

## Research Article

# An Intelligent Diagnostic System to Analyze Early-Stage Chronic Kidney Disease for Clinical Application

N. I. Md. Ashafuddula <sup>1</sup>, Bayezid Islam <sup>2</sup>, and Rafiqul Islam <sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, Dhaka University of Engineering & Technology, Gazipur 1707, Bangladesh

<sup>2</sup>Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi 6204, Bangladesh

Correspondence should be addressed to Rafiqul Islam; rafiqul.islam@duet.ac.bd

Received 15 May 2023; Revised 30 September 2023; Accepted 10 November 2023; Published 22 November 2023

Academic Editor: Nadeem Sarwar

Copyright © 2023 N. I. Md. Ashafuddula et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chronic kidney disease (CKD) is a progressive condition characterized by the gradual deterioration of kidney functions, potentially leading to kidney failure if not promptly diagnosed and treated. Machine learning (ML) algorithms have shown significant promise in disease diagnosis, but in healthcare, clinical data pose challenges: missing values, noisy inputs, and redundant features, affecting early-stage CKD prediction. Thus, this study presents a novel, fully automated machine learning approach to tackle these complexities by incorporating feature selection (FS) and feature space reduction (FSR) techniques, leading to a substantial enhancement of the model's performance. A data balancing technique is also employed during preprocessing to address data imbalance issue that is commonly encountered in clinical contexts. Finally, for reliable CKD classification, an ensemble characteristics-based classifier is encouraged. The effectiveness of our approach is rigorously validated and assessed on multiple datasets, and the clinical relevancy of the strategy is evaluated on the real-world therapeutic data collected from Bangladeshi patients. The study establishes the dominance of adaptive boosting, logistic regression, and passive aggressive ML classifiers with 96.48% accuracy in forecasting unseen therapeutic CKD data, particularly in early-stage cases. Furthermore, the effectiveness of the FSR technique in reducing the prediction time significantly is revealed. The outstanding performance of the proposed model demonstrates its effectiveness in addressing the complexity of healthcare CKD data by incorporating the FS and FSR techniques. This highlights its potential as a promising computer-aided diagnosis tool for doctors, enabling early interventions and improving patient outcomes.

## 1. Introduction

The kidneys filter about 120 to 150 quarts of blood per day to generate approximately 1 to 2 quarts of urine [1, 2]. The primary function of the kidneys is to remove waste from the body's fluids via urine. CKD starts with unexpected metabolic disorders that gradually refer to the loss of endocrine, excretory, and metabolic functions in the kidneys [3]. These unusualities are evident as the signs and symptoms of renal damage. Since the underlying cause of the disorder stays unspecified in many patients, the most common causes can be diabetes, hypertension, interstitial diseases, systemic inflammatory disorders, glomerular diseases, congenital conditions, and renovascular abnormalities [4].

In the absence of timely treatment, kidney disease progresses to end-stage renal failure (ESRF), which causes coma and even death in patients [5]. According to [6], approximately 750,000 patients annually are affected by renal failure in the United States, with an estimated 2 million people globally suffering from kidney failure, and the diagnosed patient rate rises at a 5–7% rate annually. Over the past decade, the overall CKD mortality rate has shown a substantial increase at 31.7% [7]. Studies exploit the fact that, in low- and middle-income countries, CKD is a more significant burden when compared to high-income countries [7–10]. The number of patients diagnosed with renal disease in South Asian cities is 7.2%–17.2% [11]. The regularity reports that 13% of all the available populations in

Dhaka city are aged 15 years, or older [12]. About one-third of Bangladesh's rural people have incurable renal failure risk, as suggested by another community-based report [13]. Hence, CKD poses an upright threat in a developing country like Bangladesh.

A computer-aided diagnosis process can leverage an effective CKD diagnosis for accurate detection at the primary stage. ML is now one of the most essential and prosperous areas in the healthcare sectors for analyzing and making predictions for different diseases and stages [14]. The ML models gain knowledge by exploring large datasets and their features, patterns, modes, and so on. In data analysis, the FS strategy is used to select a subset of the most relevant features in the dataset to improve the performance and interpretability of the ML models, while the FSR technique aids in simplifying the feature representation and overall complexity in the dataset by extracting the principal components [15].

Previous research has shown that choosing the most relevant and useful features can improve early-stage CKD detection. Some researchers have used FS techniques and others have used FSR techniques. However, combining both has not been fully explored, resulting in limitations in reaching a maximum accuracy while keeping the ML model's generalization capabilities for clinical CKD diagnosis. Moreover, analyzing healthcare table data related to CKD is challenging due to missing or null attributes and categorical values in the dataset. A data encoding methodology is generally well-suited for categorical values, but a suitable strategy for addressing missing or null attribute values that takes into consideration the dataset's random nature is required. Though the existing studies have used various methods to overcome these issues, their effectiveness in dealing with unseen clinical data has not been fully established. Moreover, there are still a number of issues, such as a lack of standardization of CKD, models' interpretability, generalizability, and fairness, in order to ensure their safe use in normal clinical trials [16, 17].

Therefore, this work aims to extend renal disease diagnosis in a clinical setting by effectively utilizing computer intelligence. To achieve this goal, both the FS and FSR techniques are employed in the preprocessing phase. In addition, a data balancing strategy, as well as data encoding and cleaning, is used to account for clinically unseen data that is imbalanced, missing, or noisy. Finally, multiple classification models are incorporated, with adaptive boosting, logistic regression, and passive aggressive being the recommended ML models for CKD analysis due to their ensemble capabilities.

The effectiveness of the proposed intelligent diagnostic system is evaluated on multiple datasets separately. Finally, the clinical CKD detection performances are evaluated on unseen healthcare data collected from Bangladeshi patients. This study increased the model performances in clinical CKD detection by handling missing values, imbalanced data, data encoding, feature selection, and dimension reduction effectively. To sum up, the most significant contributions of this work are as follows:

- (1) The datasets are analyzed to ensure that no data loss occurs, even in the case of missing value.

- (2) The dimension reduction methodology is investigated in order to reduce the feature space; as a result, the model training and testing time could be reduced while simultaneously improving the overall results.
- (3) This study presented a generalized intelligent diagnostic system to analyze and predict renal disease at an early stage with unseen healthcare data. To the best of our knowledge, this is the first work on CKD prediction with clinical unseen data.
- (4) A comprehensive analysis was performed on four different datasets to find the best ML models for CKD analysis.
- (5) Adaptive boosting, logistic regression, and passive aggressive techniques are recommended classifiers for CKD analysis on unseen real-life data due to their robust ensemble capabilities.

The rest of the paper is organized as follows: The related literature review is discussed in Section 2. The proposed methodology is categorized into subsections and briefly discussed in Section 3. Data encoding, balancing, cleaning, feature selection, and dimension reduction techniques are discussed in Section 3.1. The experimental analysis is discussed in Section 4.3. Dataset collection and dataset descriptions are stated in Section 4.1. In Section 4.3 performance evaluation metrics and experimental results are discussed concerning different methods and datasets. Finally, the discussion and conclusion are delivered in Sections 4.4 and 5, respectively.

## 2. Literature Review

For effective disease classification and prediction, various methodologies are designed and explored. The study [18] examined 12 ML classifiers across four distinct datasets: breast cancer, liver disorders, wine quality, and Indian liver patients. The evaluation primarily focused on accuracy and prediction speed. They concluded that the classifier's performances are disease specific. However, this study did not elaborate on how the data complexity was handled, and the clinical relevancy was not discussed. As CKD is among the life-threatening diseases that necessitate early detection to enhance patient outcomes, researchers have explored numerous ML algorithms coupled with preprocessing techniques for efficient CKD prediction. A synthetic minority oversampling technique (SMOTE) is employed in [19] to balance the CKD-15 dataset. The authors tested three different FS methods including correlation-based feature selection (CFS) as a filter method, forward feature selection (FFS) as a wrapper method, and the least absolute shrinkage and selection operator (LASSO) feature selection as an embedded feature selection method. The data balancing with SMOTE and FS with LASSO resulted in an increase of 1.39% accuracy compared to using a linear support vector machine (LSVM) with the original dataset. The authors in [20] performed an FS strategy using a genetic algorithm (GA). They achieved the highest accuracy of 99.75% from the

multilayer perceptron (MLP) classifier. Different feature-based prediction models were suggested in [21] for detecting kidney disease in which logistic regression with a Chi-square test-based model showed the highest accuracy (98.75%). Similar ML-based models but different applications were analyzed by the authors in [22, 23]. In their work, the gradient boosting-based model was utilized in which their major finding was to utilize the FS and sampling techniques (SMOTE, OneR, etc.) for achieving favorable accuracy. The fuzzy-based intelligent system that incorporated fuzzification, implication, and defuzzification was proposed in [24] for CKD analysis. They modeled an IF-THEN fashion to develop the knowledge base for a fuzzy inference system. A summary of the data imbalanced analysis was presented by the authors [25]. The study investigated 23 class imbalanced techniques (resampling and hybrid systems) with three ML classifiers including random forest (RF), logistic regression (LR), and linear support vector classifier (LinearSVC) to identify the most suitable imbalanced method for the medical dataset. They found that class imbalance learning can significantly improve classification, with random oversampling (ROS) and RF delivering the best results.

