EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms

Mirza Muntasir Nishat[1], Fahim Faisal[1,*], Rezuanur Rahman Dip[1], Sarker Md. Nasrullah[2], Ragib Ahsan[1], Fahim Shikder[1], Md. Asfi-Ar-Raihan Asif[1] and Md. Ashraful Hoque[1]

[1]Department of Electrical and Electronic Engineering, Islamic University of Technology (IUT), Dhaka, Bangladesh
[2]Department of Public Health, North South University, Dhaka, Bangladesh

Abstract

INTRODUCTION: Chronic Kidney Disease refers to the slow, progressive deterioration of kidney functions. However, the impairment is irreversible and imperceptible up until the disease reaches one of the later stages, demanding early detection and initiation of treatment in order to ensure a good prognosis and prolonged life. In this aspect, machine learning algorithms have proven to be promising, and points towards the future of disease diagnosis.

OBJECTIVES: We aim to apply different machine learning algorithms for the purpose of assessing and comparing their accuracies and other performance parameters for the detection of chronic kidney disease.

METHODS: The 'chronic kidney disease dataset' from the machine learning repository of University of California, Irvine, has been harnessed, and eight supervised machine learning models have been developed by utilizing the python programming language for the detection of the disease.

RESULTS: A comparative analysis is portrayed among eight machine learning models by evaluating different performance parameters like accuracy, precision, sensitivity, F1 score and ROC-AUC. Among the models, Random Forest displayed the highest accuracy of 99.75%.

CONCLUSION: We observed that machine learning algorithms can contribute significantly to the domain of predictive analysis of chronic kidney disease, and can assist in developing a robust computer-aided diagnosis system to aid the healthcare professionals in treating the patients properly and efficiently.

Keywords: Chronic Kidney Disease, Machine Learning Algorithms, UCI Dataset, Accuracy, Precision, Sensitivity, F1 score, ROC.

Received on 01 May 2021, accepted on 07 August 2021, published on 13 August 2021

doi: 10.4108/ _____

*Corresponding author Email: faisaleee@iut-dhaka.edu

1. Introduction

Chronic kidney disease (CKD) refers to a progressive and irreversible decline of the structure and functionalities of the kidneys, especially the deterioration of the glomerular filtration rate that develops over the course of several months or years. It begins with abnormal biochemical changes which eventually lead to gradual loss of excretory, endocrine and metabolic functions of the kidneys. These abnormalities manifest as signs and symptoms of renal failure. Although the underlying etiology of the disease remains unknown in a large number of patients, the commonest causes of the disease were listed as hypertension, diabetes, interstitial diseases, glomerular diseases, systemic inflammatory disorders, renovascular abnormalities and congenital conditions [1]. The Prognosis

of the disease is determined by monitoring the glomerular filtration rate (GFR) and quantity of albumin in urine. Decreased GFR and increased albumin in urine was found to be allied with a higher risk of all-cause mortality, mortality from cardiovascular diseases (CVD), progressive kidney diseases and acute kidney injury (AKI) [2]. Atherosclerotic calcification within the vessels followed by cholesterol crystal formation was suspected to create a high risk of developing CKD in a patient [3]. If untreated, CKD leads through a spectrum of pathological conditions eventually to end-stage renal disease (ESRD) or end-stage renal failure (ESRF) which is responsible for coma and death in patients [4]. The gradual development of CKD is either asymptomatic, or it presents with a set of non-specific symptoms like loss of weight, fatigue, poor appetite, edema, headache, muscle cramps etc. which makes it quite difficult for the patient or the physician to suspect the involvement of the kidneys. Moreover, the symptoms do not show until much later in the 3rd or 4th stage of the disease, by which time the comorbidities already set in [5]. CKD also manifests with immune dysfunction, haematological abnormalities, endocrine dysfunction, neurological symptoms and electrolyte imbalance [1]. In turn, CKD acts as a risk factor for other conditions like CVD, as mentioned above, resulting in added mortalities and morbidities [6]. As a result, CKD has become a global burden, contributing to a significant portion of deaths due to non-communicable diseases (NCD). It has risen from being the 27th leading cause of global death in 1990 to being the 18th in 2010 [7]. Approximately, 1 million people died from CKD or cause related to it in the year of 2013 [8]. The number of new cases needing renal replacement therapy has increased at a rate of 8% per year for the last decade worldwide [9].

Studies have revealed CKD to be a greater burden in the low and middle-income countries when compared to the high-income ones [10], [11]. The proportion of people diagnosed with CKD in the urban areas of South Asia ranges from 7.2% to 17.2% [12]. The prevalence was reported to be 13% among the general population of Dhaka city aged 15 years or older [13]. Another community-based research suggested that about one-third of the rural people in Bangladesh were at risk of having CKD [14]. Hence, in a developing country like Bangladesh, CKD poses an impeccable threat not only as a disease but also as a financial burden due to its demand for long-term treatment. The situation calls for the innovation of a diagnostic or, at the very least, a screening technique for the early and reliable detection of CKD in a patient to ensure an effective treatment by the doctor. Machine learning (ML) is currently one of the most notable and successful technologies in the medical industry for diagnosing and forecasting various diseases and their stages. [15-21]. As machine learning is all about the exploration of the huge dataset and their patterns, features, modes etc., the dataset of various diseases can be fed into these algorithms with a view to developing ML models [22- 28]. This introduction of algorithms in medical databases will greatly assist medical professionals in making informed decisions about illnesses, preventing mistakes, and providing a safe life to the general public [29].

2. Related Works

In this context, a lot of researchers and data scientists have executed different techniques to attain satisfactory performances in terms of the ML models. Engin et al. used a dataset extracted from UCI to apply K-star, SVM, and J48 algorithms and compared them in terms of accuracy, sensitivity, and other parameters, where J48 classifier achieved 99% accuracy [30]. Gunarathne et al., on the other hand, tested different algorithms on the same dataset and found that the Multiclass Decision Forest (MDF) algorithm outperforms the other algorithms with 99.1% accuracy [31]. However, a different approach of featuring datasets was conducted by Nusrat et al. where they pre-processed the data by root mean squared error, mean absolute error and receiver operating characteristic curve. After featuring the dataset, they implemented Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) algorithm. According to their investigation, DT offers the best accuracy which is about 98%-99% [32].

