

# Classification of chronic kidney disease based on machine learning techniques

Moataz Mohamed El Sherbiny, Eman Abdelhalim, Hossam El-Din Mostafa,  
Mervat Mohamed El-Seddik

Department of Electronics and Communication Engineering, Faculty of Engineering, Mansoura University, Mansoura, Egypt

## Article Info

### Article history:

Received Mar 9, 2023

Revised Jul 16, 2023

Accepted Aug 2, 2023

### Keywords:

Chronic kidney disease

Classifier

Feature selection

Machine learning

Performance measures

## ABSTRACT

In recent years, indescribable suffering from various kidney diseases has been experienced by people all over the world. The situation has been significantly worse because of chronic kidney disease (CKD). Only through an early diagnosis of CKD may kidney disease be hinder in its early stages from progressing. However, it is easier to detect the chronic kidney disease with the aid of machine learning (ML) classifier algorithms sooner than any other existing methods. The present work proposes an approach for potentially predicting CKD infection while considering patient health dataset information into consideration, employing nine distinct ML algorithms; random forest (RF), Naïve Bayes (NB), support vector machine (SVM), decision tree (DT), logistic regression (LR), extreme gradient boosting (XGB), adaptive boosting (ADB), k-nearest neighbors (KNN), and neural network (NN). Machine learning algorithms had been utilized and conducted in four experiments, then they were compared using five performance measures; F1-score, precision, accuracy, recall and run time are used to evaluate the performance. Results had shown that AdaBoost (ADA) outperformed other techniques with achieved accuracy of 99.17%.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Moataz Mohamed El Sherbiny

Department of Electronics and Communication Engineering, Faculty of Engineering, Mansoura University

Mansoura, Egypt

Email: moatazelshebiny@mans.edu.eg

## 1. INTRODUCTION

Kidneys are vital bean-shaped organs that excrete wastes and excess fluid from blood. The most critical stage of the kidney-damaging process is chronic kidney disease (CKD). As a result of CKD, the human kidney can lose its functionality gradually over time [1]. The CKD is a widespread disease that costs a significant number of lives worldwide. One of the most prevalent causes of mortality and suffering in the 21<sup>st</sup> century is CKD. Furthermore, the number of patients with CKD has been rising, with an estimated 843.6 million people diagnosed globally in 2017. CKD is the 11th deadliest cause of mortality worldwide with 1.2 million deaths annually [2]. Currently, it is the sixth most rapidly growing reason of death globally. Since the treatment and medication are neither accessible nor affordable in the majority of developing countries, early detection is very vital to optimize treatment effectiveness and facing the financial economic impact [3]. The main aim of using machine learning (ML) is to enable computers to learn from training data and extract information, in order to perform tasks on future data commonly known as test data [4]. The accuracy and speed of diagnosis have been enhanced with aid of ML techniques. Yet improving accuracy to minimize error rates is considered a challenge. The worldwide health issue of CKD is getting worse every day. Around 10% of the world's population is in distress. CKD threatens the lives of tens of thousands of people annually. The ML

methods have been utilized in a set of researches for CKD classification and prediction in the biomedical field [5] most of previous work used the UCI dataset that contains a lot of missing values in addition to numerous features that can be misleading. A summary of related work is presented in Table 1. The main contribution of this study is to reduce complexity of dataset by selecting the most relevant features to target classification and eliminate non relevant or redundant ones. Moreover, different missing values imputation techniques were involved including k-nearest neighbors (KNN) which is as far as we know the first time to be implemented on CKD dataset.

Abuomar and Sogbe [6] achieved classification accuracy of 98.8% and 10-fold cross validation value of 95.75% using decision tree (DT). Khan *et al.* [7] identified the important attributes using feature selection technique that eliminated the uncorrelated features. Results indicate that on a smaller dataset of 23 attributes for chronic kidney disease, naïve bayes (NB) has a maximum accuracy of 99.1%. The DT outperformed the other classifiers in terms of time complexity. Charleonnann *et al.* [8] indicated that the DT classifier and logistic regression (LR) techniques outscored the other classifiers such as SVM, random forest (RF) and KNN, reaching 98.75% and 97.5% accuracy respectively. Polat *et al.* [9] reduced the dimensions of CKD dataset through choosing two types of feature selection techniques. The results demonstrate the support vector machine (SVM) classifier's accuracy rate for the diagnosis of chronic kidney disease was higher (98.5%). Yashfi *et al.* [10] applied both RF and ANN to CKD dataset. They used feature selection technique to choose 80% of total number of predictor attributes. RF has achieved the highest accuracy of 97.12%. Nirmala *et al.* [11] applied two steps for classification. In first step, they applied LR, DT, and SVM classifier for analysis. The best outcome was DT that achieved accuracy of 95.92%. In second step, they applied an ensemble method to boost the performance of classifiers and reached the highest accuracy of 97.23%. Rady and Anwar [12] showed that probabilistic neural network (PNN) outscored SVM and multi-layer perceptron (MLP) in terms of prediction performance achieving accuracy of 96.7%. Vanaja and Kumar [13] demonstrated the prediction of chronic kidney disease through NB that have got better accuracy of 94.6%. rather than other algorithms that were tested such as KNN, SVM, DT, and ANN. Xiao *et al.* [14] utilized various data mining models in prediction of CKD including RF, SVM, extreme gradient boosting (XGB), and KNN. Linear models were discovered to reach average AUC higher than 87%. Aljaaf *et al.* [15] built a prediction model for chronic kidney disease using both NB and KNN algorithms. KNN reached the most notable precision of 97%, preceding NB that has scored 91% in terms of precision.

