Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Classifying chronic kidney disease using selected machine learning techniques

Abrahem P. Anqui *

College of Technology, Cebu Technological University, Cebu, Philippines

ARTICLE INFO

Article history:
Received 19 September 2024
Received in revised form
7 January 2025
Accepted 22 January 2025

Keywords:
Chronic kidney disease
Machine learning algorithms
Early diagnosis
Classification accuracy
Key clinical features

ABSTRACT

Chronic kidney disease (CKD) is a serious global health problem with high mortality rates, often due to late diagnosis. Early detection and classification are essential to improve treatment outcomes and slow disease progression. This study evaluates the performance of four machine learning algorithms—linear discriminant analysis (LDA), Naïve Bayes, C4.5 decision tree, and Random Forest—in classifying CKD using a Kaggle dataset containing 1,659 instances and 52 features, covering demographic, lifestyle, and clinical data. After data pre-processing, the classification accuracies of the algorithms were assessed. LDA showed the highest accuracy at 92.8%, followed by Naïve Bayes (92.1%), C4.5 (92.0%), and Random Forest (91.9%) before hyperparameter tuning. After tuning, C4.5 achieved the highest accuracy of 92.5%, followed by Random Forest (92.2%), with Naïve Bayes remaining at 92.1%. However, even after tuning, LDA remained the most accurate, demonstrating superior performance. The key features contributing to CKD classification were serum creatinine, glomerular filtration rate (GFR), muscle cramps, protein in urine, fasting blood sugar, itching, systolic blood pressure, blood urea nitrogen (BUN), HbA1c, edema, total cholesterol, body mass index (BMI), and gender. These findings confirm that LDA outperforms other algorithms in CKD classification without the need for tuning, emphasizing the value of machine learning in improving early diagnosis and management of CKD.

1. Introduction

Chronic kidney disease is a silent assassin with a high risk of fatality due to its delayed symptoms which may also vary from one patient to another (Owens et al., 2020). Early detection and diagnosis with high accuracy at the early level show promising results in terms of effective medication, and mitigation of symptoms becoming worse (Senturk, 2020).

Prediction is increasingly gaining popularity across diverse fields such as; predicting chronic disease (Debal and Sitote, 2022), cancer prognosis and prediction (Shaikh and Rao, 2022), environmental science for monitoring and predicting ecological changes (Hakim et al., 2024), and the medical and mental health sectors for diagnosing and managing conditions (Chung and Teo, 2022).

Prediction plays a pivotal role in the field of medicine, particularly in diagnosing illnesses which affects proper and accurate decision-making and its outcome. Machine learning is a technique that is highly used for disease classification and early diagnoses of disease (Anqui, 2023). It is a paradigm to obtain useful information from a humongous amount of data and eventually use this information to produce valuable outcomes in addressing certain societal issues (Delima, 2019).

Discriminant analysis, decision trees, Naïve Bayes, and random forest are among the numerous machine learning techniques that can be utilized for classification to extract valuable information from large datasets (Dennis and Strafella, 2024). Linear Discriminant Analysis (LDA) is a popular technique for classification and prediction (Cui et al., 2023), based on linear regression (Anqui, 2023). As a supervised learning method, it aims to develop a discriminant function by applying regression analysis. The derived discriminant function is then used to assign the regression value to a particular category. In discriminant analysis, the dependent variable is always categorical, whereas the independent variable is continuous by nature. To

* Corresponding Author.
Email Address: ehm20.anqui@gmail.com
https://doi.org/10.21833/ijaas.2025.02.008
Corresponding author's ORCID profile:
https://orcid.org/0009-0000-1231-569X

evaluate the classification accuracy of LDA, this study undertakes a comparative analysis. Specifically, it examines how well LDA performs by comparing its results with those generated by three other widely recognized and highly effective prediction techniques: The Naïve Bayes algorithm, C4.5, and Random Forest. These three methods are chosen for their reputation for producing optimal outcomes across a range of classification problems, making them suitable benchmarks for assessing the efficacy of LDA (Sano et al., 2023; Wang et al., 2019; Meher et al., 2024).

