penalty achieved the highest accuracy of 98.86% with all features included. Yashfi *et al.* [18] proposed a system for predicting CKD based on machine learning algorithms such as ANN and RF. They utilized the Chi-Square Test to extract the most significant features. The RF algorithm achieved an accuracy of 97.12%. CKD is prevalent and difficult to diagnose early due to its asymptomatic nature. Pal [19] focuses on developing a machine learning model for CKD's early detection. The proposed approach, integrating baseline classifiers with a majority voting method, achieved a 3% accuracy increase over existing models. Utilizing data mining techniques, the model involves three steps: classification based on categorical attributes, non-categorical attributes, and a combined approach. Machine learning classifiers such as SVM, RF, and ANN were evaluated. Their performances yielded accuracies of 91%, 93%, and 89%, respectively, highlighting their effectiveness in CKD classification. Rahman *et al.* [20] addressed CKD diagnosis, focusing on South Asia's health challenges. Employing eight ensemble learning methods, including LightGBM, they optimized classifier performance using MICE imputation for missing values and Borderline-SMOTE for data balance. Recursive feature elimination and boruta were instrumental in significant feature selection. The proposed method achieved an outstanding average accuracy of 99.75%, demonstrating its effectiveness for precise CKD diagnosis. The study of Rahman *et al.* [20] uniquely combined ensemble methods, showcasing advancements not explored collectively before. Experiments on two datasets validated the model's efficiency and contributed novel strategies to CKD analysis. For precise identification of CKD, Ghosh *et al.* [21] employed advanced techniques such as SVM, AdaBoost, linear discriminant analysis (LDA), and gradient boosting (GBoost). They utilized data collected from the UCI repository for their analysis. Notably, the GBoost model achieved an outstanding accuracy of 99.80%, outperforming the other models employed in the study. Navaneeth and Suchetha [22] used an innovative hybrid deep learning network comprising a convolutional neural network (CNN) classifier and SVM, which is introduced to overcome challenges faced by conventional data classification networks. The network incorporates a dynamic clustering approach and a feature-pruning algorithm to select the most relevant attributes for the classification task. The urea concentration in saliva samples is examined for disease detection, and a novel detection module is developed to test the samples. The CNN-SVM outperformed the traditional CNN model, achieving an average accuracy of 96.51%. For effective detection of CKD, Ebiaredoh-Mienye *et al.* [23] introduced a cost-sensitive AdaBoost classifier with a feature selection strategy based on information gain. This approach expedited and economized CKD screening, as only a small subset of clinical test parameters was required for accurate interpretation. The methodology was compared with other approaches, including classifiers such as logistic regression, DT, RF, SVM, XGBoost, and the traditional AdaBoost for CKD prediction. With a reduced set of features, the AdaBoost approach outperformed other classifiers in terms of accuracy, achieving 99.8%. Experimental results demonstrated the positive impact of feature selection on the efficiency of different classifiers.

Our study utilizes four machine learning classifiers: the Bagging classifier, AdaBoost, RF, and ANN. While numerous studies have explored machine learning models, most have focused on enhancing existing datasets without exploring new perspectives in our research domain. Furthermore, our innovative contribution lies in enhancing predictive modeling capabilities by introducing a crucial feature the GFR, essential for diagnosing and predicting the stage of CKD. Unlike traditional approaches applying AI or machine learning models to datasets common to all studies, our novel approach fills an identified gap in

previous research. To bolster the relevance of our model, we implemented two feature selection methods: PCA and LASSO. This approach aims to improve accuracy by identifying the most informative features for modeling, thereby enhancing the robustness of our approach. Improving the dataset requires effective preprocessing, and to ensure the validity of our results, we also employed a separate test dataset. Moreover, hybrid methods in our study aim to optimize the modeling process, ensuring a robust and generalizable approach. By providing a unique perspective to enhance predictive accuracy in CKD, our research opens new avenues for advancing modeling in this specific domain. Table 1 provides a summary of key studies on CKD prediction using machine learning algorithms. It highlights the datasets used, the models applied, and the feature selection methods employed.

Table 1. Summary of previous studies comparison

Previous studies	Dataset	Models	Feature selection methods	Accuracy	Limitation
Ani *et al.* [13]	UCI	LDA classifier, NN-based back propagation (BPN), Naive Bayes, RF, DT, and KNN	_	76%,81%,78%,94%, 93%, and 90%	The authors did not use feature selection methods, and they did not use data preprocessing to improve accuracy.
Xioa *et al.* [14]	551 patient's datasets	LR, Elastic Net, LASSO regression, ridge regression, SVM, RF, XGBoost, NN, and KNN	_	82%,82%,81%,81%, 81%, 80%,83%,80 and 74%	No feature selection method was used in this work to improve accuracy.
Hassan *et al.* [15]	UCI	NN, RF, SVM, RT, BTM	XGBoost	97.5%,98.75%,100 %,96,25%,97.5%	The authors used only 7 features
Baidya *et al.* [16]	UCI	ADB,DT,XGB,CatBoost ,KNN,RF,ET,GNB	_	94%,95%95%,98%, 99%,95%,99%, 96%	The researchers did not use any feature selection methods to improve the accuracy and efficiency of the machine learning models.
Chittora *et al.* [17]	UCI	ANN, C5.0, CHAID, LR, LSVM_L1,LSVM_L2,RT,KNN	LASSO,Wrapper, CFS	90,2%,88,29%,97,07%,74,15%,97,07%, 97,07%,88,78%,56, 59 (with LASSO)	While this study employed three feature selection techniques (correlation-based, Wrapper method, and LASSO regression), other potentially effective methods were not explored.
Yashfi *et al.* [18]	UCI	ANN, RF	Chi-Square	94,5%,97,12%	The research does not include performance results after utilizing the chi-square feature selection method; accuracy results are on 24 features.
Pal [19]	UCI	SVM, RF, and ANNk	_	91%, 93%, and 89%,	No feature selection methods were used in this research.
Rahman *et al.* [20]	UCI	RF, Voting, Bagging, AdaBoost, GBDT, XGBoost, LightGBM, and Stacking	Boruta and RFE	96,15%,100%,98,07%,100%,98,07%, 96,15%,98,07% and 100%	This study lacks indication or mention of the features selected by the RFE and Boruta methods. This lack of transparency hinders a thorough understanding of the selection process, limiting transparency and result interpretation.
Ghosh *et al.* [21]	UCI	SVM, AdaBoost, LDA, and GBoost	_	99,56%,97,91%,97, 91%99,8%	No feature selection methods were employed in this study, which may lead to less efficient models.
Navaneeth and Sucheta [22]	172 participants	CNN classifier, SVM, and CNN-SVM	pruning	87,32%,95,01% and 96,12%	In this study, utilizing a small dataset (172 patients) compromises the generalization of results and the model's robustness due to an elevated risk of random variations.
Ebiaredoh-Mienye *et al.* [23]	UCI	AdaBoost, LR, DT, RF, SVM, XGBoost	Information Gain	93%,94%,90,2%,95,2%,93,7%,95,8%	No data preprocessing methods were used in this study, which may introduce biases and compromise the robustness of results due to the potential vulnerability to missing.

3. METHOD

Our research used a UCI machine learning repository dataset dedicated to CKD. This dataset consists of 25 attributes. To enhance the quality and relevance of this data, we have made concerted efforts to incorporate critical medical markers relevant to the condition. A significant addition is the eGFR calculation, which is a pivotal metric for assessing kidney function. The inclusion of the eGFR is critical for improving the accuracy and precision of our predictions, as it is a key biomarker in diagnosing CKD and determining its

progression stages. This is particularly important since the stage of CKD progression influences therapeutic decisions and prognoses, thus aiding clinicians in providing timely interventions.

The choice of eGFR as a core attribute aligns with the primary objective of this study, which is to go beyond traditional predictive models by enhancing the dataset with medically significant markers. By integrating the eGFR, we address a gap noted in previous studies, where its absence limited the ability to accurately predict both the occurrence and progression of CKD. To further refine our model, we employed two well-established techniques: LASSO for feature selection and PCA for dimensionality reduction. LASSO was chosen for its ability to identify and retain the most relevant features, thereby improving model accuracy and minimizing the risk of overfitting. PCA, on the other hand, was applied to reduce the dataset's dimensionality while preserving the most significant patterns in the data. Both methods aim to enhance the model's efficiency and predictive power by focusing on the most informative attributes, leading to better generalization and interpretation of results. The dataset was then divided into 80% for training and 20% for testing to assess the performance of our machine learning models. The selected models include RF, AdaBoost, Bagging, and ANN, which were chosen for their proven ability to handle complex classification tasks. The performance of these models, with and without feature selection, was compared to evaluate the impact of eGFR and other selected features on the prediction accuracy.

Our approach to CKD prediction comprises six phases, as shown in Figure 1, and is designed to ensure the reliability and generalizability of the results. By systematically incorporating feature selection and validation steps, we aim to improve model performance and enhance the clinical relevance of our predictions.

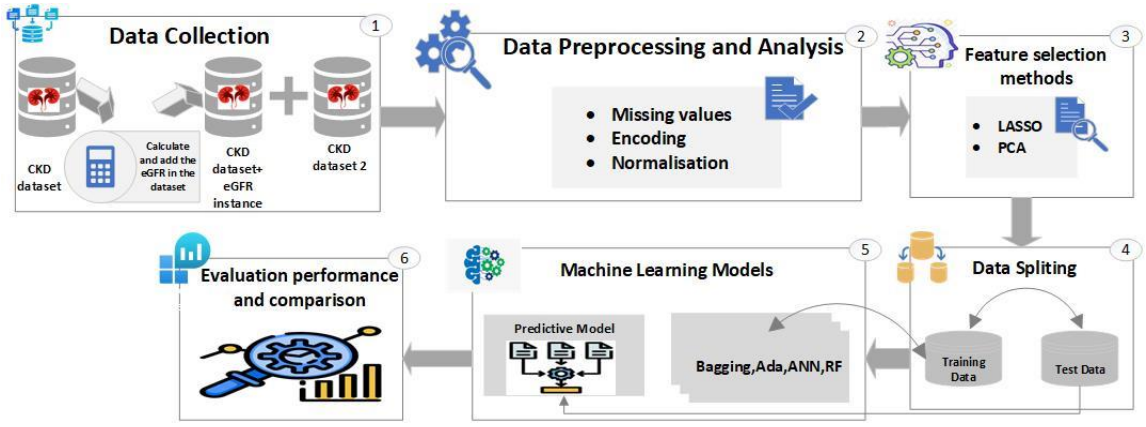

Figure 1. Description of our approach

3.1. Description of the dataset 1

The dataset used for this work is collected from the UCI machine learning repository [24], specifically focusing on CKD. Initially, the dataset comprised 25 features encapsulating various clinical measures and indicators pertinent to renal disorders. It contains 400 instances, each representing a distinct diabetic patient case, and is categorized into two target classes: "ckd" and "notckd." Given the pivotal role of the eGFR as a primary indicator of renal functionality, we elected to enhance the dataset by incorporating eGFR as an additional feature. We determined the eGFR utilizing the MDRD [25], which is widely recognized for its effectiveness in estimating the kidney's glomerular filtration rate (eGFR). The MDRD method is a specialized equation that considers various factors, including serum creatinine and age, to measure kidney function accurately. This formula has become an essential tool in nephrology because it offers a reliable estimate of eGFR, which is crucial for diagnosing and monitoring kidney diseases. The MDRD formula for calculating the eGFR is given below.

$$eGRF = 186 \times (SC)^{-1,154} \times (age)^{-0,203} \times 1\ (if\ male) \times 0{,}742\ (if\ female)$$