Several other FS methods have been explored to identify the most relevant features. The L1-regulated FS technique has been explored in [26] to classify microarray cancer data with improved performance. The authors in [27] applied L1-norm-based and chi-square-based FS strategies to classify breast cancer. In other CKD studies [28–30], principal component analysis (PCA) is utilized to extract noteworthy features from the dataset. The authors [28] extracted 19 features using PCA and achieved the highest accuracy of 98% using the support vector machine (SVM) classifier. Other classifiers such as LR, naive Bayes (NB), and k-nearest neighbor (KNN) also demonstrated noteworthy performance. The study [31] utilizes PCA, discriminant analysis (DA), and LR to extract features from the breast cancer dataset. While achieving notable accuracy with a hybrid feature extraction technique, discriminant logistic (DA-LR), the study failed to discuss the data complexity, such as data balancing and cleaning issues. The authors in [32] performed their experimental analysis on the CKD-15 dataset without employing a feature optimization strategy. Despite this, they were able to attain an accuracy of 97.25% using MLP as the classifier, 96.5% using LR, and 95.75% using NB. The highest accuracy of 98.25% was achieved using SVM as the classifier. Another study [30, 33] worked with handling nominal attributes and observed the feature selection strategy in performance analysis. The nominal attributes were transformed into binary attributes, and then they conducted a best-fit feature selection (BFFS) method. According to their findings, SVM and KNN outperformed LR and decision tree (DT) classifiers, with accuracy rates of 98.3% and 98.1%, respectively. Non-numerical data of the CKD-15 dataset were transformed into binary data in the study [34]. The authors aimed to identify the most significant clinical test attributes by using SHapley Additive exPlanations (SHAP) values and reducing the number of attributes to a minimum for optimal clinical testing and high CKD detection accuracy. Among the tested classifiers, the RF achieved the

highest accuracy of 99.5%, while gradient boosting (GB), extreme gradient boosting (XGB), LR, and SVM also performed well with high accuracy.

To handle the missing values in the CKD dataset, the authors in [28, 30, 34, 35] replaced the missing values with the mean value. The missing values are handled in [3] with the mean, median, and mode values of the attributes and also dropped the null values. The authors in [36] utilized mutual information measures (MIMs) for feature selection and replaced missing values through multiple imputations while analyzing kidney disease. The authors in [37] used the median technique to replace the missing values. Other studies in [38, 39] replaced the missing values with 0. The top accuracy of 99.1% was achieved by decision forest (DF) and 97.5% while implementing NN with an arbitrary selection of 14 attributes [38]. Other authors [35] have selected 13 out of 24 attributes for classification, and the results showed that adaptive boosting (ADAB) achieved a prediction accuracy of 99% while the extra-tree classifier (ETC) obtained 98% accuracy. The authors in [39] considered 21 attributes from the CKD-15 dataset. During the classification phase, the DF achieved the highest prediction accuracy (99.17%) to predict three different potassium zones: LR 89.17% and NN 82.15%. The authors in [28] handled the categorical variables by converting them to a corresponding numerical value utilizing the one-hot encoding technique. They found the best performance of 98.0% accuracy using an SVM classifier. In a study [1], the attributes with more than 20% missing values were removed from the dataset, and the remaining values were filled using KNN imputation. The authors then selected features based on statistical significance, medical importance, and test data availability. Eleven ML algorithms were evaluated, and four classifiers (DT, RF, ETC, and ADAB classifier) showed 100% accuracy.

The comprehensive literature review highlights various techniques and approaches employed in disease prediction, particularly early CKD diagnosis and reveals common data preprocessing techniques such as nominal-to-binary transformation and one-hot encoding for categorical variables. Handling missing data involves methods like mean imputation, median, mode, multiple imputations, or replacement with 0. Existing studies often focus solely on feature selection or reduction techniques. For CKD prediction, popular methods included CFS, FFS, LASSO, GA, Chi-square, BestFit, SHAP, MIM, and PCA. Breast cancer classification is employed, while L1-regulated and L1-norm-based feature selections are used for efficient breast cancer classification. While these studies demonstrated high accuracy on the datasets utilized for training and testing by splitting them, a critical gap emerged. None of them combined feature selection and reduction methods to improve model performance, particularly in better handling clinical CKD data complexity. In addition, they also lack the assessment of the performance of their models on real-world, unseen clinical CKD data to provide patient-centric CKD solutions at the initial phase. These raise the necessity for an improved automated diagnostic system for CKD detection. This study aims to address these gaps by introducing a novel methodology tailored to enhance CKD

diagnostic accuracy and handle data complexity in a patient-centric manner.

### 3. Proposed Methodology

Preprocessing and classification are the two parts of the proposed methodology. In the preprocessing step, data encoding, balancing, cleaning, feature selection, and dimension reduction approaches were implemented to properly train the ML algorithms. The entire block diagram of the proposed methodology is shown in Figure 1.

*3.1. Preprocessing.* The datasets contain a mixture of numerical, categorical, nominal, and missing values, so the data are preprocessed to address the issues with categorical, nominal, and missing data. Before starting the preprocessing phase, the “Affected” attribute is manually omitted from the processed data, thus the processed data could not be affected by the class variables.

*3.1.1. Categorical Variable.* Variables with two or more categories but without intrinsic ordering to the categories are known as categorical variables, often known as nominal variables [40]. Categorical variables are the types of data that may be divided into groups. For example, the categorical variables are age, sex, group, race, educational level, etc.

*3.1.2. Data Encoding.* Data encoding is the process of converting data or a given sequence of characters, symbols, alphabets, etc., into a specific format that can be processed by a computer system or application. The purpose of data encoding is to transform the data into a standard format. This study utilized the label encoding or ordinal encoding technique to complete this task. All the non-numerical (nominal categorical variables) labels are mapped to numerical labels using this encoding (Table 1).

*3.1.3. Data Balancing.* Data balancing is a procedure in which the amount of class data is equalized using different data balancing techniques. This analysis used two datasets, CKD-15, and CKD-21; both datasets were imbalanced. CKD-15 contains 250 CKD and 150 non-CKD instances, and CKD-21 contains 78 CKD and 122 non-CKD data. To address this imbalance and prevent potential bias and poor model generalization, the ROS technique was employed to increase the lower number of instances. ROS duplicates minority class examples randomly, ensuring an equal representation of CKD and non-CKD instances in both datasets.

Table 2 shows the data imbalance for both datasets. Table 3 shows the amount of data after balancing the data using the sampling technique.

*3.1.4. Data Cleaning.* Missing entries are common in clinical CKD data due to the challenges of tackling a large number of CKD patients within a limited time by medical assistants. However, simply removing instances with missing data can

pose issues for accurate classification by ML models. In addition, to ensure accurate and reliable outcomes, it is crucial to avoid bias and data distortion caused by incomplete or erroneous data. This necessitates employing a data imputation technique tailored to the specific disease characteristics. Here, the study addressed the lost data by filling up the mean value of the corresponding attribute based on how the missing values were distributed randomly. It serves to preserve the statistical properties of the dataset while ensuring accuracy and reliability in subsequent analyses.

*3.1.5. Feature Selection.* As the increasing number of features creates computation overhead and increased model overfitting possibilities, FS comes into the solution [41]. The FS strategy reduces the input variables by using only relevant data and eliminating unnecessary and noisy data [42]. It is an automatically relevant feature-choosing process. The significant advantage of using this technique is that it reduces overfitting [43]. Regularization is a useful technique for reducing model complexity and feature selection [26]. The penalty “L1” (Lasso regularization) and solver “liblinear” are used here with the “LogisticRegression” method to select essential features based on the importance weights. It employs the shrinking strategy by penalizing the least-square errors. To minimize the cost function, the model set the weights of some features to zero, and a total of 13 features are chosen for CKD-15, CKD-21, hybrid, and unseen clinical data.

*3.1.6. Dimension Reduction.* Dimension reduction is a process that reduces the feature space to the most relevant feature space [15, 44] while preserving the maximum amount of relevant information from the actual data. This technique can enormously reduce the time complexity of the ML algorithm’s training phase, and it does not degrade ML model performance [45]. Among other dimension reduction techniques including PCA, singular value decomposition (SVD), linear discriminant analysis (LDA), and generalized discriminant analysis (GDA), an unsupervised ML technique, PCA, is employed here due to its effectiveness and popularity in feature reduction particularly in CKD analysis. PCA employs mathematical principles to reduce a large number of potentially correlated variables to a smaller number of variables (lower dimension), which are referred to as principal components [46]. This investigation utilized PCA as the dimension reduction strategy in four ways to prepare 4 categories of datasets.

For effective PCA analysis, the features of datapoints “X” are standardized through mean removal and scaling to unit variance using the following equation to ensure equal feature scaling:

$$X = \frac{X - \text{mean}(X)}{\text{std}(X)}. \quad (1)$$

To determine the direction in which the features are most correlated, the covariance matrix “COV” is calculated

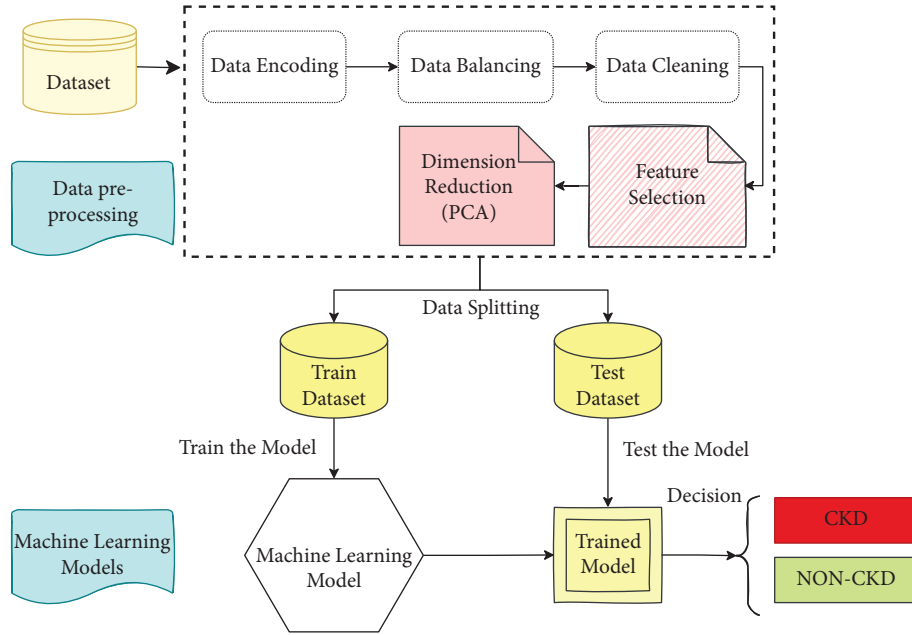


FIGURE 1: Proposed intelligent system for clinically early-stage chronic kidney disease diagnosis.