2. Review of related literature

Artificial Intelligence, often referred to as machine intelligence, is the ability of machines to mimic human thought processes to assist humans better perform in the fields of science and technology. It involves the accurate interpretation of datasets, learning from them, and applying these insights to achieve specific objectives (Kaplan and Haenlein, 2019). Machine learning, widely regarded as a subset of artificial intelligence (Jagdale et al., 2022), is a method of analyzing vast amounts of data, learning from patterns, and eventually performing predictions with high efficiency and reliability (Xue et al., 2024). One important field that requires highly accurate predictions is medical mining (Pareek et al., 2024), high accuracy prediction is pivotal to the medical practitioners to assist them in their decision-making such as when diagnosing certain illnesses (Mantelakis et al., 2021). The field of medicine is increasingly focused on developing new approaches that extract knowledge from raw data within the medical environment (Chakraborty et al., 2024), these advancements are poised to have a profound impact, benefiting not only medical practitioners but also, most importantly to the patients (Zampogna et al., 2024). By facilitating immediate access to necessary medications, these methods help reduce the risk of symptoms becoming severe (Wu et al., 2023), and mortality (Yu et al., 2024). Different machine learning techniques were used to predict and prognose diseases such as chronic diseases, breast cancer prognosis and diagnosis, and Alzheimer among other diseases. The study of Bansal et al. (2022) showed that chronic diseases like diabetes, cancer, and hypertension are crucial for early prevention, with machine learning offering predictive capabilities based on medical records or checkups. The key to accurate prediction lies in data quality, addressing challenges like outliers, missing values, and imbalanced data while selecting the best machine learning method based on performance metrics like accuracy and precision. The paper provides a systematic review of machine learning techniques, including supervised, ensemble, and deep learning, and discusses preprocessing approaches to enhance prediction performance. Moreover, Rane et al. (2020) explored that breast cancer is a leading cancer among women, particularly in developing countries where most cases are diagnosed at advanced stages. The paper compares six machine learning algorithms—Naive Bayes, Random Forest, Artificial Neural Networks, K-Nearest Neighbor, Support Vector Machine, and Decision Tree—on the Wisconsin Diagnostic Breast Cancer dataset to classify cancer as benign or malignant. The best-performing algorithm will serve as the backend for a website designed to assist in breast cancer diagnosis using MRI data. While, Rani et al. (2024) emphasized that Alzheimer's disease is a progressive neurodegenerative disorder affecting older adults, and while it has no cure, early diagnosis can reduce its impact. This study evaluates the performance of decision tree, extreme gradient boosting (XGB), and random forest (RF) algorithms using the Open Access Series of Imaging Studies dataset, with SMOTE applied for balancing. On the balanced dataset, the random forest algorithm achieved the highest accuracy of 95.03%, making it the most effective for predicting Alzheimer's.

On the other hand, Rani et al. (2024) experimented that Chronic kidney disease (CKD) is a lifelong condition that can lead to end-stage kidney failure, but early detection and proper treatment can slow its progression. This study explores the potential of machine learning and predictive modeling to identify CKD, narrowing 25 initial variables down to the most predictive subset. Among the 12 classifiers tested, XgBoost achieved the highest performance with accuracy, precision, recall, and an F1-score of 0.98, demonstrating the effectiveness of advanced machine learning techniques for early CKD diagnosis.

Discriminant Analysis is one of the popular machine learning algorithms used for prediction, gaining more attention in this big data era (Sun, 2022). Discriminant analysis is generally employed for predictive purposes by which discriminant function determines the group membership of a particular subject. The process begins with formulating the discriminant function by using the linear regression concept to determine the predicted value, the result is then classified into a specific group membership using the cut-off score computed using the data set used (Anqui, 2023). Ricciardi et al. (2020) used linear discriminant analysis to classify patients as either having a coronary artery disease or not, results show that LDA alone achieved 84.5% accuracy, thus, presenting a practical application of data mining technique that helps medical practitioners in improving decision-making.

The Naïve Bayes algorithm is also widely used for various prediction tasks, including analyzing tourism sentiments. These sentiment predictions are then utilized to create visual comparative analyses, achieving an accuracy rate of up to 80% (Ricciardi et al., 2020). Similarly, the C4.5 algorithm is applied to different classification tasks, such as product classification in vending machines and secure decision support systems. The C4.5 algorithm has demonstrated a maximum accuracy of 87%, with the lowest recorded accuracy being 67% (Li et al., 2024). Based on the findings from the literature review, predicting chronic kidney disease using various machine learning techniques offers significant advantages, such as helping to prevent the need for kidney transplants and dialysis (Chaithra et al., 2023). Furthermore, to ensure the accuracy of the classification algorithm, it is essential to compare its results with those of other well-established algorithms.

3. Methodology

3.1. Data set

The data set used in this study is outsourced from the Kaggle website[†] with 1659 instances. The dataset has 52 features where features 1 to 51 are input features (independent/predictor variables) while feature 52 is the output (dependent variable), the index description of the dataset is shown in Table 1.

3.2. Data processing

Before using the dataset as input to the algorithms: LDA, Naïve Bayes, and C4.5, multiple data processing steps were conducted and are listed as follows:

1. Removal of insignificant variables (Patient ID and Doctor in charge).
2. Converting of Age variable from days to years.
3. Recoding of categorical and binary variables.