With this inclusion, the dataset has expanded to encompass 26 features while maintaining the original 400 instances. This strategic augmentation is geared towards bolstering our analyses' accuracy and contextual relevance and ensuing predictive models for renal disorders in the diabetic population. The characteristics of the original and new "eGFR" features are shown in Table 2, which provides an overview of all features in the dataset. The target class that is mentioned in the dataset that is "classification" which is divided into two classes' ckd=250 and notckd=150, as shown in the following Figure 2:

Table 2. Represents original and new "eGFR" features

Features	Description	Type
age	age in years	Numerical
bp	blood pressure in mm/Hg	Numerical
sg	Specific gravity	Nominal
al	Albumin	Nominal
su	Sugar	Nominal
rbc	Red blood cells	Nominal
pc	Pus cells	Nominal
pcc	Pus cell clumps	Nominal
ba	Bacteria	Nominal
bgr	Blood glucose random in mgs/dl	Numerical
bu	Blood urea in mgs/dl	Numerical
sc	Serum creatinine in mgs/dl	Numerical
sod	Sodium in mEq/L	Numerical
pot	Potassium in mEq/L	Numerical
hemo	Hemoglobin (gms)	Numerical
pcv	Packed cell volume	Numerical
wc	White blood cells count (Cells/cumm)	Numerical
rc	Red blood cells count (millions/cumm)	Numerical
htn	Hypertension (yes, no)	Nominal
dm	Diabetes mellitus (yes, no)	Nominal
cad	Coronary artery disease (yes, no)	Nominal
appet	Appetite (good, poor)	Nominal
pe	Pedal Edema (yes, no)	Nominal
ane	Anemia (yes, no)	Nominal
eGFR	estimated glomerular filtration rate in mL/min/1.73m^2	Numerical
class	Class (ckd or notckd)	Nominal

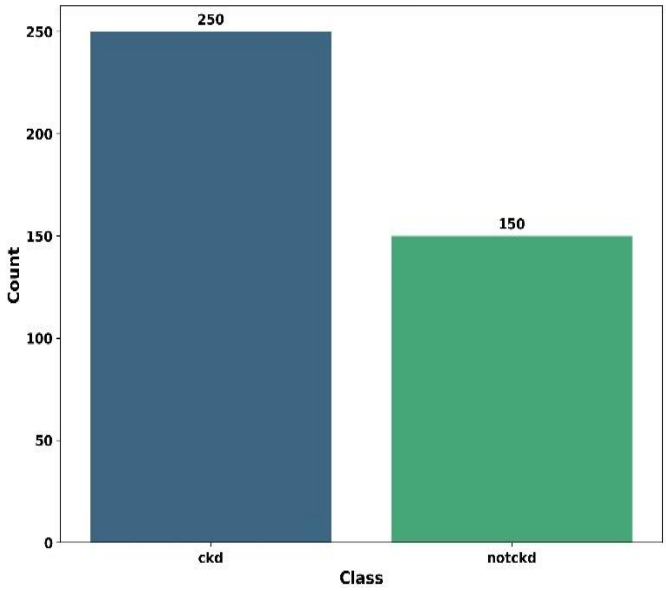

Figure 2. Distribution of patients with and without CKD

The following Figure 3 shows the correlation between CKD and diabetes among the studied population. Out of the 400 patients analyzed, 250 individuals were diagnosed with CKD. Remarkably, over half of these, precisely 136, are also diabetic. This significant proportion strongly suggests that diabetes could be a pivotal factor or a comorbidity in the development or progression of CKD in this particular cohort. On the other hand, the dataset indicates that all 150 patients without CKD are non-diabetic. This absence of any diabetic patient in the notckd group further amplifies the potential association between diabetes and CKD. In the context of our study, which aims to predict kidney disease among people with diabetes, these observations emphasize the critical nature of our research. The current dataset draws a compelling narrative about the potential risks people with diabetes face concerning CKD.

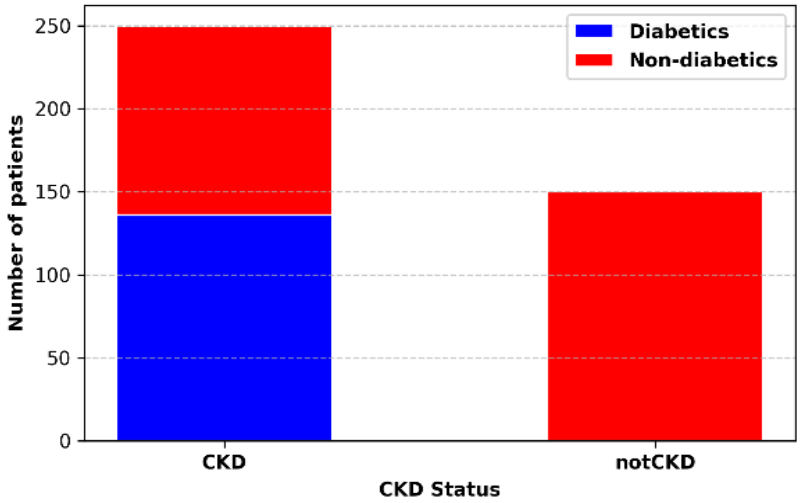

Figure 3. Distribution of diabetic patients with CKD

3.2. Description of the dataset 2

In this study, we utilized a dataset Islam *et al.* [26] collected from Enam Medical College, Savar, Dhaka, Bangladesh. This dataset was employed to validate the outcomes of our models following appropriate preprocessing steps. It comprises 200 instances and encompasses 28 distinct features. Among these features, particular emphasis is placed on the eGFR in predicting CKD. To enhance our primary dataset, we computed the eGFR and incorporated it into our initial dataset. The eGFR estimates kidney filtration rate, a pivotal indicator of renal function, thereby serving as a critical feature in CKD prediction. Furthermore, the disease stage, a significant parameter influencing CKD prediction, is also encompassed in this dataset. We opted for this dataset due to its similarity to both the training and test datasets, ensuring consistency in evaluating our model performance and reinforcing the reliability of our findings. The features of this dataset 2 are delineated in Table 3, facilitating a comprehensive analysis of their impact on our predictive models.

Table 3. Features and description of the dataset 2

Features	Description
bp (Diastolic)	blood pressure diastolic
bp limit	blood pressure
sg	Specific gravity
al	Albumin
class	Class (ckd or notckd)
rbc	Red blood cells
su	Sugar
pc	Pus cell
pcc	Pus cell clumps
ba	Bacteria
bgr	Blood glucose random
bu	Blood urea
sod	Sodium
sc	Serum creatinine
pot	Potassium
hemo	Hemoglobin
pcv	Packed cell volume
rbcc	Red blood cells count
htn	Hypertension
dm	Diabetes mellitus
cad	Coronary artery disease
appet	Appetite
pe	Pedal Edema
ane	Anemia
GRF	glomerular rate filtration
stage	CKD Stage (S1, S2, S3, S4, S5)
Affected	Affected by CKD or not (1,0)
age	Age in years

3.3. Preprocessing

The dataset from the UCI machine learning repository contains missing and null values. Therefore, the data-preprocessing step is crucial in training machine learning models, as these can lead to errors during algorithm execution and inaccurate predictions. To address this issue in our study, we used imputation methods.

The features of the dataset are initially categorized into categorical and numerical. For the categorical columns, missing values are replaced by the mode (the most frequent value) of the respective feature. Figure 4 illustrates the distribution of numerical characteristics before using imputation methods. An imputation technique is employed for the numerical features to substitute missing values with the average of the available values for that feature. Subsequently, a function is devised to detect and rectify outliers within the numerical attributes. Utilizing the interquartile range method establishes boundaries to identify upper and lower outliers. Figure 5 demonstrates the distribution of numerical characteristics after applying imputation techniques. Any detected outliers are then replaced with the mean value of their respective feature to make the data set more stable and less sensitive to these extreme fluctuations. The formula for calculating the interquartile range [24], [27] is given below, where Q3 is the third quartile (or the 75th percentile), and Q1 is the first quartile (or the 25th percentile).

$$IQR = Q3 - Q1 \tag{1}$$

The bounds for outliers are determined as upper bound and lower bound; the formula [28] of each outlier is given below:

$$Outliers\ Bound\ Upper(OU) = Q3 + 1.5 \times IQR \tag{2}$$

$$Outliers\ Bound\ Lower(OL) = Q1 - 1.5\ \times IQR \tag{3}$$

Then, we encoded the categorical variables into numerical variables and replaced the original data with the encoded data, which is necessary for machine learning algorithms that only accept numerical data using the "LabelEncoder" class. Figure 6 shows the distribution of these categorical variables.

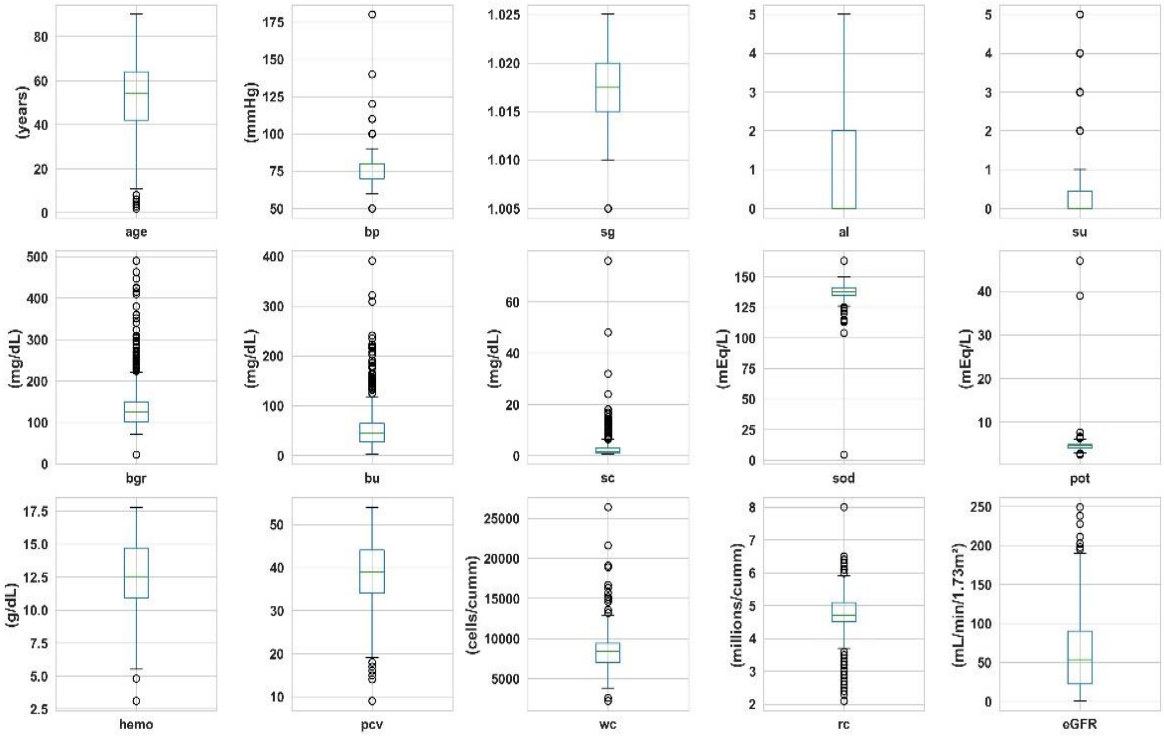

Figure 4. Distribution of numerical characteristics before using imputation methods

