Intelligent Chronic Kidney Disease Diagnosis System using Cloud Centric Optimal Feature Subset Selection with Novel Data Classification Model

Pramila Arulanthu (pramimark@gmail.com)
Alagappa University
Eswaran Perumal
Alagappa University

Research Article

Keywords: Chronic Kidney Disease, IoT, Cloud based decision support system, Feature selection, Classification,

Posted Date: June 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-380904/v1

Intelligent Chronic Kidney Disease Diagnosis System using Cloud Centric Optimal Feature Subset Selection with Novel Data Classification Model

Pramila Arulanthu[1] and Eswaran Perumal[2]

Research Scholar (Ph.D.)[1], Assistant Professor[2]
Department of Computer Applications
Alagappa University, Karaikudi, India.
pramimark@gmail.com[1], eswaran@alagappauniversity.ac.in[2]

Abstract

Internet of Things (IoT) and cloud computing offers diverse applications in the medicinal sector by the integration of sensing and therapeutic gadgets. Medical expenses are rising gradually and different new diseases also exist globally, it becomes essential to transform the healthcare facilities from a hospital to patient-centric platform. For providing effective remote healthcare services to patients, this paper introduces an optimal IoT and cloud based decision support system for Chronic Kidney Disease (CKD) diagnosis. The proposed method makes use of simulated annealing (SA) based feature selection (FS) with Root Mean Square Propagation (RMSProp) Optimizer based Logistic Regression (LR) model called SA-RMSPO-LR to classify the existence of CKD from medical data. The proposed model involves a set of four subprocesses, which include data collection, preprocessing, FS, and classification. The inclusion of SA for FS helps to improvise the classifier results of the SA-RMSPO-LR model. The effectiveness of the SA-RMSPO-LR model has been validated using a benchmark CKD dataset. The experimental results indicated that the proposed SA-RMSPO-LR model leads to effective CKD classification with the maximum sensitivity of 98.41%, specificity of 97.99%, accuracy of 98.25%, F-score of 98.60% and kappa value of 96.26%. The experimental outcome indicates that the proposed SA-RMSPO-LR model has the capability to detect and classify CKD over the compared methods proficiently.

Keywords: Chronic Kidney Disease, IoT, Cloud based decision support system, Feature selection, Classification,

1. Introduction

Internet of Things (IoT) is a significant model that concentrates on modeling and interconnection of Internet-linked things under the application of computer systems. IoT is mainly applied for diverse applications using a maximum number of low power devices such as wrist band, fridge, umbrella, and so forth, instead of using a minimum number of high power devices like computers, tablets, smartphones, etc. [1]. IoT and Cloud Computing (CC) are advantageous when it is combined for developing a novel technique [2]. An observance model has been designed under the integration of IoT and CC to observe the patient's condition and effectively collect the details even at the remote areas that may be more helpful for medical physicians. In several cases, the IoT method is often retained with the help of CC environment to enhance efficiency concerning productive resource application, data storage, computation and processing abilities. Besides, CC earns the merits of IoT by extending the value of tackling the current world issues and dynamically providing new facilities.

The combination of IoT and CC is a relied platform; it serves quite-well when compared with conventional CC based performance [3]. Some of the major applications such as: healthcare, armed forces, user appliances as well as banking sectors that exploit the concatenation of IoT and CC method. Among other domains, medical and healthcare which has two challenging works that tend to develop various techniques in medical tools and screening gadgets [4]. Usually, the medical costs are more expensive and also several diseases may exist around the world; it is significant to transform the healthcare service from a

clinic or hospital to a user-based environment. *Thibaud, M. et al.* have presented a Clinical Decision Support System (CDSS) under the application of detecting capabilities of IoT devices to forecast the existence of the severe or chronic disease in a person [5]. In this approach, an IoT and CC based CDSS have been proposed by the employment of computational sciences [6].

Under the application of IoT tools in healthcare, useful metrics were applied for collecting the data, such that continuously varying health measures within a given period of time and existence of several unusual cases at any circumstances. Also, IoT devices and healthcare sensor values are employed for the disease analysis with the base of severity level. Personal healthcare service can be offered with the help of applying IoT and CC facilitates to live a healthy life at a minimum cost. Thus, an effective healthcare mechanism is more essential for the disease analysis by employing medical IoT tools. Since more number of data has been developed by the IoT devices in medical application, data science is capable of making the IoT method still a smarter technique.

Here, data science is defined as a multidisciplinary domain, which is a combination of Data Mining (DM), Machine Learning (ML) and few more techniques that encounter the patterns and novel directions from the data [7]. ML method is one of the vital applications in CDSS that can handle large-scale data. Much number of existing data investigation methods are Neural Network (NN) [8], classification approaches, clustering schemes and few more efficient techniques. Data can be developed under diverse sources using specified data type which is more required for developing models which need to be capable of handling diverse features of data. In IoT, more amounts of data sources could develop essential data in real-time application without the problem of scalability, velocity and to compute the effective approach. Such an issue tends to various opportunities for developing novel methodologies [9]. Systematic review have presented on CKD classification and disease diagnosis system [10]. All the existing risk factors, issues and problems on disease diagnosis system have explained properly.

However, chronic kidney disease is a severe health issue; it affects 10-15% of people. The beginning level of CKD does not exhibit any signs and it is difficult to recognize it using some tests like blood or urine. By the early identification of CKD, preventive measures can be taken and useful medications can be provided for controlling it. As the CKD dataset is comprised with several features [11], it degrades the classification task because of the irregular features. To overcome these limitations, FS method is established for selecting the essential features and removing the unnecessary features, so that the processing time would be reduced with better classification outcome. The FS problem can be reported as a combinatorial optimization issue [12]. Hence, the developing variables are addition (1) or subtraction (0) of features. An extensive FS involves applying massive integrations in which (2^N, where N is the count of features). To attain an optimized FS process with rapid value, optimization methods such as Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) were employed.

This paper introduces an optimal IoT and cloud based decision support system for CKD diagnosis for providing effective remote healthcare services to patients. The proposed method makes use of a simulated annealing (SA) based feature selection (FS) with Root Mean Square Propagation (RMSProp) based Logistic Regression (LR) model called SA-RMSPO-LR to classify the existence of CKD from medical data. The proposed model involves a set of four subprocesses, includes data collection, preprocessing, FS and classification. The inclusion of SA for FS helps to improvise the classifier results of SA-RMSPO-LR model. The effectiveness of the SA-RMSPO-LR model has been validated using a benchmark CKD dataset. The experimental result indicates that the proposed SA-RMSPO-LR model has the competence to classify and diagnose CKD over the compared methods proficiently.

The upcoming sections of the paper are organized as follows. Section 2 briefs the existing works related to this proposed system. The proposed SA-RMSPO-LR model is discussed in Section 3 and simulated results are examined in Section 4. Finally section 5 presents conclusion of this proposed CKD diagnosis system with future directions.