3.3. Linear discriminant analysis

LDA is a variant of discriminant analysis that is based on linear regression, it discriminates a group membership of a specific subject. Steps of LDA are as follows:

a. Slope equation. The first step is to get the slope value to quantify the relationship between the independent and dependent variable using this equation:

$$b_1 = \frac{\sum(x_i - \hat{X})(y_i - \hat{Y})}{\sum(x_i - \hat{X})^2}$$

where, $b_1$is the computed slope value; x_i is the score of the predictor variable; $\hat{X}$ is the mean of all predictor variable; y_i is the score of the dependent variable; while $\hat{Y}$ is the mean of the output (dependent) variable.

b. Intercept Equation. The second step of the LDA process is to compute the intercept value using this equation:

$$b_0 = \tilde{y} - b_1\hat{X}$$

where, b_0 is the computed intercept value, ỹ is the mean of all dependent features, $b_1$is the computed slope value, while $\hat{X}$ is the mean of the predictor (independent) variable.

c. Linear Regression Equation. It is used to compute the predicted value of the variable (dependent) using the 52-predictor variable, this is also known as the discriminant function. To compute the discriminant function value, use the previously computed intercept and slope values using this equation:

$$Y = a + b_1X_1 + b_2X_2 \ldots b_iX_i$$

where, Y is the output value; a is the priorly computed intercept score; b_1 is the slope score; X_i is the respondent's score on the given predictor variable; i is the number of predictor features of the dataset.

d. Cut-off score. After determining the discriminant function and computing the predicted value of the dependent variable, the cut-off score must also be determined to distinguish the membership of the predicted value to a particular group by using this equation:

$$Z_c = \frac{n_a z_b + n_b z_a}{n_a + n_b}$$

where, Z_c is the centroid value (membership); n_a is the total number in group a (1); n_b is the total number in group b (0); z_a is the mean of group a; $z_b$is the mean of group b.

e. Significance of regression coefficients. When the discriminant function processes multiple variables, it is normal to check which independent variables significantly contribute to the process after the effects of other features are taken into account. This can be checked through the stepwise statistics, this will display the variables entered/removed.

3.4. Naïve Bayes classifier

Naïve Bayes algorithm that is based on Bayes' theorem with strong (Naïve) independence assumptions between the features. Primarily Naïve Bayes is used for classification tasks using the following equation:

$$P(C|X) = \frac{P(C) \cdot P(x_1|c) \cdot P(x_2|c) \ldots . . P(x_n|c)}{P(X)}$$

where, $P(C|X)$ is the posterior probability of class C given the feature vector X; $P(C)$ is the prior probability of class C; $P(x_i|c)$ is the likelihood of feature x_i given class C; $P(X)$ is the marginal likelihood of the feature vector X.

3.5. C4.5 classifier

The C4.5 algorithm is a widely used decision tree algorithm and an extension of the earlier ID3 algorithm, which is commonly applied in machine learning for classification tasks. The C4.5 algorithm

[†] https://www.kaggle.com/datasets/rabieelkharoua/chronic-kidney-disease-dataset-analysis

utilizes the concept of information gain ratio to determine the most suitable attribute for splitting the data, using the following equation:

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{Split\ Information\ (A)}$$

where, $Gain\ (A)$ is the information gain obtained by using attribute A to split the data; while $Split\ Information\ (A)$ is a measure of how uniformly the attribute A splits the data.

3.6. Random forest

Random Forest is a widely used ensemble learning method in machine learning that combines multiple decision trees to make more accurate and robust predictions. It operates by creating a collection, or forest, of decision trees, where each tree is trained on a different subset of the data and a different subset of features. The final prediction is an aggregation of the predictions made by the individual trees, it makes predictions by majority voting for classification or averaging for regression, with the general formula:

$$\hat{y} = \underset{y \in Y}{argmax} \sum_{i=1}^{m} 1(T_i(x) = y)$$

where, $\hat{y}$ is the predicted class label, Y is the set of possible class labels, while $T_i(x)$ is the prediction of the $i - th$ tree for input x.While m is the number of trees in the forest. On the other hand, $1(T_i(x) = y)$ is the indicator function that is 1 if tree T_i predicts class y, and 0 otherwise. The sum counts how many trees predicted each class, and the argmax function selects the class with the most votes.