TABLE 1: Data representation from non-numerical to numerical label using encoding operation.

Attribute name	Attribute value	Mapping value
DD/MM	01-Jan	1 (month index)
A range	2-4	3 (mean between them)
Less than	5	4 (immediate lesser decimal than the given value)
Less than zero	0	0 (same as the given value)
Less or equal	5	5 (same as the given value)
Greater than	2	3 (immediate greater decimal than the given value)
Greater or equal	5	5 (same as the given value)
Examples	1.019-1.021	1.02
	1.009-1.011	1.01
	$\geq 1.023$	2
	$< 1.007$	0
	01-Jan	1

using equation (2). It is a square matrix with dimensions equal to the number of features, and each element in the matrix represents the covariance between two features, indicating their linear association.

$$\text{COV}(X_i, X_j) = \frac{\sum (X_i - X_{\text{mean}}) * (X_j - X_{\text{mean}})}{N} \quad (2)$$

Here,  $N$  = number of samples in the dataset.

The eigenvectors and eigenvalues of the “COV” matrix are computed from equation (3). They determine the directions in which the features are most varied and the amount of variance explained by each component. The eigenvalues and their corresponding eigenvectors are sorted in descending order, with the largest eigenvalues being considered the principal components for projecting the data onto a lower-dimensional space.

$$(\text{COV} - \lambda I)\nu = 0. \quad (3)$$

Here,  $\lambda$  is the eigenvalue,  $I$  is the identity matrix, and  $\nu$  is the eigenvector.

Then, the first “ $k$ ” values are chosen as the largest eigenvalues and their corresponding eigenvectors to form a matrix for the projection step to reduce data dimensionality utilizing the following equation:

$$\nu_{\text{reduced}} = \nu[:, 0:k]. \quad (4)$$

In the experiment, the study used  $k=2$ , for CKD-15,  $k=7$ , for CKD-21,  $k=3$ , for hybrid data, and  $k=10$ , for clinical unseen data.

$$X_{\text{reduced}} = X \times \nu_{\text{reduced}}. \quad (5)$$

Finally, the new feature vectors are calculated from equation (5). In the experimental analysis, the “ $k$ ” values are chosen such that they are minimal and outperform the existing models.

TABLE 2: Imbalanced dataset CKD-15 and CKD-21.

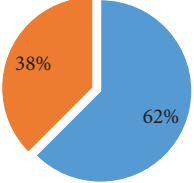
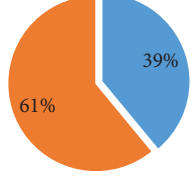
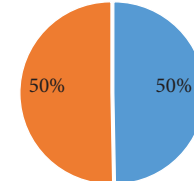
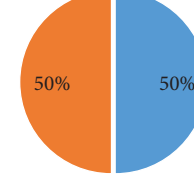
Dataset	CKD	Non-CKD	Plotting data
CKD-15	250	150	IMBALANCED CKD-15 
			<ul style="list-style-type: none"> <li><span style="color: blue;">■</span> CKD</li> <li><span style="color: orange;">■</span> NON-CKD</li> </ul>
CKD-21	78	122	IMBALANCED CKD-21 
			<ul style="list-style-type: none"> <li><span style="color: blue;">■</span> CKD</li> <li><span style="color: orange;">■</span> NON-CKD</li> </ul>

TABLE 3: Balanced dataset CKD-15 and CKD-21.

Dataset	CKD	Non-CKD	Plotting data
CKD-15	250	253	IMBALANCED CKD-15 
			<ul style="list-style-type: none"> <li><span style="color: blue;">■</span> CKD</li> <li><span style="color: orange;">■</span> NON-CKD</li> </ul>
CKD-21	128	128	IMBALANCED CKD-21 
			<ul style="list-style-type: none"> <li><span style="color: blue;">■</span> CKD</li> <li><span style="color: orange;">■</span> NON-CKD</li> </ul>

3.2. *Classification.* The advantage of the ML algorithm is that it can adapt to various cases by observing them. Throughout this paper, twelve supervised classification ML algorithms are picked and compared to detect CKD early on in different scenarios.

3.2.1. *Logistic Regression.* The logistic regression model estimates the possibility of an event within a particular class [47].

LR is commonly used for binary classification, although its title incorporates “regression.” A decision boundary is

a value that is set to predict the data class. The sigmoid activation function is used here to compute this classification probability. The mathematical model of the algorithm can be denoted as in the following equation:

$$P_i = \frac{1}{1 + e^{-\sum_{j=0}^m \beta_j x_{ij}}}, \quad (6)$$

where  $i=1$  to  $N$  (number of observations),  $j=1$  to  $M$  (number of individual variables),  $P_i$  = probability of “1” at observation  $i$ ,  $\beta_j$  = regression coefficient, and  $x_{ij}$  = the  $j^{\text{th}}$  variable at observation  $i$ .

3.2.2. *Decision Tree.* The basic goal of the decision tree algorithm is to generate a prediction model from a set of training data sets to predict classes or values of target variables. The DT algorithm is structured like a tree, with leaves, branches, and roots. When compared to other classification algorithms, the DT algorithm is simple to grasp.

3.2.3. *Random Forest.* This algorithm creates multiple decision trees during training and provides an output class of individual trees [48]. This method incorporates the decorrelated tree by building a substantial range of decision trees on bootstrapped samples from the training dataset. It screens a few feature columns among all feature columns throughout bootstrapping. Gini impurity is used in the experiment with ten maximum depths of the tree. The tree grows with ten maximum leaf nodes. Predictions for unknown data after training can be defined as in the following equation:

$$f' = \frac{1}{B} \sum_{b=1}^B f_b(x'), \quad (7)$$

where  $B$  = optimal number of trees and  $f_b(x')$  = prediction from the  $i$ -th decision tree for the unknown sample  $x'$ .

Also, the uncertainty ( $\sigma$ ) of the prediction is defined by the following equation:

$$\sigma = \sqrt{\frac{\sum_{b=1}^B f_b(x' - f')^2}{B-1}}. \quad (8)$$

3.2.4. *Passive Aggressive Classifier.* The passive aggressive classifier (PAC) is one of the online learning algorithms in ML. It responds passively to correct classifications and aggressively to any miscalculation. Generally, large-scale learning works better. In contrast to batch learning, where the entire training dataset is utilized at once, input in online ML algorithms is received sequentially, and the ML model is gradually updated. Fifty passes over the training data are used in the experiment.

3.2.5. *Support Vector Machines.* The SVM is built upon a statistical learning framework, providing solutions for both

regression and classification problems [49]. SVM can categorize both linear and nonlinear datasets by using the kernel trick. As a subset of training points in the decision function, it is also computationally efficient (called support vectors). The prediction function of an SVM classifier can be described by the following equation (9). The “rbf” kernel and regularization parameter “1” are used in the experiment.

$$f(x) = \beta_o + \sum_{i \in S} a_i K(x_i, x'_i), \quad (9)$$

where  $x$  = new data point,  $\beta_o$  = bias,  $S$  = set of support vectors,  $a_i$  = corresponding weights of the training data  $x_i$ , and  $x'_i$  are support vectors in the training data.

**3.2.6. K-Nearest Neighbor.** KNN is the most straightforward supervised ML algorithm [50]. A distance is calculated to determine similarities with other instances. For example, the closest data point to the point under observation is thought to be the most appropriate for the data point. There are numerous distance metrics for calculating the nearest point, such as Euclidean, Hamming, Manhattan, Cosine, Jaccard, and Minkowski distances. In the experiment, 7 neighbors are used with the Euclidean distance 10. Here,  $p$  and  $q$  are two points in the space,  $p_i$  and  $q_i$  are the  $i$ <sup>th</sup> dimensions of points  $p$  and  $q$ , and  $n$  is the number of dimensions.

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}. \quad (10)$$

**3.2.7. Gradient Boosting.** This classifier [51] is also operated to estimate the prediction performance as a boosting algorithm. The primary stages of a GB classifier are computing the error residual, learning a regression predictor, and memorizing to predict the residual. Additive models are usually utilized, and weak learners are counted to optimize the loss function. For weak learners, decision trees (regression trees) are employed.

**3.2.8. Naive Bayes.** The NB is a probabilistic supervised algorithm while classifying data imposes independence of features [52]. The method works effectively for datasets with a significant number of input variables. It assumes all the features available, including weak features, in the final prediction. The probabilistic naive Bayes ML model can be stated as the following equation (11) where  $A$  and  $B$  are two independent events.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}. \quad (11)$$

**3.2.9. Stochastic Gradient Descent.** The word “stochastic” denotes a system or process connected with a random probability. Hence, for each iteration, a few samples are selected randomly instead of the whole data in stochastic gradient descent (SGD). To perform each iteration, SGD

uses only a sample, i.e., a batch size of one. The sample is shuffled randomly and picked for executing the iteration. To train the model, L1 regularization and 20 epochs are used.

**3.2.10. Multilayer Perceptron.** A multilayer perceptron is considered to be the most significant class of feed-forward artificial neural networks (ANNs) that is made up of several layers of perceptron [52]. The network contains three layers where at least one hidden layer is required, and others are the input and the output layer. This experiment used the sigmoid activation function and “lbfgs” solver which is an optimizer in the family of quasi-Newton methods.

**3.2.11. Adaptive Boosting.** The adaptive boosting algorithm also known as AdaBoost is an ensemble ML technique that merges a number of weak classifiers to form a stronger classifier to increase the classification performance [53]. The performance of this model is improved by using extra copies of the classifier on the same dataset; for incorrectly classified samples, weights are adjusted to represent the final output of the boosted classifier.