2. Related works

Diverse type of methods has been deployed for the efficient detection of CKD under the application of patient's medical information. Cuckoo Search trained Neural Network (NN-CS) approach is projected for identifying the existence of CKD [13]. The main goal of the proposed system is to solve the problems

involved in local search based learning techniques. The CS approach is more applicable in selecting the input weight vector of NN in order to provide an appropriate training for the data. The classification process in a deployed model concentrates on providing optimal performance. Modified approaches of NN-CS (NN-MCS) [14] have been established for resolving the issue exist in local optimum of NN-CS technique. The primary weights of neuron link handles the function of NN, and the projected model applies MCS scheme to reduce the Root Mean Square Error (RMSE) measure that is used at the time of NN training. The attained results show that NN-MCS model reached an optimized function when compared with NN-CS approach.

Chen, Z. et al. [15] have introduced two fuzzy classification models such as fuzzy rule-building expert system (FuRES) and Fuzzy Optimal Associative Memory (FOAM) it has been applied to find the presence of CKD. FuRES produces a classification tree that includes a minimal NN. It tends to develop the classification rules for computing the weight vector using lower fuzzy entropy. The two fuzzy classifiers were utilized to diagnose the CKD patients. Additionally, FuRES is related to FOAM in case of training, and forecasting task, that has same intensity of noise. FuRES and FOAM has accomplished an optimal function in the CKD analysis; simultaneously, FuRES is more important when compared with FOAM.

Arasu, S. D., & Thirumalaiselvi, R has proposed a new technique termed as Weighted Average Ensemble Learning Imputation (WAELI) [16]. The missing values present in a dataset minimize the accuracy level of CKD. Since the traditional approaches are applied with data preprocessing process, the data cleaning task is essential to occupy the missing values and eliminate the ineffective scores. A revaluation strategy is projected for every CKD phases in which missing values have been estimated and placed in corresponding locations. Even though the conventional models are productive, it still requires a professional disease diagnosis system in healthcare dataset to assure the CKD values.

Here, FS task is treated as a vital portion in the region of data classification that is applied for identifying tiny set of rules from a training dataset with permanent objectives. There are various models such as AI (Artificial Intelligence), bio-inspired mechanism that is utilized in FS process. *Tan, K. C., et al.* projected a wrapper technique which is mainly used for hybridization of GA using Support Vector Machine (SVM) known as GA-SVM approach that selects the feature subset in an optimized manner [17]. The minimization of repeated features of presented system enhances the classification task that has been verified under the application of five various disease dataset. In addition, *Chetty, N.,* et al. deployed a wrapper scheme for identifying CKD using 3 phases, a technique is produced from DM, Wrapper subset attribute calculator as well as best first search model is applied for selecting attributes and classifier [18]. The experimental results stated that, the accuracy has been incremented for lower dataset than the actual dataset. *P. Arulanthu and E. Perumal* have suggested classifiers for effective CKD classification and prediction with reduced attribute information [19].

Wibawa, M. S., et al. have introduced an approach to improve the superiority of CKD [20]. This model contributes in 3 steps like FS, ensemble learning as well as classification process. The combination of Correlation-based FS (CFS) and k-nearest neighbour (kNN) classification concludes in maximum classification accuracy. *Polat, H., et al.* have applied alternate CKD identification model under the application of filter and wrapper methodologies [21]. Therefore, the results attained from this method reveal that, the reduced number of features could not ensure the efficiency of classification task. *Pramila Arulanthu and Eswaran Perumal* have proposed efficient online CKD diagnosis method using IoT and cloud support system for easy identification [22].

An intelligent prediction and classification model for healthcare sector using Density based Feature Selection (DFS) with Ant Colony based Optimization (D-ACO) model for CKD has been developed [11]. The presented model makes use of DFS to select features and applies ACO for data classification.

IoT enabled cloud based disease diagnosis model for CKD has been presented in [23]. Deep Neural Network (DNN) classifier is used to predict CKD with its severity level. Besides, PSO algorithm is applied to increase the classification results by selecting the required features.

Set of two ensemble models such as Bagging and Random Subspace methods on three base-learners – k Nearest Neighbours, Naïve Bayes and Decision Tree has been presented in [24] to improve the classifier outcome. The presented model involves data preprocessing for handling the missing values and data scaling for the normalization of the range of independent variables.

3. Proposed CKD Diagnosis Model

Entire process of the proposed SA-RMSPO-LR model has been shown in Figure 1. As shown in figure 1, data collection process takes place in diverse ways. Followed by data preprocessing takes place and then the preprocessed data is provided to the SA-FS model. The SA-FS model will choose the optimal subset of features and then classification process is carried out by the RMSPO-LR model. These processes have been discussed in the following subsections.