Table 1: Data set

No.	Variable	Type
1	Age	Int - years
2	Gender	Categorial: 0: Male, 1: Female
3	Ethnicity	Categorical; 0: Caucasian 1: African American, 2: Asian, 3: Other
4	Socioeconomic status	Categorical; 0: Low, 1: Middle, 2: High
5	Education level	Categorical: 0: None, 1: High School, 2: Bachelors, 3: Higher
6	BMI	Int - continuous
7	Smoking	Binary
8	Alcohol consumption	Int - continuous
9	Physical activity	Int - continuous
10	Diet quality	Int - continuous
11	Sleep quality	Int - continuous
12	Family history of kidney disease	Binary
13	Family history hypertension	Binary
14	Family history diabetes	Binary
15	Previous acute kidney injury	Binary
16	Urinary tract infection	Binary
17	Systolic BP	Int - continuous
18	Diastolic BP	Int - continuous
19	Fasting blood sugar	Int - continuous
20	HbA1c	Int - continuous
21	Serum creatine	Int - continuous
22	Blood urea nitrogen (BUN) levels	Int - continuous
23	Glomerular filtration rate (GFR)	Int - continuous
24	Protein in urine	Int - continuous
25	ACR	Int - continuous
26	Serum electrolytes sodium	Int - continuous
27	Serum electrolytes potassium	Int - continuous
28	Serum electrolytes calcium	Int - continuous
29	Serum electrolytes phosphorus	Int - continuous
30	Hemoglobin levels	Int - continuous
31	Cholesterol total	Int - continuous
32	Cholesterol LDL	Int - continuous
33	Cholesterol HDL	Int - continuous
34	Cholesterol triglycerides	Int - continuous
35	ACE inhibitors	Binary
36	Diuretics	Binary
37	NSAIDs use	Int - continuous
38	Statins	Binary
39	Antidiabetic medication	Binary
40	Edema	Binary
41	Fatigue levels	Int - continuous
42	Nausea vomiting	Int - continuous
43	Muscle cramps	Int - continuous
44	Itching	Int - continuous
45	Quality of life score	Int - continuous
46	Heavy metal exposure	Binary
47	Occupational exposure chemicals	Binary
48	Water quality	Categorical: 0: Good, 1: Poor
49	Medical checkups frequency	Int - continuous
50	Medication adherence	Int - continuous
51	Health literacy	Int - continuous
52	Diagnosis (absence or presence of CKD)	Binary

3.7. Hyperparameter tuning

Hyperparameter tuning is the process of adjusting an algorithm's hyperparameters to achieve optimal performance. These hyperparameters, which include settings like the learning rate, number of layers, or regularization strength, are external configurations that control how the model learns which may also vary from one algorithm to another. Selecting the right hyperparameter values is crucial, as they directly influence the model's accuracy, speed, and ability to generalize to new data. One commonly used technique for hyperparameter tuning is Random Search, where hyperparameter combinations are randomly sampled from predefined ranges, rather than exhaustively tested as in grid search. This approach offers a more efficient exploration of the hyperparameter space. However, even when using Random Search, hyperparameters can still be manually adjusted for further fine-tuning, allowing for flexibility in optimizing the model's performance.

4. Results and discussion

The classification analysis of CKD was conducted using four well-known machine learning algorithms: LDA, Naïve Bayes, C4.5, and Random Forest. The objective was to determine which algorithm performs best in classifying CKD based on a given set of predictor variables. To ensure reliability, hyperparameter tuning was applied to optimize the classification accuracy and validate the performance of each algorithm.

The dataset comprises 52.4% male and 47.6% female individuals. Smoking is prevalent, with 56.3% of the population identified as smokers, while 43.7% are non-smokers. Additionally, 55.7% of the population consumes alcohol, whereas 44.3% do not. The average Body Mass Index (BMI) in the dataset is 27.62, which falls into the overweight category. These factors are commonly associated with an increased risk of CKD. The dataset includes a diverse range of symptoms and contributing factors, making it suitable for CKD classification. A total of 1,659 instances were analyzed to evaluate and compare the accuracy of the LDA, Naïve Bayes, C4.5, and Random Forest algorithms in classifying CKD. The results, presented in Table 2, indicate that LDA achieved an accuracy of 92.8% on the test data, with a slightly reduced accuracy of 92.6% after cross-validation. The analysis was conducted using Statistical Package for the Social Sciences (SPSS) software.

