International Journal of Artificial Intelligence
p-ISSN: 2407-7275, e-ISSN: 2686-3251

Original Research Paper

A Proposed Multilayer Perceptron Model and Kernel Principal Component Analysis for the Prediction of Chronic Kidney Disease

Iliyas Ibrahim Iliyas[1*], Souley Boukari[2], Abdulsalam Ya'u Gital[2]

[1] *Department of Computer Science, University of Maiduguri. Borno State, Nigeria.*

[2] *Department of Computer Science, Abubakar Tafawa Balewa University. Bauchi State, Nigeria.*

Article History
Received:
03.11.2024

Revised:
22.11.2024

Accepted:
06.12.2024

***Corresponding Author:**
Iliyas Ibrahim Iliyas
Email
iliyasibrahimiliyas@unimaid.edu.ng

Abstract: Chronic Kidney Disease (CKD) is a stage in which the kidney cannot filter waste from the blood that circulates in the body; unfortunately, this stage is mostly detected at a late stage, leading to dialysis or transplantation. Early detection is important for the effective management of CKD. ML has shown success in the early prediction of CKD by using an algorithm that learns and predicts without being programmed. ML requires appropriate datasets for this process, and one of the aspects is dimensionality reduction, which addresses the challenges of unnecessary tests, high-cost tests and the use of redundant tests. Principal Component Analysis (PCA) is a widely used method for dimensionality reduction; however, it relies on linear transformation to identify relationships within features. Medical datasets such as CKD exhibit complex nonlinear features, which is important for exploring alternative dimensionality reduction methods that can rely on nonlinear transformation. This study aims to propose an ML approach that utilises kernel PCA to reduce dimensionality based on nonlinearity structures and enhance the prediction of CKD. We evaluated seven ML models on the different kernel functions of PCA. The ML models included random forest (RF), decision tree (DT), multilayer perceptron (MLP), support vector machine (SVM), extreme gradient boosting (XgBoost), adaptive boosting (AdaBoost), logistic regression (LR), and gradient boosting. The kernel functions used for dimensionality reduction are cosine principal component analysis (CPCA), polynomial principal component analysis (PPCA), radial basis principal component analysis (RPCA), sigmoid principal component analysis (SPCA) and linear principal component analysis (LPCA). The results of the study revealed that the MLP with RPCA, SPCA and CPCA achieved good performance in predicting CKD, with an accuracy score of 99% on DB1, and that the MLP with RPCA and SPCA achieved good performance in predicting CKD, with an accuracy score of 100% on DB2. The study showed how kernel PCA, which effectively reduces high dimensionality-based nonlinearity relationships, can positively affect the performance of predictive models and the power of dimensionality reduction toward disease prediction.

Keywords: Chronic Kidney Disease, Dimensionality Reduction, Kernel Function, Multilayer Perceptron, Principal Component Analysis.

1. Introduction

The kidney is an organ that filters waste from the blood and passes it out through urine. It also has other functions, such as maintaining fluid and electrolyte balance, regulating blood pressure, controlling red blood cell production, producing vitamin D, and managing pH levels [1]. However, when the kidney gradually loses its ability to filter waste and accumulate waste and fluids in the body, chronic kidney disease (CKD) can occur. CKD is defined based on glomerular filtration removal (GFR) and albuminuria, where the GFR is used to assess excretion while albuminuria crosses the renal barrier [2].

In 2016, CKD affected 336 million men and 417 million women, amounting to 753 globally. Due to the high cost of frequent dialysis or kidney replacement surgery, more than 1 million people in 112 developing nations pass away from renal failure each year. Therefore, early detection and treatment are vital to reduce the burden of CKD on public health [3]. Machine learning (ML) uses algorithms to learn and find patterns from enormous amounts of data recently used in healthcare on electronic medical records (EMRs). ML algorithms can effectively predict CKD and enable early treatment at a lower cost [4].

ML requires other methods, such as feature selection, to improve performance. By introducing dimensionality reduction from a dataset, ML reduces the number of unnecessary tests, lowers the financial burden, and minimises redundant testing while improving the effort of diagnosing CKD with great accuracy [5]. Techniques such as PCA have achieved success in reducing the dimensionality of the dataset by considering linear relationships between features [6]. Standard principal component analysis (PCA) identifies linear subsets in data that capture the highest variation but cannot detect nonlinear patterns. Kernel-based PCA addresses this limitation by mapping the input space nonlinearly, effectively reducing dimensionality in high-dimensional spaces. This allows adjustments to class boundaries via the kernel, resulting in a nonlinear transformation of data points [7].

This study contributes to the development of a modified PCA feature reduction approach with an MLP to predict CKD. The proposed study explored how a kernel PCA can reduce high dimensionality in features while considering the nonlinear correlation between features. This study highlighted how this approach can also improve the performance of the ML model in predicting CKD.

2. Literature Review

This section presents existing studies conducted on CKD prediction via machine learning algorithms and feature selection techniques.

Research in [6] shows the power of the genetic algorithm and PCA in enhancing the MLP in the prediction of CKD. The model predicts CKD based on 20 components after applying PCA to reduce the dimensionality of the dataset. The model achieved 98.34% and 98.54% accuracy during training and testing, respectively [1] highlighted the capacity of ML models to detect CKD with fewer tests or features. The author proposed a hybrid feature selection method comprising a chi-square test (Chi2) and mutual information with an extra tree classifier to predict CKD with 98.00% accuracy. This study explored how the combination of two feature selections in selecting the most impactful features based on correlation scores can improve the accuracy of ML.

In a study by [8], a CNN was developed to forecast the occurrence of CKD within the next 6 and 12 months and achieved rates of 89.00% and 88.00%, respectively. This study also explored the most prominent features for the prediction of CKD. A study by [9] proposed an XgBoost classifier for the early detection of CKD; the proposed model achieved the highest accuracy of 98.30%, and the study also highlighted how PCA can improve the accuracy of the ML model after reducing the high dimensionality of the dataset. Abdullahi et al (2019) also evaluated the performance of different feature selection methods in classifying CKD, and random forest feature selection achieved the best accuracy, with an RF of 98.82%. Disease datasets such as CKD datasets have many features that lead to high-dimensional data, which affects the performance of the ML algorithm.

Table 1 shows some of the reviewed works carried out on the prediction of CKD.