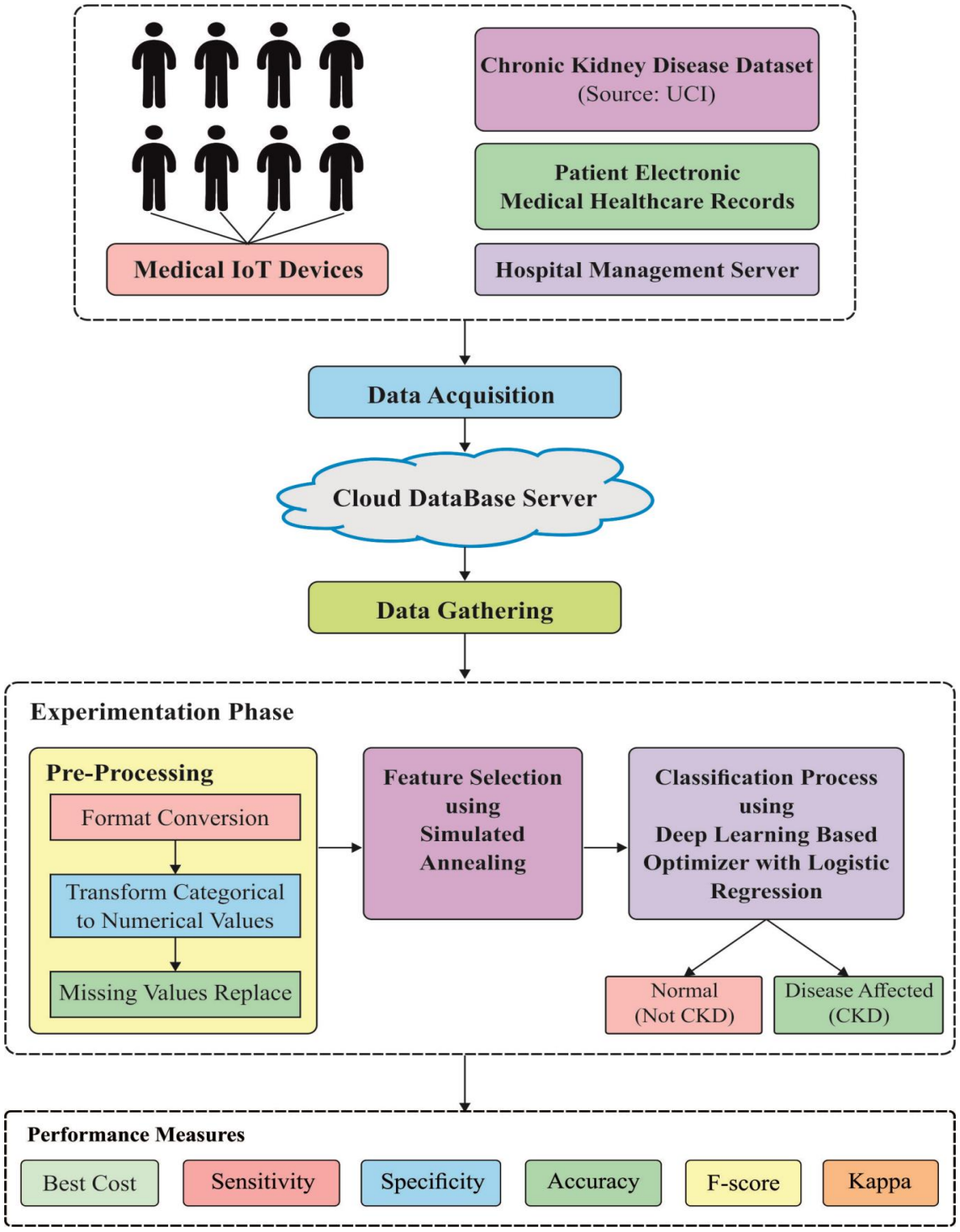


Figure 1. Overall process of the proposed CKD diagnosis system.

3.1. Data Collection

In the first stage, data collection process takes place where the data is collected using IoT gadgets which are linked to patients, standard benchmark medical dataset, patient health data and hospital management server. The medical data collected by IoT tools linked to a person are mostly required. In general, a sensor which is linked to a human being assembles the specific medical data in a regular time interval. The deployed SA-RMSPO-LR model exploits the 4G network for transmitting monitored data for CDSS. Finally won benchmark CKD real-time dataset has been applied for disease diagnosis [26]. The medicinal dataset is composed with the patient data which are gathered from hospitals and saved in hospital management server. Then, a data collection tool represents the capability of collecting required data and transmits it to the CDSS.

3.2. Pre-processing

In this phase, CKD data has been transformed into suitable form under three various steps. At the initial stage, format transforming task is carried out in which actual data is converted as .csv file format. At the same time, few categorical values in a dataset like Yes, No, Absent or Present are converted as arithmetic scores like '0' and '1'. Finally, the missing values in a data would be occupied using median process.

3.3. SA based FS

3.3.1. Simulated Annealing (SA): SA is defined as one of the random global optimization approaches, which expect the differentiability and multimodality of an objective function. It refers the annealing operation of external system present in statistical models. Figure 2 shows the general process of SA method. The physical task is that, a substance has been melted and cools with a minimum speed to ensure the process of reaching thermal equilibrium at every temperature, and if the temperature is equal to *0*, then it obtains the crystalline lattices of lower energy which is named as ground state. When a maximum temperature is lower when compared to the melting point then, it solidifies as sub-optimal configuration that is not comprised with lower energy. In addition, the emergence of a substance at a fixed temperature relied on Monte Carlo approaches. Provided with a current state l of substance with energy E_l, the subsequent state m with energy E_m have been produced under a tiny random perturbation to state l. While E_m is less than or equal to E_l, then the state m could be approved as current state. Else, the state m has been consumed with a probability as provided below:

$$\exp\left((E_l - E_m)/(k_B T)\right) \quad (1)$$

where T denotes the present temperature and k_B implies Boltzmann constant. The acceptance rule mentioned is named as Metropolis criterion, and a model which has been operated along with it is called as Metropolis technique. Since the temperature has been decreasing gradually, the substance attains a thermal equilibrium at every temperature. Hence, thermal equilibrium undergoes characterization using Boltzmann distribution that offers the probability of a substance in state l with energy E_l at temperature T. It can be expressed as:

$$P_T(X = l) = \exp\left(-E_l/(k_B T)\right) \Big/ \sum_{m} \exp\left(-E_m/(k_B T)\right) \quad (2)$$

where the denominator is the addition of energy of every feasible state at temperature T.

3.3.2. Optimal Feature Subset Selection Problem: The FS issue for CKD is assumed to be a combinatorial optimization problem. The count of viable feature integration from a feature set which has 24 features that are computed as 2^{24}–1. It refers that, it is not possible to process the estimation of every feature combinations. Hence, optimal feature subset selection issue can be described in the following way:

Definition 1 Optimal feature subset selection problem

Given a feature set $f = \{f_1, f_2, f_3, \ldots, f_n\}$ and a cost function $: f \rightarrow q\,(0 \leq q)$, identify feature subset (s) f' where the measure of cost function has been reduced.