However, another work was conducted by Huseyin et al. where they improvised on the feature selection of the dataset before applying the algorithm. Hence, they applied the filter, wrapper and embedded feature selection methods on the dataset and then passed them through the SVM algorithm. According to their work, the filter scheme subset evaluation achieved the best accuracy of 98.5% [33]. Furthermore, Devika et al. focused on Naïve Bayes (NB), K-Nearest Neighbour (KNN) and Random Forest (RF) algorithms in their research to predict CKD. Among these classifiers, RF has outperformed other algorithms with an accuracy of 99% [34]. Besides, Merve et. al. has accomplished better accuracy (99.5%) by deploying AdaBoost ensemble learning approach. In their work, they have evaluated the performances of the ML models by utilizing mean absolute error (MAE), root mean squared error (RMSE) and area under curve (AUC) [35].

In addition, Amanah et al. has implied PSO algorithms to optimize their result more precisely and has obtained an accuracy of 99.5%. After applying AdaBoost and PSO feature selection algorithm combined, they were able to increase their average accuracy by 36.20% [36]. On the other hand, Chittora et. al carried out six different methods of feature selection and implemented seven machine learning algorithms where 99.6% accuracy was attained by deep learning network [37]. Moreover, Sobrinho et. al conducted a research where they have analyzed how machine learning approaches can help in the early detection of CKD in underdeveloped countries. The study findings indicate that the J48 decision tree is a good machine learning technique for such screening in developing nations due to the ease of comprehension of its classification results, with 95.00% accuracy [38]. In our previous study, boosting

algorithms were deployed in the same dataset where we achieved 99.75% accuracy AdaBoost algorithm [39]. Hence, it is evident that machine learning algorithms open windows of identifying chronic kidney diseases at an early stage so that better treatment can be ensured for the patient. In this research, we focused on an investigative approach in terms applying supervised ML algorithms in UCI dataset pertaining to CKD and comparing the performances of the ML models in a comprehensive manner so that a vivid idea is portrayed in terms of developing a computer aided diagnosis system to detect CKD at an early stage and ensure proper treat for the patients [40].

3. Methodology

3.1. Data Processing

One of the most widely used and accurate datasets for implementing machine learning algorithms is the UCI dataset repository. The dataset contains 400 instances and 25 attributes. Hence, the description of the attributes with necessary information is presented in Table 1. In order to apply machine learning algorithms, data must be reliable and well-structured. There are two types of data in this dataset: (i) numerical values and (ii) categorical values. Categorical values were replaced with dummy values for the implementation of ML algorithms. Since there are many missing values in various attributes in this data set, four separate data frames have been created to apply the algorithms and extract the results. The missing values have been first filled with the mean values of the corresponding attributes. Moreover, the missing values were then imputed using median and mode values. Finally, missing values were omitted in which we were left with 158 instances from 400 instances.

Each of the data frames was split into two portions: (i) training set and (ii) testing set. The training set has been comprised with 60% of the data and the rest of the data has been used for testing purpose. The splitting of each data frames has been cross-validated and hyperparameter tuning was accomplished. Hence, the performance parameters are observed and tabulated for both 'with tuning' and 'without tuning' case. The correlation heatmaps for all the data frames like mean, mode, median and no null are illustrated in Figure 1. Finally, a detailed step-by-step workflow diagram is presented in Figure 2 which provides a clear idea of the overall approach. Jupyter Notebook from Anaconda navigator was utilized as simulation platform for this research. However, this analysis was executed on a computer with Intel Core i5 9th generation processor with 16 GB RAM.

Table 1. Description of Attributes with Information

Attributes	Information
age- Age	Discrete Integer Values
bp- Blood pressure	Discrete Integer Values
al- Albumin	Nominal Values (0,1,2,3,4,5)
su- Sugar	Nominal Values (0,1,2,3,4,5)
rbc- Red blood cells	Nominal Values (Normal, Abnormal)
pc- Pus cell	Nominal Values (Normal, Abnormal)
pcc- Pus cells clumps	Nominal Values (Present, Not-Present)
ba-Bacteria	Nominal Values (Present, Not-Present)
bgr- Blood Glucose Random	Numerical Values in mgs/dl
bu- Blood Urea	Numerical Values in mgs/dl
sc- Serum creatinine	Numerical Values
sod- Sodium	Numerical Values in mEq/L
pot- Potassium	Numerical Values in mEq/L
hemo- Hemoglobin	Numerical Values in gms
pcv- Packed Cell Volume	Numerical Values
wc- White blood cell count	Discrete Integer Values
rc- Red blood cell count	Numeric Values
htn- Hypertension	Nominal Values (Yes, No)
dm- Diabetes Mellitus	Nominal Values (Yes, No)
cad- Coronary Artery Disease	Nominal Values (Yes, No)
appet- Appetite	Nominal Values (Good, Poor)
pe- Pedal Edema	Nominal Values (Yes, No)
ane- Anemia	Nominal Values (Yes, No)
class- Classification	Nominal Values (CKD, Not CKD)

3.2. Study of Machine Learning Algorithms

Machine learning is a subset of Artificial Intelligence (AI) that studies various algorithms and adapts to different scenarios through experience. The experience is gained by training with a collection of data which is called training data. After going through the training, machine learning algorithms predict or classify data without explicit programming. In this paper, eight supervised classification learning algorithms are chosen to identify Chronic Kidney Diseases (CKD) and their results are extensively compared under different criteria.

3.2.1. Logistic Regression (LR)

Logistic Regression is a statistical classification model which estimates the probability of an event existing within a certain class [41]. Despite the fact that its name includes the word "regression," logistic regression is a commonly used binary classifier.