3. Methodology

This section discusses the various materials and methods used in this study. Data preprocessing, feature reduction and MLP were applied to predict CKD, the methodology of this study is illustrated in Figure 1.

Table 1. Background Study on the Existing Works on CKD Prediction

Reference	Contribution	Result
[1]	A proposed hybrid Chi-squared test (Chi2) and Mutual Information (MI) based feature selection method with Extra Trees classifier for prediction of CKD	The hybrid Chi-squared test (Chi2) and Mutual Information (MI) based feature selection method achieved 98% accuracy.
[6]	CKD prediction using optimized MLP and GA with the implementation of PCA for feature reduction	The optimized MLP and the feature reduction approach achieved 98.34% and 98.54% accuracy during training and testing, respectively.
[8]	Development of a machine learning model to forecast the occurrence of CKD within the following 6 or 12 months	The proposed CNN achieved 89.00% and 88.00% accuracy for the 6-month and 12-month predictions.
[10]	Development of ML model with feature selection to classify CKD	The study showed that RF with RF feature selection can classify CKD with 98.825 accuracy
[9]	A Proposed machine learning model for early detection of CKD	The study XGBoost classifier for prediction of CKD with 98.30% accuracy
[11]	Development of a machine learning model for chronic renal disease prognosis	The developed decision tree achieved 97.00% accuracy.
[12]	A proposed predictive model for detecting CKD using 30% of original features	With PCA applied to reduce the original features to 30% and the XgBoost classifier, the model achieved 98.30% accuracy.

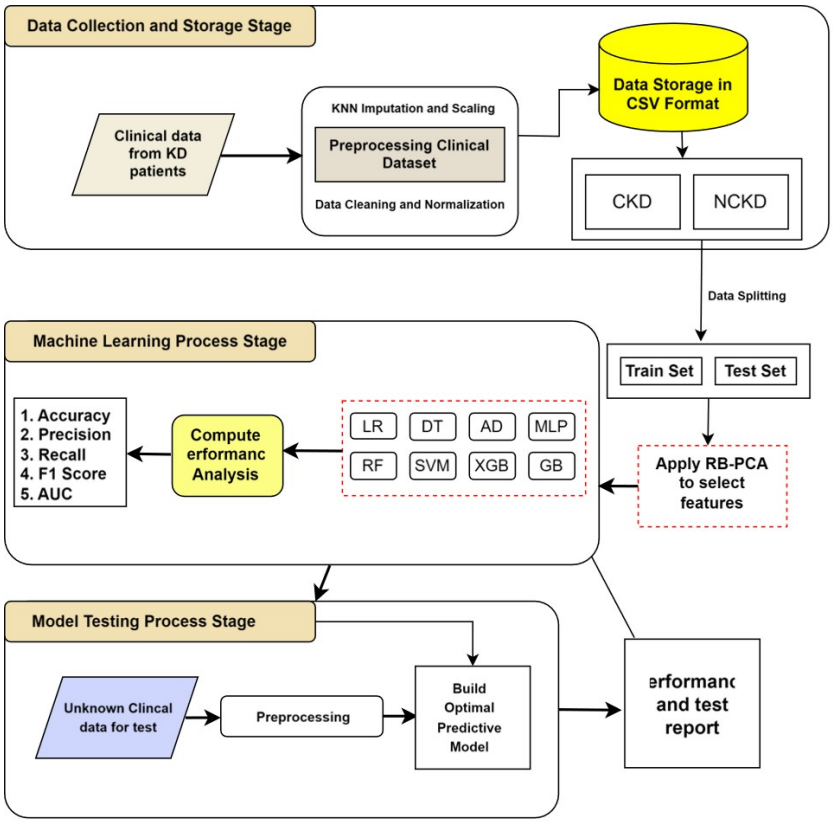

Figure 1. Proposed Methodology

3.1. Dataset Description

The CKD dataset used in the study was obtained from two sources. The first CKD dataset was from Apollo Hospital, Tamil Nadu, in July 2015 and is available online at the University of California, Irvine Machine Learning Repository (UCI) [13], within 2 months and a year ranging from 2 to 90 years of age. The public dataset has 400 instances and 25 features, including both input and output features.

The input features have 11 categorical and 14 numerical values, whereas the output feature has CKD and NCKD values representing the presence and absence of CKD. The dataset has 250 and 150 instances of CKD and NCKD, respectively.

A description of the features is provided in Table 2, and an overview of the dataset features is shown in Table 3.

Table 2. Description of Features from DB2

S/No	Feature	Type	Full Name	Units
1	age	Number	Individual Age	2-90 years
2	bp	Number	Blood Pressure measured	50-180 mm/Hg
3	sg	Number	Specific Gravity	1.005-1.025
4	al	Number	Albumin	0-5
5	su	Number	Sugar	0-5
6	rbc	Categorical	Red Blood Cells	Normal/abnormal
7	pc	Categorical	Pus cell	Normal/abnormal
8	pcc	Categorical	Pus cell clumps	Present/not present
9	ba	Categorical	Bacteria	Present/notpresent
10	bgr	Number	Blood Glucose Random	22-490 mgs/dl
11	bu	Number	Blood urea	1.5-391 mgs/dl
12	sc	Number	Serum Creatinine	0.4-76 mgs/dl
13	sod	Number	Sodium	4.5-163 mEq/L
14	pot	Number	Potassium	2.5-47 mEq/L
15	hemo	Number	Haemoglobin	3.1-17.8 gms
16	pcv	Number	Packed Cell Volume	9-54
17	wbcc	Number	White blood Cells Count	2200-26400 cells/cmm
18	rbcc	Number	Red Blood Cells Count	2.1-6.5 millions/cmm
19	htm	Categorical	Hypertension	Yes/no
20	dm	Categorical	Diabetes Mellitus	Yes/no
21	cad	Categorical	Coronary Artery Disease	Yes/no
22	appet	Categorical	Appetite	Good/poor
23	pe	Categorical	Pedal Edema	Yes/no
24	ane	Categorical	Anemia	Yes/no
25	classification	Class	Classification	CKD/NotCKD

The second dataset is the CKD dataset obtained from the UCI Repository and collected from Enam Medical College, Savar, Dhaka, Bangladesh [14]. The dataset has 200 instances consisting of 26 input features and 1 input feature. The feature "affected" is the input feature with 1 and 0 classes.

A summary of the dataset and feature descriptions are shown in Table 4 and Table 5.