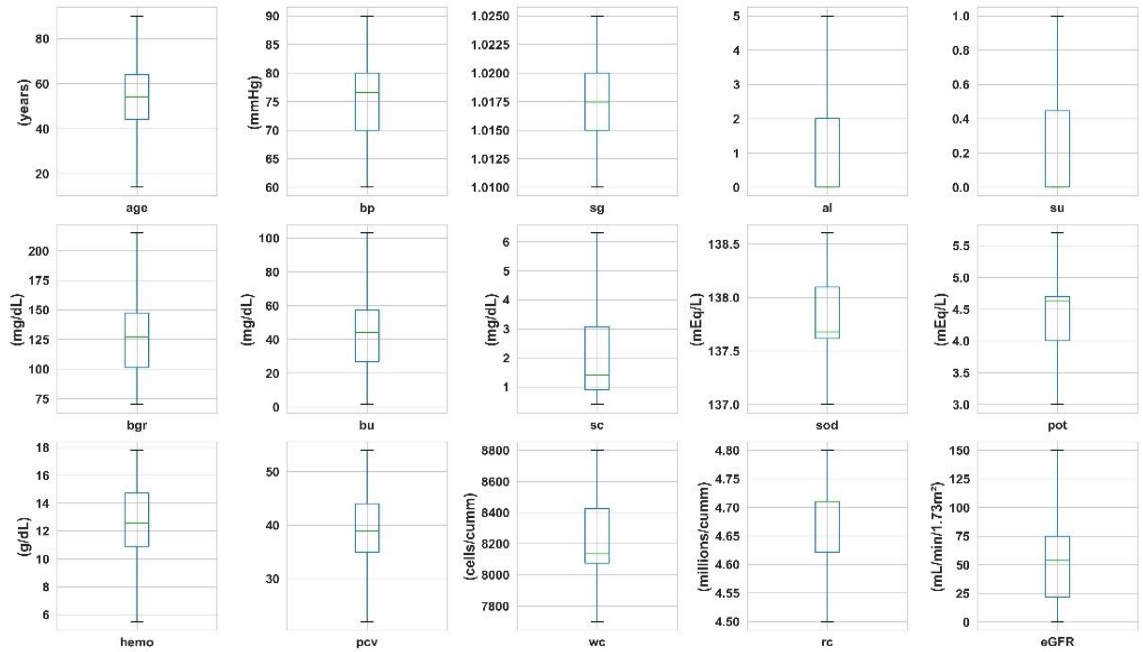

Figure 5. Distribution of numerical characteristics after using imputation methods

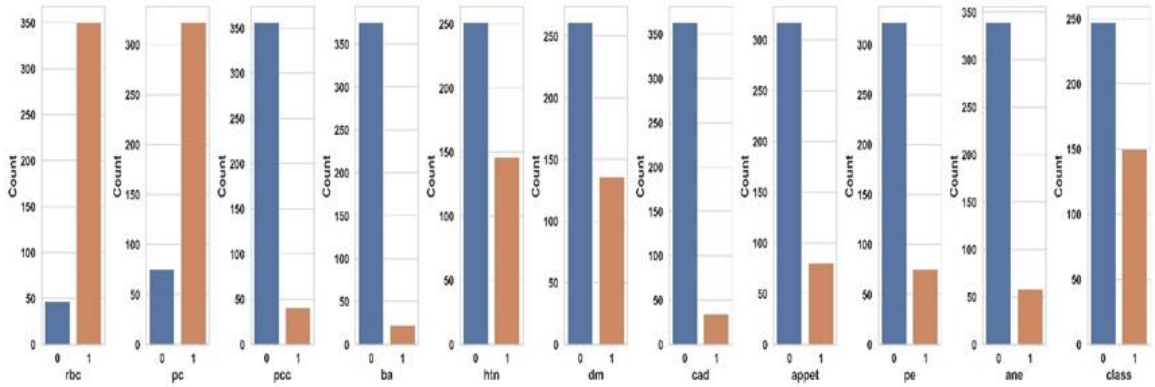

Figure 6. Distribution of categorical variables

3.4. Feature selection and dimensionality reduction methods

In machine learning, feature selection and dimensionality reduction play a crucial role in enhancing model accuracy and efficiency. Feature selection focuses on identifying and retaining only the most informative attributes, which helps to prevent overfitting, reduce execution time, and improve model interpretability. By emphasizing significant features and eliminating redundant ones, models become more robust and generalize better to unseen data. Dimensionality reduction, on the other hand, transforms the data into a smaller set of components while preserving its variability. This process simplifies the dataset and facilitates more efficient model training, without explicitly selecting the original features. In this study, we employed two complementary techniques:

3.4.1. PCA

PCA is a linear technique to reduce data dimensionality, ensuring the focus remains on the directions with the most variability [29]. The core of PCA revolves around computing the eigenvalues and eigenvectors of a data set's covariance matrix. Once these are determined, the eigenvectors are systematically arranged based on the descending sequence of their associated eigenvalues. With this order established, the data is projected onto these sorted eigenvector directions. The process starts by determining the covariance matrix for the signal samples, using both the signal matrix, which comprises M data points of N

dimensions, and its corresponding mean vector. After deriving the covariance matrix, one determines its eigenvectors and places the eigenvalues along the diagonal of another matrix. Once sorted by descending eigenvalue, the data undergoes a projection phase through a dot product operation with the sorted eigenvectors. Finally, only the leading principal components are selected to ensure optimal data representation, capturing a specified percentage of total variability, such as 95% or 98%. This methodology provides the retention of significant data patterns while minimizing dimensionality [29], [30].

3.4.2. LASSO

LASSO is a method that seeks to minimize the mean square error with a constraint on the sum of the absolute values of the regression coefficients [31]. It is particularly effective for feature selection as it can shrink the coefficients of less important features to zero, effectively eliminating them [32]. This unique capability of LASSO to select and regularize features allows for identifying the most critical features in a dataset. The reliability of feature selection can be enhanced through the randomized LASSO, which involves repeated applications to determine the most frequently occurring features indicative of their importance [33], [34].

3.5. Machine learning models

Four machine-learning models are applied in this work: Bagging classifier, RF, AdaBoost, and ANN.

3.5.1. Bagging classifier

Bagging is an ensemble technique that boosts classification algorithm performance [35]. It operates by training individual classifiers on random subsets of the dataset drawn with replacement [36]. These classifiers' predictions are then combined, typically through voting or averaging, to produce a unified outcome [35], [36]. The primary goal of Bagging is to diminish overfitting, counterbalancing any rise in bias with a decrease in variance [36]. This method leverages the strength of multiple models to produce a more robust and accurate prediction [37].

3.5.2. Random forest

RF is an ensemble method, which means that it brings together several algorithms to form an optimal model [33], [38]. More precisely, it comprises several DT built on randomly chosen and uncorrelated datasets during the training phase [39]. In the RF working process, we first select K data points from the training set and then create DT for these points [19], [40]. This procedure is repeated several times while deciding on a number N for the DT. When predicting new data, RF collects the predictions from each DT and assigns the new data to the category receiving the most votes [19], [40]. Although computational complexity may increase with RF compared to a standalone DT due to the use of multiple features, it generally exhibits better accuracy when facing previously unseen datasets [30]. In summary, RF leverages the training set to produce better results than an individual model by integrating predictions from multiple DT [33], [19], [40].

3.5.3. AdaBoost

AdaBoost is a Boosting algorithm designed for binary classification that amalgamates multiple weak classifiers to formulate a stronger and more robust classifier [33], [41]. Initiated with a predetermined weight on the training dataset instances, its intent is to enhance accuracy predictions based on a large number of samples, such as 1000 [33]. The AdaBoost technique used in certain studies seeks to establish a potent classifier through a training approach often referred to as weak learning [33], [35]. The essence of weak learning is to identify the optimal weak classifier that adeptly differentiates between positive and negative samples. During this process, the optimal threshold value for every feature is determined to minimize the misclassification of instances [35].

3.5.4. Artificial neural network

An ANN is a supervised learning technique composed of interconnected simple units called perceptrons. Structurally, an ANN typically consists of an input layer, which receives data, one or more hidden layers, which perform computations, and an output layer that delivers the final prediction [35], [42]. These perceptrons, when organized across multiple layers, form what is known as a multilayer perceptron (MLP). In an MLP, signals are transmitted from the input layer through to the hidden layer, with each perceptron having its unique weight. Every neuron in the input layer connects to every neuron in the hidden layer, and computations primarily take place in the hidden layer, where input weights are multiplied and processed [42]. The processed data is then conveyed to the output layer, which provides the final prediction. The efficiency of the model can be assessed by determining the error, which is the discrepancy between the actual and predicted outcomes [36].

4. RESULTS AND DISCUSSION

4.1. Identification and distribution of chronic kidney disease stages among diagnosed patients

In this section, we sought to identify and classify the stages of CKD among diagnosed patients using their eGFR. Based on current clinical guidelines, we categorized patients into one of six stages: CKD1 (eGFR ≥ 90), CKD2 (eGFR ≥ 60), CKD3a (eGFR ≥ 45), CKD3b (eGFR ≥ 30), CKD4 (eGFR ≥ 15), and CKD5 (eGFR < 15) [43]. The distribution of patients across these stages is depicted in the Figure 7. Importantly, our study was adept at early prediction, identifying a substantial number of patients in the preliminary stages of CKD, especially within the CKD1 and CKD2 stages. This early predictive capability is paramount as it allows for prompt intervention and tailored therapeutic approaches for these patients before the disease advances further. Our findings indicate that the CKD2 stage is the most prevalent with 110 patients, closely followed by the CKD1 stage with 69 patients. As illustrated in Figure 8, which shows the distribution of CKD stages for patients, some patients were initially not classified as having CKD, but after eGFR calculation and CKD staging, they were categorized into CKD1 and CKD2. This observation highlights that despite the absence of an initial CKD diagnosis, certain patients were identified as having CKD upon further analysis.

It is crucial to note that, even with this initial diagnosis, a significant number of patients are in the advanced stages of the disease, with 68 patients in CKD4 and 66 in CKD5. The CKD3b and CKD3a stages have 45 and 42 patients, respectively. These findings underscore the diversity of CKD stages within our patient cohort, the importance of accurate screening methods, and the need for stage-specific medical intervention to ensure optimal patient care.