The author identified in the stepwise statistics shown in Table 3 that there are 13 key variables that significantly affect the diagnosis of chronic kidney disease viz; SerumCreatine, GFR, MuscleCramps, ProteinInUrine, FastingBloodSugar, Itching, SystolicBP, BUNLevels, HbA1c, Edema, CholesterolTotal, BMI, and Gender. LDA is not limited to classification; it can also perform feature extraction to identify which features statistically contribute to the prediction or classification process. The stepwise process utilizes Wilks' Lambda to select features that best discriminate between groups in predicting CKD. Lower Lambda values indicate better discriminators, with key variables like Serum Creatinine, Glomerular Filtration Rate, Muscle Cramps, and Protein in Urine significantly reducing Lambda, enhancing the model's accuracy. Additional features such as Fasting Blood Sugar, Itching, Systolic BP, BUN Levels, and HbA1c further refine the model, highlighting their metabolic and cardiovascular relevance to CKD. The inclusion of features like Edema, Cholesterol, BMI, and Gender in later steps continues to improve prediction, though with smaller incremental gains. The largest reductions in Wilks' Lambda occur in the initial steps, with Serum Creatinine reducing it from 1 to 0.960 and GFR lowering it further to 0.929, indicating their strong predictive power for diagnosing CKD. As more variables are added, the reductions become smaller, such as Muscle Cramps to 0.918 and Protein in Urine to 0.909, showing that the initial features contribute the most to improving model accuracy. This systematic selection process ensures the most relevant features are used, optimizing the model's predictive power. This capability allows the LDA classifier to achieve more accurate results, even in the absence of feature optimization (Hossain et al., 2022).

Table 2: Classification result of LDA

Classification results					
		Diagnosis	Predicted group membership		Total
			0	1	
Original	Count	0	27	108	135
		1	11	1513	1524
	%	0	20.0	80.0	100.0
		1	.7	99.3	100.0
Cross-validated	Count	0	23	112	135
		1	11	1513	1524
	%	0	17.0	83.0	100.0
		1	.7	99.3	100.0

92.8% of original grouped cases correctly classified; Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case; 92.6% of cross-validated grouped cases correctly classified

Table 4 presents the classification results without hyperparameter tuning of the Naïve Bayes, C4.5 algorithms, and Random Forest. Naïve Bayes correctly classified 92.1% of instances with a precision of 90.9, recall of 92.0 and F1-Score of 91.3, while C4.5 correctly classified 92.0% with a precision of 89.9, recall of 92.1 and F1-Score of 89.4, and Random Forest correctly classified 91.9% with a precision of 92.6, a recall of 91.9 and an F1-Score of 88.1. In contrast, 7.90% of instances were misclassified by Naïve Bayes, 8.0% were misclassified by the C4.5 algorithm, and 8.1% were misclassified by the Random Forest. Without hyperparameter tuning, Naïve Bayes offers the highest accuracy and recall, Random Forest excels in precision, and C4.5 provides the best balance between precision and recall based on the F1-Score. All three models perform similarly, with minor differences in classification results.

Table 3: Stepwise statistics

Variables entered/removed									
Step	Entered	Wilks' lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	Serum creatinine	.960	1	1	1657.000	69.853	1	1657.000	.000
2	GFR	.929	2	1	1657.000	63.454	2	1656.000	.000
3	Muscle cramps	.918	3	1	1657.000	48.962	3	1655.000	.000
4	Protein in urine	.909	4	1	1657.000	41.369	4	1654.000	.000
5	Fasting blood sugar	.899	5	1	1657.000	37.022	5	1653.000	.000
6	Itching	.890	6	1	1657.000	33.932	6	1652.000	.000
7	Systolic BP	.881	7	1	1657.000	31.802	7	1651.000	.000
8	BUN levels	.875	8	1	1657.000	29.556	8	1650.000	.000
9	HbA1c	.871	9	1	1657.000	27.100	9	1649.000	.000
10	Edema	.869	10	1	1657.000	24.934	10	1648.000	.000
11	Cholesterol total	.866	11	1	1657.000	23.136	11	1647.000	.000
12	BMI	.864	12	1	1657.000	21.595	12	1646.000	.000
13	Gender	.862	13	1	1657.000	20.308	13	1645.000	.000

At each step, the variable that minimizes the overall Wilks' lambda is entered; Maximum number of steps is 102; Minimum partial F to enter is 3.84; Maximum partial F to remove is 2.71

Table 5 presents the hyperparameter tuning process conducted for the Naïve Bayes, C4.5, and Random Forest algorithms. It details the tuning techniques employed, the manually adjusted hyperparameters, and the 10-fold cross-validation method used to evaluate model performance.