Table 3. Overview of DB1

Item	Description
Name	CKD dataset
Data source	UCI ML Repository
Total records	400
Total columns	26
Input features	24
Output features	1
Class	CKD/NCKD
Categorical values	11
Numerical values	14
Class distribution	250/150(CKD/NCKD)
Missing values	1009
Missing rows	242
Years	2 to 90 years
Period	2 months

Table 4. Description of Features from DB2

S/No	Feature	Type	Full Name	Units
1	bp (Diastolic)	Number	blood pressure (Diastolic)	mm/Hg
2	bp limit	Number	Blood Pressure limit	mm/Hg
3	sg	Categorical	Specific Gravity	1005 to 10025
4	al	Categorical	Albumin	0-5
5	class	Binary	Classification	CKD/NCKD
6	rbc	Number	Red Blood Cells	Normal/abnormal
7	su	Categorical	Sugar	0-5
8	pc	Number	Pus cell	Normal/abnormal
9	pcc	Number	Pus cell clumps	Present/not present
10	ba	Number	Bacteria	Present/notpresent
11	bgr	Categorical	Blood Glucose Random	mgs/dl
12	bu	Categorical	Blood urea	mgs/dl
13	sod	Categorical	Sodium	mEq/L
14	sc	Categorical	Serum Creatinine	0.4-76 mgs/dl
15	pot	Categorical	Potassium	mEq/L
16	hemo	Categorical	Haemoglobin	3.1-17.8 gms
17	pcv	Categorical	Packed Cell Volume	9-54
18	rbcc	Categorical	Red Blood Cells Count	2.1-6.5 millions/cmm
19	wbcc	Categorical	White blood Cells Count	2200-26400 cells/cmm
20	htn	Number	Hypertension	yes or no
21	dm	Number	Diabetes Mellitus	Yes/no
22	cad	Number	Coronary Artery Disease	Yes/no
23	appet	Number	Appetite	Good/poor
24	pe	Number	Pedal Edema	Yes/no
25	ane	Number	Anemia	Yes/no
26	grf	Categorical	Glomerular filtration rate	mL/min
27	stage	Categorical	Stage	1-4
28	affected	Number	Affected	Yes/No
29	age	Categorical	Age	<74

Table 5. Overview of DB2

Item	Description
Name	CKD dataset
Data source	UCI ML Repository
Total records	200
Total columns	28
Input features	26
Output features	1
Class	1/0
Categorical values	12
Numerical values	14
Class distribution	0/1
Missing values	0
Missing rows	0
Years	<74 years
Period	NA

3.2. Data Preparation

DB2 has no missing values, but DB1 has some missing values, and out of 400 instances, 242 instances have missing values. Additionally, out of the 25 features, all the features have missing values except for the classification feature, and the total number of cells with missing cells is 1009. For columns with numeric values, we used KNN Imputer to fill in the missing values of the CKD dataset by identifying the closest data point. The technique calculates the similarity between rows based on the available data in other columns, and the techniques find the average of the values from the 5 neighbours and uses the average to fill in the missing cells. For columns with categorical values, missing cells are filled with the mode in the dataset; by calculating the most frequent value of the column and filling the missing cell with it, the equation of the KNN imputer is shown in Equation 1.

$$KNN = \hat{x}_i = \frac{1}{k}\sum_{j \in N_k(i)} x_j \quad (1)$$

The datasets have categorical columns such as 'rbc', 'pc', 'pcc', 'ba', 'htn', 'dm', 'cad', 'appet', 'pe', 'ane', and 'classification', where the values are not represented as numbers but as groups of categories. This study uses LabelEncoder to assign a unique category in each categorical column and assigns numerical values to a unique category. To standardise the range of values of the independent features, we used min-max scaling to standardise the range of values so that the model would not be biased based on the inconsistent range of the independent features.

3.3. Dimensionality Reduction

The process of selecting the most critical risk features in healthcare helps to remove redundant features and can be used to minimise model training time, improve data quality, and enhance prediction performance. Recent studies have paved the way for different techniques that help in choosing the best features and removing less important features [1]. Dimensionality reduction is the method used to convert high-dimensional data into low-dimensional data while maintaining the original structure and meaning of the data. Dimensionality reduction is widely used to minimise noise and irrelevant data, which enhances the ability of models to work more efficiently and improves accuracy [15]. Principal component analysis is a dimensional reduction technique that is used to reduce the high dimensionality of a dataset by recognizing a small set of principal components that capture the most important information from the dataset.

PCA helps improve the model's efficiency and computational complexity, but it also has several limitations, such as the process of identifying data based on their linear relation to the principal components when high dimensionality is reduced [16]. Kernel PCA extends PCA to nonlinear dimensionality reduction by employing a kernel function to map the data into a higher-dimensional feature space. It allows for capturing complex relationships between variables and is particularly useful when the data have nonlinear structures [16]. Kernel PCA has five functions: cosine-principal component (CPCA), polynomial principal component analysis (PPCA), radial basis function (RPCA),

sigmoid principal component analysis (SPCA) and PCA, which use a linear function [7]. The functions are defined as follows:

The linear function is described below:

$$f(x_i, x_j) = x_i.x_j \tag{2}$$

The sigmoid function is described below:

$$\mathrm{f}(x, y) = \tanh\,(axy + c) \tag{3}$$

The polynomial function is described below:

Polynomial $f(x_i, x_j) = (x_i \cdot x_j + C)^2$ (4)

The radial basis function is described below:

$$f(x_i, x_j) = \exp\left(-\gamma \|x_i|| \cdot ||x_j\|^2\right) \tag{5}$$

The cosine function is described below:

$$f(x_i, x_j) = \frac{(x_i \cdot x_j)}{\|x_i\| \cdot \|x_j\|} \tag{6}$$

3.4. Machine Learning Models

Machine learning (ML) is a technique that uses algorithms to predict or classify data without being programmed after passing through a training process [17]. The training process of ML algorithms is based on supervised and unsupervised methods. When the algorithm is trained on a labelled dataset, the process is called supervised learning, whereas if the algorithm is trained on an unlabelled dataset, the process is called unsupervised learning. Supervised learning involves classification and regression tasks, whereas unsupervised learning involves clustering tasks. This study is based on classification tasks, and the ML algorithms used in the study are explained in this section.