This study aims to improve the prediction model's accuracy for CKD by integrating multiple missing data imputation strategies as well as employing feature selection methods. The proposed method is discussed in detail in section 2. In section 3 delves into the evaluation and discussion of experimental results. Finally, work conclusions are given in section 4.

Table 1. Summary of related work

Author	Technique	Accuracy
Abuomar and Sogbe [6]	Decision tree	98.8%
Khan <i>et al.</i> [7]	Naïve bayes and feature selection of 23 attributes	99.1%
Charleonnann <i>et al.</i> [8]	Logistic regression	97.5%
Polat <i>et al.</i> [9]	Support vector machine and feature selection using wrapper and filter techniques	98.5%
Yashfi <i>et al.</i> [10]	Random forest with and feature selection of 20 attributes	97.12%
Nirmala <i>et al.</i> [11]	Decision tree	95.92%
Rady and Anwar [12]	Neural network	96.7%
Xiao <i>et al.</i> [14]	Naïve bayes	94.6%
Aljaaf <i>et al.</i> [15]	K-nearest neighbors	97%

## 2. METHODS

This section contains the proposed system and framework of this study. It consists of three main blocks. Firstly, dataset description, then data preprocessing, ML algorithms and finally evaluation methods as shown in Figure 1.

### 2.1. Dataset description

The data source used in this paper is obtained from University of California Irvine ML respiratory which was collected from hospital in India over two months. The dataset consists of 400 sample record and 25 features including 14 numerical features, 10 categorical ones in addition to the target classification property labeled as ckd or not CKD. The dataset consists of 250 cases of individuals who have CKD and 150 instances of individuals who do not have CKD. The number of complete instances without any missing values in the dataset is 158.

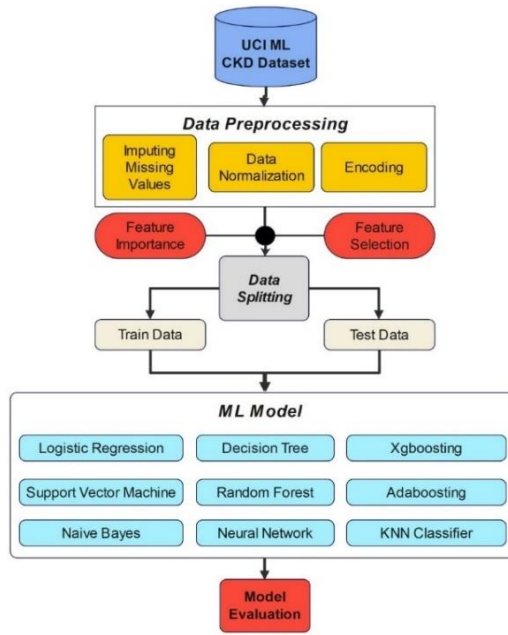


Figure 1. Block diagram of the proposed system

**2.2. Data preprocessing**

Preprocessing is a very crucial component of developing the prediction model in our study. Since the inconsistent data alter the accuracy of the model, considering that the gathered data includes missing values and nominal variables. Therefore, cleaning noisy data is done by handling outliers, the values that are beyond the range of the remaining values. The dataset has a significant number of missing values as shown in Figure 2. The dataset contains 158 complete instances only. Patients frequently overlook several measures for a variety of reasons. As a result, missing values will show up in the data if the diagnostic categories of the samples are unidentified, necessitating the use of an appropriate imputation method. Data preprocessing is implemented in four different stages which are imputation, encoding, scaling and feature selection.

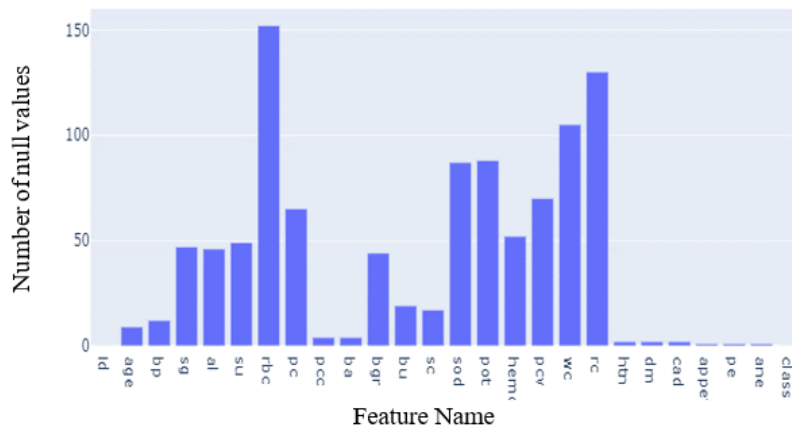


Figure 2. The dataset missing values

**2.2.1. Imputation**

There are various iterative strategies for dealing with null values such as replacing with constant “zero”, mean, median and most frequency techniques. Replacing missing feature values with zero has no effective biasing in prediction but this assumption is medically impossible. In this paper, two methods were implemented. First one is simple imputer by replacing null values with the mean value, while second one k-nearest neighbor imputer with different k-values. Mistyped data are replaced by correctly typed ones.