**3.2.12. Extreme Learning Machine.** An extreme learning machine (ELM) is a single hidden layer feed-forward neural network that solves problems by finding the minimum norm least-square (MNLS) solution of a system [54]. It provides good generalization performance by solving problems in a single iteration at an extremely fast speed. The model Moore–Penrose generalized inverse is used to set its weights. In this experiment, 150 hidden nodes with the sigmoid activation function are used. The output of this model is calculated using the following equation:

$$f_L(x) = \sum_{i=1}^L \beta_i g_i(x). \quad (12)$$

Here,  $x$  represents the input feature vector, and the prediction is made by summing the product of the weights “ $\beta_i$ ” and the activation function “ $g_i(x)$ ” for each hidden node “ $i$ ” in the hidden layer.

## 4. Performance Evaluation

Statistically, finding the best ML classifiers is difficult because it relies on the type of application and the data format. Therefore, the focus of this work is on experimentally validating all ML models in terms of CKD analysis. Based on the data, both balanced and imbalanced conclusions can be drawn about the most effective models for the application.

**4.1. Dataset Description.** To substantiate the clinical relevancy of this study and demonstrate the effectiveness of ML techniques enhanced by feature selection and reduction, this work employed two distinct datasets: the chronic kidney disease dataset, 2015 (CKD-15) [55] and the chronic kidney disease dataset, 2021 (CKD-21) [56]. These datasets

represent diverse patient clinical data and were obtained from the “UCI Machine Learning Repository.”

*4.1.1. CKD-15 Dataset.* The CKD-15 dataset [55] comprises clinical data collected from the southern part of India with an age range of patients between 2 and 90 years. This dataset encompasses 400 instances, which are classified into two distinct categories: “ckd” (chronic kidney disease) and “notckd” (without chronic kidney disease). Each sample has 25 features, with 24 being predictive variables (11 numeric and 14 nominal). Notably, the dataset exhibits a significant class imbalance, with 250 instances classified as CKD and 150 as not-CKD.

*4.1.2. CKD-21 Dataset.* The CKD-21 dataset [56] comprises real-world patient data collected from Enam Medical College, Savar, Dhaka, Bangladesh. It consists of 200 samples, including 78 cases classified as “ckd” and 122 cases as “notckd.” The dataset contains a total of 29 attributes, which are of three types: (i) numerical values, (ii) categorical values, and (iii) nominal values. Within these 29 feature sets, the target values are represented in two specific features, denoted as “class” and “affected.”

Both datasets have a significant number of missing values, especially in the CKD-15 dataset. Tables 4 and 5 contain a description of the attributes with the necessary information for the CKD-15 and CKD-21 datasets, respectively. To apply ML algorithms, data must be well structured and reliable.

*4.2. Training and Testing.* To train and test the proposed model, two datasets, namely, CKD-15 and the real-world clinical dataset CKD-21 are used in four ways. The extensive experimentation on different datasets ((i) CKD-15, (ii) CKD-21, (iii) hybrid, and (iv) unseen clinical cases) with the combination of multiple evaluation metrics strengthens the validity of the work and demonstrates the proposed model’s generalization capability for early CKD prediction in a clinical setting. Furthermore, validating the model on clinically unseen data highlights its clinical relevance for CKD detection. In the CKD-21 dataset, “Affected” and “Class” attributes have the same meaning.

As both datasets have different dimensions, PCA helps here to bring them to the same number of dimensions for hybrid and unseen cases.

- (1) For both the CKD-15 and CKD-21, the model was trained with 70% of the data and tested with the rest 30% of the data, as depicted in Figure 2(a).
- (2) A hybrid dataset is created by utilizing both the CKD-15 and CKD-21 datasets. To make a hybrid dataset, all the datasets must be in the same space. As the datasets contain different feature spaces, this analysis transformed the dimensions of the two datasets into a particular dimension utilizing PCA. Here, for both datasets, 3 feature spaces are chosen to carry out the research by configuring PCA. Then, the

vertical (row-wise) concatenation is performed on the transformed CKD-15 and CKD-21 datasets to create a new dataset. The diversity inherent in hybrid datasets significantly enhances the generalization capabilities of ML models, which is a crucial aspect when tackling real-world applications. The ML models are trained on 70% of the sample data and tested on the remaining 30%, as shown in Figure 2(a).

- (3) The study transformed the existing feature space of both datasets to 10 feature spaces by using PCA for evaluating the ML models on clinically unknown patient data. As Figure 2(b) shows, in the experiment, the model is trained using dataset CKD-15 (i.e., 503 samples) and tested the model with a clinical real dataset CKD-21 (i.e., 256 samples) for clinical analysis of the unseen data.

All three datasets (CKD-15, CKD-21, and hybrid) were additionally split using a random state argument to ensure a nonoverlapping and unbiased evaluation of the proposed approach on all datasets. This approach helps maintain the integrity of the testing process and ensures the generalizability of the model’s performance.

*4.3. Experimental Analysis.* This work utilized the PCA as a dimension reduction technique that addresses the issue of overfitting in ML models, improves computational efficiency, and enhances the model’s generalization capability, thereby reinforcing its clinical relevance. The use of multiple metrics and datasets provides a holistic assessment and reduces the likelihood of biased results. The previous studies suggest evaluating multiple classifiers comprehensively on multiple datasets using considerable evaluation metrics, recall, true negative ratio (TNR), positive predictive value (PPV), *f1*-score, area under the receiver operating characteristic (ROC-AUC) curve, and accuracy metrics that are appropriate and relevant to evaluate ML models’ performance in the context of early-stage CKD detection. These metrics are commonly used in medical and healthcare-related studies to understand each classifier’s performance in different aspects, particularly in the context of early-stage CKD detection, where sensitivity, specificity, and diagnostic accuracy are critical. Though cross-validation is a common and widely used technique for evaluating ML models, it may not be feasible for our specific datasets (unseen and hybrid) due to their unique characteristics. For instance, cross-validation on clinical unseen datasets might not provide meaningful insights as this experiment aims to simulate real-world clinical scenarios by testing the model on entirely unseen data. Similarly, for the hybrid dataset, it may introduce biases due to the combination of datasets with varying characteristics. The work fully operated on Google’s cloud platform using “Colab Notebook.”

As Table 6 recites, eleven ML models (i.e., ADAB, DT, ELM, GB, KNN, LR, MLP, PAC, RF, SGD, and SVM) with PCA performed with 100% test accuracy, and ROC-AUC value was exactly 1 in the experiment for CKD-15 dataset. Although three ML algorithms (ADAB, DT, and RF) achieve



TABLE 4: Description of different attributes of CKD-15 dataset.

Attribute	Information (values)	Missing ratio (%)
age (age)	Numerical	2.25
bp (blood pressure)	Numerical	3.0
Wbcc (white blood cell count)	Numerical	26.5
sg (specific gravity)	Nominal	11.5
al (albumin)	Nominal (0, 1, 2, 3, 4, 5)	11.5
Su (sugar)	Nominal (0, 1, 2, 3, 4, 5)	12.25
Rbc (red blood cells)	Nominal (normal, abnormal)	38.0
Pc (pus cell)	Nominal (normal, abnormal)	16.25
Pcc (pus cells clumps)	Nominal (present, not present)	1.0
ba (bacteria)	Nominal (present, not present)	1.0
bgr (blood glucose)	Random numerical in mgs/dl	11.0
bu (blood urea)	Numerical in mgs/dl	4.75
sc (serum creatinine)	Numerical	4.25
Pcv packed cell volume)	Numerical	17.75
sod (sodium)	Numerical in mEq/L	21.75
pot (potassium)	Numerical in mEq/L	22.0
hemo (hemoglobin)	Numerical in mEq/L	13.0
rc (red blood cell count)	Numeric	32.75
htn (hypertension)	Nominal (yes, no)	0.5
dm (diabetes mellitus)	Nominal (yes, no)	0.5
Cad (coronary artery disease)	Nominal (yes, no)	0.5
pe (pedal edema)	Nominal (yes, no)	0.25
ane (anemia)	Nominal (yes, no)	0.25
appet (appetite)	Nominal (good, poor)	0.25
class (classification)	Nominal (CKD, not CKD)	0

TABLE 5: Description of different attributes of CKD-21 dataset.

Attribute	Information	Missing ratio (%)
bp (blood pressure) (diastolic)	Numeric	0.0
bp (blood pressure) (limit)	Numeric	0.0
rbc (red blood cells)	Numeric	0.0
pc (pus cell)	Numeric	0.0
pcc (pus cells clumps)	Numeric	0.0
ba (bacteria)	Numeric	0.0
htn (hypertension)	Numeric	0.0
dm (diabetes mellitus)	Numeric	0.0
cad (coronary artery disease)	Numeric	0.0
appet (appetite)	Numeric	0.0
pe (pedal edema)	Numeric	0.0
ane (anemia)	Numeric	0.0
stage CKD stage	Numeric	0.0
affected (duplicate of class variable)	Numeric	0.0
sg (specific gravity)	Categorical, ranging (1.009–1.021), (<1.007)	20.5
al (albumin)	Categorical, ranging (<0)	42.5
su (sugar)	Categorical, ranging (<0)	15.0
bgr (blood glucose)	Categorical, ranging (112–448), (<112)	1.5
bu (blood urea)	Categorical, ranging (48.1–276.7), (<48.1)	0.5
sod (sodium)	Categorical, ranging (118–153), (<118)	0.5
sc (serum creatinine)	Categorical, ranging (3.65–19.4), (<3.65)	0.5
pot (potassium)	Categorical, ranging (7.31–42.59), (<7.31)	0.5
hemo (hemoglobin)	Categorical, ranging (6.1–16.5), (<6.1)	7.5
rbcc (red blood cell count)	Categorical, ranging (2.69–6.82), (<2.69)	0.5
wbcc (white blood cell count)	Categorical, ranging (4980–21640), (<4980)	0.5
grf (glomerular filtration rate)	Categorical, ranging (26.6175–227.944), (<26.6175)	1.5
age (patient's age)	Categorical, ranging (26.6175–227.944), (<26.6175)	5.0
pcv (packed cell volume)	Categorical, ranging (20–74), (<12)	9.5
Class	Nominal (CKD or not CKD)	0.0

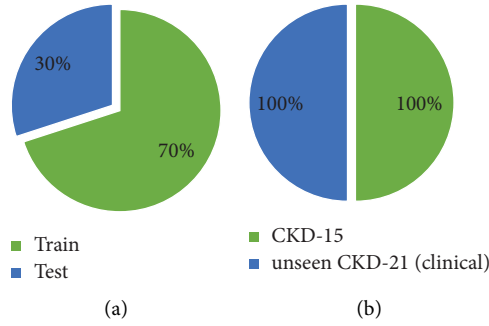


FIGURE 2: Dataset splitting for (a) CKD-15, CKD-21, and hybrid case and (b) unseen case.