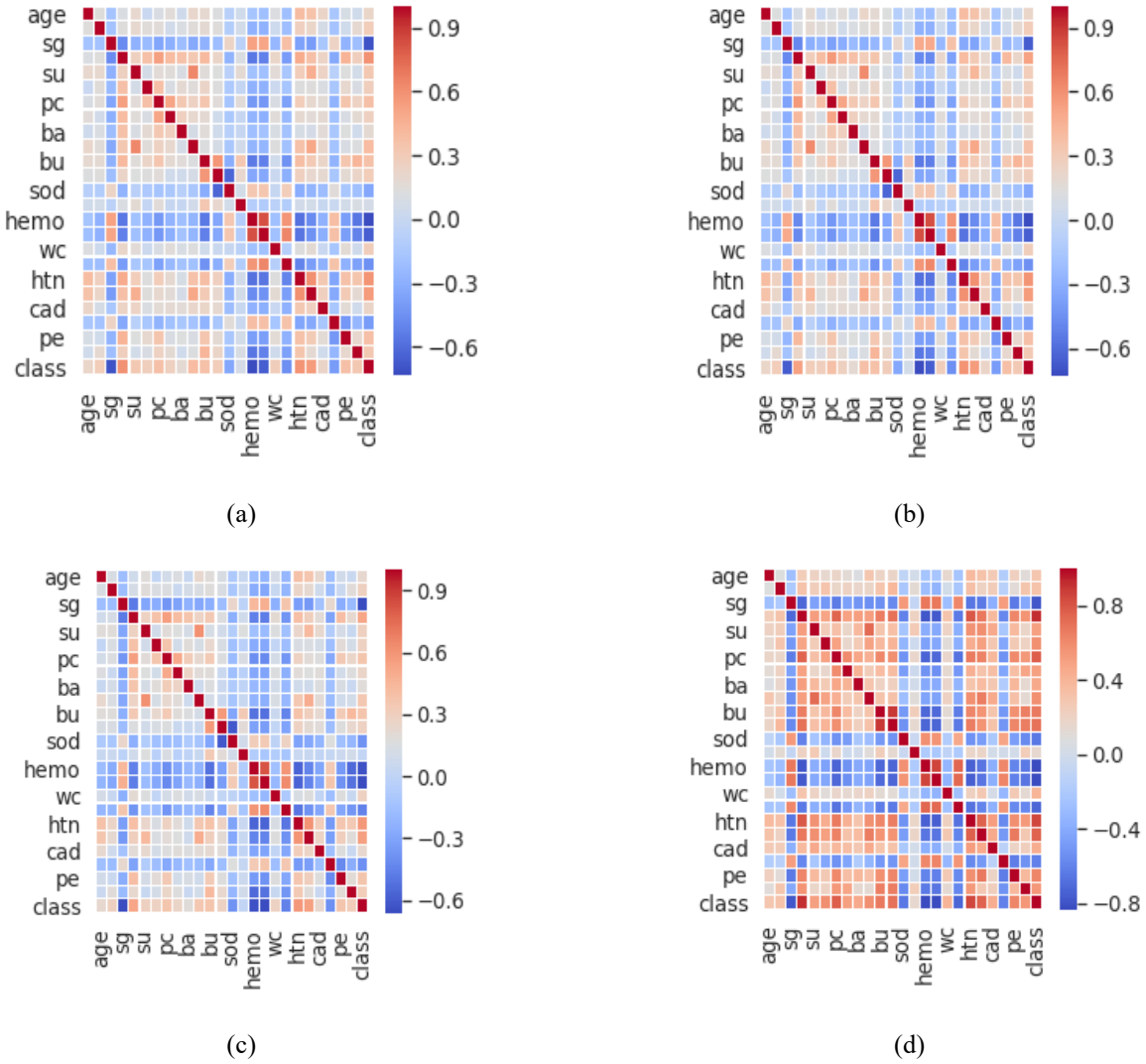


Figure 1. Correlation heatmap of different data frames (a) mean (b) median (c) mode and (d) no null

A threshold is set to predict in which class does a data belong, which is called a decision boundary. This classification probability is calculated by the logistic function which is actually a sigmoid function. The mathematical model of the algorithm can presented like below:

$$p_i = \frac{1}{1 + e^{-\sum_{j=0}^{M} \beta_j x_{ij}}} \quad (1)$$

Here,
i= 1…N (number of observations)
j= 1…M (number of individual variables)
p_i = Probability of a '1' at observation i
β_j= Regression Coefficient
x_{ij}= The j^{th} variable at observation i

3.2.2. K-Nearest Neighbours (KNN)

K-Nearest Neighbors is one of the simplest and most used supervised machine learning algorithms [42]. Technically it does not train any dataset; instead an observation is predicted to fall under those classes which have the largest proportion of k-nearest neighbors around it. Distance is considered to be a metric to determine similarity.

For instance, the closet data point around the point under observation can be considered most similar to the data point. There are a large variety of distance-metrics like Euclidean distance (d),