### 2.2.2. Encoding

Nominal values need to be converted into numbers to make ML algorithm able to understand data it receives in order to facilitate processing. Categorical variables were encoded using ordinal encoder. While class target value was encoded using label encoder.

### 2.2.3. Scaling

Before fitting any models, it is usually vital to scale numeric descriptive features since several significant classes of techniques require it such as SVM and other ML Algorithms since scaling facilitates the ability of model to learn and comprehend the problem [16]. Although there are various methods for scaling data, standard scaling, which adjusts the attribute to 0 mean and 1 standard deviation, was employed in this work. Normalization and standardization are the two most effective scaling techniques. Each data point is resized in a certain range during normalization using the (1);

$$y = \frac{x - \text{Min.Value}}{x - \text{Max.Value}} \quad (1)$$

where;  $x$  is the value before normalization, and  $y$  is the value after normalization. While standardization shifts the distribution to have a mean of zero and a standard deviation of one by the (2);

$$y = \frac{x - \mu}{\sigma} \quad (2)$$

where;  $\mu$  is the mean, and  $\sigma$  is the standard deviation.

### 2.2.4. Feature selection

It is a mechanism for minimizing the number of irrelevant input variables that do not have a significant contribution on the target variable. It focuses on the issue of high dimensionality to achieve better relative accuracy [17]. Consequently, the model runs more quickly and the complexity of the data is reduced. Feature selection is an approach is independent of the learning algorithm. Therefore, it is used for elimination of irrelevant features and selection of relevant ones. Feature selection was implemented in this research using the ANOVA for continuous numerical features while chi-square was used for categorical nominal ones;

$$F = \frac{MST}{MSE} \quad (3)$$

where;  $F$  is the ANOVA coefficient,  $MST$  is the mean squares of treatments. And  $MSE$  is the mean squares of errors;

$$X^2 = \sum \frac{(\text{Observed Value} - \text{Expected Value})^2}{\text{Expected Value}} \quad (4)$$

where;  $X^2$  is chi-square test.

## 2.3. Machine learning algorithms

The ML algorithms classify or predict data without explicit programming after going through the training phase. Eight supervised ML techniques had been utilized in this study. In order to determine the best ML technique that provides the highest classification performance thorough a comparative analysis of the tested algorithms. Method that have been tested includes RF, NB, SVM, DT, LR, extreme gradient boosting (XGB), Adaboost (ADA), KNN and neural network (NN).

### 2.3.1. Random forest

A learning algorithm that develops numerous DT during the training phase and provides output class of those individual trees. Regression and classification can both be employed. This model makes a slight adjustment that makes use of the de-correlated tree by bagging, which is the development of numerous DTs from training data using bootstrapped samples. A specified number of feature columns are removed from the total number of feature columns during bootstrapping. Bootstrap modelling increases bias while minimizing variance.

### 2.3.2. Naive bayes

A probability-based model is a supervised algorithm that necessitates feature independence for classifying data. This model works well for datasets with a large number of input attributes. It encompasses every feature that is provided, even some that have minor effect on the outcome of the prediction [18].

The probabilistic model can be expressed as the following equation, where X and Y are two independent events;

$$P(X|Y) = \frac{P(Y|X)*P(X)}{P(Y)} \tag{5}$$

where;  $P(X|Y)$  is the posterior probability of target class given predictor attribute.  $P(Y|X)$  is the likelihood which is the probability of predictor given class.  $P(X)$  is the prior probability of class and  $P(Y)$  is the prior probability of predictor.

**2.3.3. Support vector machine**

A decision plane-based model is one of the most robust statistical learning framework-based algorithms that provides a solution for both regression and classification problems as well as both linear and non-linear datasets [19]. Every data point is regarded as an n-dimensional vector, and a (n-1) hyper plane divides the datasets. A hyper plane is a line that splits a plane into two halves in a two-dimensional space. A support vector classifier can be explained by the (6):

$$f(x) = \beta_o + \sum_{i \in S} a_i K(x_i, x'_i) \tag{6}$$

where;  $\beta_o$  is bias, and S is set of observation.

**2.3.4. Decision tree**

A supervised learning approach, whose purpose is to comprehend basic chained decision rules from prior input variables in order to train a model to classify a target variable. A set of impurity criteria is applied to recursively separate the variables until a set of stopping requirements are met. Gini impurity is chosen for the model from a variety of impurity measuring techniques;

$$G(t) = 1 - \sum_{i=1}^c p_i^2 \tag{7}$$

where; G(t) is Gini impurity at node t, and  $p_i$  is proportion of observation at class c of node t.