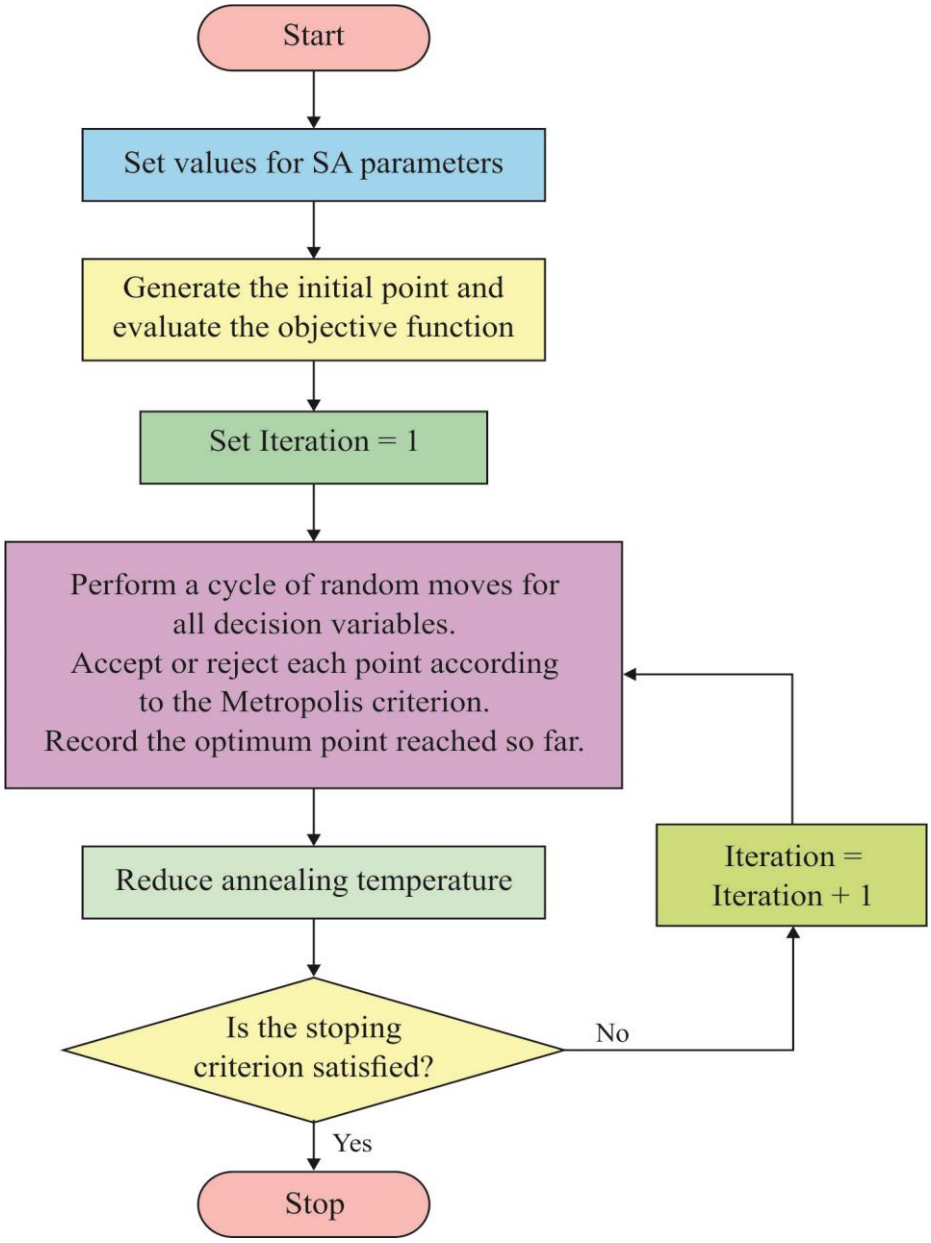


Figure 2. General process of the simulated annealing model.

3.3.3. Feature Selection Approach based on SA: Here, it has been developed with the FS technique for CKD diagnosis. The presented FS method depends upon SA method that has been vastly applied for combinatorial optimizing task. SA model is assumed to be major searching techniques. But, if a naïve local search algorithm applies a greedy approach to identify the best solution, SA is one of the probabilistic scheme which activates to leave the local optimum for identifying better solutions. For this purpose, SA model has the nature of obtaining optimal solution when compared with naïve local search algorithm at any circumstances.

i) Solutions

A solution applied for FS method is shown by a binary vector f with a length of 41 as shown in Eqn. (3). The value 1 is allocated for selected feature, whereas 0 is declared for unselected feature.

$$f =< f_1, f_2, f_3, \ldots, f_{24} >, where\ f_l \in \{0, 1\}, 0 \leq l \leq 24 \quad (3)$$

Various types of search modules were employed for handling the optimization issue which requires a primary solution. The possible solution is selected in a random manner and applied as initial solution. Simultaneously, the neighboring solutions are provided for binary vectors with a single bit, which have varied from the predefined solution. For instance, the neighboring solution applies a set of every 24 features, which has a set of 24 binary vectors with single bit of value $0. < 0, 1, 1, 1, \ldots, 1 >$ which is a single function among others.

ii) Cost function

Major aspect on the process of optimization of heuristic approaches like SA is based on the cost function which helps to estimate the single solutions. Besides, the working function of a technique is highly based on the definition of cost function and the way it was defined. In addition, the cost φ for the provided record x from a training data can be estimated by

$$\varphi(X) = \begin{cases} 1 & if\ p(x) = q(x) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$\varphi(x)$ refers the set to 1 while class $p(x)$ is computed by clustering model that is equal to actual class $q(x)$ where the record $\mathbf{x}$ is constrained, else, it is assigned to 0. The cost $C(f)$ for a given solution f has been evaluated across the training data set with the application of Eqn. (5).

$$C(v) = \frac{1}{\sum_{l=1}^{N} \varphi\,(x_l)} \quad (5)$$

iii) Some other parameters

The SA model applies the cooling procedure for discovering the best solutions to eliminate local optima at the time of exploring the solution space. Basically, a cooling method is defined as that, to define the way of searching an optimal solution. Some of the attributes which are initial temperature, a temperature reducing function as well as stopping criteria has to be employed.

The initial temperature T should be high to enable the transition which has to be approved. A measure of 100,000 was allocated as initial value T, that is higher when compared with size of a training data set. The temperature reduction function is a simple iteration that is a product of T combined with a constant r.

$$T \leftarrow r \times T \quad (6)$$

where the value of $\mathbf{r}$ is declared to 0.9. Finally, a termination criteria is that, when a score of T is lower than 0.001 then, the method terminates at 0.001 and computed with various process.

iv) Procedure of algorithm

Generally, an initial solution has been selected in a random manner as it has been considered as optimal solution. Consecutively, the cost of an initial solution can be estimated under the application of cost function. When temperature T is not capable to meet the termination criteria, a neighboring solution of recent optimal solution were chosen and evaluate the cost. When a cost of selected neighboring solution is less than or equal to recent optimal solution, then current optimal solution can be interchanged with novel neighbour solution. While a cost of neighboring solution is higher when compared with current optimal solution, an arbitrary value q has been selected at the range of (0, 1). At this point, the replacement of best solution is activated when random value $\mathbf{q}$ is minimum when compared with $e^{-\frac{Cost(v_n)-Cost(v_b)}{T}}$. Once the temperature $\mathbf{T}$ is decreased by Eq. (6), it is repeated until T meets the termination criteria.

Pseudo-code for the FS technique depends upon SA as given below:

Algorithm 1: SA Based Feature Selection

Input: Training data set

Output: Combination of features: v_b

$v_b \leftarrow$ Null;

$T \leftarrow 100000$;

$r \leftarrow 0.9$;

Produce an initial solution, v_l;

$v_b \leftarrow v_l$;

Compute the cost of initial solution, *Cost* (v_b);.

while $(T > 0.001)$ do

Begin

arbitrarily choose a nearby solution, v_n, of v_b that has one bit varied from v_b;

if (Cost(v_b) = Cost(v_n)):

$v_b \leftarrow v_n$;

else

Create an arbitrary number q regularly in the range $(0,1)$;

if $(q < e^{-\frac{Cost(v_n)-Cost(v_b)}{T}})$

$v_b \leftarrow v_n$;

$T \leftarrow r \times T$

End

3.4. LR- based classification model

Classification task tends to develop a presentation which undergoes mapping with data items of provided classes that depends upon recent data. The elimination of required data substance from a method is applied for detecting the nature of data. In most of the cases, the LR model has been relied on a variable to perform the binary classification. In order to specify LR approaches, it is concerned with 2 types of difficulties. The key intention of this technique is, to predict the presence of CKD that is often processed by a binary classification approach. In addition, the LR methods are often applied for identifying diseases, DM and classification of health care data.