$$d_{euclidean} = \sqrt{\sum_{i=1}^{n} (x_i^2 - y_i^2)} \quad (2)$$

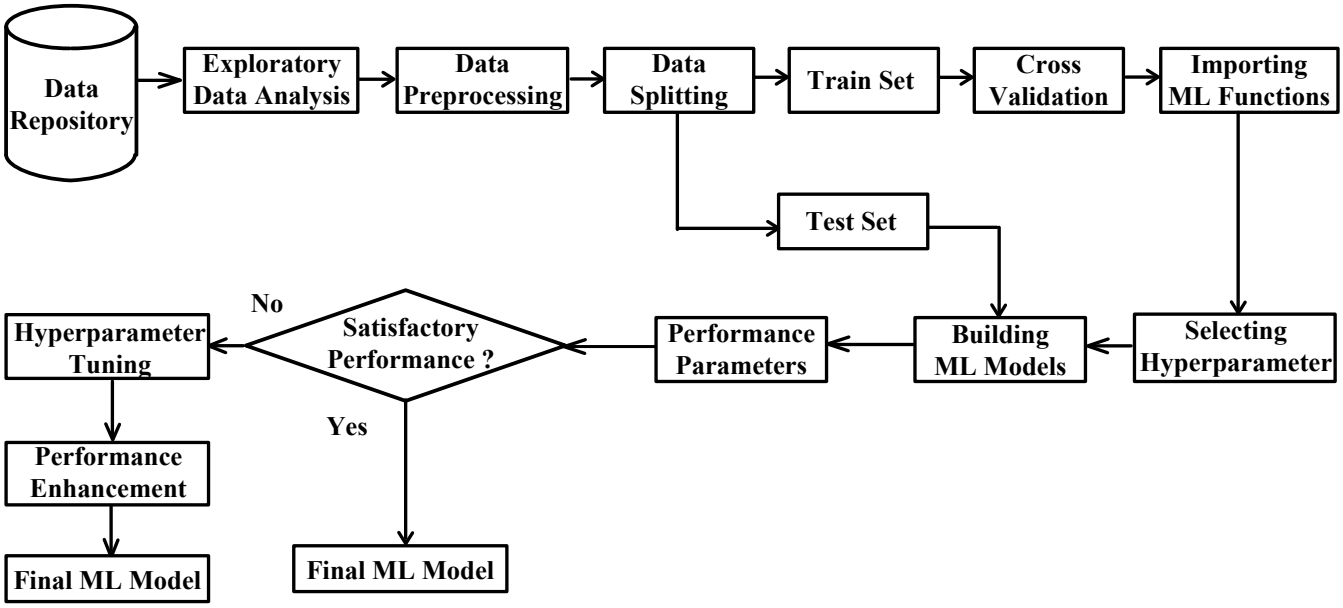


Figure 2. Overall Workflow Diagram

3.2.3 Support Vector Machine (SVM)

Support Vector Machine is one of the most robust algorithms based on the statistical learning framework which offers solution for both regression and classification problems [43]. Using the kernel trick, SVM can classify both linear and non-linear datasets. The datasets are separated by a (n-1) hyper plane, where every data point is considered to be an n-dimensional vector. For a two-dimensional space, hyper plane is a line separating a plane in two parts. A support vector classifier can be defined by the following terms:

$$f(x) = \beta_o + \sum_{i \in S} a_i K(x_i, x_{i'}) \qquad (3)$$

Here,
β_o= Bias
S= Set of observations
α= Model parameters that have to be learned

3.2.4 Decision Tree (DT)

Decision Tree is another supervised learning algorithm whose goal is to train a model to classify a target variable by learning simple chained decision rules from previous input variables [44]. The variables are split recursively based on a set of impurity criteria until some stopping criteria are reached.

The decision tree model resembles an upside-down tree, with the first decision rule at the top and subsequent decision rules dispersed across the tree like branches. Among many impurity measurement systems, Gini impurity is selected for the used model.

$$G(t) = 1 - \sum_{i=1}^{c} p_i^2 \qquad (4)$$

Here,
G(t)= Gini impurity at node t
p_i = Proportion of observation at class c of node t

3.2.5 Random Forest (RF)

Random Forest is a learning algorithm which operates by creating multiple decision trees at training time and providing output class of individual trees [45]. It is applicable for both regression and classification. This model does a small tweak that utilizes the de-correlated tree by building a multitude of decision trees on bootstrapped samples from training data, this process is known as bagging. During bootstrapping, it filters a few numbers of feature columns out of all feature columns. Bootstrap modelling decreases the variance and increases the bias. Predictions of unknown inputs after training can be written as:

$$f' = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \qquad (5)$$

Where,
B= Optimal number of trees

Also, the uncertainty (σ) of the prediction can be written as the folowing:

$$\sigma = \sqrt{\frac{\sum_{b-1}^{B}(f_b(x') - f')^2}{B-1}} \quad (6)$$

3.2.6 Naïve Bayes (NB)

Naïve Bayes is a supervised algorithm which imposes independence of features while classifying data [46]. This model is an effective tool for datasets which have a high number of input features. It considers all the features available including some of the features that have weak effects on the final prediction. The probabilistic model of the Naïve Bayes algorithm can be written as the following equation where A and B are two independent events

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)} \quad (7)$$

3.2.7. Multilayer Perceptron (MLP)

A multilayer perceptron (MLP) is a feed-forward artificial neural network made up of several layers of perceptron [47]. It contains nodes of at least three layers named input node, hidden layer and the output node. This network uses a non-linear activation function which maps weighted inputs to each neuron outputs. In this paper, we used sigmoid functions as the activation functions.

$$y(v_i) = \tanh(v_i) \quad (8)$$

$$y(v_i) = (1 + e^{-v_i})^{-1} \quad (9)$$

The range of the first hyperbolic tangent is -1 to 1 and the second hyperbolic tangent is a logistic function. Learning in the perceptron is carried out by back propagation. Minimized error function (ε) at the output node j, after performing gradient descent, can be written as:

$$\varepsilon(n) = \frac{1}{2}\sum_j e_j^2(n) \quad (10)$$

3.2.8. Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis is a statistical classifier which disjoints two or more classes of data by using a quadratic decision surface [48]. This classifier is used on those cases where there exists a difference between the covariance matrices.

In this classifier, the mean (μ_k) and the covariance matrix $\sum_k$ are estimated separately for each classifier. For a particular input, the objective function is derived where the function is quadratic in 'x' and so the decision boundaries are 0s of quadratic functions.

$$\int \delta_k(x) = \log \pi_k - \frac{1}{2}\mu_k^T \sum_k^{-1} \mu_k + x^T \sum_k^{-1} \mu_k - \frac{1}{2} x^T \sum_k^{-1} x - \frac{1}{2}\log|\sum_k| \quad (11)$$

Where,

$\sum_k$ = Covariance Matrix,

μ_k = Mean,

δ = Objective Function

4. Results

After implementing different machine learning algorithms, necessary simulations have been performed extensively in Python. The confusion matrices are tabulated for each ML model which are depicted from Table 2 to Table 9 consecutively. However, the performances are enhanced by tuning the hyperparameters by random search cross validation.

Table 2. Confusion Matrix for K-Nearest Neighbour

Confusion Matrix KNN			Predicted	
			False	True
Mean	Actual	False	143	7
		True	27	223
Median	Actual	False	143	7
		True	35	215
Mode	Actual	False	144	6
		True	31	219
No Null	Actual	False	115	0
		True	9	34

Table 3. Confusion Matrix for Logistic Regression

Confusion Matrix LR			Predicted	
			False	True
Mean	Actual	False	148	2
		True	4	246
Median	Actual	False	145	5
		True	17	233
Mode	Actual	False	128	22
		True	17	239
No Null	Actual	False	115	0
		True	1	42

Table 4. Confusion Matrix for Decision Tree

Confusion Matrix DT			Predicted	
			False	True
Mean	Actual	False	145	5
		True	11	239
Median	Actual	False	143	7
		True	35	215
Mode	Actual	False	145	6
		True	6	244

No Null	Actual	False	115	0
		True	3	40

Table 5. Confusion Matrix for Random Forest

Confusion Matrix RF			Predicted	
			False	True
Mean	Actual	False	150	0
		True	1	249
Median	Actual	False	150	0
		True	1	249
Mode	Actual	False	150	0
		True	1	249
No Null	Actual	False	115	0
		True	1	42