**2.3.5. Logistic regression**

In the healthcare system, LR is a well-known supervised learning algorithm [20]. The LR predicts the probability of the class output using a set of independent features. Assuming that p is the probability of a subject belongs to the CKD class, therefore 1-p is the probability of a subject belongs to the non-CKD class. Decision boundary is the threshold set to determine which data belongs to certain class This classification probability is calculated using the logistic sigmoid function. This algorithm can be described mathematically as:

$$P_i = \frac{1}{1 + e^{-\sum_{j=0}^M \beta_j x_{ij}}} \tag{8}$$

where; i is the number of observations, j is the number of individual variables,  $P_i$  is the probability of '1' at observation I,  $\beta_j$  is the regression coefficient, and  $x_{ij}$  is the J<sup>th</sup> variable at observation i.

**2.3.6. Xgboost**

An XGB is a tree-based sequential DTs algorithms [21]. It is regarded as one of the most efficient methods for performing classification and predictions on small to medium-sized structured or tabular datasets. It uses a gradient descent architecture to accurately estimate a target variable or feature, through integrating relatively weaker and simpler models. One of XGBoost's most significant aspects is scalability, where it directs abrupt learning through parallel and distributed computing as well as provides well-structured memory usage [22].

$$L(t) = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{9}$$

Where;  $f_t$ : the term added to minimize objective function,  $\hat{y}_i^t$ : Prediction of the i-th instance at t-th iteration,  $\Omega$ : Penalize complexity of model function, and  $l$ : Differentiable convex loss function.

**2.3.7. ADA**

The ADA is an iterative ML algorithm that is less prone to over-fitting of data. Where dataset is split into two partitions for each iteration, the features used in the first iteration will be given less weight, and the incorrectly classified data are given more weight in the next iteration. When all iterations are finally completed,

they are merged with appropriate weights to yield a powerful and effective classifier that predicts the classes of the unseen data. The output of classifier can be written as:

$$F_T(x) = \sum_{t=1}^T f_t(x) \quad (10)$$

the error function after adaptive boosting (ADB) can be calculated as:

$$E_t = \sum_i E [F_{t-1}(x_i) + \alpha_t h(x_i)] \quad (11)$$

where;  $f_i$ : Weak feature,  $F_{t-1}$ : Boosted classifier, and  $\alpha_i$ : Coefficient assigned to weak classifier.

### 2.3.8. K-nearest neighbors

A distance-based supervised ML algorithm that can be used for regression as well as for classification which is done by most votes to its neighbors [23]. It is a lazy learner algorithm because it doesn't immediately apply what it has learned from the training set. Instead, it stores the dataset then applies what it has learned when it comes to classify.

### 2.3.9. Neural network

NNs would be developed for classification, recognition as well as numerous other uses based on their layers and neurons. Its fundamental methodology is underpinned by biological neural systems. The two key features of ANN are its ability to generalize and produce an acceptable response to unobserved data, as well as its ability to learn how to carry out its functions once it has been properly trained [24]. Initially the best subset of features is fed to the NN as the inputs. A weighted sum is calculated at each neuron for each feature subset. Then, a transfer function is applied to this weighted sum to determine the output value of neuron.

## 2.4. Evaluation methods

The most prominent performance measurements are precision, F1-score, sensitivity (recall), and accuracy. True positives (TP), false positives (FP), true negatives (TN), and false negatives are the four variables needed by the evaluation methods (FN).

Accuracy: This is the percentage of cases that were correctly identified out of all the cases.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (12)$$

Precision: It is the ratio of correctly predicted positive outcomes to all positive outcomes.

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

Recall: It is the proportion of correctly predicted events among the foreseen data.

$$Recall = \frac{TP}{TP+FN} \quad (14)$$

F1-Score: It is the weighted average of precision and recall.

$$F1 - Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (15)$$

Sensitivity: It is the mean proportion of actual TP that are correctly identified.

$$Sensitivity = \frac{TP}{TN+FP} \quad (16)$$

Specificity: It is used to measure the fraction of negative values that are correctly classified.

$$Specificity = \frac{TN}{TN+FP} \quad (17)$$

## 3. RESULTS AND DISCUSSION

Dataset is split into two parts for training and testing purposed. The model learns to recognize relationships in between the data during training, so that it can make accurate predictions on new data. The testing set is used to evaluate the performance of the trained model on new unseen data. In this study, 70% for training and 30% for testing. Training and testing have been applied using kaggle on a processor core i7-10<sup>th</sup> generation and 16 GB RAM.

**3.1. Experiment 1**

Simple imputation approach is used to handle with null values in the CKD dataset. There are several imputation techniques including median, mean and most frequent. This experiment utilized the mean strategy to deal with missing values. Results are summarized in Table 2.