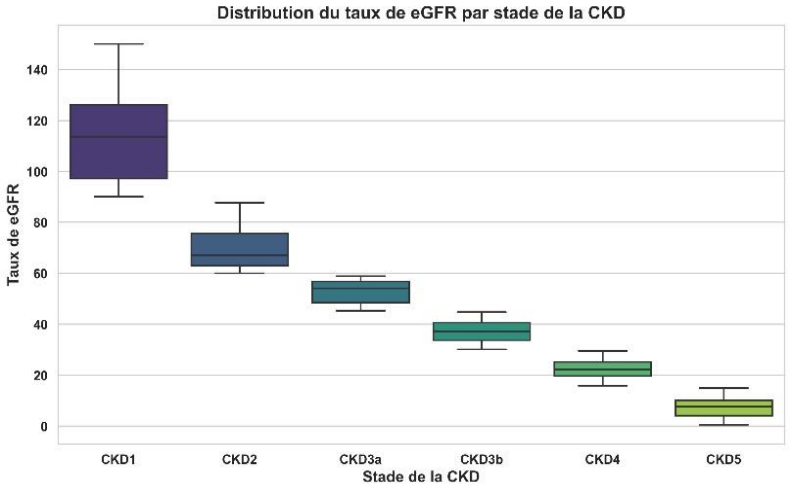

Figure 8. Distribution of eGFR rate by CKD stage

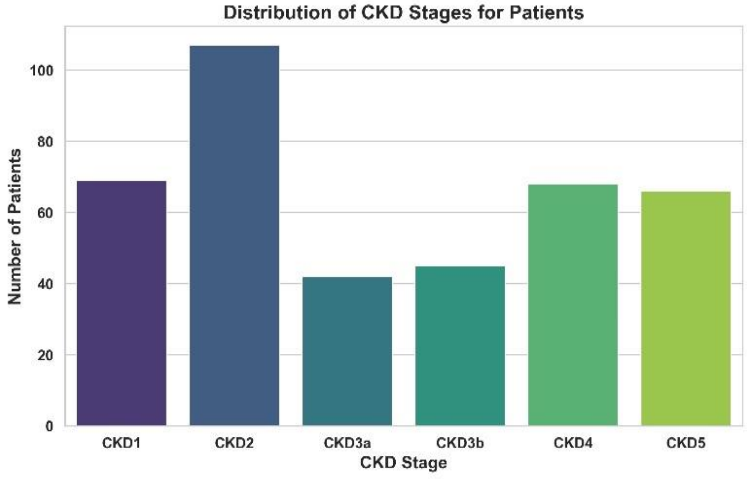

Figure 7. Distribution of CKD stages

4.2. Results of feature selection techniques

Figure 9 offers crucial insights into the most influential factors in CKD prediction. The specific gravity of urine, denoted as 'sg,' has a strong positive association with CKD, suggesting that its concentration is a notable indicator of kidney function. Albumin in urine, or 'al,' displays a negative relationship, emphasizing its role as a symptom of kidney damage when present in higher concentrations. Serum creatinine, 'sc,' another crucial marker, also shows a negative association, underscoring its well-established connection with kidney dysfunction. Hemoglobin levels and packed cell volume, 'hemo' and 'pcv' respectively, have positive coefficients, indicating their significance in CKD, particularly as markers of anemia, a condition commonly associated with kidney disease. Hypertension, 'htn,' and diabetes mellitus, 'dm,' both show negative coefficients, reinforcing their adverse effects on kidney health, with diabetes being a leading cause of CKD globally. Appetite changes, represented by 'appet,' correlate negatively with CKD progression. Finally, the 'eGFR' showcases a positive relationship, highlighting its value in representing kidney function for CKD prediction. These features and their associated coefficients provide a comprehensive understanding of the factors influencing CKD prognosis.

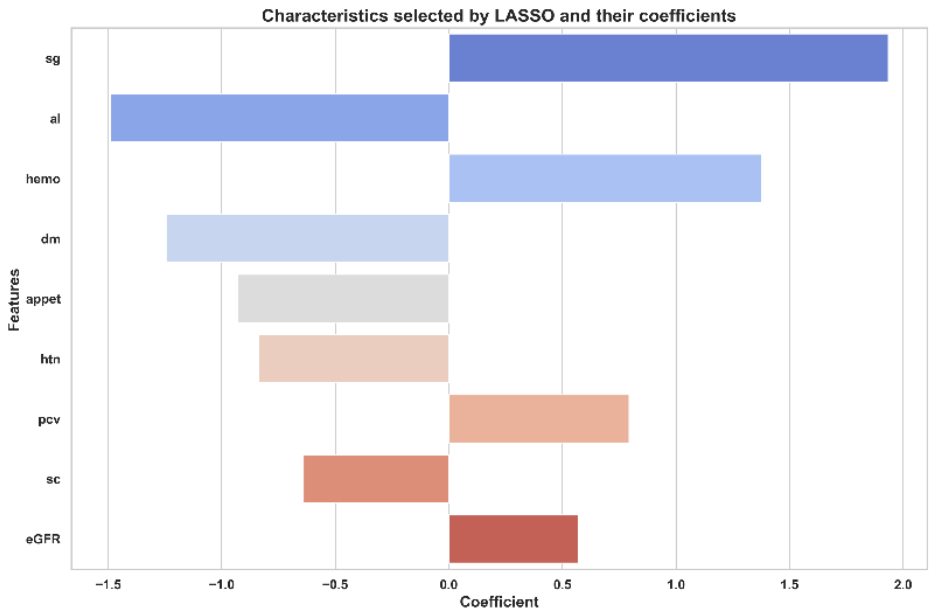

Figure 9. Features selected by LASSO

The PCA results provide valuable insights into the relationships between the principal components (PCs) and the original features. PC-1 exhibits a strong positive correlation with 'sc' (serum creatinine) and a negative correlation with 'hemo' (hemoglobin), implying its sensitivity to changes in serum creatinine levels and susceptibility to fluctuations in hemoglobin levels. PC-2 prominently aligns with 'sod' (sodium) and inversely associates with 'rc' (red blood cell count), indicating its significant dependence on sodium levels and responsiveness to fluctuations in red blood cell counts. PC-3 demonstrates a predominant positive link with 'wc' (white blood cell count) and an adverse connection with 'pot' (potassium), highlighting its dependence on shifts in white blood cell counts and vulnerability to potassium level variations. PC-4 is positively influenced by 'bu' (blood urea) and negatively affected by 'bp' (systolic blood pressure), emphasizing its responsiveness to variations in blood urea levels and sensitivity to changes in systolic blood pressure. PC-5 displays a strong positive association with 'pot' and a negative association with 'age,' indicating its sensitivity to potassium level variations and its vulnerability to age-related changes. PC-6 is characterized by a strong positive correlation with 'age' and a negative correlation with 'bp' (blood pressure), illustrating its dependence on age-related variations and susceptibility to fluctuations in blood pressure. PC-7 highlights its responsiveness to changes in blood glucose levels and vulnerability to fluctuations in white blood cell counts. PC-8 is positively influenced by 'age' and negatively affected by 'sod' (sodium), emphasizing its dependence on age-related variations and sensitivity to changes in sodium levels. PC-9 primarily correlates positively with 'al' (albumin) and negatively with 'bgr,' indicating its responsiveness to variations in albumin levels and vulnerability to changes in blood glucose levels. PC-10 is positively influenced by 'sg' (urine specific gravity) and negatively affected by 'sod,' highlighting its dependence on

variations in urine specific gravity and sensitivity to changes in sodium levels. PC-11 exhibits a strong positive correlation with 'bu' and a negative correlation with 'rc,' emphasizing its responsiveness to changes in blood urea levels and vulnerability to fluctuations in red blood cell counts. PC-12 is mainly positively influenced by 'eGFR' and negatively affected by 'sc,' indicating its responsiveness to changes in eGFR and sensitivity to fluctuations in serum creatinine levels. PC-13 primarily correlates positively with 'pcv' (packed cell volume) and negatively with 'hemo,' illustrating its dependence on variations in packed cell volume and sensitivity to fluctuations in hemoglobin levels. PC-14 shows a strong positive association with 'sc' and a negative association with 'dm' (diabetes mellitus), emphasizing its responsiveness to changes in serum creatinine levels and sensitivity to the presence of diabetes mellitus. PC-15 is positively influenced by 'sc' and negatively affected by 'pcv,' highlighting its dependence on variations in serum creatinine levels and sensitivity to changes in packed cell volume. The following Figure 10 provides a visualization of each principal component's contribution to the total variance in the data and shows how the total variance accumulates as additional principal components are included in the PCA method.

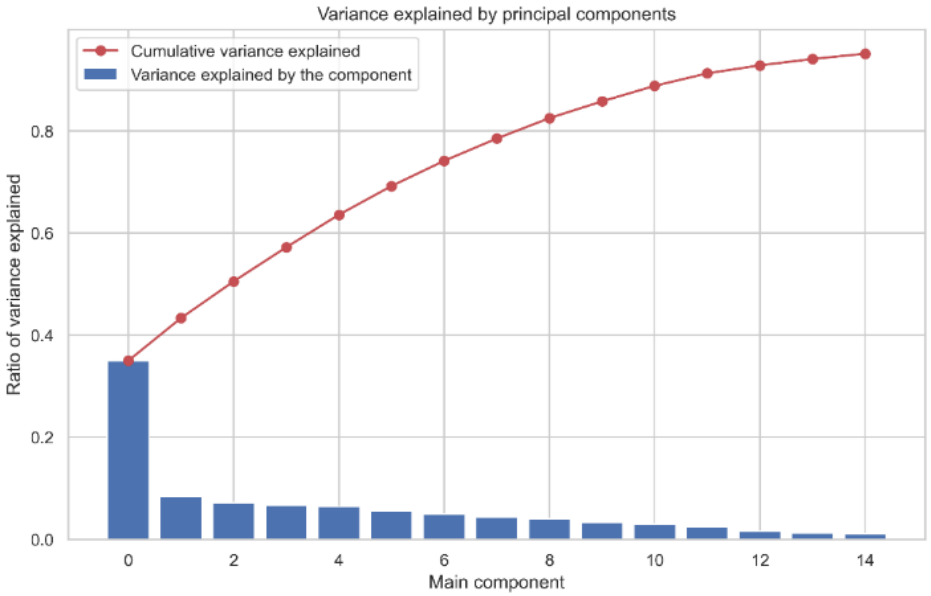

Figure 10. Variance explained by principal components

4.3. Comparison of results for each model of dataset 1

We further emphasize the rigorous assessment of our models by incorporating cross-validation with 5-fold validation, a widely accepted practice for robust model evaluation. To rigorously assess the performance of our models, we utilized several key metrics: accuracy, precision, recall, and the F1-score. Accuracy provides a general measure of the overall correctness of the model, indicating the proportion of all predictions that were correct. Precision gives insight into how many of the positive identifications were actually correct, making it crucial when the cost of a false positive is high. Recall, on the other hand, tells us about the model's ability to identify all relevant instances, proving vital when the cost of a false negative is significant. The F1-score harmoniously combines both precision and recall into a single metric, providing a balanced measure, especially useful for datasets with uneven class distributions. It's crucial to highlight that these evaluations were performed on the transformed dataset using the PCA method. This dimensionality reduction technique ensures that only the most significant features are considered, potentially enhancing the model's performance while reducing computational costs. In addition to PCA, we also employed the LASSO method for feature selection. LASSO is particularly effective in reducing model complexity by selecting the most relevant features while simultaneously regularizing the model, helping to avoid overfitting and improving generalization on new data.

4.3.1. Comparison between different models based on accuracy using PCA

Accuracy is a cornerstone metric in the machine-learning domain, providing a panoramic view of the model's overall efficiency by representing the ratio of accurate predictions to the total predictions made.

When it comes to predicting CKD, this metric assumes paramount significance. The ability to pinpoint correct diagnoses has a reverberating impact on patient outcomes, shaping medical interventions and treatment pathways. Among the various models evaluated, RF emerged as the unequivocal frontrunner, charting an impeccable accuracy of 98.3%. This metric speaks volumes about the model's discerning prowess in navigating the intricacies of the dataset and delivering near-perfect predictions.

The AdaBoost algorithm delivered an impressive performance with an accuracy rate of 99.1%, reinforcing its robustness and reliability in CKD detection. The Bagging model also demonstrated its capabilities, achieving an accuracy of 97.5%. While this figure is slightly lower than that of the RF and AdaBoost models, it still highlights the model's proficiency in CKD detection. The ANN model, designed to replicate neural pathways and processes, recorded an accuracy of 96.67%. Although this is a commendable score, it also indicates potential areas for improvement that, if addressed, could further enhance the model's predictive accuracy.

Our experimental exploration of hybrid models yielded the ANN_AdaBoost model, which is a fusion of the intricate neural pathways of ANN and the algorithms of AdaBoost, achieving an accuracy of 99.5%. This score not only validates the effectiveness of combining different algorithmic strengths but also suggests opportunities for further refinement and enhancement, as detailed in Figure 11.