Table 6. Confusion Matrix for Support Vector Machine

Confusion Matrix SVM			Predicted	
			False	True
Mean	Actual	False	144	6
		True	7	243
Median	Actual	False	143	7
		True	5	245
Mode	Actual	False	144	6
		True	11	239
No Null	Actual	False	115	0
		True	1	42

Table 7. Confusion Matrix for Naïve Bayes

Confusion Matrix NB			Predicted	
			False	True
Mean	Actual	False	150	0
		True	17	233
Median	Actual	False	150	0
		True	14	236
Mode	Actual	False	150	0
		True	14	236
No Null	Actual	False	115	0
		True	15	28

Table 8. Confusion Matrix for Multilayer Perceptron

Confusion Matrix MLP			Predicted	
			False	True
Mean	Actual	False	135	15
		True	20	230
Median	Actual	False	121	29
		True	20	230
Mode	Actual	False	149	1
		True	52	198
No Null	Actual	False	115	0
		True	9	34

Table 9. Confusion Matrix for Quadratic Discriminant Analysis

Confusion Matrix QDA			Predicted	
			False	True
Mean	Actual	False	150	0
		True	26	224
Median	Actual	False	150	0
		True	23	227
Mode	Actual	False	150	0
		True	24	226
No Null	Actual	False	115	0
		True	43	0

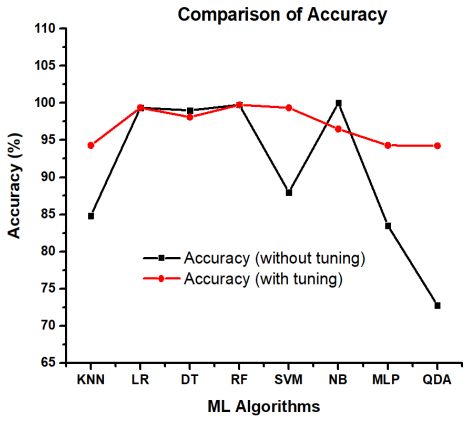
After tweaking the hyperparameters, the machine learning models were trained to have as little bias as feasible in order to minimize overfitting. They were then evaluated via cross-validation to eliminate the possibility of data leakage while maintaining the variance as low as possible. Hence, the performance parameters like accuracy, precision, sensitivity, specificity, F1-score, and the area under the receiver operating characteristic (ROC-AUC) curve are calculated and evaluated accordingly. The graphical comparative presentation of all the performance metrics is displayed in Figure 3(a), 3(b), 3(c), 3(d) and 3(e). The corresponding equations of the parameters are depicted below:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

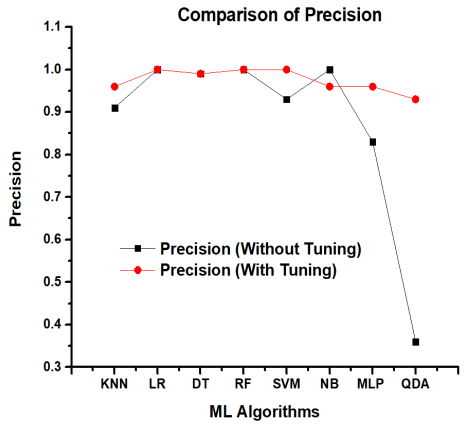
$$Precision = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

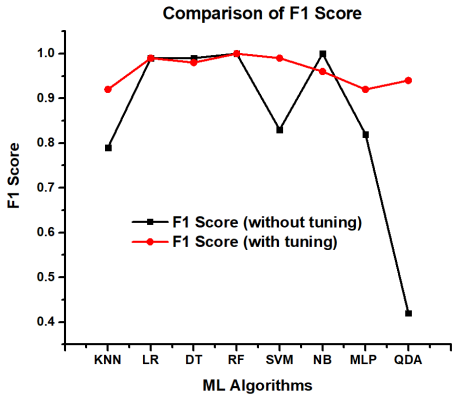
$$F1 - Score = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity}$$



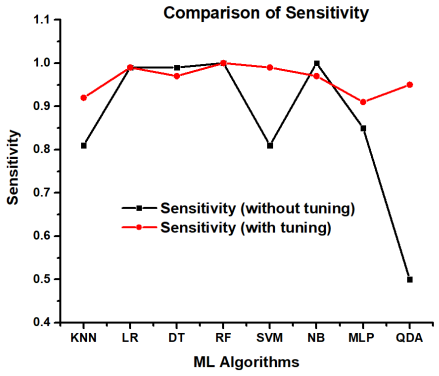
(a)



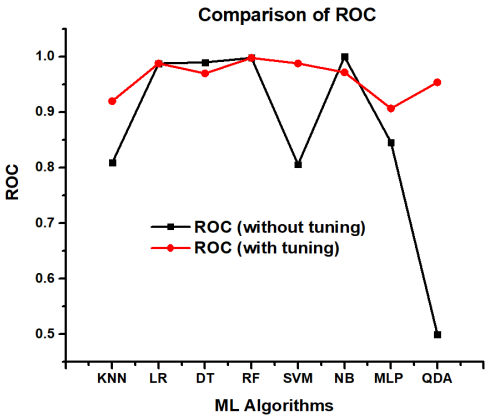
(b)



(c)



(d)



(e)

Figure 3. Comparison of (a) accuracy (b) precision (c) F1 score (d) sensitivity and (e) ROC of different ML algorithms for both without and with tuning of hyperparameters

5. Discussions

A comparative analysis of all the performance metrics for both 'tuning' and 'without tuning' of the hyperparameters are shown in this section. The results are portrayed for four different data frames. Furthermore, Random Search Cross Validation (*RandomizedSearchCV*) was utilized for hyperparameter tuning which uses random hyperparameter combinations to discover the optimum solution for the built model. The value of hyperparameters has a substantial impact on model performance. It is important to note that there is no way to predict the optimum values for hyperparameters in advance, thus all possible values must be attempted to get the optimal values. As doing this manually could take a significant amount of time and resources, *RandomizedSearchCV* is brought into action to automate hyperparameter tuning. Hence, accuracy is calculated for different ML algorithms for both before and after hyperparameter tuning and the results are then tabulated in Table 10. The other performance parameters are presented in Table 11.