LR represents the function of predicting the existence or absence of CKD. LR model is mainly depends upon linear regression approach as represented in Eqn. (7):

$$P = \alpha + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_m x_m \tag{7}$$

Classification issues are same as linear regression issues which requires the nonstop measures. The expected value of classification model is from the range of [0, 1]. The result might be 1, if the values are more than a threshold value; else, it is 0. Hence, the parameter sequence of LR comes under the sequence of [0, 1]. According to LR, it uses a sigmoid function level. A feature is defined as total linear model that detects the disease under the application of sigmoid function.

$$\Pr(P = +1|Q) \sim \beta.Q and \Pr(P = -1|Q) = 1 - \Pr(P = +1|Z) \tag{8}$$

$$\downarrow \sigma(x) := \frac{1}{1 + e^{-x}} \in [0,1] \tag{9}$$

$$\Pr(P = +1|Q) \sim \sigma(\beta.Q) and \Pr(P = -1|Q) = 1 - \Pr(P = +1|Q) \tag{10}$$

Here, a classification is implemented at negative and positive class. Where, P is the presence of CKD, Q shows the autonomous variables for the collection of eight elements. Each dependent variable Q has been allocated to a coefficient value named as β that specifies the weight. As it is identified by LR model, the database values are comprised with weight values. For diverse weights, it specifies different links among

Q and P. Parameters in LR can be modified to reach optimized classification outcome. At this point, RMSProp model is used to choose the parameters present in LR.

3.5. RMSProp model

RMSProp model depends on developing weighted average of gradients like Gradient Descent (GD) with momentum with the diverse upgrading parameters. By considering the instance, to optimize a cost function that is comprised with contours where red dot is a position of local optima. The initial GD begins from point 'A' and the iteration of GD has been completed at point 'B', and the alternate side of an ellipse is illustrated in figure 3. Then, the following step of GD is concluded with point 'C'. Among all other iteration of GD, it moves towards the local optima in up and down directions. While there is an application of higher learning value, then a vertical oscillation has high magnitude, and vertical oscillation lowers the GD and removed from the employment of higher learning measure. The bias is responsible for vertical oscillations while 'weight' implies a motion present in a horizontal direction. After upgrading the bias, it decreases the vertical oscillation and if 'weights' are extended with higher values. In case of backward propagation (BP), dW and db parameters were employed to update W and b is predefined one.

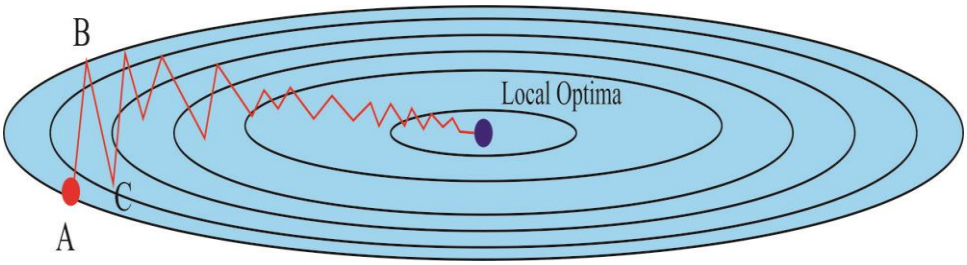


Figure 3. General prgress of the RMSProp model.

$$W = W - learning\ rate * dW \tag{11}$$

$$b = b - learning\ rate * db \tag{12}$$

In RMSprop, rather the application of dW and db, which is not based on every epoch, exponentially weighted averages of a square of dW and db has been utilized.

$$SdW = \beta * SdW + (1 - \beta) * dW^2 \tag{13}$$

$$Sdb = \beta * Sdb + (1 - \beta) * db^2 \tag{14}$$

Where βdenotes alternate hyperparameter that consumes the value of 0 and 1. It assumes the weight from average of preceding values as well as square of current values to calculate the new weighted average. After the estimation of exponentially weighted averages, it results in the variables updating process,

$$W = W - learning\ rate * \frac{dW}{sqrt(SdW)} \tag{15}$$

$$b = b - learning\ rate * \frac{dW}{sqrt(Sdb)} \tag{16}$$

SdW is relatively minimum and Sdb is comparatively maximum.

4. Performance validation

This section discusses the effectiveness of the proposed SA-RMSPO-LR model on the benchmark CKD dataset. A detailed comparative examination with the existing methods takes place to verify the superiority of the presented model. The proposed model is simulated using MATLAB tool. The parameter

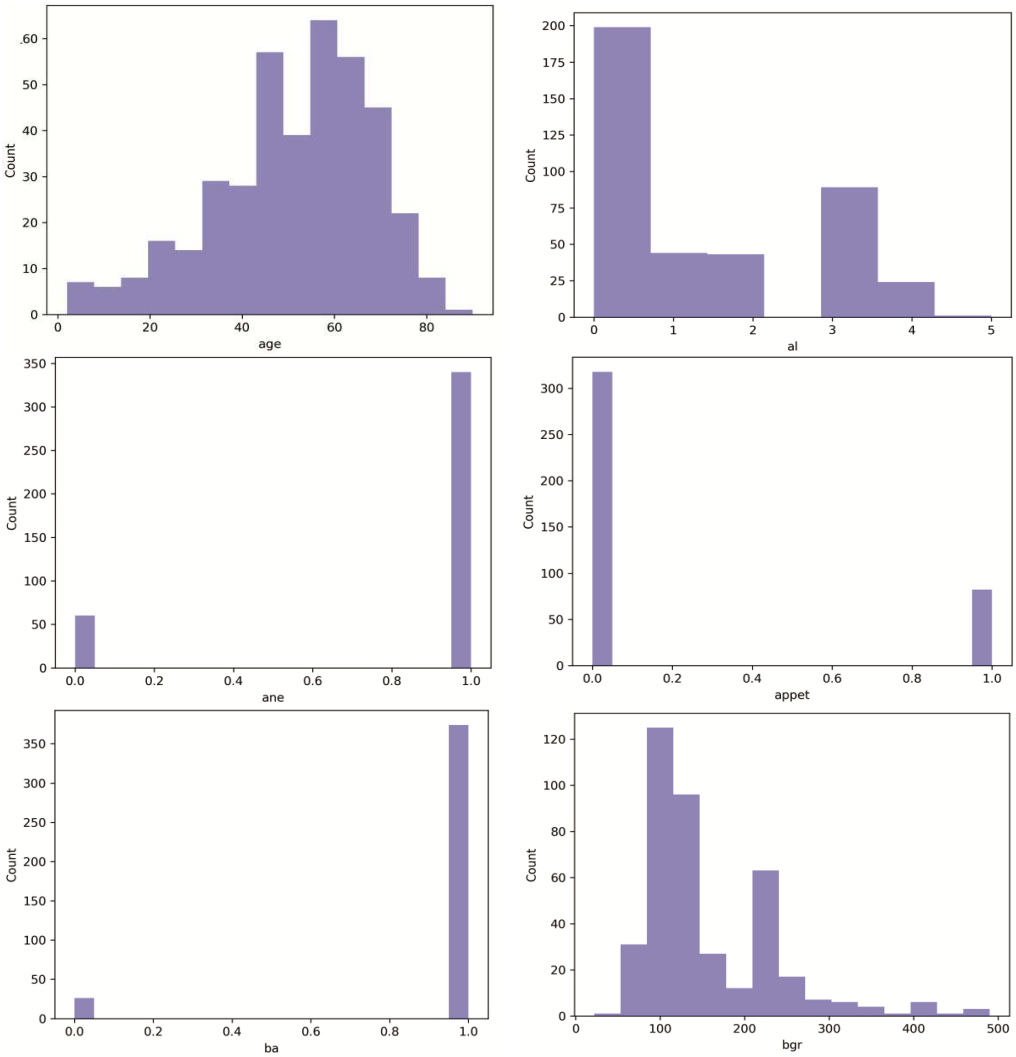
settings of SA are Maximum Number of Iterations: 20, Maximum Number of Sub-iterations: 5, Initial Temperature T_0=10, and Temp. Reduction Rate alpha=0.99.