Table 2. Results of experiment 1

Model name	Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)	Run time (ms)
LR	99.17	100	99.35	98.7	2,706
SVM	100	100	100	100	2,411
KNN	95.83	100	96.64	93.51	0.692
DT	95.83	97.37	96.73	96.1	1,924
RF	99.17	98.72	99.35	100	156,298
NB	93.33	97.26	94.67	92.21	1,863
XGB	98.33	97.47	98.72	100	385.03
ADA	100	100	100	100	89,881

**3.2. Experiment 2**

People with similar physical conditions should have consistent physiological data, which is the reason the KNN-based technique was selected to perform imputation of the uncomplete records of dataset. KNN technique is used to handle null values with different k values of 3, 5, 7, and 9. Performance results are illustrated in Table 3.

Table 3. Results of experiment 2

K-value	Model name	Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)	Run time (ms)
K = 3	LR	98.33	100	98.68	97.4	3,532
	SVM	99.17	100	99.35	98.7	1,896
	KNN	96.67	100	97.33	94.81	0.679
	DT	97.5	97.44	98.06	98.7	2,677
	RF	99.17	98.72	99.35	100	156,596
	NB	95	96.1	96.1	96.1	1,922
	XGB	99.17	100	99.35	98.7	376,257
	ADA	97.5	98.68	98.04	97.4	87,891
K = 5	LR	99.17	100	99.35	98.7	3,751
	SVM	99.17	100	99.35	98.7	2.31
	KNN	95	100	95.95	92.21	0.648
	DT	96.67	97.4	97.4	97.4	2,122
	RF	99.17	100	99.35	98.7	155,343
	NB	925.5	95.95	94.04	92.21	1,657
	XGB	97.5	98.68	98.04	97.4	405,916
	ADA	98.33	100	98.68	97.4	97,951
K = 7	LR	99.17	100	99.35	98.7	2,815
	SVM	99.17	100	99.35	98.7	2,516
	KNN	95	100	95.95	92.21	0.882
	DT	97.5	98.68	98.04	97.4	2.03
	RF	99.17	100	99.35	98.7	146,251
	NB	93.33	96	94.74	93.51	1,867
	XGB	99.17	100	99.35	98.7	374.66
	ADA	99.17	100	99.35	98.7	95,541
K = 9	LR	99.17	100	99.35	98.7	2.57
	SVM	99.17	100	99.35	98.7	2,156
	KNN	94.17	100	95.24	90.91	0.711
	DT	98.33	98.7	98.7	98.7	1,555
	RF	99.17	100	99.35	98.7	146,168
	NB	94.17	96.05	95.42	94.81	1,407
	XGB	99.17	100	99.35	98.7	369,967

**3.3. Experiment 3**

Feature selection is a crucial component of ML so it is needed to select the most relevant features to build the model [25]. The dataset used in this research has quite a lot of features for relatively small size. The chi-square and ANOVA tests examine the connection between the features. The feature importance scores are shown in Figure 3 where continuous (numerical) and categorical (nominal) variables are shown in Figures 3(a) and 3(b) respectively. Results due to different number of attributes selected for prediction of target class are illustrated in Table 4.

Table 4. Results of experiment 3

Model name	Number of features			Accuracy %	Precision %	F1-score %	Recall %	Run time (ms)
	Total	Numeric	Nominal					
LR				99.17	100	99.35	98.7	2,564
SVM				98.33	100	98.68	97.4	1,765
KNN	20	10	10	98.33	100	98.68	97.4	0.738
DT				95.83	97.37	96.73	96.1	1,559
RF				99.17	98.72	99.35	100	120,453
NB				90.83	97.14	92.52	88.31	0.906
XGB				98.33	97.47	98.72	100	232,541
ADA				100	100	100	100	67,065
LR				98.33	100	98.68	97.4	2,951
SVM				98.33	98.7	98.7	98.7	1,978
KNN	17	10	7	98.33	100	98.68	97.4	1,163
DT				95.83	97.37	96.73	96.1	1,377
RF				99.17	98.72	99.35	100	116,721
NB				90.83	97.14	92.52	88.31	0.805
XGB				98.33	97.47	98.72	100	263,891
ADA				98.33	98.7	98.7	98.7	82,859
LR				98.33	100	98.68	97.4	1,735
SVM				99.17	100	99.35	98.7	1,151
KNN	14	7	7	97.5	100	98.01	96.1	0.798
DT				96.67	98.67	97.37	96.1	1,314
RF				97.5	98.68	98.04	97.4	120,382
NB				90	97.1	91.78	87	1,487
XGB				98.33	98.7	98.7	98.7	237,532
ADA				100	100	100	100	67,242
LR				98.33	100	98.68	97.4	1,849
SVM				99.17	100	99.35	98.7	1,361
KNN	12	7	5	97.5	100	98.01	96.1	1,161
DT				96.67	98.67	97.37	96.1	1,398
RF				99.17	98.72	99.35	100	125,476
NB				90	97.1	91.78	87.01	1.39
XGB				98.33	98.7	98.7	98.7	232,277
ADA				99.17	100	99.35	98.7	72,441
LR				98.33	100	98.68	97.4	2,145
SVM				99.17	100	99.35	98.7	1,855
KNN	12	6	6	98.33	100	98.68	97.4	1,298
DT				96.67	98.67	97.37	96.1	1,396
RF				98.33	100	98.68	97.4	113.3
NB				90	97.1	91.78	87.01	1,422
XGB				98.33	97.47	98.72	100	836,308
ADA				99.17	100	99.35	98.7	68.11
LR				98.33	100	98.68	97.4	1,774
SVM				98.33	98.7	98.7	98.7	1,833
KNN	10	5	5	98.33	100	98.68	97.4	1.8
DT				98.3	100	98.68	97.4	1,036
RF				99.17	100	99.35	98.7	113,443
NB				90	97.1	91.78	87.01	1,052
XGB				98.33	98.7	98.7	98.7	228,421
ADA				100	100	100	100	71,856