After tuning the hyperparameters best accuracy was shown by the Random Forest algorithm which is 99.75%. However, the other algorithms like LR (99.36%), DT (99%), SVM (99.36%), NB (99%) also showed promising results in terms of accuracy. In case of precision, all the models performed more than 95% except QDA. However, it is observed that all the models showed sensitivity more than 95% except MLP and KNN. On the other hand, the all models depicted F1 score and ROC-AUC more than 90%. But Random Forest outperforms all other models with highest value in terms of every performance parameter.

Table 10. Comparison of Accuracy of different ML algorithms

Algorithm		KNN	LR	DT	RF	SVM	NB	MLP	QDA
Accuracy (without Tuning)	Mean	79.25%	98.25%	98.75%	99%	85.5%	95.75%	83.5%	37.5%
	Median	78%	97.75%	98.75%	99.5%	83.25%	96.5%	83.5%	37.5%
	Mode	78%	98.25%	**99%**	99.75%	85%	96.5%	81.25%	37.5%
	Dropping Null	84.81%	99.36%	96.83%	98.73%	87.97%	**99%**	72.78%	72.78%
Accuracy (with Tuning)	Mean	91.5%	98.5%	96%	99.75%	96.75%	95.75%	91%	93.5%
	Median	89.5%	98.75%	96.75%	99.75%	97%	96.5%	87.75%	**94.25%**
	Mode	90.75%	98.5%	97.25%	98%	95.75%	96.5%	86.75%	94%
	Dropping Null	**94.3%**	**99.36%**	98.10%	98.73%	**99.36%**	90.50%	**94.3%**	72.78%

Table 11. Performance Parameters of different ML algorithms

Algorithm		KNN	LR	DT	RF	SVM	NB	MLP	QDA
Precision (without Tuning)	Mean	0.79	0.98	0.98	0.99	0.91	0.95	0.83	0.19
	Median	0.78	0.98	0.98	0.99	0.89	**0.96**	0.83	0.19
	Mode	0.78	0.98	**0.99**	1.0	0.90	**0.96**	0.83	0.19
	Dropping Null	0.91	1.0	0.98	0.99	0.93	1.0	0.36	0.36
Precision (with Tuning)	Mean	0.91	0.98	0.95	1.0	0.96	0.95	0.90	0.93
	Median	0.89	**0.99**	0.96	**1.0**	0.97	**0.96**	0.87	0.93
	Mode	0.90	0.98	0.97	0.97	0.95	**0.96**	0.86	**0.93**
	Dropping Null	**0.96**	1.0	0.99	0.99	**1.0**	0.94	**0.96**	0.36
Sensitivity (without Tuning)	Mean	0.81	0.98	**0.99**	0.99	0.81	0.97	0.81	0.50
	Median	0.80	0.98	0.99	1.0	0.78	0.97	0.82	0.50
	Mode	0.80	0.99	0.99	**1.0**	0.80	0.97	0.85	0.50
	Dropping Null	0.72	0.99	0.94	0.98	0.78	1.0	0.50	0.50
Sensitivity (with Tuning)	Mean	0.92	**0.99**	0.96	1.0	**0.97**	**0.97**	**0.91**	**0.95**
	Median	0.91	0.99	0.97	1.0	0.97	0.97	0.86	0.95
	Mode	**0.92**	0.99	0.97	0.98	0.96	0.97	0.88	0.95
	Dropping Null	0.90	0.99	0.97	0.98	0.99	0.83	0.90	0.50
F1 (without Tuning)	Mean	0.79	0.98	0.99	0.99	0.83	0.96	0.82	0.27
	Median	0.78	0.98	0.99	0.99	0.80	0.96	0.82	0.27
	Mode	0.78	0.98	**0.99**	**1.0**	0.82	0.96	0.81	0.27
	Dropping Null	0.76	**0.99**	0.96	0.98	0.82	**1.0**	0.42	0.42
F1 (with Tuning)	Mean	0.91	0.98	0.96	1.0	0.97	0.96	0.90	0.93
	Median	0.89	0.99	0.97	1.0	0.97	0.96	0.87	**0.94**
	Mode	0.90	0.98	0.97	0.98	0.95	0.96	0.86	0.94
	Dropping Null	**0.92**	0.99	0.98	0.98	**0.99**	0.86	**0.92**	0.42
ROC (without Tuning)	Mean	0.80	0.98	0.99	0.99	0.81	0.97	0.81	0.50
	Median	0.80	0.98	0.99	0.99	0.78	0.97	0.82	0.50
	Mode	0.80	0.99	**0.99**	**0.99**	0.80	0.97	0.85	0.50
	Dropping Null	0.72	0.99	0.94	0.98	0.78	1.0	0.50	0.50
ROC (with Tuning)	Mean	**0.92**	**0.99**	0.96	0.99	0.97	0.97	**0.91**	**0.95**
	Median	0.90	0.99	0.97	0.99	**0.97**	0.97	0.86	0.95
	Mode	0.92	0.99	0.97	0.98	0.96	**0.97**	0.88	0.95
	Dropping Null	0.90	0.99	0.97	0.97	0.99	0.82	0.90	0.50

Finally, the best performing ML models in terms of different performance metrics are tabulated in Table 12. However, our proposed ML model (Random Forest, accuracy 99.75%) is compared with other related research works in Table 13.

Table 12. Best Performing ML Models

Performance Metric	ML Models
Accuracy	Random Forest (99.75%)
Precision	Logistic Regression, Random Forest Support Vector Machine (1.0)
Sensitivity	Random Forest (1.0)
F1-Score	Random Forest (1.0)
ROC	Random Forest (0.998)

Table 13. Comparison with other research works

Research Paper	Algorithm	Best Accuracy
[30]	J48 Classifier	99.00%
[31]	MDF	99.10%
[32]	Decision Tree	98-99%
[33]	SVM	98.50%
[34]	RF	99.00%
[35]	Adaboost	99.50%
[36]	PSO-Adaboost	99.50%
[37]	LSVM	99.60%
[38]	J48 Classifier	95.00%
[39]	Adaboost	99.75%
Proposed	RF	99.75%