1) Random Forest
   The random forest (RF) algorithm is a type of ML algorithm that creates multiple decision trees during the training process as its working principles; it can be used for regression or classification tasks, and it generates different decision trees from bootstrapped samples via the bagging technique during the training process. During the training process, some columns are randomly removed to reduce the variance of the features [17].
2) Decision Tree
   A decision tree (DT) is an ML algorithm that can be used for classification or regression tasks. DT splits data into smaller portions based on certain conditions. This splitting process starts through a series of decisions, which are called decisions. Each node tests a condition, and based on the results of the condition, the data follow through. It continues until classification is achieved at the end, and the DT is based on how trees are, with different branches that are represented as nodes [18].
3) Multilayer Perceptron
   The multilayer perceptron (MLP) is another type of ML but is an improved type of ML called deep learning. The model mimics how the brain works; it has an input layer, one or more hidden layers and an output. Like neurons in our brains, an MLP has nodes that are in every layer, and each node is connected to the nodes in the next layers. Each node has its own bias and weight. The MLP model uses a function called the activation function to make every node produce an output, which can be called active or not activated. The MLP model has now been regarded as a method with good performance in classification tasks [19].

4) Logistic regression
Logistic regression (LR) is a type of supervised learning algorithm that is widely used in the healthcare sector for predicting the probability of class outcomes based on independent variables. LR assumes p as the probability of a positive outcome, while 1-p is the probability of a negative outcome. LR uses a decision boundary to set the threshold that classifies data into positive and negative outcomes, and the classification probability is calculated via the logistic sigmoid function [17].
5) Support Vector Machine
Support vector machine (SVM) is a type of supervised learning algorithm that is used for classification tasks. SVM works by identifying a hyperplane that best separates data into classes, maximizing the margin between them to ensure clear boundaries. SVMs work well in handling complex, nonlinear relationships and overfitting [20].
6) Gradient boosting
Gradient boosting (GB) is a type of ML algorithm that can be used for regression or classification tasks, where it builds a predictive model by combining multiple weak learners, starting with a simple regression tree. GB enhances its performance when these weak learners are combined sequentially. This method reduces the loss function, which measures the difference between actual predicted values through a step-by-step sampling procedure [21].

3.5. Performance Metrics

This section discusses the various matrices used to evaluate the performances of the ML algorithms used.

1) Precision
Precision determines the ratio of correctly predicted positive outcomes to all positive outcomes [17].

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (7)$$

2) Accuracy
Accuracy is defined as the percentage of the total number of correct predicted data points out of all the data points [22].

$$Accuracy - \frac{TN+TP}{TN+TP+FN+FP} \times 100\% \quad (8)$$

3) Recall
Recall that the ratio of true positive (TP) to all the actual positive outcomes captures all the positive outcomes as positive [23].

$$Recall = \frac{TP}{TN+TP+FN+FP} \times 100\% \quad (9)$$

4) F1 score
The FI score is used to balance the precision and recall for model evaluation [9].

$$F1\ score = 2 \times \frac{P \times R}{P+R} \times 100\% \quad (10)$$

4. Finding and Discussion

This section highlights the results obtained from this study. The datasets used were split into training and testing sets, and necessary data preprocessing was carried out, such as handling missing values with the KNN imputation technique, LabelEncoder for handling categorical values and Min–Max for transforming the values within the range of 0--1. Table 1 highlights the environment setup utilized during the execution of this study. The study implementation was conducted via Jupyter Notebook with the Python programming language. The libraries used are pandas, NumPy, Sklearn, Matplotlib, and Seaborn. The ML algorithms evaluated are GB, MLP, SVM, LR, RF, XgBoost, AdaBoost and DT, and kernel-based PCA functions are used to explore the impact of those functions on the performance of the ML model in predicting CKD in Database 1 (DB1) and Database 2 (DB2). The

functions include sigmoid, cosine, polynomial, radial basis and linear functions. The evaluation metrics used are accuracy, precision, F1 score, and recall score.

As illustrated in Figure 2, seven ML algorithms were evaluated for the prediction of CKD with PCA kernel functions. On the basis of their accuracies on DB1, DT with PPCA achieved its highest accuracy of 99%, AdaBoost with PPCA and SPCA achieved its highest accuracy of 96%, the GB model achieved its highest accuracy of 96% with PPCA, LR achieved its highest accuracy of 96% with PCA, RF achieved its highest accuracy of 99% with PPCA, SVM achieved the highest accuracy of 96% for all kernels, XGB achieved its highest accuracy of 97% with SPCA, and MLP achieved the best accuracy of 100% with SPCA, followed by RF with 99% with RPCA. MLP achieved its lowest accuracy of 97% with LPCA.

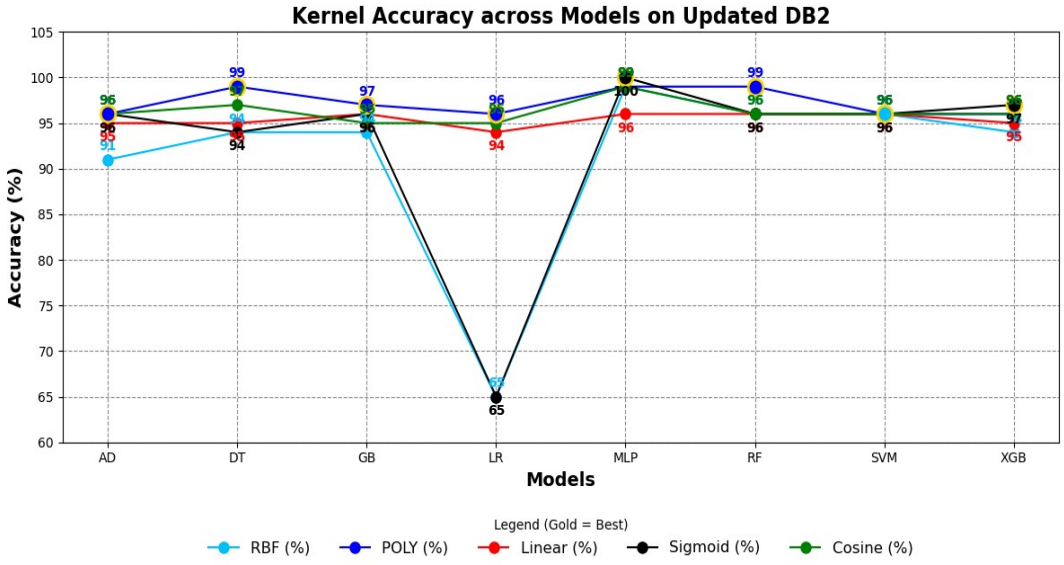

Figure 2. Comparative Accuracy of The ML Models on PCA and Kernel PCA on DB1

As illustrated in Figure 3, seven ML algorithms were evaluated for the prediction of CKD with PCA kernel functions on DB2 based on accuracy.