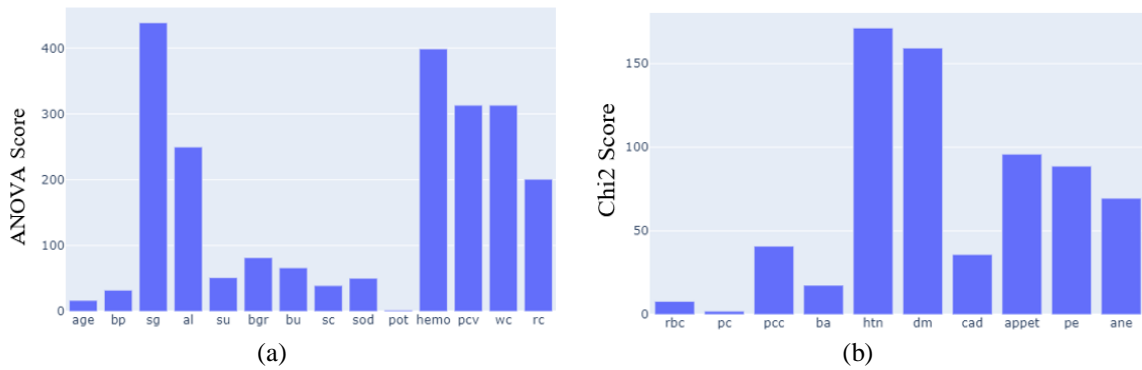


Figure 3. Feature importance scores for (a) numeric features and (b) nominal features



**3.4. Experiment 4**

The NN is known for higher accuracy in prediction and classification of medical datasets of large size. In this experiment, NN with a hidden layer is applied. Batch size is 16, learning rate is 0.0001, activation function is sigmoid and using Adams solver. Accuracy and test time shown in Table 5. While accuracy and loss versus epochs is shown in Figure 4. Different number of epochs such as 100 and 120 were executed in this experiment and their results are shown in Figures 4(a) and 4(b). However, accuracy did not improve at higher computational cost.

**Table 5. Results of experiment 4**

No. of epochs	Accuracy	Time in S
100	95.83	6.47
120	97.5	7.34

Four experiments were conducted in this study on the chronic kidney dataset which contains 62.5% of CKD patients and 37.5% as non-ckd patients. After shuffling and splitting data into 70% train size and 30% test size performance were evaluated and compared in terms of accuracy, precision, F1-score, recall, ROC curves and run time. ADA and SVM achieved an accuracy of 100% when mean imputation is used to handle missing values. The highest accuracy was 99.17% when missing data was imputed using KNN. The chi-square and ANOVA tests examine the connection between the attribute to perform feature selection. ADA reached 100% when the top 14 features were selected. The DT achieved the highest accuracy of 99.17% among other models when 17 features were chosen. While the NN performance for this small size dataset stands at 97.5% accuracy.