6. Conclusion

Kidneys are not only required to filter the toxic substances from the body but also vital for maintaining acid-base balance, electrolyte balance and blood pressure of the body. Malfunction of the kidneys is responsible for mild to fatal diseases as well as dysfunction of other organs of the body. That is why researchers around the world have devoted themselves to seeking ways of precise diagnosis and effective treatment of kidney diseases. As machine learning techniques have become more prevalent in the medical field for diagnosing, chronic kidney disease (CKD) is now on the list of diseases that can be predicted leveraging machine learning algorithms. All the researches to identify CKD using ML algorithms has improved the process and result accuracy from day to day. In our work, we have proposed the random forest algorithm (accuracy 99.75%) as the most efficient algorithm among all other algorithms. In this investigation, the data are processed efficiently as the missing values are handled with four different criteria like mean, mode, median and null dropping method. Moreover, the study also focuses on measuring the performances of the ML models for both tuning and without tuning of the hyperparameters. Significant improvements in performances of the ML models are witnessed which are presented graphically. Overall, the study explores the applicability of the supervised machine learning algorithms in bioinformatics and presents their compatibilities in diagnosing various fatal diseases like CKD at an early stage. This research will guide future researches on predictive analysis of other health conditions with machine learning algorithms where applicable, and help to polish and correct the techniques further. We plan to collect datasets from local health care facilities to estimate the regional parameters in order to develop a diagnostic model in the context of Bangladesh. Furthermore, we will explore deep learning and neural networking methods, and apply them to hone the process to near perfection.

Conflict of Interest

No author has any conflict of interest.

References

[1] Cohen, S., Kamarck, T, and Mermelstein, R., "A global measure of perceived stress," *J. Health Soc. Behav.*, vol. 24, no. 4, 1983, 385-396.

[2] Rayan, Z., Alfonse, M., and Salem, A. B. M., "Machine Learning Approaches in Smart Health," *Procedia Computer Science*, vol. 154, 2018, 361–368

[3] Mccullough, P. A., Agrawal, V., Danielewicz, E., and Abela, G. S., "Accelerated Atherosclerotic Calcification and Mö nckeberg's Sclerosis: A Continuum of Advanced Vascular Pathology in Chronic Kidney Disease," *Clin J Am Soc Nephrol*, vol. 3, 2008, 1585–1598

[4] "Chronic Kidney Disease | Harrison's Principles of Internal Medicine, 20e | AccessMedicine | McGraw-Hill Medical."[Online].Available:https://accessmedicine.mhmedical.com/content.aspx?bookid=2129&sectionid=186950702. [Accessed: 30-Dec-2020].

[5] "Chronic Kidney Disease Clinical Presentation: History, Physical Examination." [Online]. Available: https://emedicine.medscape.com/article/238798-clinical. [Accessed: 30-Dec-2020].

[6] Vanholder, R., Van Laecke, S., Glorieux, G., Verbeke, F., Castillo-Rodriguez, E., and Ortiz, A., "Deleting death and dialysis: Conservative care of cardio-vascular risk and kidney function loss in chronic kidney disease (CKD)," *Toxins*, 10 (6), 2018, 237-300.

[7] Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., Saran, R., Wang, A. Y. M., and Yang, C. W., "Chronic kidney disease: Global dimension and perspectives," *The Lancet*, vol. 382, no. 9888, 2013, 260–272

[8] M. Naghavi *et al.*, "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013," *Lancet*, vol. 385, no. 9963, 2015, 117–171.

[9] Schieppati, A., and Remuzzi, G., "Chronic renal diseases as a public health problem: Epidemiology, social, and economic implications," *Kidney International*, 68 (98), 2005, S7-S10.

[10] Stanifer, J. W., Muiru, A., Jafar, T. H., and Patel, U. D., "Chronic kidney disease in low- and middle-income countries," *Nephrology Dialysis Transplantation*, 31 (6), 2016, 868–874.

[11] Mills, K.T., Xu, Y., Zhang, W., Bundy, J. D., Chen, C. S., Kelly, T. N., Chen, J., and He, J., "A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010," *Kidney Int.*, vol. 88, no. 5, 2015, 950–957.

[12] "Chronic kidney disease prevalence among health care providers in Bangladesh - PubMed." [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20639836/. [Accessed: 30-Dec-2020].

[13] Anand, S., Khanam, M. A., Saquib, J., Saquib, N., Ahmed, T., Alam, D. S., Cullen, M. R., Barry, M., and Chertow, G. M., "High prevalence of chronic kidney disease in a community survey of urban Bangladeshis: A cross-sectional study," *Global. Health*, 10 (1), 2014, 1-7.

[14] Huda, M. N., Alam, K. S., and Harun-Ar-Rashid, "Prevalence of chronic kidney disease and its association with risk factors in disadvantageous population," *Int. J. Nephrol.*, vol. 2012, 2012, 1-7.

[15] Asif, M. A. A. R., Nishat, M. M., Faisal, F., Shikder, M. F., Udoy, M. H., Dip, R. R., and Ahsan, R., "Computer Aided Diagnosis of Thyroid Disease using Machine Learning Algorithms." In *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*, pp. 222-225. IEEE, 2020

[16] Nishat, M. M., Faisal, F., Mahbub, M. A., Mahbub, M. H., Islam, S., and Hoque, M. A., "Performance Assessment of Different Machine Learning Algorithms in predicting Diabetes Mellitus", *Biosc. Biotech. Res. Comm*., vol. 14, no. 1, 2021, 74-82.

[17] Faisal, F., and Nishat, M. M., Faisal, "An Investigation for Enhancing Registration Performance with Brain Atlas by Novel Image Inpainting Technique using Dice and Jaccard Score on Multiple Sclerosis (MS) Tissue", *Biomedical and Pharmacology Journal*, 12 (3), 2019, 1249-1262.

[18] Farazi, M. R., Faisal, F., Zaman, Z., & Farhan, S., "Inpainting multiple sclerosis lesions for improving registration performance with brain atlas", In *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec,)* pp. 1-6. IEEE, 2016.