Table 4: Naïve Bayes, C4.5 results, and random forest results without hyperparameter tuning

Criteria	C4.5	Naïve Bayes	Random forest
Accuracy	92.0%	92.1%	91.9%
Precision	90.9	89.9	92.6
Recall	92.0	92.1	91.9
F1-score	91.3	89.4	88.1
Correctly classified	1527	1528	1525
Incorrectly classified	132	131	134

Table 5: Hyperparameter tuning

Machine learning algorithm	Hyperparameter tuning technique	Manually adjusted hyperparameter	Cross-validation

C4.5	Random search	minNumObj = 9
		unpruned = false
			10-Folds Cross-validation

Naïve Bayes	Random search	useSupervisedDiscretization=false	10-Folds Cross-validation

Random forest	Random search	Max-depth = 20
		NumFeatures = 10
		NumIterations = 1000
			10-Folds Cross-validation

Table 6 presents the classification results of the machine learning algorithms after hyperparameter tuning, the performance results show that C4.5 has the highest accuracy at 92.5%, followed by Random Forest with 92.2%, and Naïve Bayes slightly behind at 92.1%. Random Forest excels in precision with 92.8%, meaning it is the most accurate in predicting positive instances, while C4.5 and Naïve Bayes lag behind at 91.4 and 89.8, respectively. C4.5 has the best recall at 92.5%, capturing the most actual positive cases, while Random Forest and Naïve Bayes have similar recall rates of 92.2% and 92.1%. C4.5 also has the highest F1-Score at 91.8%, indicating the best balance between precision and recall, whereas Naïve Bayes and Random Forest score lower at 89.5 and 88.7, respectively. In terms of classification counts, C4.5 correctly classified 1535 instances, Random Forest 1529, and Naïve Bayes 1528, with a slightly higher number of misclassifications for Naïve Bayes and Random Forest. To evaluate the effectiveness of machine learning algorithms in predicting CKD, a comparison was conducted between LDA, Naïve Bayes, C4.5, and random forest as illustrated in Fig. 1. The results indicate that LDA outperforms the other three algorithms in terms of accuracy without hyperparameter tuning. Although all three algorithms demonstrate promising results in predicting CKD, LDA proved to be the most effective among them (Anqui, 2023).

Table 6: Naïve Bayes, C4.5 results, and random forest results with hyperparameter tuning

Criteria	C4.5	Naïve Bayes	Random forest
Accuracy	92.5%	92.1%	92.2%
Precision	91.4	89.8	92.8
Recall	92.5	92.1	92.2
F-1 score	91.8	89.5	88.7
Correctly classified	1535	1528	1529
Incorrectly classified	124	131	130

Fig. 2 shows the classification accuracy of four machine learning algorithms after hyperparameter tuning. LDA, which did not undergo hyperparameter tuning but included feature selection, achieved the highest accuracy at around 92.6%. C4.5 follows closely at just above 92.5%, while Random Forest reaches 92.2%. Naïve Bayes, despite undergoing hyperparameter tuning, has the lowest accuracy at just over 92.1%. Overall, LDA outperforms the other algorithms in terms of accuracy (Anqui, 2023; Díaz-Navarro et al., 2024).