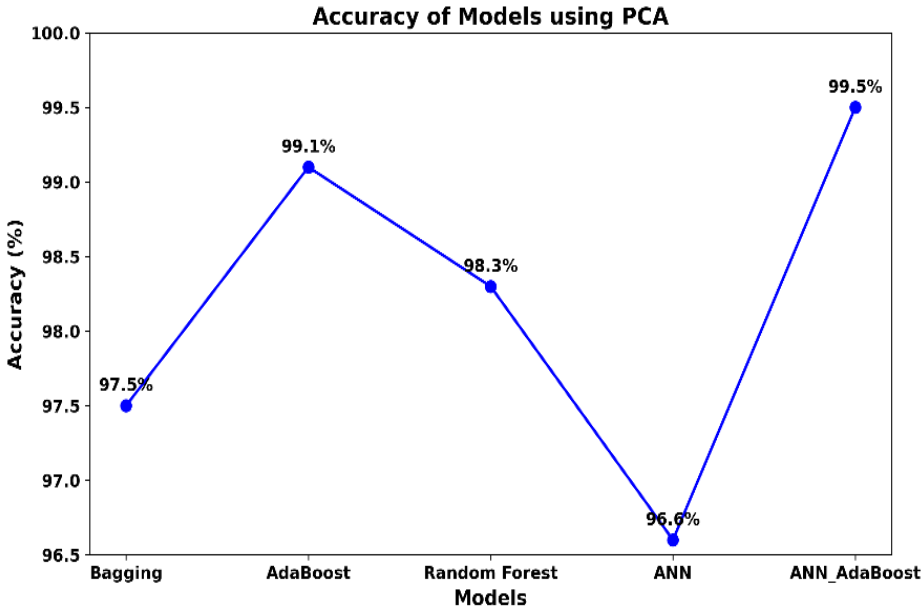

Figure 11. Results of accuracy scores using PCA

4.3.2. Comparison between different models based on precision using PCA

Precision stands as an instrumental metric in machine learning, underscoring the credibility of positive predictions made by a model. In the medical landscape, the gravity of precision is accentuated. A high precision implies that a model has a minimal tendency to raise false alarms incorrectly diagnosing a patient with CKD when they are, in fact, unaffected. Such misdiagnoses can have severe implications, making the precision metric indispensable.

Diving into our diverse range of models, the RF algorithm rose as the gold standard in terms of precision, achieving a near-perfect score of 98.3%. This stellar result echoes the model's prowess when it flags a case as CKD, the diagnosis is very likely spot-on. Close on its trail, the AdaBoost model demonstrated precision at 97.9%, reflecting its acute discernment and robustness in positively identifying CKD cases. The Bagging model also carved its niche, registering a precision of 97.7%, indicating its steadfastness in CKD detection. The ANN model, with its intricate architecture and algorithms, clinched a precision score of 96%. While commendable, this score simultaneously hints at potential avenues to refine and enhance its predictive accuracy further. Venturing into the domain of hybrid models, our ANN_AdaBoost model unveiled its formidable capabilities. Achieving a precision of 97.8%, this model stands as a testament to the synergistic amalgamation of the intricacies of neural networks and the robustness of boosting mechanisms. This score, especially when juxtaposed with other models, underscores the model's consistent and reliable performance. The outcomes for precision are illustrated in Figure 12.

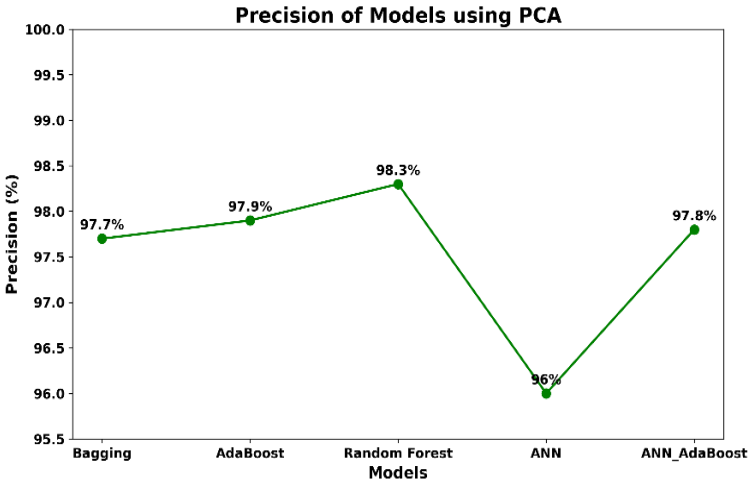

Figure 12. Results of precision scores using PCA

4.3.3. Comparison between different models based on recall using PCA

Recall, within the medical diagnostic sphere, stands as a defining metric, illuminating a model's capability to accurately recognize every case of a specific medical condition in this study, CKD. The implications of recall are profound, especially in the medical landscape. A missed true positive CKD case, resulting from lower recall, can lead to overlooked treatments and escalated medical complications, amplifying the repercussions of inaccurate model predictions. Among the ensemble of machine learning models we evaluated, AdaBoost emerged as the paragon of precision, achieving a perfect recall score of 100%. This exceptional performance reaffirms the algorithm's unparalleled prowess in CKD detection, ensuring that every CKD case was accurately flagged, leaving no patient undiagnosed.

The RF model showcased exemplary proficiency with a recall score of 98.3%. The architectural foundation of RF a compilation of numerous DTs enhances its ability to meticulously sift through data, ensuring comprehensive CKD detection and making instances of overlooked CKD patients exceedingly rare. Our experiment with hybrid models yielded the ANN_AdaBoost model, which performed impressively, achieving a recall rate of 97.8%, as illustrated in Figure 13. This score underscores the hybrid model's adeptness in CKD detection, drawing from the strengths of both ANN and AdaBoost. The ANN model, despite its intricate neural pathways, achieved a recall of 96%. While commendable, this metric highlights potential areas for improvement, suggesting that refinements could enhance the model's CKD detection efficiency. However, the Bagging model should not be overlooked. With a recall of 95.7%, it firmly positions itself among the top performers, further underscoring the power and potential of ensemble techniques in medical diagnostics. In essence, while each model exhibited its strengths and demonstrated commendable recall metrics, the collective results illuminate the monumental potential of machine learning in revolutionizing CKD diagnostics.

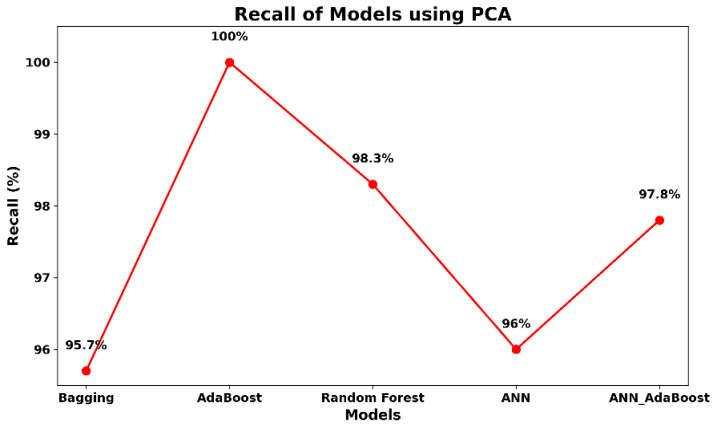

Figure 13. Results of recall scores using PCA

4.3.4. Comparison between different models based on F1-score using PCA

The F1-score emerges as a crucial metric, particularly in critical scenarios where the repercussions of misclassifications can be profound. It provides a harmonized understanding of a model's performance by encapsulating the combined potency of precision and recall, making it especially significant in datasets where class imbalances could skew evaluations. The RF model, underpinned by its intricate ensemble mechanism, exemplifies this balance with an impressive F1-score of 98.3%. This remarkable score attests to the model's unmatched ability to strike an optimal balance between precision and recall, marking it as a gold standard in CKD prediction. Close behind, the AdaBoost algorithm demonstrated robust capabilities with an F1-score of 98.9%, highlighting its efficacy in harmonizing the strengths of both precision and recall, thereby minimizing both false positives and negatives. Turning to hybrid models, our ANN_AdaBoost creation an intricate blend of neural architectures and boosting techniques achieved an F1-score of 97.8%, as shown in Figure 14. This score reflects the model's stellar performance in CKD prediction, showcasing its proficiency in integrating the complexities of neural processes with the robustness of boosting mechanisms to yield reliable predictions. The ANN model, with its unique neural pathways, achieved an F1-score of 96%. While this score establishes the model as a reliable tool for CKD predictions, it also suggests areas for improvement to ensure a more harmonious balance between precision and recall in future iterations. However, it's important not to overlook the Bagging model, which showcased a formidable F1-score of 96.7%. This score underscores the model's consistency and adeptness in delivering balanced CKD diagnostic outputs.

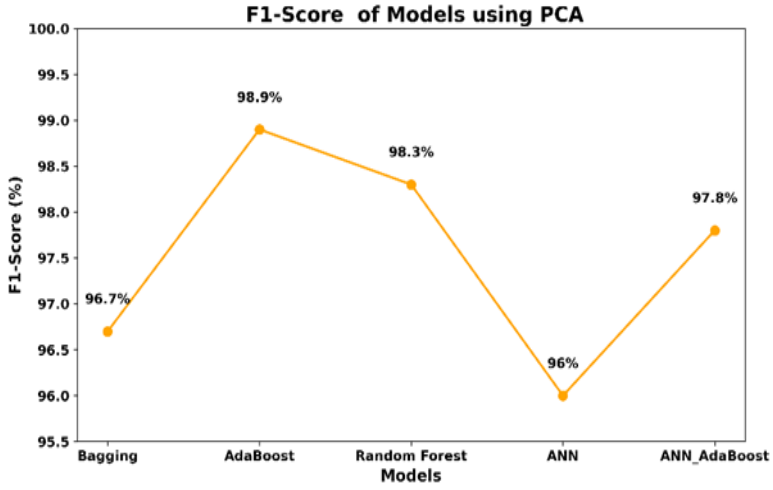

Figure 14. Results of F1-score scores using PCA

4.3.5. Hybrid model's AUC analysis in CKD detection using PCA

The area under the curve (AUC) represents a crucial metric in evaluating the efficacy of a classifier, especially in binary classification tasks. As illustrated in Figure 15, an AUC of 0.999, as achieved by our hybrid model, is particularly striking. It suggests a near-perfect ability of the model to discern between the two classes, in this case, distinguishing those with CKD from those without. In more tangible terms, the high AUC value indicates that the model is 99.9% confident in its classifications, an exceptionally rare and commendable feat in predictive modeling. In the spectrum of AUC values, a score of 0.5 suggests no discriminative power akin to a random guess. On the other hand, a perfect AUC of 1.0 denotes that the model has an impeccable discriminative capability. Our model's AUC of 0.999 positions it very closes to this pinnacle of perfection. Such an outstanding result underscores the model's robustness, reliability, and heightened sensitivity and specificity in identifying CKD cases. While this metric alone paints a compelling picture of the model's prowess, it is essential to recognize the broader implications. In medical diagnostics, where the stakes are inherently high, misclassification can have profound consequences. The exceptional AUC value provides assurance that the hybrid model can be entrusted with the critical task of CKD detection, reducing the chances of diagnostic errors to a minimum.