[19] Nishat, Mirza Muntasir, Tasnimul Hasan, Sarker Md. Nasrullah, Fahim Faisal, Md. Asfi-Ar-Raihan Asif & Md. Ashraful Hoque. "Detection of Parkinson's Disease by Employing Boosting Algorithms." *2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 5th International Conference on Imaging, Vision & Pattern Recognition (ICIVPR*), IEEE, 2021, in press

[20] Nishat, Mirza Muntasir, Fahim Faisal, Tasnimul Hasan, Sarker Md. Nasrullah, Afsana Hossain Bristy, Md. Minhajul Islam Shawon & Md. Ashraful Hoque. "Detection of Autism Spectrum Disorder by Discriminant Analysis Algorithm." *2021 International Conference on Big Data, IOT & Machine Learning (BIM)*, September 23-25, Springer, 2021, in press

[21] Faisal, Fahim, Mirza Muntasir Nishat, Md. Ashif Mahbub, Md. Minhajul Islam Shawon, and Md. Mahbub-Ul-Huq Alvi. "Covid-19 and its impact on school closures: a predictive analysis using machine learning algorithms." *In 2021 International Conference on Science and Contemporary Technologies (ICSCT)*, IEEE, 2021, in press

[22] Battineni, G., Chintalapudi, N., and Amenta, F., "Performance analysis of different machine learning algorithms in breast cancer predictions", *EAI Endorsed Transactions on Pervasive Health and Technology*, *6*(23), 2020

[23] Sahu, B., Mohanty, S. and Rout, S., "A hybrid approach for breast cancer classification and diagnosis", *EAI Endorsed Transactions on Scalable Information Systems*, *6*(20), 2019.

[24] Asif, M. A. A. R., Nishat, M. M., Faisal, F., Dip, R. R., Udoy, M. H., Shikder, M. F., and Ahsan, R., "Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease." *Engineering Letters,* vol. 29, no. 2, 2021, 731-741.

[25] Chakraborty, C., and Abougreen, A. N., "Intelligent Internet of Things and Advanced Machine Learning Techniques for COVID-19." *EAI Endorsed Transactions on Pervasive Health and Technology,* 7(26), 2021, e1

[26] Kaur, P., and Sharma, M., "Analysis of data mining and soft computing techniques in prospecting diabetes disorder in human beings: a review." *Int. J. Pharm. Sci. Res,* vol. 9, 2018, 2700-2719

[27] Gautam, R., Kaur, P., and Sharma, M., "A comprehensive review on nature inspired computing algorithms for the diagnosis of chronic disorders in human beings." *Progress in Artificial Intelligence,* vol. 8, no. 4, 2019, 401-424

[28] Sharma, S., and Singh, G., "Diagnosis of cardiac arrhythmia using Swarm-intelligence based Metaheuristic Techniques: A comparative analysis." *EAI Endorsed Transactions on Pervasive Health and Technology,* 6(23), 2020

[29] Bharadwaj, Hemantha Krishna, et al. "A Review on the Role of Machine Learning in Enabling IoT Based Healthcare Applications." *IEEE Access,* vol. 9, 2021, 38859-38890

[30] Avci, E., Karakus, S., Ozmen, O., & Avci, D. "Performance comparison of some classifiers on Chronic Kidney Disease data." In *2018 6th International Symposium on Digital Forensic and Security (ISDFS),* pp. 1-4. IEEE, 2018.

[31] Gunarathne, W. H. S. D., Perera, K. D. M., and Kahandawaarachchi, K. A. D. C. P., "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)." In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE),* pp. 291-296. IEEE, 2017.

[32] Tazin, N., Sabab, S. A., and Chowdhury, M. T., "Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique." In *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec),* pp – 1-6. IEEE, 2016.

[33] Polat, H., Mehr, H. D., and Cetin, A., "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods." *Journal of medical systems,* vol. 41, no. 4, 2017, 41-55.

[34] Devika, R., Avilala, S. V., and Subramaniyaswamy, V., "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest." In *2019 3rd International Conference on*

Computing Methodologies and Communication (ICCMC), pp. 679-684. IEEE, 2019.

[35] Başar, M. D., Sarı, P., Kılıç, N., and Akan, A., "Detection of chronic kidney disease by using Adaboost ensemble learning approach." In *2016 24th Signal Processing and Communication Application Conference (SIU),* pp. 773-776. IEEE, 2016.

[36] Indriani, A. F., & Muslim, M. A. "SVM Optimization Based on PSO and AdaBoost to Increasing Accuracy of CKD Diagnosis." *Lontar Komputer*, vol. 10, no. 02, 2019, 119-127.

[37] Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., Jasiński, M., Jasiński, Ł., Gono, R., Jasińska, E., and Bolshev, V., "Prediction of chronic kidney disease-a machine learning perspective," *IEEE Access*, vol. 9, 2021, 17312-17334.

[38] Sobrinho, A., Queiroz, A. C. M. D. S., Dias Da Silva, L., De Barros Costa, E., Eliete Pinheiro, M., and Perkusich, A., "Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries: A Comparative Analysis of Machine Learning Techniques," *IEEE Access,* vol. 8, 2020, 25407–25419

[39] Nishat, M. M., Faisal, F., Dip, R. R., Shikder, M. F., Ahsan, R., Asif, M. A. A. R., and Udoy, M. H., "Performance Investigation of Different Boosting Algorithms in Predicting Chronic Kidney Disease." In *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI),* pp. 1-5. IEEE, 2020

[40] https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[41] Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y., and Cheng, C. Y., "Logistic regression was as good as machine learning for predicting major chronic diseases," *Journal of clinical epidemiology,* vol. 122, 2020, 56-69

[42] Li, W., Chen, Y., and Song, Y., "Boosted K-nearest neighbour classifiers based on fuzzy granules," *Knowledge-Based Systems,* vol. 122, 2020.

[43] Wang, M., and Chen. H., "Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis," *Applied Soft Computing,* vol. 88, 2020.

[44] Ghiasi, M. M., Zendehboudi, S. and Mohsenipour, A. A., "Decision tree-based diagnosis of coronary artery disease: CART model," *Computer Methods and Programs in Biomedicine*, vol. 192, 2020

[45] Jiang, N., Fu, F., Zuo, H., Zheng, X., and Zheng, Q., "A Municipal PM2. 5 Forecasting Method Based on Random Forest and WRF Model," *Engineering Letters,* vol. 28, no. 2, 2020, 312-321.

[46] Jackins, V., Vimal, S., Kaliappan, M., and Lee, M. Y., "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *The Journal of Supercomputing*, vol. 77, no. 5, 2021: 5198-5219.

[47] Theerthagiri, P., "Prognostic Analysis of Hyponatremia for Diseased Patients Using Multilayer Perceptron Classification Technique." *EAI Endorsed Transactions on Pervasive Health and Technology,* vol. 7, no. 26, 2021, e5.

[48] Silitonga, P., Bustamam, A., Muradi, H., Mangunwardoyo, W., and Dewi, B. E., "Comparison of Dengue Predictive Models Developed Using Artificial Neural Network and Discriminant Analysis with Small Dataset," *Applied Sciences*, vol. 11, no. 3, 2021