TABLE 6: Experimental result analysis of various classifiers for with and without PCA on dataset-1 (CKD-15).

Classifier	Recall		TNR		PPV		F1-score		Train accuracy		Test accuracy		ROC-AUC	
	With PCA	Without PCA	With PCA	Without PCA	With PCA	Without PCA	With PCA	Without PCA	With PCA	Without PCA	With PCA	Without PCA	With PCA	Without PCA
ADAB	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>1</b>	<b>1</b>
DT	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>1</b>	<b>1</b>
ELM	<b>100</b>	88	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	94	<b>100</b>	96	<b>100</b>	94.04	<b>1</b>	0.990
GB	<b>100</b>	99	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	99	<b>100</b>	99.72	<b>100</b>	99.34	<b>1</b>	0.993
KNN	<b>100</b>	43	<b>100</b>	97	<b>100</b>	94	<b>100</b>	59	<b>100</b>	76.99	<b>100</b>	70.2	<b>1</b>	0.7
LR	<b>100</b>	95	<b>100</b>	96	<b>100</b>	96	<b>100</b>	95	<b>100</b>	93.75	<b>100</b>	95.36	<b>1</b>	0.985
MLP	<b>100</b>	88	<b>100</b>	97	<b>100</b>	97	<b>100</b>	92	<b>100</b>	94.03	<b>100</b>	92.72	<b>1</b>	0.927
PAC	<b>100</b>	83	<b>100</b>	25	<b>100</b>	52	<b>100</b>	64	<b>100</b>	58.24	<b>100</b>	53.64	<b>1</b>	0.540
RF	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>1</b>	<b>1</b>
SGD	<b>100</b>	48	<b>100</b>	99	<b>100</b>	97	<b>100</b>	64	<b>100</b>	77.84	<b>100</b>	73.51	<b>1</b>	0.748
SVM	<b>100</b>	43	<b>100</b>	83	<b>100</b>	71	<b>100</b>	54	<b>100</b>	60.8	<b>100</b>	62.91	<b>1</b>	0.628
NB	<b>92</b>	88	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>96</b>	94	<b>97.16</b>	96.31	<b>96.03</b>	94.04	0.991	<b>1</b>

The best results are indicated in bold.

100% accuracy without PCA, the other nine ML models degrade their performance. Inferior experimental performances are noticed in the KNN, PAC, SGD, and SVM classifiers without using PCA, ranging the test accuracy from 53.64% to 70.2% where PAC shows the least test accuracy of 53.64%. For the CKD-21 dataset, Table 7 shows seven ML models (i.e., ADAB, DT, ELM, GB, KNN, SVM, and RF) with PCA performing 100% accurately on test data and the lowest accuracy (94.81%) achieved by the PAC classifier. Though the SGD model's accuracy is not perfect, it has a perfect ROC-AUC curve for the CKD-21 dataset, whereas the ELM classifier's ROC-AUC value is degraded to 0.729. The proposed model without PCA could not fit the ELM, SVM, and PAC ML learning models well; hence, the overall model's performance has degraded and ranged in test accuracy from 49.35% to 54.55%. The MLP model performs the best for the hybrid dataset. Table 8 describes the best test accuracy of 99.12% for the MLP model, and the best ROC-AUC value is 0.996 for the LR model though it achieves 96.93% of test accuracy. The ML classifiers DT, GB, RF, and SGD show an equal amount of test accuracy of 97.81% and ROC-AUC of 0.978. In the clinical unseen dataset, the ADAB, LR, and PAC classifiers achieve the highest accuracy of 96.48%, and the best ROC-AUC of 0.984 is achieved by the LR model. Among the other ML models, DT and GB produce the least results (97.27% test accuracy and 0.93 ROC-AUC), as shown in Table 9. Though the NB model's

accuracy is not the best, its ROC-AUC value of 0.981 was the closest to the LR's ROC-AUC value, establishing it as the second-best well-fitted model whereas with 96.01% accuracy, RF and SGD acquire the second-best performing models for unseen clinical data.

The proposed model with a dimension reduction technique (PCA) achieves the final predicted value for a classifier in an average of 1.93 seconds for CKD-15 datasets and 6.57 seconds for CKD-21 datasets. The model without PCA takes 2.33 seconds for the CKD-15 dataset and 15.9 seconds for the CKD-21 dataset, as shown in Table 10. The hybrid dataset model takes 7.7 seconds, while the unseen dataset model takes 7.95 seconds. An ML model needs to have the same dimension of datasets to create a hybrid dataset, and it also needs to have the same dimension for training and testing the model. Hence, the average required time without considering PCA could not be calculated for hybrid and unseen cases.

**4.4. Results and Discussion.** Our innovative machine learning approach is fully automated and integrates feature selection through L1 regularization and feature space reduction using PCA during the preprocessing phase. These techniques were specifically designed to address the complexities of therapeutic data in CKD diagnosis, with a primary focus on enhancing early-stage prediction accuracy.

TABLE 7: Experimental result analysis of different classifiers for with and without PCA on dataset-2 (CKD-21).

Classifier	Recall		TNR		PPV		F1-score		Train accuracy		Test accuracy		ROC-AUC	
	With PCA	Without PCA	With PCA	Without PCA	With PCA	Without PCA	With PCA	Without PCA	With PCA	Without PCA	With PCA	Without PCA	With PCA	Without PCA
ADAB	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>1</b>	<b>1</b>
DT	<b>100</b>	97	<b>100</b>	97	<b>100</b>	97	<b>100</b>	97	<b>100</b>	<b>100</b>	<b>100</b>	97.4	<b>1</b>	0.974
ELM	<b>100</b>	53	<b>100</b>	56	<b>100</b>	54	<b>100</b>	53	98	100	<b>100</b>	54.55	<b>0.729</b>	0.53
GB	<b>100</b>	97	<b>100</b>	95	<b>100</b>	95	<b>100</b>	96	<b>99.44</b>	98.88	<b>100</b>	96.1	<b>1</b>	0.961
KNN	<b>100</b>	76	<b>100</b>	95	<b>100</b>	94	<b>100</b>	84	<b>96.65</b>	86.59	<b>100</b>	85.71	<b>1</b>	0.856
SVM	<b>100</b>	76	<b>100</b>	38	<b>100</b>	55	<b>100</b>	64	<b>97.77</b>	61.45	<b>100</b>	57.14	<b>1</b>	0.574
RF	<b>100</b>	<b>100</b>	<b>100</b>	97	<b>100</b>	97	<b>100</b>	98	99.44	<b>100</b>	<b>100</b>	98.7	<b>1</b>	0.99
LR	97	<b>100</b>	<b>97</b>	92	<b>97</b>	93	<b>97</b>	96	<b>98.32</b>	96.09	<b>97.4</b>	96.1	<b>0.995</b>	0.99
NB	<b>97</b>	84	97	<b>100</b>	97	<b>100</b>	<b>97</b>	91	<b>96.65</b>	93.85	<b>97.4</b>	92.21	<b>0.993</b>	0.99
SGD	<b>100</b>	<b>100</b>	<b>95</b>	0	<b>95</b>	49	<b>97</b>	66	<b>98.32</b>	50.28	<b>97.4</b>	49.35	<b>1</b>	0.86
MLP	92	<b>92</b>	<b>100</b>	97	<b>100</b>	97	<b>96</b>	93	<b>100</b>	89.39	<b>96.1</b>	94.81	<b>0.987</b>	0.95
PAC	95	<b>100</b>	<b>95</b>	0	<b>95</b>	49	<b>95</b>	66	<b>98.32</b>	50.28	<b>94.81</b>	49.35	<b>0.987</b>	0.5

The best results are indicated in bold.

TABLE 8: Experimental result analysis of different classifiers for the hybrid dataset (mixture of CKD-15 and CKD-21).

Classifier	Recall	TNR	PPV	F1-score	Train accuracy	Test accuracy	ROC-AUC
MLP	0.98	1	1	0.99	100	<b>99.12</b>	0.991
DT	0.96	1	1	0.98	100	97.81	0.978
GB	0.96	1	1	0.98	100	97.81	0.978
RF	0.96	1	1	0.98	99.62	97.81	0.978
SGD	0.97	0.98	0.98	0.97	98.49	97.81	0.978
ADAB	0.95	1	1	0.97	100	97.37	0.974
KNN	0.95	1	1	0.97	98.31	97.37	0.974
SVM	0.96	0.99	0.99	0.97	98.12	97.37	0.974
LR	0.95	0.99	0.99	0.97	98.12	96.93	<b>0.996</b>
ELM	0.97	0.96	0.96	0.96	0.98	96.49	0.985
PAC	0.97	0.96	0.96	0.96	98.31	96.49	0.965
NB	0.9	1	1	0.95	96.99	95.18	0.992

The best results are indicated in bold.

TABLE 9: Experimental result analysis for clinical unseen dataset, i.e., training the model with CKD-15 and testing with CKD-21.