4.3.6. Comparison between different models based on accuracy using LASSO

In the pursuit of accurately predicting CKD using select features derived via the LASSO method-sg, al, sc, hemo, pcv, htn, dm, appet, and eGFR several models were rigorously evaluated. The RF model stood out distinctly, achieving an outstanding accuracy of 98.3%, attesting to its adeptness at effectively processing

and interpreting the intricacies of the selected features. Following closely, the AdaBoost model achieved an accuracy of 96.6%, underlining its robust capability in detecting CKD with commendable reliability. Notably, as illustrated in Figure 16, the Bagging model demonstrated a notable accuracy of 95.8%, further asserting its proficiency as a formidable tool in the CKD diagnostic arsenal. The ANN model, inspired by its intricate neural network architecture, posted an accuracy of 96.67%. While commendable, this figure suggests potential avenues for refinement to harness even greater predictive precision. Interestingly, our exploration with the hybrid ANN_AdaBoost model showcased a synergistic melding of the two algorithms, yielding an accuracy of 96.6%. This result highlights not only the merit of integrating diverse algorithms but also the room for further fine-tuning to elevate this hybrid model's performance.

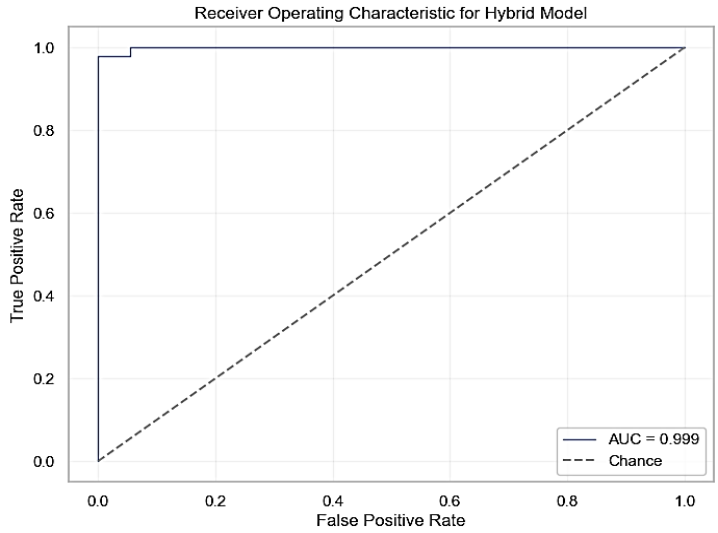

Figure 15. ROC analysis of the hybrid model for CKD prediction

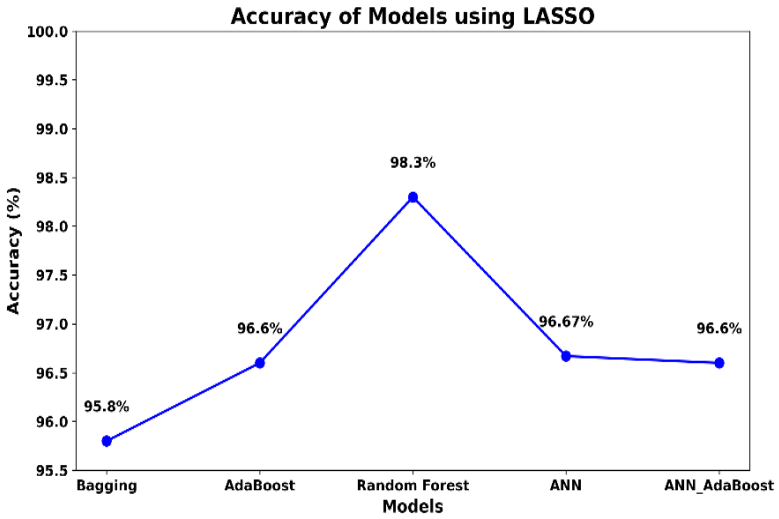

Figure 16. Results of accuracy scores using LASSO

4.3.7. Comparison between different models based on precision using LASSO

In predictive modeling, precision elucidates the trustworthiness of identifications, making it paramount when the implications of false positives are significant. The RF model stands out in predicting CKD using the nine selected features, achieving a flawless precision score of 100%. This impeccable score underscores the model's adeptness in identifying true CKD cases while minimizing false alerts. Following closely is the ANN model, with a precision of 97%. While it hasn't reached the perfection of RF, it still demonstrates commendable discernment in its predictions. The Bagging and ANN_AdaBoost models

showcased precision scores of 95.8% and 92%, respectively, reinforcing their capacity to make largely accurate positive predictions. AdaBoost, with a precision of 92%, reflects strong performance but also hints at areas for refinement to further reduce false positives. In an era where precision can directly influence medical decisions and patient care pathways, these findings underscore the importance of selecting the right predictive model. The detailed precision scores offer valuable insights for clinicians, researchers, and data scientists striving to improve CKD prediction algorithms. Figure 17, which displays the precision scores using LASSO, illustrates these results.

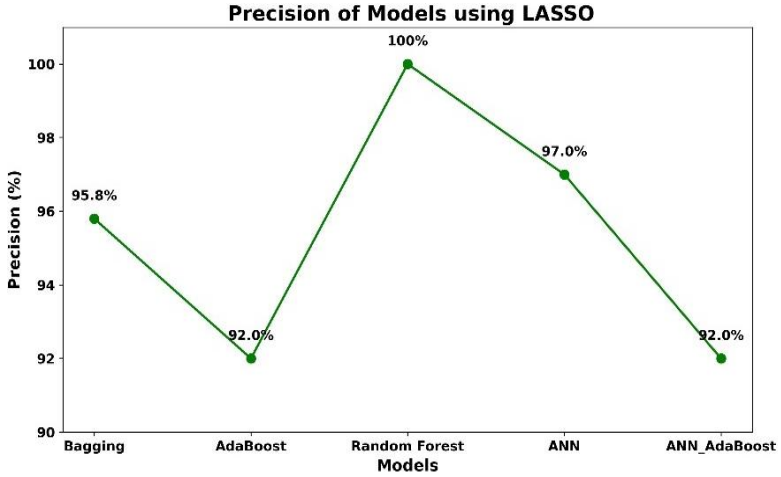

Figure 17. Results of precision scores using LASSO

4.3.8. Comparison between different models based on recall using LASSO

Recall, also known as sensitivity, is a pivotal metric in machine learning, particularly in medical diagnostics like predicting CKD. It reflects the model's ability to correctly identify all relevant cases, making it crucial to capture all actual CKD cases while minimizing missed diagnoses. Using the nine features selected by LASSO, both AdaBoost and ANN_AdaBoost demonstrated exceptional proficiency, each achieving a perfect recall score of 100%. This flawless performance underscores their ability to detect every instance of CKD without omission. Close behind, the ANN model achieved a commendable recall of 97%, showcasing its strong capability in reducing false negatives. RF, a model known for its robustness, reached a recall rate of 95.65%, further emphasizing its effectiveness in this critical diagnostic task. Meanwhile, the Bagging model, with a recall of 93.47%, demonstrated potential in CKD prediction, although there is room for further improvement. In CKD prediction, where missed diagnoses can have significant implications, these results highlight the importance of selecting a model with high recall. The outlined recall rates offer valuable insights for healthcare professionals and data practitioners striving to enhance CKD diagnostic tools. Figure 18 illustrates these findings.

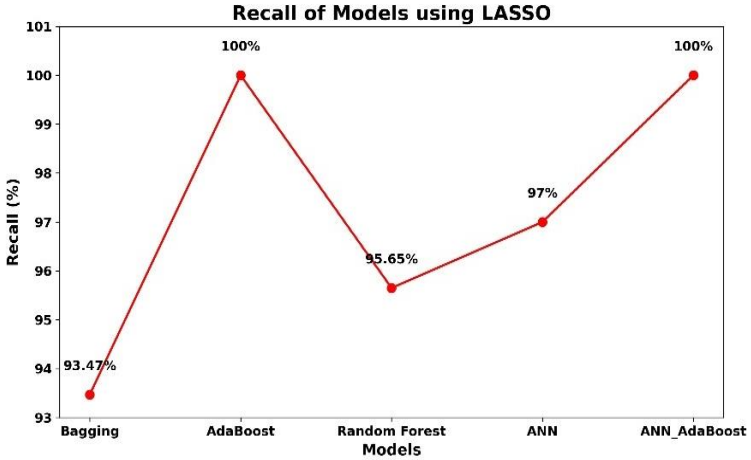

Figure 18. Results of recall scores using LASSO

4.3.9. Comparison between different models based on F1-score using LASSO

The F1-score, a harmonic mean of precision and recall, provides a holistic measure of a model's performance, effectively balancing false positives and false negatives. This metric is especially pertinent in medical diagnostics, where both over-diagnosis and under-diagnosis can have significant consequences. Utilizing the selected features: sg, al, sc, hemo, pcv, htn, dm, appet, and eGFR for CKD prediction, the RF model stood out, achieving an impressive F1-score of 97.7%. This score highlights its optimal blend of precision and sensitivity in detecting CKD. Close behind, the ANN model recorded an F1-score of 97%, affirming its balanced capability in CKD detection while maintaining a minimal margin of error. The AdaBoost and ANN_AdaBoost models both posted F1-scores of 95.8%, further validating their robustness in harmonizing precision and recall, which is crucial in medical diagnostics. Meanwhile, the Bagging model, with its F1-score of 94.5%, still demonstrates promise in CKD prediction, indicating potential for further refinement to enhance its performance. As efforts continue to perfect CKD diagnostic tools, insights into F1-scores are indispensable. These scores underscore the critical importance of balancing precision and recall, guiding healthcare professionals and researchers in refining their diagnostic algorithms, as illustrated in the following Figure 19.

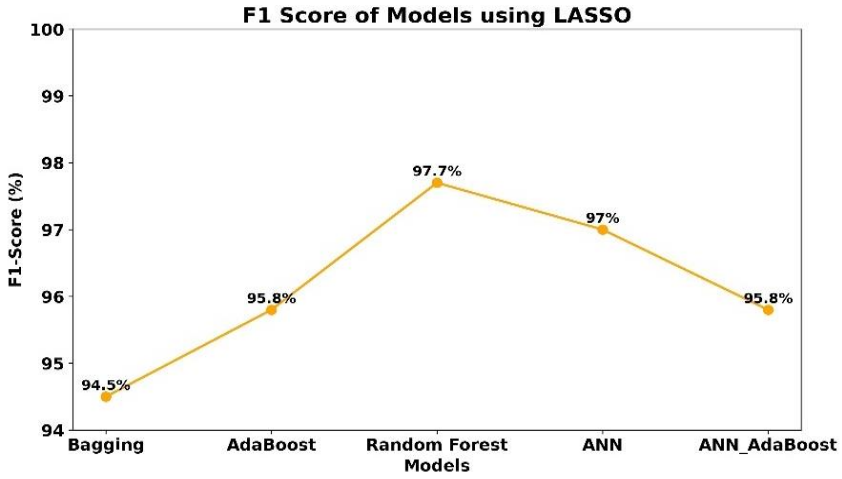

Figure 19. Results of F1-score scores using LASSO

4.3.10. Comparative analysis of accuracy with LASSO and PCA feature selection

In the intricate realm of CKD prediction, feature selection is pivotal in enhancing machine learning models' performance. LASSO and PCA, two distinct methodologies, cater to this by focusing on feature regularization and dimensionality reduction. Using the LASSO methodology, the RF model tops the list, posting an impressive accuracy of 98.3%. However, when adopting the PCA approach for feature selection, its performance remains at the same level, registering an accuracy of 98.3%. The consistency suggests that PCA did not offer any additional advantage for the RF model in this case.