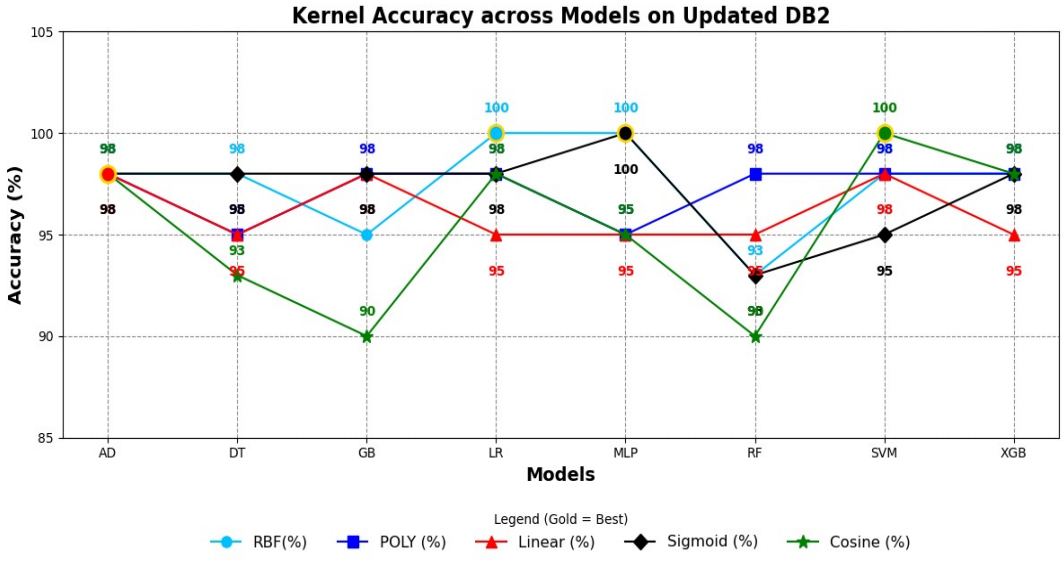

Figure 3. Comparative Accuracy of The ML Models on PCA and Kernel PCA on DB2

DT with RBF and SPCA achieved its highest accuracy of 98%, AdaBoost achieved an accuracy of 98% for all kernels, the GB model achieved its highest accuracy of 98% with PPCA and LPCA and SPCA, RF achieved its highest accuracy of 98% with PPCA, and LR achieved its highest accuracy of 100% with RBF. The SVM achieves its highest accuracy of 100% with CPCA, and XGB achieves the highest accuracy of 98% with PPCA, RPCA, SPCA, and CPCA, except for LPCA, which achieves the lowest accuracy of 95%. MLP achieved the best accuracy at 100% with both RPCA and SPCA, followed by LR and SVM, which achieved 100% accuracy with RPCA and CPCA, respectively. The MLP achieved its lowest accuracy of 95% with RPCA and LPCA.

Figures 4 and Figure 5 provide insights into the performances and classifications of the MLP model when trained with the kernel functions on DB1. Figure 4 visualizes the data distribution across the two components mapped by the five-kernel PCA used in this study with the MLP on DB1. While Figure 5 visualizes the data distribution across the two components mapped by the five-kernel PCA used in this study with the MLP on DB2.

The mapping illustrates how each kernel function influences the separation of data points.

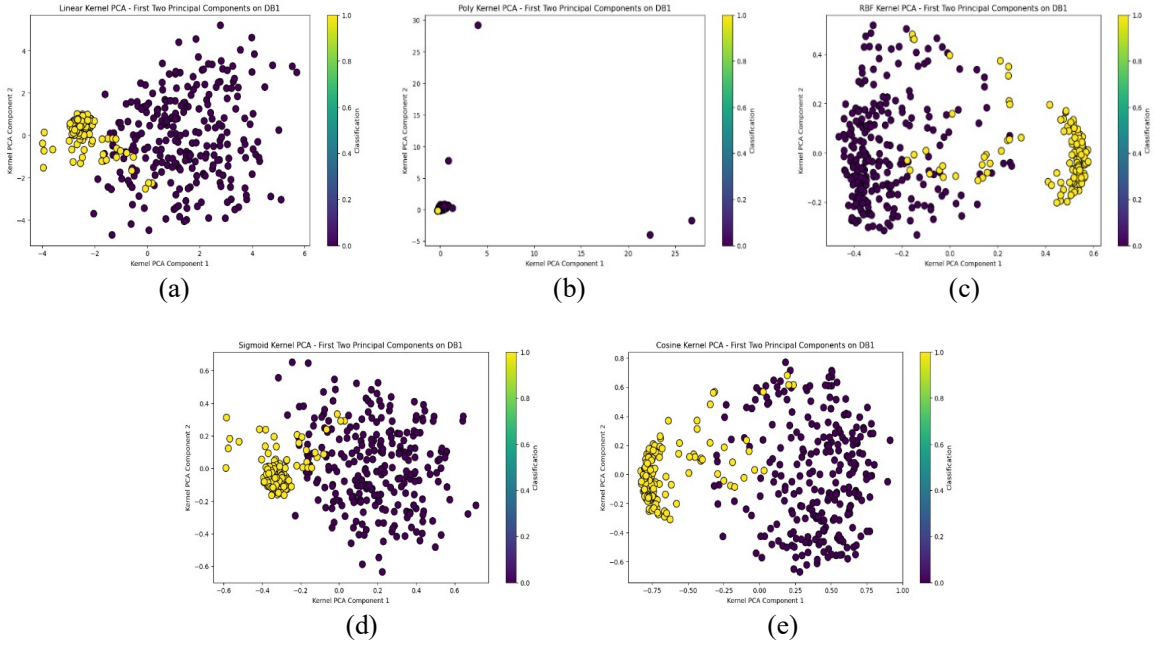

Figure 4. Kernel Function PCA-Based Dimensional Space Reduction by Using Five Functions on DB1: (a) Linear (b) Polynomial, (c) Radial Basis, (d) Sigmoid, (e) Cosine

Figure 6 presents the confusion matrix results, where further clarification was made on the impact of each kernel on the classification accuracy.