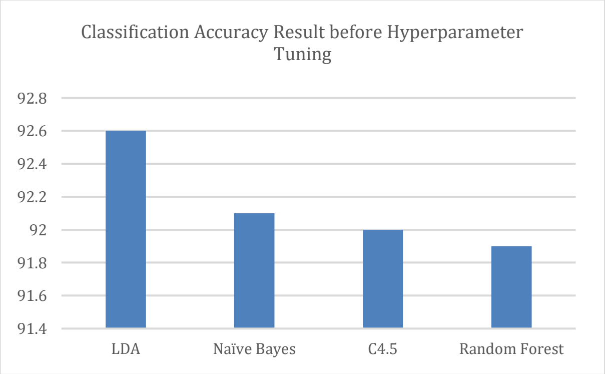

Fig. 1: Comparative classification accuracy before hyperparameter tuning

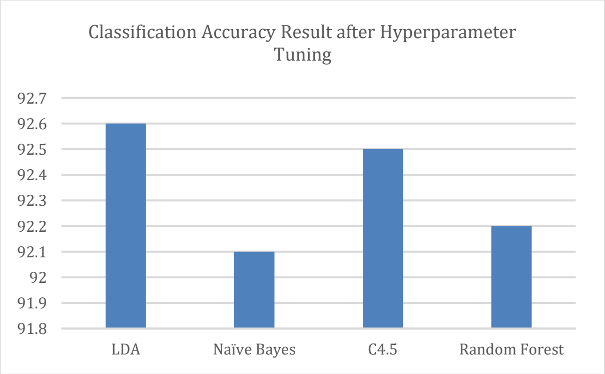

Fig. 2: Comparative classification accuracy after hyperparameter tuning

5. Conclusion and recommendation

The primary objective of this study is to evaluate the performance of three machine learning algorithms in diagnosing CKD, using a specific set of predictor variables. Simultaneously, the study aims to leverage LDA to determine which features significantly contribute to the accuracy of the prediction, offering deeper insights into the factors that most impact CKD diagnosis. The results of the experiment indicate that LDA surpasses the Naïve Bayes, C4.5 classifiers, and Random Forest achieving an impressive prediction accuracy of 92.6%, compared to 92.1%, 92.0%, and 91.9%, respectively before hyperparameter tuning, despite tuning the hyperparameter LDA still achieved the highest accuracy of 92.6% followed by C4.5 with an accuracy rate of 92.5%, while random forest achieved 92.2% accuracy rate and Naïve Bayes with 92.1% accuracy. The findings highlight the effectiveness of LDA as a powerful tool for predicting diseases. Future research should consider applying LDA to a broader range of datasets beyond those used in this study. Additionally, integrating optimization techniques, such as genetic algorithms and ant colony optimization, may improve the model's performance and robustness. Exploring these areas could significantly advance the field of disease prediction and classification through advanced computational methods.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Anqui AP (2023). Respiratory disease classification using selected data mining techniques. International Journal of Advanced and Applied Sciences Journal, 10(7): 219-223. https://doi.org/10.21833/ijaas.2023.07.024

Bansal M, Goyal A, and Choudhary A (2022). Stock market prediction with high accuracy using machine learning techniques. Procedia Computer Science, 215: 247-265. https://doi.org/10.1016/j.procs.2022.12.028

Chaithra AS, Chandana DK, Chetana SM, and Greeshma N (2023). Risk prediction of chronic kidney disease using machine learning algorithms. In: Kumar A, Gunjan VK, Hu YC, and Senatore S (Eds.), International conference on data science, machine learning and applications: 333-338. Springer Nature, Singapore, Singapore. https://doi.org/10.1007/978-981-99-2058-7_30

Chakraborty C, Bhattacharya M, Pal S, and Lee SS (2024). From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare. Current Research in Biotechnology, 7: 100164. https://doi.org/10.1016/j.crbiot.2023.100164

Chung J and Teo J (2022). Mental health prediction using machine learning: Taxonomy, applications, and challenges. Applied Computational Intelligence and Soft Computing, 2022: 9970363. https://doi.org/10.1155/2022/9970363

Cui H, Deng Y, Zhong R, Li W, Yu C, Danyushevsky LV, Belousov I, Li Z, and Wang H (2023). Determining the ore-forming processes of Dongshengmiao Zn-Pb-Cu deposit: Evidence from the linear discriminant analysis of pyrite geochemistry. Ore Geology Reviews, 163: 105782. https://doi.org/10.1016/j.oregeorev.2023.105782

Debal DA and Sitote TM (2022). Chronic kidney disease prediction using machine learning techniques. Journal of Big Data, 9: 109. https://doi.org/10.1186/s40537-022-00657-5

Delima AJP (2019). Predicting scholarship grants using data mining techniques. International Journal of Machine Learning and Computing, 9(4): 513-519. https://doi.org/10.18178/ijmlc.2019.9.4.834

Dennis AGP and Strafella AP (2024). The role of ai and machine learning in the diagnosis of Parkinson's disease and atypical Parkinsonisms. Parkinsonism and Related Disorders, 126: 106986. https://doi.org/10.1016/j.parkreldis.2024.106986 **PMid:38724317**

Díaz-Navarro S, Díez-Hermano S, Rojo-Guerra MA, Maurandi JL, Valdiosera C, Gunther T, and Uriarte MH (2024). Sex estimation using long bones in the largest burial site of the Copper Age: Linear discriminant analysis and random forest. Journal of Archaeological Science: Reports, 58: 104730. https://doi.org/10.1016/j.jasrep.2024.104730

Hakim DK, Gernowo R, and Nirwansyah AW (2024). Flood prediction with time series data mining: Systematic review.

Natural Hazards Research, 4(2): 194-220. https://doi.org/10.1016/j.nhres.2023.10.001

Hossain MM, Swarna RA, Mostafiz R, Shaha P, Pinky LY, Rahman MM, Rahman W, Hossain MS, Hossain ME, and Iqbal MS (2022). Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease. Machine Learning with Applications, 9: 100330. https://doi.org/10.1016/j.mlwa.2022.100330

Jagdale KR, Shelke CJ, Achary R, Wankhede DS, and Bhandare TV (2022). Artificial intelligence and its subsets: Machine learning and its algorithms, deep learning, and their future trends. Journal of Emerging Technologies and Innovative Research, 9(5): 112-117.