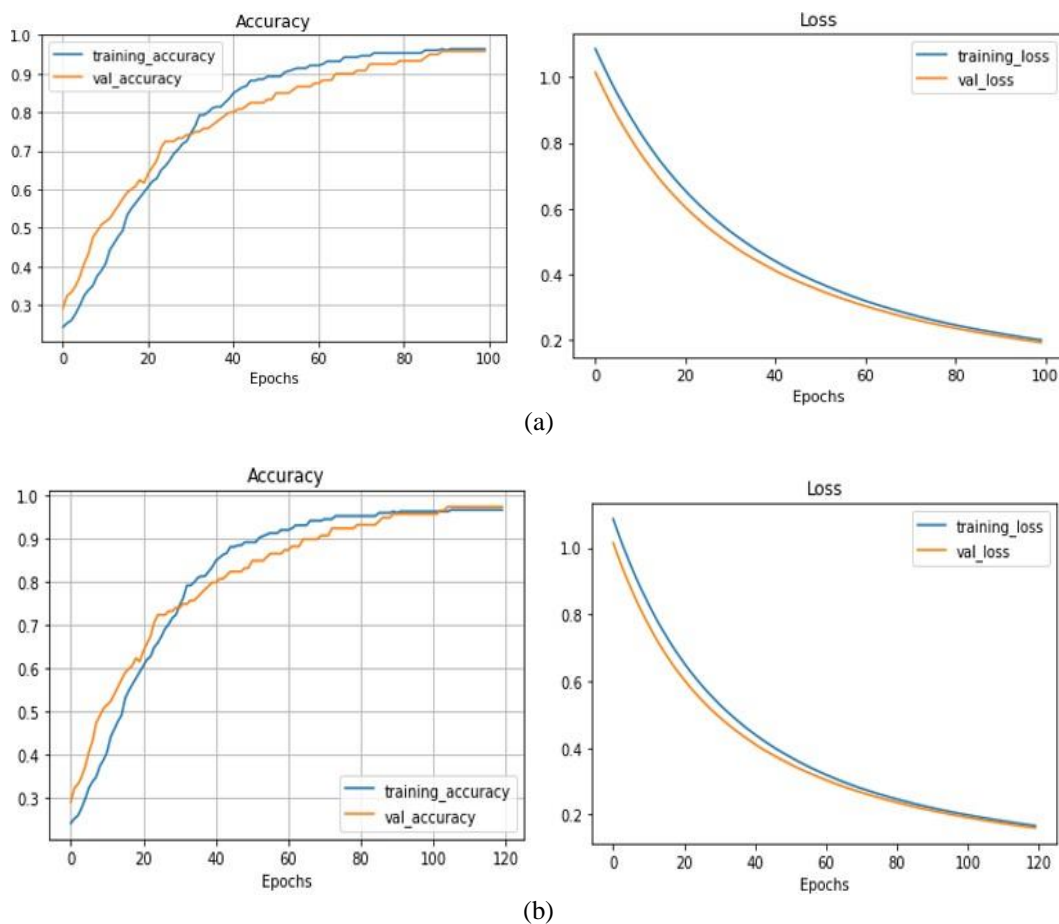


Figure 4. Accuracy and loss curves of NN versus number of epochs (a) 100 epochs and (b) 120 epochs

#### 4. CONCLUSION

The accurate prediction of chronic kidney disease considered to be one of the challenging biomedical research topics nowadays. This research has resulted in the development of a ML-based pipeline to successfully identify chronic kidney disease using a data set of 400 sample instants with 24 features. Quite a lot of features for a relatively small dataset which led us to applying feature selection method. Consequently, we met our goal by utilizing and analyzing various ML algorithms such as RF, DT, ADA, KNN, XGB, and gaussian NB in addition to artificial NN, then compared the performance of these algorithms. The proposed model demonstrated an accuracy, sensitivity, specificity, and ROC of 100%, 100%, and 100%, 100%, respectively using ADA with feature selection technique in less running time. Validation and testing were performed. The reliability of the method has been confirmed using ROC analysis. In future work, more advanced ML and DL algorithms will be applied on different datasets either statistical or medical images so that the efficiency and effectiveness of CKD prediction can be boosted at earlier stages.