The matrix in Figure 6 shows how the MLP achieved zero misclassifications with the kernel function in predicting CKD on DB1, except for RPCA and LPCA, which had one misclassification of NCKD with RPCA and one misclassification of CKD and NCKD with LPCA.

Figure 7 presents the confusion matrix results, where further clarification was made on the impact of each kernel on the classification accuracy.

The matrix in Figure 7 shows how the MLP achieved zero misclassifications with the kernel function in predicting CKD on DB2 except for the PPCA, SPCA and LPCA, which had one misclassification that was not affected by the PPCA, one misclassification that was not affected by the LPCA and one misclassification that was affected by the SPCA.

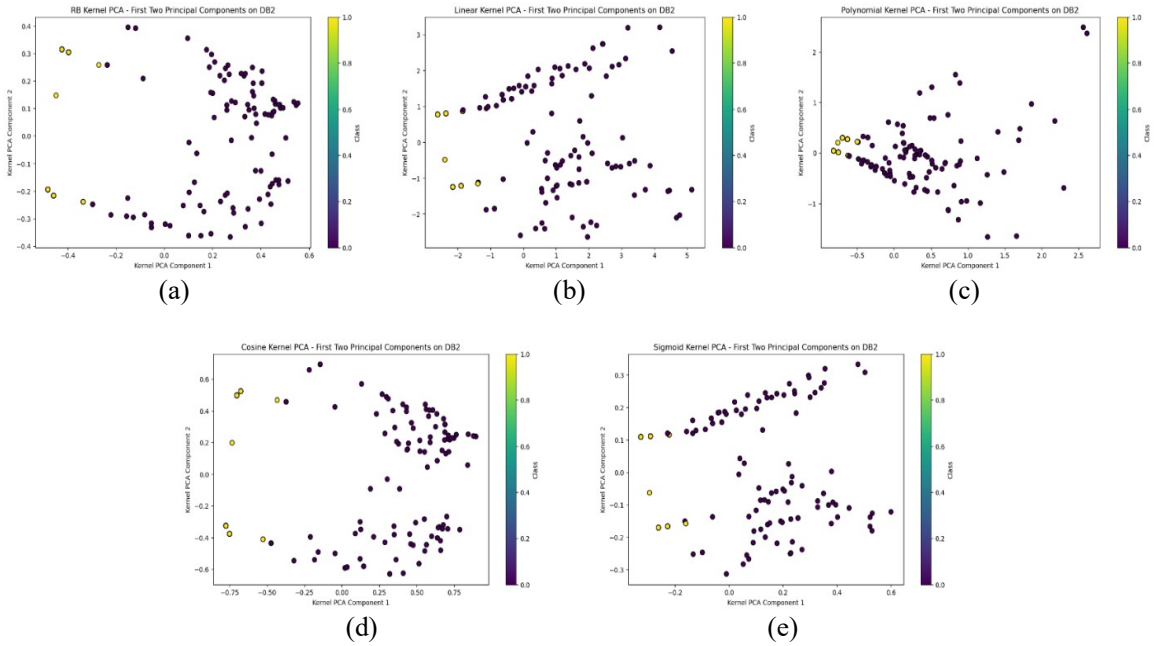

Figure 5. Kernel Function PCA-Based Dimensional Space Reduction via five Functions on DB2: (a) Radial Basis (b) Linear, (c) Polynomial, (d) Cosine, (e) Sigmoid

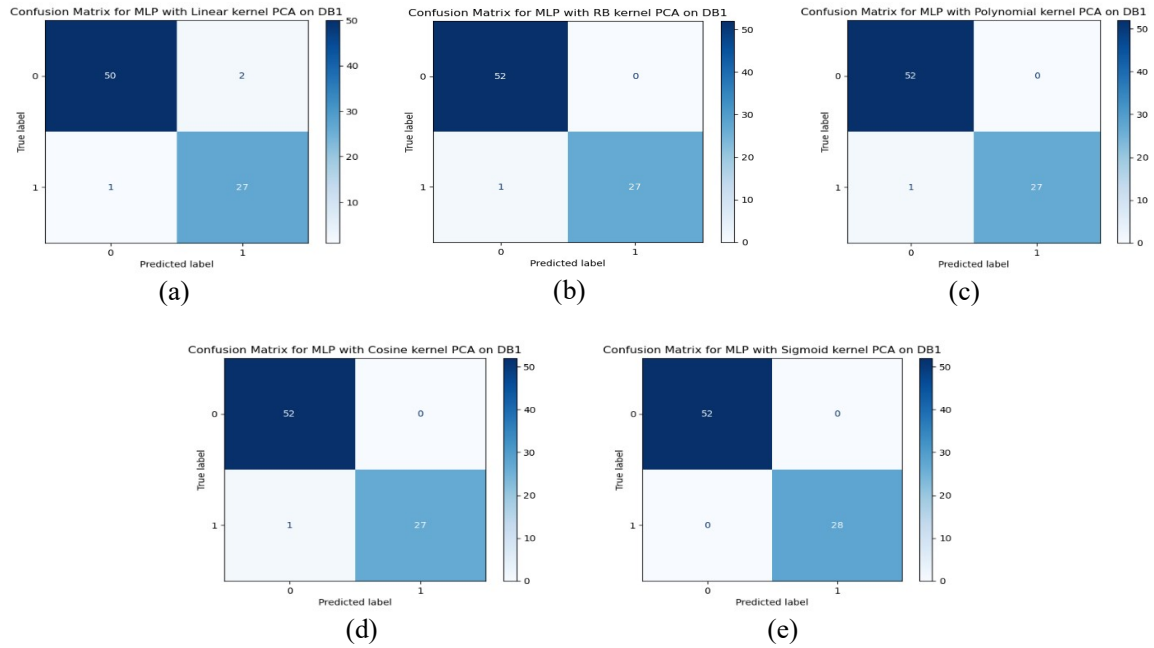

Figure 6. Confusion Matrix for All the Kernel Functions of PCA with MLP on DB1: (a) Radial Basis (b) Polynomial, (c) Sigmoid, (d) Linear, (e) Cosine