Similarly, the AdaBoost model experienced a significant boost when transitioning from LASSO (96.6% accuracy) to PCA (99.1% accuracy). This uptick accentuates PCA's prowess in underscoring salient features that align with AdaBoost algorithmic structure. Conversely, the Bagging model, with accuracies of 95.8% (LASSO) and 97.5% (PCA), and the ANN model, consistently marking accuracy of 96.67% across both methodologies, elucidate the nuanced interplay between feature selection techniques and underlying model architectures. As depicted in Figure 20, the ANN_AdaBoost hybrid model, a fusion of neural networks and boosting, portrayed slight variability. While the LASSO approach yielded an accuracy of 96.6%, the PCA technique enhanced its performance, taking it to 99.5%. This improvement showcases the effective synergy between PCA-based feature selection and the intricacies of the hybrid model structure.

4.4. Performance comparison of models for the test dataset

In our pursuit of achieving precise CKD predictions on the test dataset, various models were evaluated, as depicted in Figure 21. The AdaBoost model emerged as the most accurate, attaining an impressive accuracy of 99.3%, showcasing its exceptional ability to handle this dataset effectively. Following this, the ANN_AdaBoost hybrid model demonstrated an equally strong performance, achieving the highest accuracy of 99.6%. The RF model, known for its robustness, achieved a commendable accuracy of 98.3%,

reinforcing its reliability in accurately predicting CKD. Similarly, the Bagging model demonstrated a strong performance with an accuracy of 98%, reaffirming its effectiveness in ensemble learning for this task. Finally, the ANN, despite its intricate architecture, yielded a slightly lower accuracy of 97.1%, but still highlighted its capacity to model complex relationships in the data. A key strength of our approach lies in the use of a distinct second dataset for validation, enhancing the generalizability and robustness of our results. This dataset allowed for a more comprehensive evaluation of our models' performance, ensuring that our results are not overly optimized for a single dataset, and providing a more reliable assessment of how these models perform in diverse real-world scenarios. This approach of testing on a separate dataset confirms the broader applicability of our findings and the models' predictive power beyond the initial data.

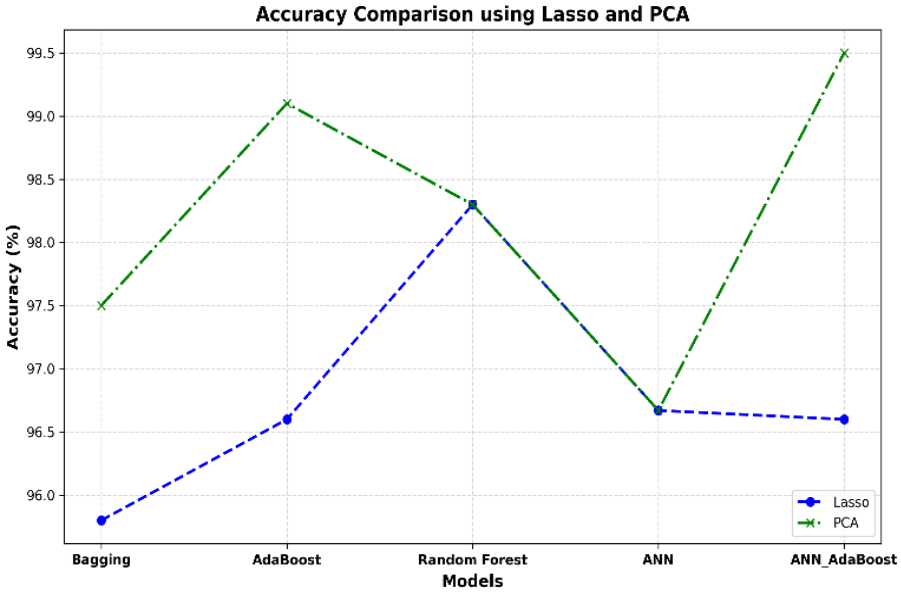

Figure 20. Results accuracy models with LASSO and PCA

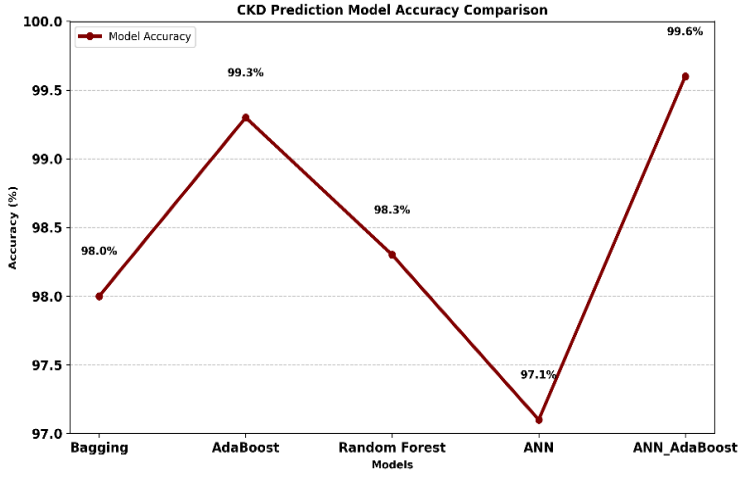

Figure 21. The results accuracy of our models on the test dataset

4.5. Comparison table between the accuracy of the proposed models and existing techniques

Our study explored two datasets for training, validation, and testing for CKD. In working with real-world data, it is common to encounter gaps for various reasons, such as errors, difficulties in collecting information, or incomplete extractions. Managing these missing values is crucial for building robust data models. A common method for this is data imputation, where missing values are replaced with simple statistics, such as the mean or mode of the respective column. This approach is effective as it is easy to

implement and often yields satisfactory results. However, more advanced approach involving imputation methods has recently been proposed, showing a significant improvement in accuracy compared to previous methods. Unlike most studies in this field, such as those conducted by Xioa *et al.* [14], Hassan *et al.* [15], Baidya *et al.* [16], and Chittora *et al.* [17], the authors of this article did not employ the approach of replacing missing data with the mean or median for all features. Instead, advanced imputation techniques were used, which resulted in a well-prepared dataset before applying machine learning algorithms. Additionally, unlike these studies, we used another dataset to validate the results of our research. Our proposed models demonstrated superior accuracies across multiple metrics compared to previous studies. While prior research by Xioa *et al.* [14] reported accuracies of 80% for both RF and ANN models, our AdaBoost model achieved an accuracy of 97.5%, showcasing a significant enhancement. Similarly, the RF model in our study surpassed the performance reported by Yashfi *et al.* [18], achieving an accuracy of 99.1% compared to their reported 97.12%. Additionally, our Bagging model outperformed the results of Rahman *et al.* [20], achieving an accuracy of 98.3% compared to their reported 98.07%.

Furthermore, our ANN model demonstrated notable improvement over the findings of Pal [19], with an accuracy of 96.6% compared to their reported 89%. Lastly, our ANN_AdaBoost hybrid model achieved an impressive accuracy of 99.5%, surpassing the accuracies reported in previous studies. Overall, our models exhibited significant advancements in accuracy compared to existing research, validating the efficacy of our approach in predicting CKD. As illustrated in Table 4 below, the results of our classifiers demonstrate accuracies that surpass those reported in other studies, underscoring the robustness of our approach.

Table 4. Comparing the accuracy of our study with existing work

Works	Models	Accuracy	Dataset 2 for validating results	Feature eGFR
Xioa *et al.* [14]	RF	80%	_	_
	ANN	80%		
Hassan *et al.* [15]	ANN	97.5%	_	_
	RF	98.75%		
Islam *et al.* [16]	AdaBoost	94%	_	_
	RF	95%		
Chittora *et al.* [17]	ANN	90.2%	_	_
Yashfi *et al.* [18]	ANN	94.5%	_	_
	RF	97.12%		
Pal [19]	RF	93%	_	_
	ANN	89%		
Rahman *et al.* [20]	RF	96.15%	yes	_
	Bagging	98.07%		
	AdaBoost	100%		
Ghosh *et al.* [21]	AdaBoost	97.91%		
Ebiaredoh-Mienye *et al.* [23]	AdaBoost	93%		
	RF	95.2%		
Islam *et al.* [42]	AdaBoost	98.3%	_	_
	RF	97.5%		
	ANN	60%		
Our proposition	AdaBoost	97.5%	yes	yes
	RF	99.1%		
	Bagging	98.3%		
	ANN	96.6%		
	ANN_AdaBoost	99.5%		

5. CONCLUSION

In this research, we pioneered an innovative approach for predicting and diagnosing CKD through the application of advanced machine learning and deep learning models. The linchpin of our methodology was the calculation of eGFR using the MDRD formula. a well-established equation for estimating kidney function. Our study conducted an in-depth analysis of CKD risk factors, emphasizing the critical role of eGFR as a key biomarker for both disease diagnosis and staging. This enhancement allows clinicians to develop more effective therapeutic strategies, ultimately improving patient outcomes.

To enhance predictive accuracy, we employed feature selection techniques, specifically LASSO, and dimensionality reduction through PCA, ensuring that our model focused on the most relevant predictors. Among the models evaluated, the hybrid ANN_AdaBoost demonstrated outstanding performance, achieving an accuracy of 99.5% when using PCA. These promising results highlight the potential of our approach in advancing CKD diagnosis and staging.

However, this study is not without limitations. The dataset used was relatively small, comprising only 400 samples, which may impact the generalizability of our findings. A larger and more representative

dataset would provide a more robust and comprehensive analysis. As part of our future work, we aim to enhance both the quantity and quality of the dataset, just as we improved diagnostic accuracy by incorporating eGFR. By expanding our dataset and refining our predictive models, we seek to further advance early CKD detection, enabling more precise and reliable clinical decision-making.

ACKNOWLEDGMENTS

We extend our sincere gratitude to the Algerian General Directorate of Research (DGRSTD) and the Laboratory of Mathematics, Informatics, and Systems (LAMIS) at Echahid Cheikh Larbi Tebessi University of Tébessa for their invaluable support in providing resources and facilitating this research. We also wish to express our heartfelt thanks to our families for their unwavering support and encouragement throughout this work.

FUNDING INFORMATION

This research did not receive any financial support from public, commercial, or non-profit funding agencies.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Safa Boughougal	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Mohamed Ridda Laouar	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Abderrahim Siam		✓		✓	✓	✓	✓				✓	✓	✓	✓
Sean Eom		✓		✓	✓	✓	✓		✓	✓	✓	✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease. This dataset was originally collected and made publicly available by UCI and has been used for the purposes of this study.

REFERENCES

[1] Q. Yang *et al.*, "Efficacy and safety of drugs for people with type 2 diabetes mellitus and chronic kidney disease on kidney and cardiovascular outcomes: a systematic review and network meta-analysis of randomized controlled trials," *Diabetes Research and Clinical Practice*, vol. 198, p. 110592, Apr. 2023, doi: 10.1016/j.diabres.2023.110592.

[2] H. Jiang, C. Xia, J. Lin, H. A. Garalleh, A. Alalawi, and A. Pugazhendhi, "Carbon nanomaterials: A growing tool for the diagnosis and treatment of diabetes mellitus," *Environmental Research*, vol. 221, p. 115250, Mar. 2023, doi: 10.1016/j.envres.2023.115250.

[3] E. Kelepouris, W. St. Peter, J. J. Neumiller, and E. E. Wright, "Optimizing multidisciplinary care of patients with chronic kidney disease and Type 2 diabetes mellitus," *Diabetes Therapy*, vol. 14, no. 7, pp. 1111–1136, Jul. 2023, doi: 10.1007/s13300-023-01416-2.

[4] J. M. Krzesinski and A. J. Scheen, "La maladie rénale diabétique: prise en charge actuelle et perspectives d'avenir," *Revue Médicale Suisse*, vol. 11, no. 483, pp. 1534–1542, 2015.