#### REFERENCES

- [1] M. M. Hossain *et al.*, "Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease," *Machine Learning with Applications*, vol. 9, p. 100330, Sep. 2022, doi: 10.1016/j.mlwa.2022.100330.
- [2] K. J. Jager, C. Kovesdy, R. Langham, M. Rosenberg, V. Jha, and C. Zoccali, "A single number for advocacy and communication-worldwide more than 850 million individuals have kidney diseases," *Kidney International*, vol. 96, no. 5, pp. 1048–1050, Nov. 2019, doi: 10.1016/j.kint.2019.07.012.
- [3] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2020, doi: 10.1109/ACCESS.2019.2963053.
- [4] S. Alghunaim and H. H. Al-Baity, "On the scalability of machine-learning algorithms for breast cancer prediction in big data context," *IEEE Access*, vol. 7, pp. 91535–91546, 2019, doi: 10.1109/ACCESS.2019.2927080.
- [5] G. M. Ifraz, M. H. Rashid, T. Tazin, S. Bourouis, and M. M. Khan, "Comparative analysis for prediction of kidney disease using intelligent machine learning methods," *Computational and Mathematical Methods in Medicine*, vol. 2021, pp. 1–10, Dec. 2021, doi: 10.1155/2021/6141470.
- [6] O. Abuomar and P. Sogbe, "Classification and detection of chronic kidney disease (CKD) using machine learning algorithms," in *International Conference on Electrical, Computer, and Energy Technologies, ICECET 2021*, Dec. 2021, pp. 1–8, doi: 10.1109/ICECET52533.2021.9698666.
- [7] M. Khan, M. S. I. Prottasha, T. Nasim, A. Mehedi, M. A. M. Pranto, and N. Asad, "Performance analysis of various machine learning classifiers on reduced chronic kidney disease dataset," *International Journal of Recent Research and Review*, vol. XIII, pp. 16–21, 2020.
- [8] A. Charleonnann, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *2016 Management and Innovation Technology International Conference, MITICON 2016*, Oct. 2017, pp. MIT80–MIT83, doi: 10.1109/MITICON.2016.8025242.
- [9] H. Polat, H. D. Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," *Journal of Medical Systems*, vol. 41, no. 4, p. 55, Apr. 2017, doi: 10.1007/s10916-017-0703-x.
- [10] S. Y. Yashfi *et al.*, "Risk prediction of chronic kidney disease using machine learning algorithms," in *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*, Jul. 2020, pp. 1–5, doi: 10.1109/ICCCNT49239.2020.9225548.
- [11] M. B. Nirmala, D. K. Priyamvada, P. R. Shetty, and S. P. Singh, "Chronic kidney disease prediction using machine learning techniques," *12th International Conference on Advances in Computing, Control, and Telecommunication Technologies, ACT 2021*, vol. 2021-August, pp. 185–190, Aug. 2021, doi: 10.1007/s44174-022-00027-y.
- [12] E. H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, p. 100178, 2019, doi: 10.1016/j.imu.2019.100178.
- [13] S. Vanaja and K. R. Kumar, "Analysis of feature selection algorithms on classification: a survey," *International Journal of Computer Applications*, vol. 96, no. 17, pp. 29–35, Jun. 2014, doi: 10.5120/16888-6910.
- [14] J. Xiao *et al.*, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," *Journal of Translational Medicine*, vol. 17, no. 1, p. 119, Dec. 2019, doi: 10.1186/s12967-019-1860-0.
- [15] A. J. Aljaaf *et al.*, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in *2018 IEEE Congress on Evolutionary Computation, CEC 2018 - Proceedings*, Jul. 2018, pp. 1–9, doi: 10.1109/CEC.2018.8477876.
- [16] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification models," *Journal of Machine Learning Research*, vol. 8, pp. 1625–1657, 2007.
- [17] C. W. Chen, Y. H. Tsai, F. R. Chang, and W. C. Lin, "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results," *Expert Systems*, vol. 37, no. 5, Oct. 2020, doi: 10.1111/exsy.12553.
- [18] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *Journal of Supercomputing*, vol. 77, no. 5, pp. 5198–5219, May 2021, doi: 10.1007/s11227-020-03481-x.
- [19] M. Wang and H. Chen, "Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis," *Applied Soft Computing Journal*, vol. 88, p. 105946, Mar. 2020, doi: 10.1016/j.asoc.2019.105946.
- [20] S. Nusinovi *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *Journal of Clinical Epidemiology*, vol. 122, pp. 56–69, Jun. 2020, doi: 10.1016/j.jclinepi.2020.03.002.
- [21] A. Ogunleye and Q. G. Wang, "XGBoost model for chronic kidney disease diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 6, pp. 2131–2140, Nov. 2020, doi: 10.1109/TCBB.2019.2911071.
- [22] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-August-2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [23] J. Gola *et al.*, "Objective microstructure classification by support vector machine (SVM) using a combination of morphological parameters and textural features for low carbon steels," *Computational Materials Science*, vol. 160, pp. 186–196, Apr. 2019, doi: 10.1016/j.commatsci.2019.01.006.




- [24] N. A. Almansour *et al.*, "Neural network and support vector machine for the prediction of chronic kidney disease: a comparative study," *Computers in Biology and Medicine*, vol. 109, pp. 101–111, Jun. 2019, doi: 10.1016/j.compbimed.2019.04.017.
- [25] 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, pp. 89–96, 2019, doi: 10.14569/ijacsa.2019.0100813.

## BIOGRAPHIES OF AUTHORS






**Moataz Mohamed El Sherbiny**    received the B.Sc. degree in communication and information engineering from Mansoura University, in 2017, and the M.Sc. degrees in electronics and communication engineering from Mansoura University, in 2023. He is currently an assistant lecturer at electronics and communication department, faculty of engineering, Mansoura University. His research interests are in area of artificial intelligence and biomedical engineering. He can be contacted at email: moatazelsherbiny@mans.edu.eg.






**Eman Abdelhalim**    assistant Professor, electronics and communications department, faculty of engineering. Her research interests cover several aspects of communications, cloud computing, big data analytics, and medical imaging. She has been working since 2008 on graduation projects in the field of applications of deep learning in communications and medical imaging fields. She can be contacted at email: eman-haleim@mans.edu.eg.



**Hossam El-Din Mostafa**    professor at the department of electronics and communications engineering, the founder and former executive manager of Biomedical Engineering Program (BME) at the Faculty of Engineering, Mansoura University. He is an IEEE senior member. Research interests include biomedical imaging, image processing applications, and bioinformatics. He can be contacted at email: hossam-moustafa@hotmail.com.



**Mervat Mohamed El-Seddek**    received the B.Sc. degree in electronics and communications from the Electronic and Communication Department, Faculty of Engineering, Mansoura University, in 1999, and the M.Sc. and Ph.D. degrees in electrical communications from the Faculty of Engineering, Mansoura University, in 2009 and 2015, respectively. She was appointed as an Assistant Professor in the Department of Communications and Electronics Engineering at the Mansoura Higher Institute of Engineering and Technology. Image processing, medical imaging, and machine learning are among the primary research areas. Member at IEEE. She can be contacted at email: mervat.elseddek@ieee.org.