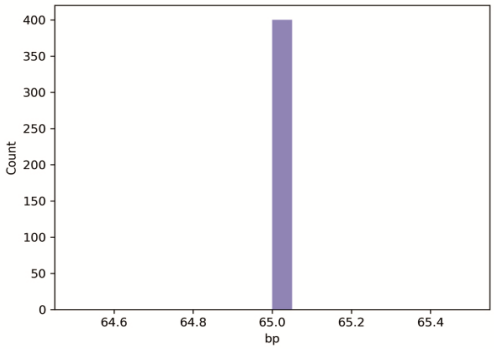
4.1. Dataset Description

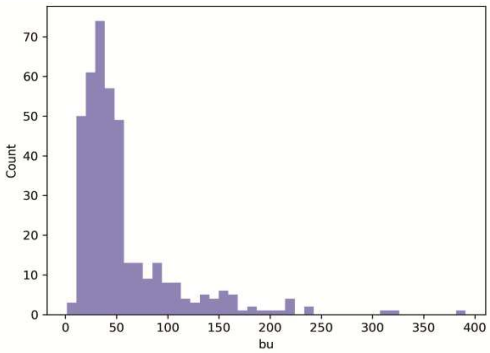
For applying the classification task of SA-RMSPO-LR approach, a standard CKD dataset has been applied [25]. The dataset description as well as accessible features is depicted in Table 1. The CKD dataset is comprised with a sum of 400 instances with 24 features. Among the 400 instances, 250 instances are labeled as CKD present and the remaining 150 instances are labeled with the non-existence of CKD. The sample frequency distribution and class distribution of 24 features are illustrated in Figure 4. On the other side, the features which affects on CKD are given in Figure 5. For experimentation, 10-fold cross validation method is employed to assess the efficiency of the projected technique.

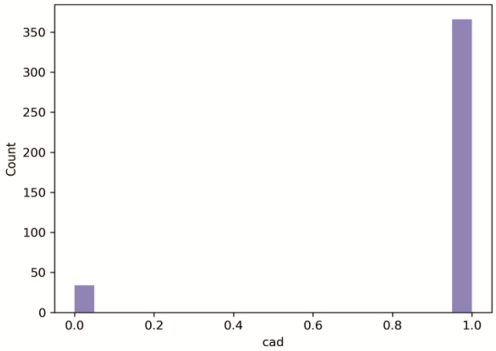
Table 1. Dataset Description.

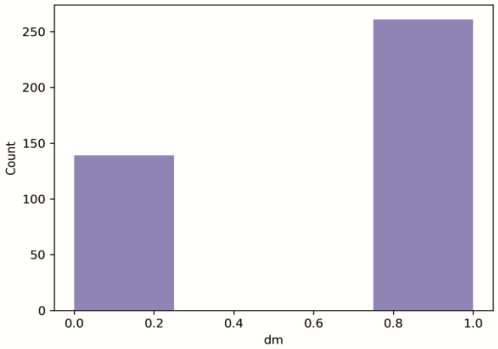
Data Set Description	Values of CKD Real Clinical Data
No. of Instances	400
No. of Features	24
No. of Class	2
Percentage of Positive Samples	62.50%
Percentage of Negative Samples	37.50%
Data source	Own data set from UCI Data Repository