Classifier	Recall	TNR	PPV	F1-score	Train accuracy	Test accuracy	ROC-AUC
ADAB	0.98	0.95	0.95	0.96	100	<b>96.48</b>	0.965
LR	0.95	0.98	0.98	0.96	99.01	<b>96.48</b>	<b>0.984</b>
PAC	0.95	0.98	0.98	0.96	99.2	<b>96.48</b>	0.961
RF	0.97	0.95	0.95	0.96	100	96.09	0.961
SGD	0.97	0.95	0.95	0.96	99.8	96.09	0.938
MLP	0.96	0.95	0.95	0.95	100	95.7	0.957
SVM	0.97	0.94	0.94	0.95	99.2	95.31	0.953
NB	0.9	0.99	0.99	0.94	97.22	94.53	0.981
ELM	0.89	0.98	0.98	0.93	0.98	93.75	0.964
KNN	0.88	0.99	0.99	0.93	98.41	93.75	0.938
DT	0.97	0.89	0.9	0.93	100	92.97	0.93
GB	0.97	0.89	0.9	0.93	100	92.97	0.93

The best results are indicated in bold.

TABLE 10: Average model processing time comparison for the proposed model with and without PCA.

Dataset	With PCA (s)	Without PCA (s)
CKD-15	1.93	2.33
CKD-21	6.57	15.9
CKD-hybrid	7.7	—
CKD-unseen	7.95	—

TABLE 11: Experimental result (accuracy (%)) comparison for CKD-15 with state-of-the-art methodologies.

Methodology	NB	SVM	LR	KNN	PAC	RF	DT	GB	SGD	ADAB	MLP
Nishat et al. [3]	90.50	99.36	99.36	94.3	—	—	98.10	—	—	—	94.3
Gupta et al. [57]	—	—	99.24	—	—	94.16	98.48	—	—	—	—
Herath et al. [1]	93.33	—	95.00	98.33	—	<b>100</b>	<b>100</b>	—	—	<b>100</b>	—
Chu et al. [58]	95	—	96	—	—	99	98	—	98	96	98
Gokiladevi et al. [59]	—	73.75	94.68	67.5	—	98.75	96.25	—	—	—	—
Chittora et al. [19]	—	94.63	71.71	64.39	—	98.75	96.25	—	—	—	—
Islam et al. [30]	88.33	96.66	—	59.00	—	97.50	97.50	97.50	97.50	98.30	—
Dritsas et al. [60]	98.40	—	97.40	98.40	—	98.90	97.40	—	—	—	—
Kaur et al. [61]	—	—	—	74.00	—	96.00	95.00	—	—	—	—
Proposed model without PCA	94.04	62.91	95.36	70.2	54.3	<b>100</b>	<b>100</b>	99.34	73.51	<b>100</b>	92.72
Proposed model	<b>96.03</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>

The best results are indicated in bold.

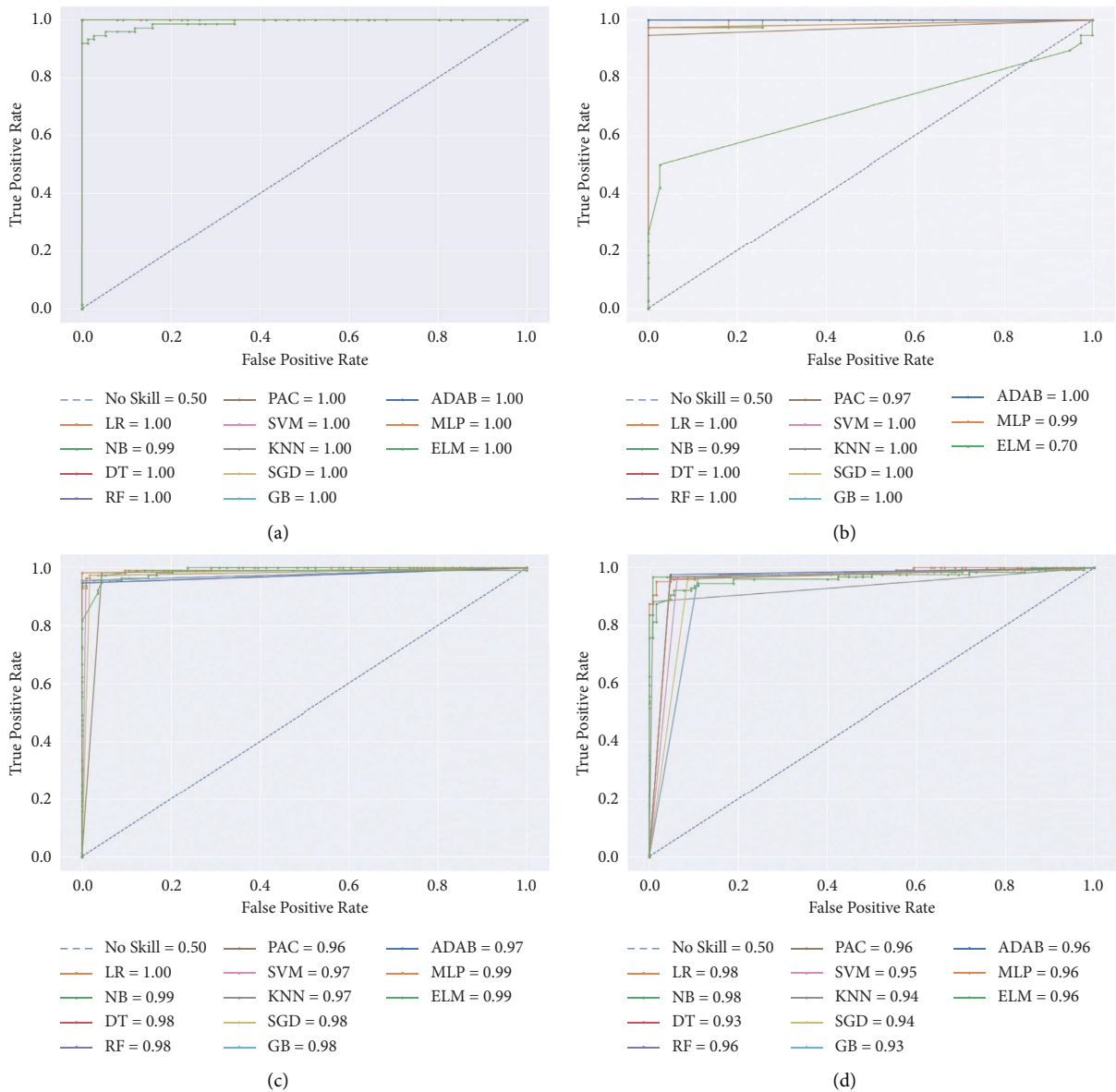


FIGURE 3: ROC-AUC curve analysis of ML models using PCA for (a) CKD-15, PCA = 2; (b) CKD-21, PCA = 7; (c) CKD-15 and CKD-21, PCA = 3; and (d) CKD-15 and CKD-21, PCA = 10.

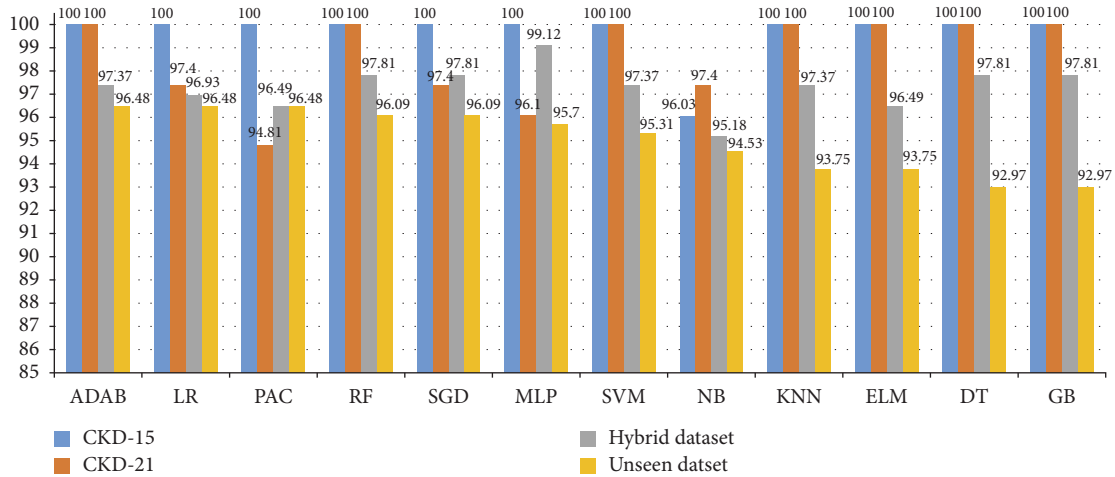


FIGURE 4: Accuracy (%) comparison graph for the four types of data using PCA.

TABLE 12: Average performance analysis of ML models on all four types of dataset.

S/L	Classifier	Training accuracy (%)	Testing accuracy (%)	ROC-AUC
1	RF	99.77	<b>98.48</b>	0.98
2	ADAB	<b>100</b>	98.46	0.98
3	SVM	98.77	98.17	0.98
4	SGD	99.15	97.83	0.98
5	KNN	98.34	97.78	0.98
6	MLP	<b>100</b>	97.73	<b>0.99</b>
7	DT	<b>100</b>	97.7	0.91
8	GB	99.86	97.7	0.98
9	LR	98.86	97.7	<b>0.99</b>
10	ELM	98.5	97.56	<b>0.99</b>
11	PAC	98.96	96.95	0.98
12	NB	97.01	95.79	<b>0.99</b>

The best results are indicated in bold.

Consequently, the intelligent model surpasses all other existing methodologies. A comprehensive analysis of the performance metrics for the four types of datasets, namely CKD-15, CKD-21, hybrid, and unseen clinical cases, is presented in Tables 6–9 subsequently. The datasets (CKD-21, hybrid, and unseen clinical cases) employed here are novel and unique for early-stage CKD detection, and there were no previous works available that used these datasets for CKD diagnosis. While conducting a direct comparison with state-of-the-art methods for these specific datasets (CKD-21, hybrid, and unseen clinical cases) was not feasible, this study thoroughly evaluated the proposed approach on the CKD-15 dataset in Table 11. The outcomes demonstrated the superiority of our approach over previous works by a wide margin for the CKD-15 dataset, establishing its effectiveness in CKD detection.