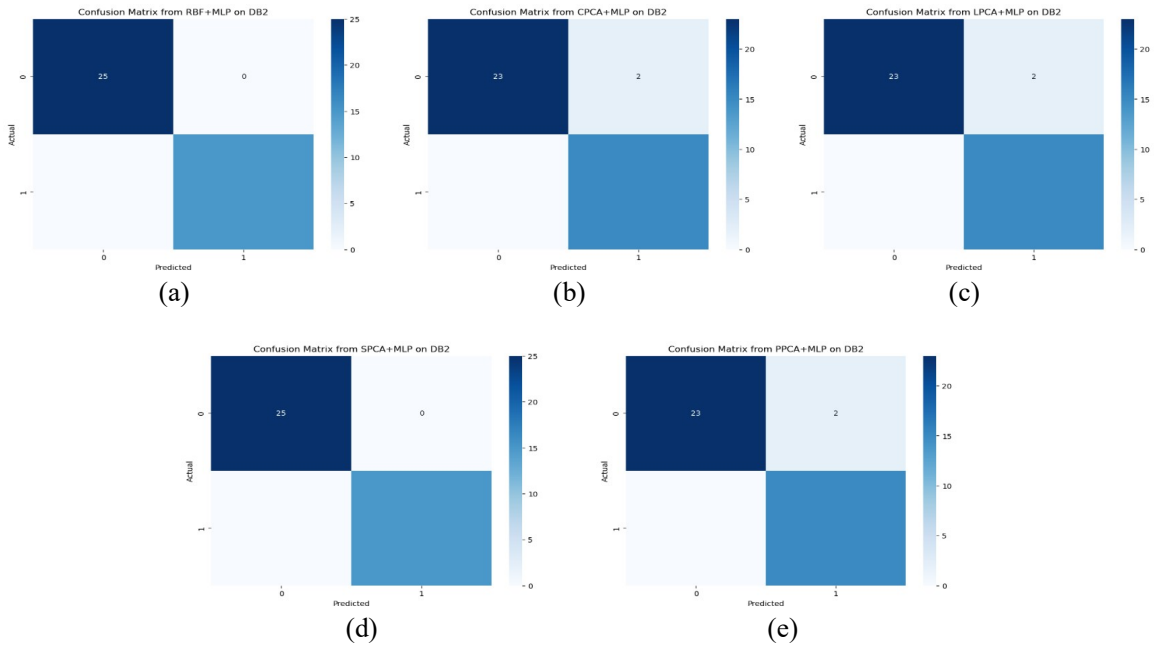

Figure 7. Confusion Matrix for All the Kernel Functions of PCA with MLP on on DB2: (a) Radial Basis (b) Cosine, (c) Linear, (d) Sigmoid, (e) Polynomial

Table 6 summarizes the performance achieved by the MLP in predicting CKD with the kernel function of PCA on DB1. The MLP achieved a good performance of 100 in terms of accuracy, precision, recall and F1 scores with SPCA and achieved the lowest accuracy of 97% with the linear PCA, which uses a linear function.

Table 6. Performance Metrics of The MLP With Kernel Functions on DB 1

DB1	PCA Kernels	Accuracy	Precision	Recall	F1 Score
MLP	RBF	99%	0.98	0.98	0.98
	Poly	99%	0.98	0.98	0.98
	Cosine	99%	0.98	0.98	0.98
	Sigmoid	100%	1.00	1.00	1.00
	Linear	96%	0.96	0.96	0.96

Table 7 summarizes the performance achieved by the MLP in predicting CKD with the kernel function of PCA on DB2. The MLP achieved a good performance of 100 in terms of accuracy, precision, recall and F1 scores with both PPCA, CPCA, and SPCA.

Table 7. Performance Metrics of The MLP With Kernel Functions on DB 2

DB2	PCA Kernels	Accuracy	Precision	Recall	F1 Score
MLP	RBF	99%	0.99	0.98	0.98
	Poly	100%	1.00	1.00	1.00
	Cosine	100%	1.00	1.00	1.00
	Sigmoid	100%	1.00	1.00	1.00
	Linear	97%	0.97	0.97	0.97

Table 8 presents a comparison of the proposed model with existing methods. Some of the studies achieved reasonable performance; however, some of the studies did not perform feature selection, whereas the remaining studies adopted an approach that did not consider an approach that reduces high dimensionality based on nonlinearity, which is a limitation that requires improvement. However, none of these studies have attempted to accommodate data with nonlinear structures during dimensionality reduction, even though nonlinear relationships between features are common in medical datasets.

Table 8. Comparative Analysis of The Proposed Study with Existing Studies

Reference	Accuracy	AUC	Feature Selection	Model
[6]	98.54%	0.99	PCA	MLP
[24]	96.48%	0.98	PCA	Passive aggressive Classifier (PCA)
[25]	99%	0.98	PCA	RF
[1]	98%	NA	A hybrid Chi-squared test (Chi2) and Mutual Information (MI)	Extra Trees
[26]	98%	NA	-	ANN
[27]	99%	100	-	Rotation Forest
Proposed Model	**100%**	**1.00**	**KPCA**	**MLP**

According to the results of this study, PCA with kernel functions can enhance PCA to reduce the dimensionality of CKD features with nonlinear structures and improve the performance of predictive models. Kernel PCA has improved the MLP model in the prediction of CKD from 96% to 100% and 97% to 100% on DB1 and DB2, respectively, and across other performance metrics, such as precision, recall and F1 score. Owing to the inclusion of nonlinear structures among features, this study has addressed the limitation of PCA and has shown how the accuracy of models can enhance the prediction of disease in the real-world field, providing an efficient expansion for including important features even if their relationship is nonlinear. Moreover, the dataset was prepared by managing missing values with a KNN imputer and transforming the dataset for efficient training, and the dataset was split into training and testing sets. The datasets used in the study are obtained from the UCI repository on two datasets with 400 and 200 instances on DB1 and DB2, respectively. The study proposed a kernel PCA with 20 components that were used for training the MLP and achieved good performance, and eight (8) different MLs were compared and evaluated for the prediction of CKD.

5. Conclusion

This study revealed the ability of the ML model to predict CKD using a low-dimensional dataset with features with nonlinear structures. This study employed eight ML models: RF, DT, MLP, MLP, RF, LR, SVM and GB. A kernel function with the PCA technique was applied to two CKD datasets, and dimensionality reduction was conducted to predict CKD. The proposed PCA with MLP performed well and produced the best result scores by outperforming the other algorithms. Based on the results achieved in this research, the model can be utilized for the early prediction of CKD. The future work of this study involves exploring the new enhanced PCA to other datasets for generalization and the development of an application that can be used in the real world for CKD prediction from patient records and, finally, the use of RPCA on other models in different settings.