Kaplan A and Haenlein M (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1): 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

Li P, Xiong F, Huang X, and Wen X (2024). Construction and optimization of vending machine decision support system based on improved C4.5 decision tree. Heliyon, 10(3): e25024. https://doi.org/10.1016/j.heliyon.2024.e25024 **PMid:38318033 PMCid:PMC10838796**

Mantelakis A, Assael Y, Sorooshian P, and Khajuria A (2021). Machine learning demonstrates high accuracy for disease diagnosis and prognosis in plastic surgery. Plastic and Reconstructive Surgery–Global Open, 9(6): e3638. https://doi.org/10.1097/GOX.0000000000003638 **PMid:34235035 PMCid:PMC8225366**

Meher BK, Singh M, Birau R, and Anand A (2024). Forecasting stock prices of fintech companies of India using random forest with high-frequency data. Journal of Open Innovation: Technology, Market, and Complexity, 10(1): 100180. https://doi.org/10.1016/j.joitmc.2023.100180

Owens E, Tan KS, Ellis R, Del Vecchio S, Humphries T, Lennan E, Vesey D, Healy H, Hoy W, and Gobe G (2020). Development of a biomarker panel to distinguish risk of progressive chronic kidney disease. Biomedicines, 8(12): 606. https://doi.org/10.3390/biomedicines8120606 **PMid:33327377 PMCid:PMC7764886**

Pareek A, Karlsson J, and Martin RK (2024). Machine learning/artificial intelligence in sports medicine: State of the art and future directions. Journal of ISAKOS, 9(4): 635-644. https://doi.org/10.1016/j.jisako.2024.01.013 **PMid:38336099**

Rane N, Sunny J, Kanade R, and Devi S (2020). Breast cancer classification and prediction using machine learning. International Journal of Engineering Research and Technology, 9(2): 576-580. https://doi.org/10.17577/IJERTV9IS020280

Rani P, Lamba R, Sachdeva RK, Kumar K, and Iwendi C (2024). A machine learning model for Alzheimer's disease prediction. IET Cyber-Physical Systems: Theory and Applications, 9(2): 125-134. https://doi.org/10.1049/cps2.12090

Ricciardi C, Valente AS, Edmund K, Cantoni V, Green R, Fiorillo A, Picone I, Santini S, and Cesarelli M (2020). Linear discriminant analysis and principal component analysis to predict coronary artery disease. Health Informatics Journal, 26(3): 2181-2192. https://doi.org/10.1177/1460458219899210 **PMid:31969043**

Sano AVD, Stefanus AA, Madyatmadja ED, Nindito H, Purnomo A, and Sianipar CP (2023). Proposing a visualized comparative review analysis model on tourism domain using Naïve Bayes classifier. Procedia Computer Science, 227: 482-489. https://doi.org/10.1016/j.procs.2023.10.549

Senturk ZK (2020). Early diagnosis of Parkinson's disease using machine learning algorithms. Medical Hypotheses, 138: 109603. https://doi.org/10.1016/j.mehy.2020.109603 **PMid:32028195**

Shaikh FJ and Rao DS (2022). Prediction of cancer disease using machine learning approach. Materials Today: Proceedings, 50: 40-47. https://doi.org/10.1016/j.matpr.2021.03.625

Sun W (2022). Data mining in the big data era. Advances in Social Science, Education and Humanities Research, 664: 2107-2111. https://doi.org/10.2991/assehr.k.220504.381 **PMid:36572217**

Wang X, Zhou C, and Xu X (2019). Application of C4.5 decision tree for scholarship evaluations. Procedia Computer Science, 151: 179-184. https://doi.org/10.1016/j.procs.2019.04.027

Wu Y, Li L, Xin B, Hu Q, Dong X, and Li Z (2023). Application of machine learning in personalized medicine. Intelligent Pharmacy, 1(3): 152-156. https://doi.org/10.1016/j.ipha.2023.06.004

Xue J, Alinejad-Rokny H, and Liang K (2024). Navigating micro- and nano-motors/swimmers with machine learning: Challenges and future directions. ChemPhysMater, 3(3): 273-283. https://doi.org/10.1016/j.chphma.2024.06.001

Yu YP, Liu S, Geller D, and Luo JH (2024). Serum fusion transcripts to assess the risk of hepatocellular carcinoma and the impact of cancer treatment through machine learning. The American Journal of Pathology, 194(7): 1262-1271. https://doi.org/10.1016/j.ajpath.2024.02.017 **PMid:38537933 PMCid:PMC11220925**

Zampogna B, Torre G, Zampoli A, Parisi F, Ferrini A, Shanmugasundaram S, Franceschetti E, and Papalia R (2024). Can machine learning predict the accuracy of preoperative planning for total hip arthroplasty, basing on patient-related factors? An explorative investigation on Supervised machine learning classification models. Journal of Clinical Orthopaedics and Trauma, 53: 102470. https://doi.org/10.1016/j.jcot.2024.102470 **PMid:39045495**