Table 6 depicts that overall, the ADAB, DT, and RF classifiers achieve better performance than other models regardless of PCA usage, while the GB classifier performs better when PCA is utilized. The performance of other models steadily decreased when PCA was not considered.

The four ML models (i.e., ELM, SVM, SGD, and PAC) perform worst without PCA for the CKD-21 dataset depicted in Table 7. To the best of our knowledge, this is the first work

done on this dataset to detect CKD from the non-CKD class. A few works have been conducted on the CKD-21 dataset but they are limited to identifying renal disease risk factors only. Furthermore, no works on CKD-hybrid and clinical unseen data are found to compare with our model outputs.

Figure 3 shows the ROC-AUC curves, and Figure 4 shows the result comparison with all the 12 ML models on four types of datasets (i.e., CKD-15, CKD-21, hybrid, and unseen case) using PCA. For the CKD-15 dataset, all models perform with 100% of accuracy except for the NB model. The PAC model performs least for the CKD-21 dataset. In aggregate, ADAB, DT, ELM, GB, KNN, SVM, and RF models perform best for both datasets.

MLP performs best for the hybrid dataset (99.12% test accuracy), and NB performs the least (95.18% test accuracy). The average performance on the hybrid dataset was relatively good and is in a steady state, whereas for the unseen clinical data, ML model performances steadily degraded from 96.48% (ADAB, LR, and PAC) to 92.97% (DT and GB) test accuracy.

To evaluate the overall ML model’s performance on the four types of the dataset, this exploration presents the average performance analysis in Table 12 and plots the average train-test accuracy and ROC-AUC value in Figures 5 and 6,

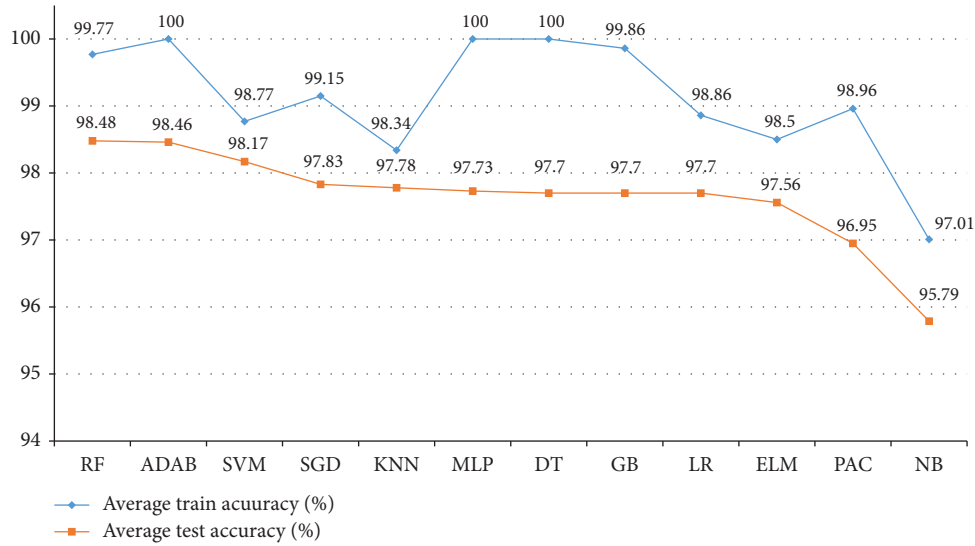


FIGURE 5: Average train-test accuracy (%) comparison of ML models using PCA on chronic kidney disease data.

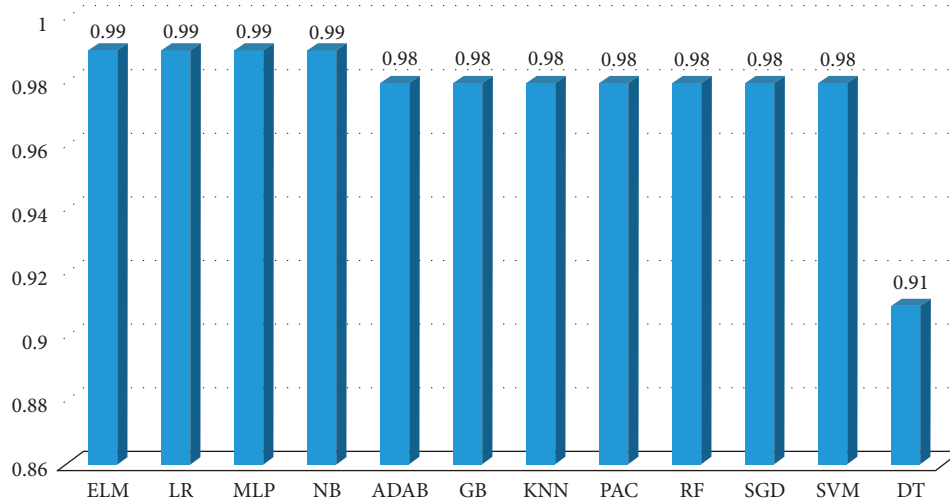


FIGURE 6: Average ROC-AUC comparison of ML models using PCA on chronic kidney disease data.

respectively. The ELM, LR, MLP, and NB show the best ROC-AUC performances at 0.99 whereas DT shows the least ROC-AUC value at 0.91, and other classifiers achieve 0.98 ROC-AUC performance in Figure 6. Figure 5 describes that on the CKD dataset, the RF classifier averagely performs the best (98.48% test accuracy and 0.98 average ROC-AUC), and ADAB acquires the second-best position with 98.46% test accuracy. Though the train-test accuracy gaps are less for the SVM, KNN, and ELM classifiers, overall, in kidney disease prediction, the RF model could be the best choice for CKD-15, CKD-21, and unseen clinical data considering the accuracy and ROC-AUC performances, and the MLP model could be the best model for hybrid renal data.

### 5. Conclusion

Kidney failure causes diseases ranging from mild to severe, has significant health implications, and demands accurate diagnosis, especially in rural areas of developing countries

where specialists are limited. To address these issues, this work suggests an intelligent diagnostic system for early CKD detection in a clinical environment with high accuracy and in a time-efficient manner. The suggested model was evaluated from four distinct perspectives to enhance its real-life clinical performance and credibility. To optimize the model’s performance, necessary corrections were made to the datasets. It outperforms previous studies for the CKD-15 dataset and exhibits impressive accuracy for test data. This positions it as a valuable novel solution and establishes its validity. The kidney disease prediction has been improved effectively by employing both the logistic regression method with the “L1” penalty as feature selection and PCA as feature space reduction technique alongside an ensemble characteristic-based classifier. It also shows notable performance for CKD-21 and hybrid datasets.

To validate the significance of this study and clinical relevancy, the proposed intelligent diagnostic system was finally evaluated on clinically unseen complex data and

achieved impressive performance, demonstrating its potential as a valuable patient-centric solution for early CKD diagnosis in clinical practice. The implementation of the model in local healthcare systems would allow for a swift assessment of patients for early-stage CKD identification. Incorporating PCA into the model improved the CKD detection performance and significantly decreased the analysis time, specifically by 0.4 seconds for the CKD-15 dataset and 9.33 seconds for the CKD-21 dataset. This states the real-life clinical applicability of the suggested model.

A future investigation might include performing statistical tests on more patient-centric data. The study acknowledges the necessity of more work on clinical benchmark data to facilitate thorough comparisons with state-of-the-art methods, especially for novel datasets like CKD-21, hybrid, and unseen clinical cases. Furthermore, validation by domain experts is a necessary step prior to clinical implementation.

### Data Availability

The data used to support the findings of this study are included within the article.

### Conflicts of Interest

The authors declare that they have no conflicts of interest.

### Acknowledgments

The authors wish to thank the Department of Computer Science and Engineering of Dhaka University of Engineering & Technology, Gazipur, for supporting research to continue the research work.