References

[1] 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," *Intell. Syst. with Appl.*, vol. 16, no. September, p. 200144, 2022.

[2] Y. Singh, M. Srivastava, S. Mahajan, M. Kukreja, and S. Dev, "Evaluation of Boosted Random Forest and Multi-Objective Artificial Neural Network for the Diagnostic Severity of Chronic

Kidney Disease (CKD) Patients," *Nanotechnoloy Perceptions*, vol. 4, pp. 629–638, 2024.

[3] W. Wang, G. Chakraborty, and B. Chakraborty, "applied sciences Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm," pp. 1–17, 2021.

[4] A. Farjana *et al.*, "Predicting Chronic Kidney Disease Using Machine Learning Algorithms," *2023 IEEE 13th Annu. Comput. Commun. Work. Conf. CCWC 2023*, no. April, pp. 1267–1271, 2023.

[5] M. D. R. Al-Mahfuz, A. Haque, A. K. M. Azad, S. A. Alyami, J. M. W. Quinn, and M. A. L. I. Moni, "Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening," *Healthcare*, vol. 9, p. 4900511, 2021.

[6] P. Ranga, V. Terlapu, D. Jayaram, S. Rakesh, and M. V. Gopalachari, "Optimizing Chronic Kidney Disease Diagnosis in Uddanam : A Smart Fusion of GA-MLP Hybrid and PCA Dimensionality Reduction," *Procedia Comput. Sci.*, vol. 230, no. 2023.

[7] Z. Mushtaq, M. F. Qureshi, M. J. Abbass, and S. M. Q. Al-fakih, "Effective Kernel-Principal Component Analysis Based Approach for Wisconsin Breast Cancer Diagnosis," *Electron. Lett.*, vol. 59, no. 2, pp. 1–4, 2023.

[8] S. Krishnamurthy, K. Ks, E. Dovgan, M. Luštrek, and B. G. Pileti˘, "Machine Learning Prediction Models for Chronic Kidney Disease Using National Health Insurance Claim Data in Taiwan," *Healthcare*, vol. 9, no. 546, pp. 1–13, 2021.

[9] L. Barivi, N. Rao, and R. Sowmya, "Chronic kidney disease prediction based on machine learning," *Int. J. Inf. Technol. Comput. Eng.*, vol. 10, no. 4, 2022.

[10] A. A. Abdullahi, A. S. Hafidz, and W. Khairunizam, "Performance Comparison of Machine Learning Algorithms for Classification of Chronic Kidney Disease (CKD)," *J. Phys. Conf. Ser.*, vol. 1529, no. 5, p. 052077, 2019.

[11] C. Kaur *et al.*, "Chronic Kidney Disease Prediction Using Machine Learning," *J. Adv. Inf. Technol.*, vol. 14, no. 2, pp. 384–391, 2023.

[12] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," *J. Pathol. Inform.*, vol. 14, no. September 2022, p. 100189, 2022.

[13] L. Rubini, P. Soundarapandian, and P. Eswaran, "Chronic Kidney Disease," UC Irvine Machine Learning Repository, 2023.

[14] 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), Thoothukudi, India*, 2020, pp. 952–957.

[15] R. R. Zebari, M. A. Abdulazeez, Q. D. Zeebaree, A. D. Zebari, and N. J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," *J. Appl. Sci. Technol. Trends*, vol. 01, no. 02, pp. 56–70, 2020.

[16] J. P. Bharadiya, "A Tutorial on Principal Component Analysis for Dimensionality Reduction in Machine Learning," *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 8, no. 5, pp. 2028–2032, 2023.

[17] M. M. El Sherbiny, E. Abdelhalim, H. El-Din Mostafa, and M. M. El-Seddik, "Classification of chronic kidney disease based on machine learning techniques," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 32, no. 2, pp. 945–955, 2023.

[18] O. I. Obaid *et al.*, "Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer," *downloadmaghaleh.com*, vol. 7, no. 4, pp. 160–166, 2018.

[19] D. H. Lubis, S. Sawaluddin, and A. Candra, "Machine Learning Model for Language Classification: Bag-of-words and Multilayer Perceptron," *J. Informatics Telecommun. Eng.*, vol. 7, no. 1, pp. 356–365, 2023.

[20] M. T. Ahmed and A. Karmakar, "Analysis of Wisconsin Breast Cancer original dataset using data mining and machine learning algorithms for breast cancer prediction," *J. Sci. Technol. Environ. Inf.*, vol. 09, no. 02, pp. 2409–7632, 2020.

[21] D. Baidya, U. Umaima, N. Islam, M. F. M. J. Shamrat, A. Pramanik, and S. Rahman, "A Deep Prediction of Chronic Kidney Disease by Employing Machine Learning Method," in *6th International Conference on Trends in Electronics and Informatics (ICOEI 2022) Tirunelveli, India*, 2022, pp. 28–30.

[22] I. I. Iliyas, I. R. Saidu, A. D. Baba, and S. Tasiu, "Prediction Of Chronic Kidney Disease Using Deep Neural Network," *FUDMA J. Sci.*, vol. 4, no. 4, pp. 34–41, 2020.

[23] S. C. Saha, "Reduced Feature based Prediction of Chronic Kidney Disease by Using Machine Learning Classifiers in A Comparative Way," 2019.

[24] N. I. Ashafuddula, B. Islam, and R. Islam, "An Intelligent Diagnostic System to Analyze Early-Stage Chronic Kidney Disease for Clinical Application," *Appl. Comput. Intell. Soft Comput.*, vol. 2023, 2023.

[25] M. U. Emon, A. M. Imran, R. Islam, S. M. Keya, R. Zannat, and Ohidujjaman, "Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches," *2021 6th Int. Conf. Inven. Comput. Technol.*, pp. 713–719, 2021.

[26] A. K. Rashid, "Diagnosing Chronic Kidney disease using Artificial Neural Network (ANN)," *J. Inf. Technol. Comput.*, vol. 4, no. 1, pp. 37–45, 2023.

[27] E. Dritsas and M. Trigka, "Machine Learning Techniques for Chronic Kidney Disease Risk Prediction," *Big Data Cogn. Comput.*, vol. 6, no. 3, p. 98, 2022.