[5] N. Borisagar, D. Barad, and P. Raval, "Chronic kidney disease prediction using back propagation neural network algorithm," In *Proceedings of International Conference on Communication and Networks: ComNet 2017*, pp. 295–303, doi: 10.1007/978-981-10-2750-5_31.

[6] N. A. ElSayed *et al.*, "11. chronic kidney disease and risk management: standards of care in diabetes—2023," *Diabetes Care*, vol. 46, no. Supplement_1, pp. S191–S202, Jan. 2023, doi: 10.2337/dc23-S011.

[7] A. Malkina, "Maladie rénale chronique," 2022. https://www.msdmanuals.com/fr/professional/troubles-génito-urinaires/maladie-rénale-chronique/maladie-rénale-chronique

[8] 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. 100189, 2023, doi: 10.1016/j.jpi.2023.100189.
[9] A. Nishanth and T. Thiruvaran, "Identifying important attributes for early detection of chronic kidney disease," *IEEE Reviews in Biomedical Engineering*, vol. 11, pp. 208–216, 2018, doi: 10.1109/RBME.2017.2787480.
[10] S. Samet, M. R. Laouar, and I. Bendib, "Use of machine learning techniques to predict diabetes at an early stage," in *2021 International Conference on Networking and Advanced Systems (ICNAS)*, Oct. 2021, pp. 1–6, doi: 10.1109/ICNAS53565.2021.9628903.
[11] A. Khediri and M. R. Laouar, "Deep-belief network based prediction model for power outage in smart grid," in *Proceedings of the 4th ACM International Conference of Computing for Engineering and Sciences*, Jul. 2018, pp. 1–6, doi: 10.1145/3213187.3287611.
[12] J. Mishra and S. Tarar, "Chronic disease prediction using deep learning," *In Advances in Computing and Data Sciences: 4th International Conference, ICACDS 2020, Valletta, Malta*, 2020, pp. 201–211, doi: 10.1007/978-981-15-6634-9_19.
[13] R. Ani, G. Sasi, U. R. Sankar, and O. S. Deepa, "Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sep. 2016, pp. 1287–1292, doi: 10.1109/ICACCI.2016.7732224.
[14] J. Xiao *et al.*, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," *Journal of Translational Medicine*, vol. 17, no. 1, p. 119, Dec. 2019, doi: 10.1186/s12967-019-1860-0.
[15] M. M. Hassan *et al.*, "A comparative study, prediction and development of chronic kidney disease using machine learning on patients clinical records," *Human-Centric Intelligent Systems*, vol. 3, no. 2, pp. 92–104, Feb. 2023, doi: 10.1007/s44230-023-00017-3.
[16] D. Baidya, U. Umaima, M. N. Islam, F. M. J. M. Shamrat, A. Pramanik, and M. S. Rahman, "A deep prediction of chronic kidney disease by employing machine learning method," *2022 6th International Conference on Trends in Electronics and Informatics, ICOEI 2022 - Proc.*, no. Icoei, pp. 1305–1310, 2022, doi: 10.1109/ICOEI53556.2022.9776876.
[17] P. Chittora *et al.*, "Prediction of chronic kidney disease - a machine learning perspective," *IEEE Access*, vol. 9, pp. 17312–17334, 2021, doi: 10.1109/ACCESS.2021.3053763.
[18] S. Y. Yashfi *et al.*, "Risk prediction of chronic kidney disease using machine learning algorithms," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2020, pp. 1–5, doi: 10.1109/ICCCNT49239.2020.9225548.
[19] S. Pal, "Prediction for chronic kidney disease by categorical and non_categorical attributes using different machine learning algorithms," *Multimedia Tools and Applications*, Apr. 2023, doi: 10.1007/s11042-023-15188-1.
[20] M. M. Rahman, M. Al-Amin, and J. Hossain, "Machine learning models for chronic kidney disease diagnosis and prediction," *Biomedical Signal Processing and Control*, vol. 87, p. 105368, Jan. 2024, doi: 10.1016/j.bspc.2023.105368.
[21] P. Ghosh, F. M. J. M. Shamrat, S. Shultana, S. Afrin, A. A. Anjum, and A. A. Khan, "Optimization of prediction method of chronic kidney disease using machine learning algorithm," in *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Nov. 2020, pp. 1–6, doi: 10.1109/iSAI-NLP51646.2020.9376787.
[22] B. Navaneeth and M. Suchetha, "A dynamic pooling based convolutional neural network approach to detect chronic kidney disease," *Biomedical Signal Processing and Control*, vol. 62, p. 102068, Sep. 2020, doi: 10.1016/j.bspc.2020.102068.
[23] S. A. E. Mienye, T. G. Swart, E. Esenogho, and I. D. Mienye, "A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease," *Bioengineering*, vol. 9, no. 8, p. 350, Jul. 2022, doi: 10.3390/bioengineering9080350.
[24] "UCI machine learning repository: early stage of Chronic kidney disease dataset," https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease.
[25] A. S. Levey and J. H. Eckfeldt, "Estimating glomerular filtration rate using serum creatinine," *Clinical Chemistry*, vol. 63, no. 6, pp. 1161–1162, Jun. 2017, doi: 10.1373/clinchem.2016.262352.
[26] S. Islam, Md. Ashiqul and Akter, "Risk factor prediction of chronic kidney disease," *UCI Machine Learning Repository*, 2023.
[27] N. Cosgrove *et al.*, "Predictive modelling of response to neoadjuvant therapy in HER2+ breast cancer," *npj Breast Cancer*, vol. 9, no. 1, p. 72, Sep. 2023, doi: 10.1038/s41523-023-00572-9.
[28] S. Demir and E. K. Sahin, "Application of state-of-the-art machine learning algorithms for slope stability prediction by handling outliers of the dataset," *Earth Science Informatics*, vol. 16, no. 3, pp. 2497–2509, Sep. 2023, doi: 10.1007/s12145-023-01059-8.
[29] D. Giri *et al.*, "Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform," *Knowledge-Based Systems*, vol. 37, pp. 274–282, Jan. 2013, doi: 10.1016/j.knosys.2012.08.011.
[30] A. Subasi and M. I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8659–8666, Dec. 2010, doi: 10.1016/j.eswa.2010.06.065.
[31] J. Lv, M. Pawlak, and U. D. Annakkage, "Prediction of the transient stability boundary using the Lasso," *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 281–288, Feb. 2013, doi: 10.1109/TPWRS.2012.2197763.
[32] S. A. Fitriani, Y. Astuti, and I. R. Wulandari, "Least absolute shrinkage and selection operator (LASSO) and k-nearest neighbors (k-NN) algorithm analysis based on feature selection for diamond price prediction," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, Jan. 2022, pp. 135–139, doi: 10.1109/ISMODE53584.2022.9742936.
[33] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.
[34] S. Boughougal, M. R. Laouar, and A. Siam, "Early prediction of nephropathy chronic with supervised machine learning algorithms and feature selection method," in *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, Apr. 2024, pp. 1–8, doi: 10.1109/PAIS62114.2024.10541274.
[35] S. K. Dey, K. M. M. Uddin, H. M. H. Babu, M. M. Rahman, A. Howlader, and K. M. A. Uddin, "Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease," *Intelligent Systems with Applications*, vol. 16, p. 200144, Nov. 2022, doi: 10.1016/j.iswa.2022.200144.
[36] K. Yashwanth *et al.*, "Identification of chronic kidney disease (CKD) using artificial neural networks algorithms," *JES*, vol. 13, no. 0377–9254, [Online]. Available: https://jespublication.com/upload/2022-V13I405.pdf
[37] S. Boughougal, M. R. Laouar, and A. Siam, "Efficient chronic kidney disease prediction: a comparative analysis using feature selection and machine learning models," in *Intelligent Technologies and Robotics Intelligent Technologies and Robotics*, 2024, pp. 191–206, doi: 10.1007/978-3-031-60591-8_16.
[38] P. Muthulakshmi, M. Parveen, and P. Rajeswari, "Prediction of heart disease using ensemble learning," *Indian Journal of Science and Technology*, vol. 16, no. 20, pp. 1469–1476, May 2023, doi: 10.17485/IJST/v16i20.2279.

[39] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science and Control Engineering*, vol. 2, no. 1, pp. 602–609, Dec. 2014, doi: 10.1080/21642583.2014.956265.
[40] D. C. Yadav and S. Pal, "An ensemble approach for classification and prediction of diabetes mellitus disease," *In Emerging Trends in Data Driven Computing and Communications: Proceedings of DDCIoT 2021*, 2021, pp. 225–235, doi: 10.1007/978-981-16-3915-9_18.
[41] A. U. Islam and S. H. Ripon, "Rule induction and prediction of chronic kidney disease using boosting classifiers, ant-miner and J48 decision tree," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Feb. 2019, pp. 1–6, doi: 10.1109/ECACE.2019.8679388.
[42] M. A. Islam, S. Akter, M. S. Hossen, S. A. Keya, S. A. Tisha, and S. Hossain, "Risk factor prediction of chronic kidney disease based on machine learning algorithms," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Dec. 2020, pp. 952–957, doi: 10.1109/ICISS49785.2020.9315878.
[43] Fondation du Rein, "What is chronic kidney disease?," *Fondation du Rein*, [Online]. Available: https://fondation-du-rein.org/la-maladie-renale-chronique/. [Accessed: May-2024].

BIOGRAPHIES OF AUTHORS

Safa Boughougal is a Ph.D. student in Information Systems at Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria, where she has been affiliated with the Department of Mathematics and Informatics and the Laboratory of Mathematics, Informatics, and Systems (LAMIS) since 2022. She earned her Master's degree in Information Systems and Web Technologies from Abdelhamid Mehri Constantine 2 University, Algeria, in 2021, and her Bachelor's degree in Information Systems from the same university. Her research interests include machine learning, artificial intelligence, deep learning, classification, and data analysis. She is actively pursuing her research in these fields. She can be contacted via email at: safa.boughougal@univ-tebessa.dz.

Mohamed Ridda Laouar is a Full Professor of Computer Science at Tebessa University, Tebessa, Algeria. He received his Ph.D. in Industrial and Human-Computer Science from the University of Valenciennes, France, in 2005. He also received a Master's degree in Urban Information Systems from the University of Artois, France, in 2001. His research areas include artificial intelligence, quantum computing, information systems (IS), decision support systems (DSS), e-library systems, and other related topics. He has contributed to such journals as Hi-Tech Library and Human Systems Management. He is the editor of the IJIST journal and the proceedings of several conferences. He was the Chair of the ICIST and ICSENT conferences from 2011 to 2024. He can be contacted at email: ridda.laouar@univ-tebessa.dz.

Abderrahim Siam received his Ph.D. in computer science from the University of Constantine, Algeria. He is working as a professor in the Department of Mathematics and Computer Science at the University of Khenchela, Algeria. He is currently the Vice-Rector of the University of Khenchela, Algeria. His research interests include software engineering, fuzzy logic, formal methods, multiagent, and complex systems. He can be contacted at email: siamabderrahim@gmail.com

Sean Eom is a Professor Emeritus of Management Information Systems (MIS) at the Harrison College of Business of Southeast Missouri State University. He received his Ph.D. in Management Science from the University of Nebraska–Lincoln. He also received an MS in international business from the University of South Carolina at Columbia. His research areas include decision support systems, bibliometrics, e-learning systems, machine learning, and deep learning. He is the author/editor of fourteen books and has published over 90 refereed journal articles and 130 articles in encyclopedias, book chapters, and conference proceedings. He can be contacted at email: sbeom@semo.edu.