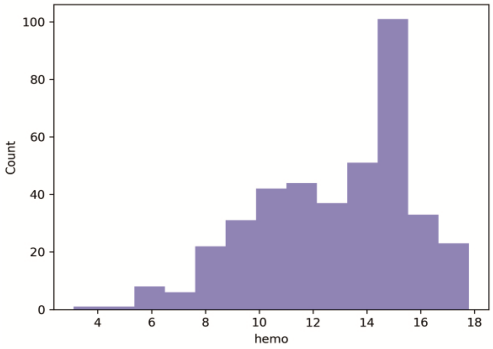


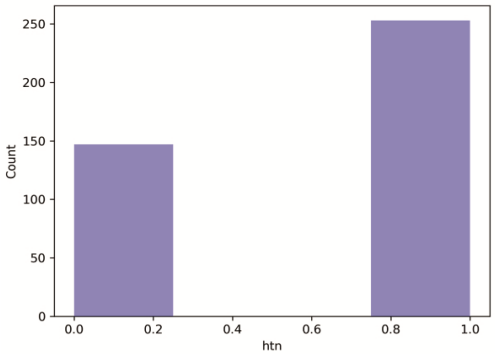

Count
bp

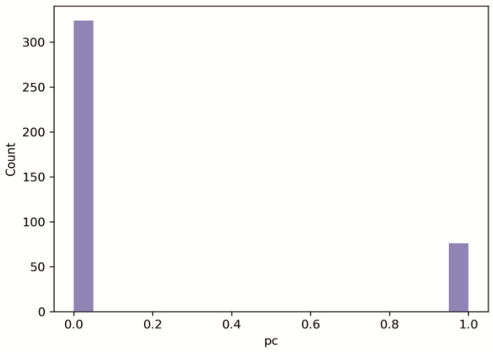

Count
bu

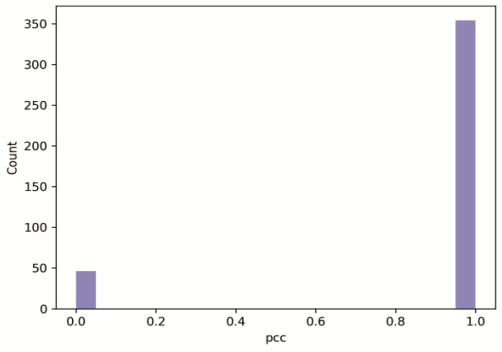

Count
cad

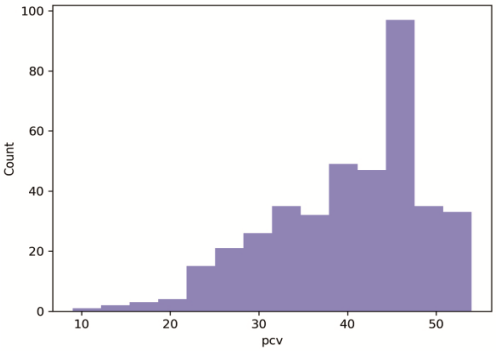

Count
dm

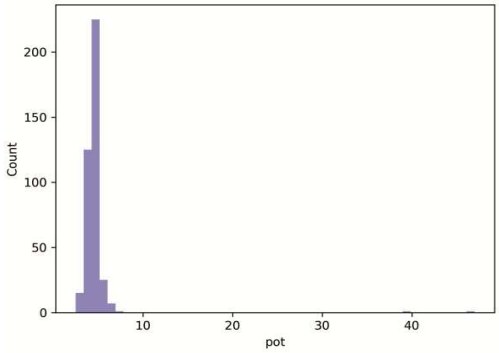

Count
hemo


Count
htn


Count
pc


Count
pcc


Count
pcv


Count
pot

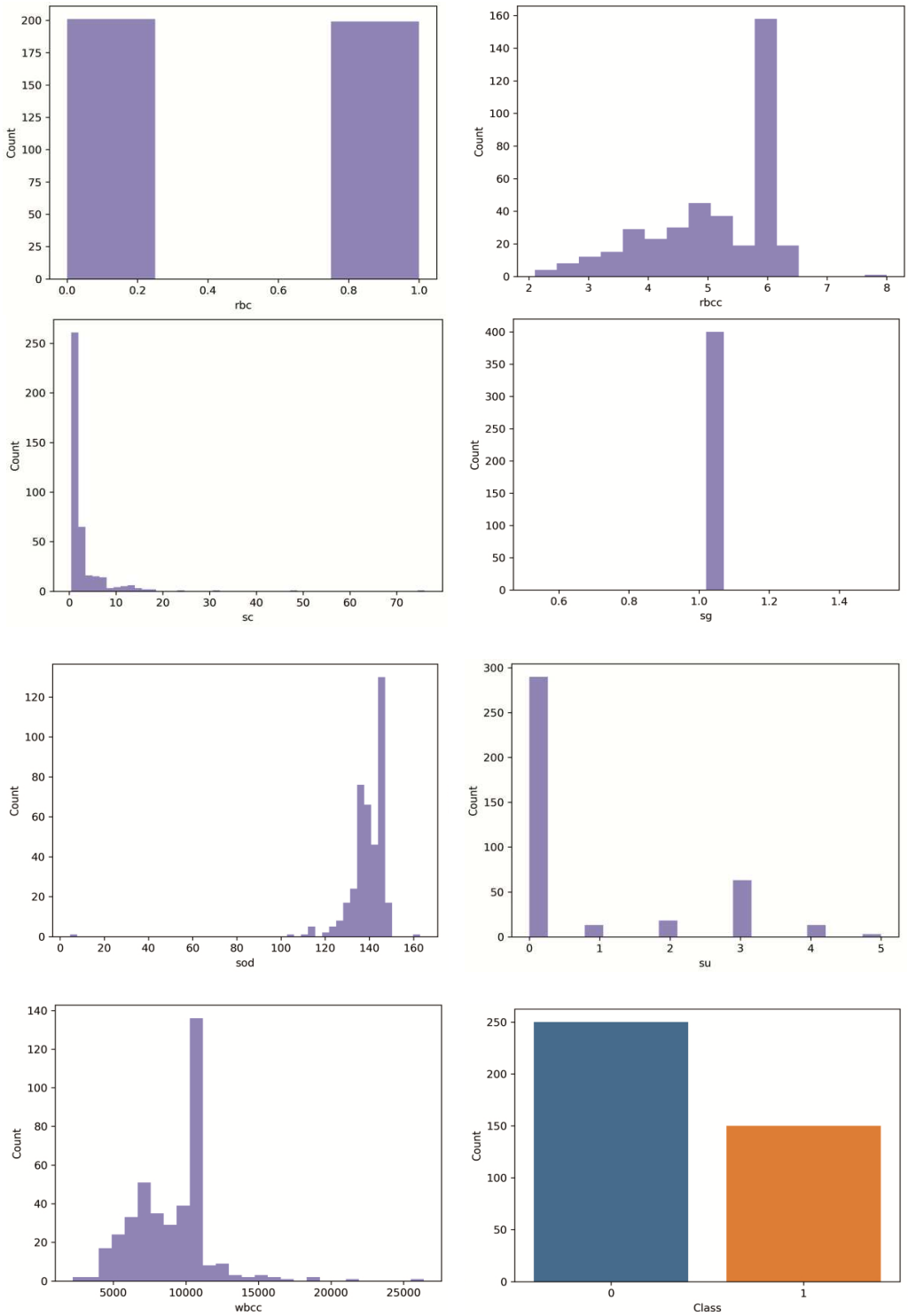


Figure 4. Frequency Distribution of Attributes in CKD.