### References

- [1] I. U. Ekanayake and D. Herath, "Chronic kidney disease prediction using machine learning methods," in *Proceedings of the 2020 Moratuwa Engineering Research Conference (MERCOn)*, pp. 260–265, IEEE, Moratuwa, Sri Lanka, July 2020.
- [2] Y. D. S. Raju, K. S. Murthy, G. Vatsa, R. M. Kingston, R. Agrawal, and A. Joshi, "A novel machine learning approach chronic kidney disease prediction," in *Proceedings of the 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pp. 887–892, IEEE, Tashkent, Uzbekistan, October 2022.
- [3] M. M. Nishat, F. Faisal, R. R. Dip et al., "A comprehensive analysis on detecting chronic kidney disease by employing machine learning algorithms," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, no. 29, Article ID 170671, 2018.
- [4] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *Journal of Health and Social Behavior*, vol. 24, no. 4, pp. 385–396, 1983.
- [5] J. M. Bargman and K. L. Skorecki, *Chronic Kidney Disease*, McGraw-Hill Education, New York, NY, USA, 2018.
- [6] The Kidney Project, "Creating a bioartificial kidney as a permanent solution to kidney failure," 2019, <https://pharm.ucsf.edu/kidney/need/statistics>.
- [7] M. Mostafi, "Managing kidney diseases in Bangladesh," *Bangladesh Journal of Medicine*, vol. 33, no. 3, pp. 233–234, 2022.
- [8] J. W. Stanifer, A. Muir, T. H. Jafar, and U. D. Patel, "Chronic kidney disease in low-and middle-income countries," *Nephrology Dialysis Transplantation*, vol. 31, no. 6, pp. 868–874, 2016.
- [9] K. T. Mills, Y. Xu, W. Zhang et al., "A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010," *Kidney International*, vol. 88, no. 5, pp. 950–957, 2015.
- [10] N. M. Hustrini, "Chronic kidney disease care in Indonesia: challenges and opportunities," *Acta Medica Indonesiana*, vol. 55, no. 1, pp. 1–3, 2023.
- [11] S. Das and P. Dutta, "Chronic kidney disease prevalence among health care providers in Bangladesh," *Mymensingh Medical Journal: MMJ*, vol. 19, no. 3, pp. 415–421, 2010.
- [12] S. Anand, M. A. Khanam, J. Saquib et al., "High prevalence of chronic kidney disease in a community survey of urban bangladeshis: a cross-sectional study," *Globalization and Health*, vol. 10, no. 1, pp. 9–7, 2014.
- [13] M. N. Huda, K. S. Alam, and H. U. Rashid, "Prevalence of chronic kidney disease and its association with risk factors in disadvantaged population," *International journal of nephrology*, vol. 2012, Article ID 267329, 7 pages, 2012.
- [14] M. A.-A.-R. Asif, M. M. Nishat, F. Faisal et al., "Computer aided diagnosis of thyroid disease using machine learning algorithms," in *Proceedings of the 2020 11th International Conference on Electrical and Computer Engineering (ICECE)*, pp. 222–225, IEEE, Dhaka, Bangladesh, December 2020.
- [15] X. Yuan, Z. Chen, and Y. Wang, "Probabilistic nonlinear soft sensor modeling based on generative topographic mapping regression," *IEEE Access*, vol. 6, pp. 10445–10452, 2018.
- [16] M. Evans, R. D. Lewis, A. R. Morgan et al., "A narrative review of chronic kidney disease in clinical practice: current challenges and future perspectives," *Advances in Therapy*, vol. 39, no. 1, pp. 33–43, 2022.
- [17] S. Pal, "Chronic kidney disease prediction using machine learning techniques," *Biomedical Materials & Devices*, 2022.
- [18] J. Nasir, A. Ahsan, N. Sarwar et al., "Classification and prediction analysis of diseases and other datasets using machine learning," in *Proceedings of the Intelligent Technologies and Applications: Second International Conference, INTAP 2019*, pp. 432–442, Springer, Bahawalpur, Pakistan, November 2020.
- [19] P. Chittora, S. Chaurasia, P. Chakrabarti et al., "Prediction of chronic kidney disease-a machine learning perspective," *IEEE Access*, vol. 9, pp. 17312–17334, 2021.
- [20] W. Wang, G. Chakraborty, and B. Chakraborty, "Predicting the risk of chronic kidney disease (ckd) using machine learning algorithm," *Applied Sciences*, vol. 11, no. 1, p. 202, 2020.
- [21] R. C. Poonia, M. K. Gupta, I. Abunadi et al., "Intelligent diagnostic prediction and classification models for detection of kidney disease," *Healthcare*, vol. 10, no. 2, p. 371, 2022.
- [22] S. Shabbir, M. S. Asif, T. M. Alam, and Z. Ramzan, "Early prediction of malignant mesothelioma: an approach towards non-invasive method," *Current Bioinformatics*, vol. 16, no. 10, pp. 1257–1277, 2021.
- [23] Z. Ali, M. F. Hayat, K. Shaikat et al., "A proposed framework for early prediction of schistosomiasis," *Diagnostics*, vol. 12, no. 12, p. 3138, 2022.
- [24] T. I. Ahmed, T. I. Ahmed, J. Bhola et al., "Fuzzy logic-based systems for the diagnosis of chronic kidney disease," *BioMed*

- Research International*, vol. 2022, Article ID 2653665, 15 pages, 2022.
- [25] M. Khushi, K. Shaikat, T. M. Alam et al., "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021.
- [26] B. Shekar and G. Dagnev, "L1-regulated feature selection and classification of microarray cancer data using deep learning," in *Proceedings of the 3rd International Conference on Computer Vision and Image Processing*, pp. 227–242, Springer, Singapore, September 2020.
- [27] Z. Mushtaq, A. Yaqub, S. Sani, and A. Khalid, "Effective k-nearest neighbor classifications for Wisconsin breast cancer data sets," *Journal of the Chinese Institute of Engineers*, vol. 43, no. 1, pp. 80–92, 2020.
- [28] M. Rahman, L. Islam, M. Rana, M. Tazim, J. F. Sorna, and S. T. Alvi, "A predictive analysis of chronic kidney disease by exploring important features," *International Journal of Computing and Digital System*, vol. 11, no. 1, 2021.
- [29] I. Saha, M. K. Gourisaria, and G. Harshvardhan, "Classification system for prediction of chronic kidney disease using data mining techniques," in *Proceedings of the Advances in Data and Information Sciences: Proceedings of ICDIS 2021*, pp. 429–443, Springer, Singapore, February 2022.
- [30] 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, Article ID 100189, 2023.
- [31] T. Iqbal, A. Farooq, N. Sarwar, M. Ashraf, and A. Irshad, "Prediction of breast cancer using machine learning techniques," *BioScientific Review*, vol. 4, no. 1, pp. 59–75, 2022.
- [32] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An empirical evaluation of machine learning techniques for chronic kidney disease prophecy," *IEEE Access*, vol. 8, pp. 55012–55022, 2020.
- [33] A. Charleonnann, T. Fufaung, T. Niyomwong, W. Chokchueyattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *Proceedings of the 2016 Management and Innovation Technology International Conference (MITicon)*, IEEE, Bang-San, Thailand, October 2016.
- [34] M. Rashed-Al-Mahfuz, A. Haque, A. Azad, S. A. Alyami, J. M. Quinn, and M. A. Moni, "Clinically applicable machine learning approaches to identify attributes of chronic kidney disease (ckd) for use in low-cost diagnostic screening," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1–11, 2021.
- [35] K. Zubair Hasan and Z. Hasan, "Performance evaluation of ensemble-based machine learning techniques for prediction of chronic kidney disease," in *Emerging Research in Computing, Information, Communication and Applications*, pp. 415–426, Springer, Singapore, 2019.
- [36] M. Almasoud and T. E. Ward, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, 2019.
- [37] P. Ghosh, F. 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 *Proceedings of the 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pp. 1–6, IEEE, Bangkok, Thailand, November 2020.
- [38] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (ckd)," in *Proceedings of the 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 291–296, IEEE, Washington, DC, USA, October 2017.
- [39] M. Wickramasinghe, D. Perera, and K. Kahandawaarachchi, "Dietary prediction for patients with chronic kidney disease (ckd) by considering blood potassium level using machine learning algorithms," in *Proceedings of the 2017 IEEE Life Sciences Conference (LSC)*, pp. 300–303, IEEE, Sydney, Australia, December 2017.
- [40] S. Valcheva, "Categorical data examples and definitionaccessed," 2022, <https://www.intellspot.com/categorical-data-examples/>.
- [41] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, Article ID 022022, 2019.
- [42] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: a review," *Data Classification: Algorithms and applications*, vol. 37, 2014.
- [43] A. Alsaafin and A. Elnagar, "A minimal subset of features using feature selection for handwritten digit recognition," *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 4, pp. 55–68, 2017.
- [44] S. Arya, N. Pratap, and V. Kumar, "Enhancement in security by reducing dimensions of hyperspectral face images for face recognition," *African Journal of Computing & ICTs*, vol. 8, no. 2.
- [45] G. T. Reddy, M. P. K. Reddy, K. Lakshmana et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.
- [46] M. A. Carreira-Perpinán, "A review of dimension reduction techniques," Technical Report CS-96-09, Department of Computer Science. University of Sheffield, England, UK, 1997.
- [47] M. M. Ghiasi, S. Zendeheboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: cart model," *Computer Methods and Programs in Biomedicine*, vol. 192, Article ID 105400, 2020.
- [48] N. Jiang, F. Fu, H. Zuo, X. Zheng, and Q. Zheng, "A municipal pm2. 5 forecasting method based on random forest and wrf model," *Engineering Letters*, vol. 28, no. 2.
- [49] M. Wang and H. Chen, "Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis," *Applied Soft Computing*, vol. 88, Article ID 105946, 2020.
- [50] W. Li, Y. Chen, and Y. Song, "Boosted k-nearest neighbor classifiers based on fuzzy granules," *Knowledge-Based Systems*, vol. 195, Article ID 105606, 2020.
- [51] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [52] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "Ai-based smart prediction of clinical disease using random forest classifier and naive bayes," *The Journal of Supercomputing*, vol. 77, no. 5, pp. 5198–5219, 2021.
- [53] 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 *Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–6, Cox'sBazar, Bangladesh, February 2019.
- [54] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *Proceedings of the 2004 IEEE International Joint*



- Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, pp. 985–990, Ieee, Budapest, Hungary, July 2004.
- [55] L. J. Rubini, *UCI Machine Learning Repository*, Algappa University, Department of Computer Science and Engineering, karaikudi, India, 2023.
- [56] 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 *Proceedings of the 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 952–957, IEEE, Thoothukudi, India, December 2020.
- [57] R. Gupta, N. Koli, N. Mahor, and N. Tejashri, “Performance analysis of machine learning classifier for predicting chronic kidney disease,” in *Proceedings of the 2020 International Conference for Emerging Technology (INCET)*, pp. 1–4, IEEE, Belgaum, India, June 2020.
- [58] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, “Data cleaning: overview and emerging challenges,” in *Proceedings of the 2016 International Conference on Management of Data*, pp. 2201–2206, San Francisco, CA, USA, June 2016.
- [59] M. Gokiladevi, S. Santhoshkumar, and V. Varadarajan, “Machine learning algorithm selection for chronic kidney disease diagnosis and classification,” *Malaysian Journal of Computer Science*, pp. 102–115, 2022.
- [60] E. Dritsas and M. Trigka, “Machine learning techniques for chronic kidney disease risk prediction,” *Big Data and Cognitive Computing*, vol. 6, no. 3, p. 98, 2022.
- [61] C. Kaur, M. S. Kumar, A. Anjum, M. Binda, M. R. Mallu, and M. S. A. Ansari, “Chronic kidney disease prediction using machine learning,” *Journal of Advances in Information Technology*, vol. 14, no. 2, pp. 384–391, 2023.