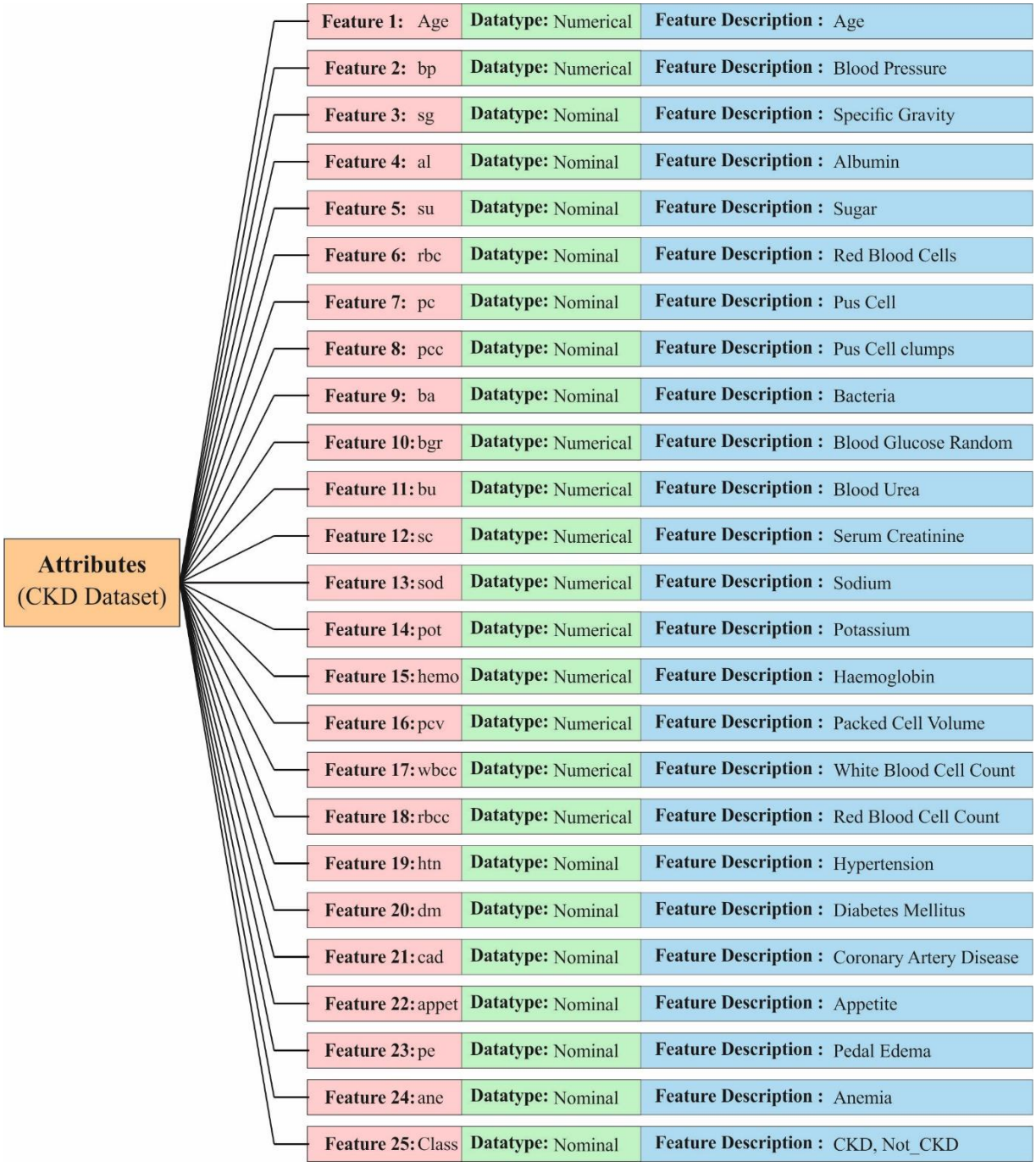


Figure 5. Attribute information of the CKD dataset.

4.2. Result analysis

Table 2 shows the outcome of the FS models on the applied CKD dataset. Figure 6 shows the best cost analysis of the presented SA-FS model. The table values indicated that the CFS model has exhibited an inferior FS results with the best cost of 0.79. Besides, it is demonstrated that the Principal component analysis (PCA) model has offered slightly lower best cost of 0.04570 over CFS, but not than other models. It is also noted that the GA-FS and PSO-FS models have outperformed the earlier models and attained near identical best cost of 0.03440 and 0.03656 respectively. However, the proposed SA-FS model has chosen a set of 13 features with the best cost of 0.01053. This minimum best cost value offered by the SA-FS model clearly ensured the effective performance over its existing models.

Table 2 Result analysis of feature selection methods for CKD Diagnosis.

Methods	Best Cost	Selected Features
SA-FS	0.01053	6, 15, 8, 10, 20, 7, 5, 14, 18, 12, 3, 2, 4
PSO-FS	0.03656	15, 12, 24, 23, 13, 20, 11, 8, 18, 3, 9, 1, 14, 5, 2, 6, 17, 19
GA-FS	0.03440	16, 24, 13, 9, 14, 17, 22, 19, 2, 15, 23, 18, 12, 6, 4, 10, 3, 20
PCA	0.04570	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18
CFS	0.79000	4,6,7,10,15,17,19,22

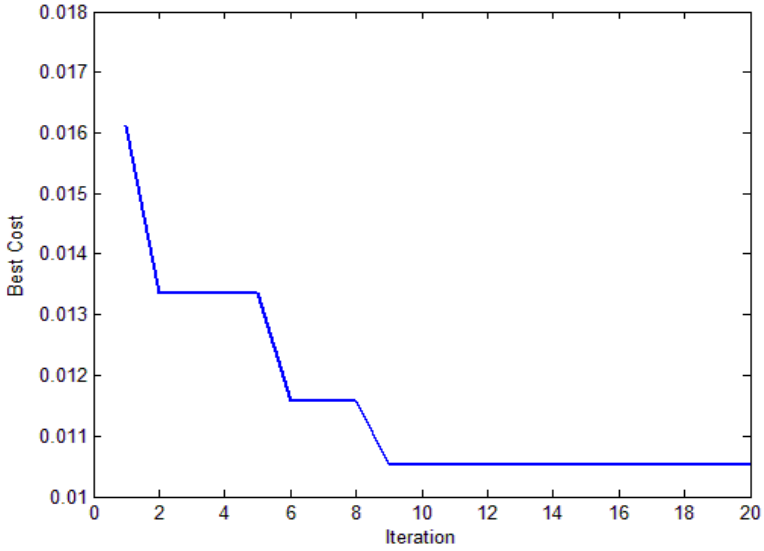


Figure 6. Best cost analysis of SA-FS model.

Figure 7 describes the different confusion matrix generated at the execution of the RMSPO-LR model without FS under different epoch count. Under the epoch count of 400, it is shown that the RMSPO-LR model has classified a set of 182 instances as present and 150 instances as absent. Similarly, under the epoch count of 800, it is depicted that the RMSPO-LR model has classified a set of 117 instances as present and 150 instances as absent. Likewise, under the epoch count of 1200, it is noticed that the RMSPO-LR model has classified a set of 237 instances as present and 147 instances as absent. In the same way, under the epoch count of 1600, it is observed that the RMSPO-LR model has classified a set of 167 instances as present and 150 instances as absent. Concurrently, under the epoch count of 2000, it is demonstrated that the RMSPO-LR model has classified a set of 242 instances as present and 146 instances as absent.

Figure 8 shows the diverse confusion matrix produced at the implementation of SA-RMSPO-LR technique with FS under the diverse epoch count. With the application of different epoch count of 400, it is pointed that the SA-RMSPO-LR approach has classified a collection of 243 instances as present and 145 instances as absent. Likewise, with the epoch count of 800, it is demonstrated that the SA-RMSPO-LR scheme classifies a set of 243 instances as present and 145 instances as absent. Similarly, using the epoch count of 1200, it is evident that the SA-RMSPO-LR framework has classified a set of 244 instances as present and 146 instances as absent. In line with this, under the epoch count of 1600, it is monitored that the SA-RMSPO-LR approach has classified a set of 247 instances as present and 146 instances as absent. At the same time, under the epoch count of 2000, it is depicted that the SA-RMSPO-LR technology has classified a set of 246 instances as present and 146 instances as absent.

After examining the classifier results offered by SA-RMSPO-LR model under varying epoch count, it is observed that the SA-RMSPO-LR model leads to maximum classification outcome under the epoch

count of 1600 rounds prior to 2000 rounds. At the particular round, it is noted that the SA-RMSPO-LR model classifies a set of 247 instances as present and 146 instances as absent as shown in Figure 9.

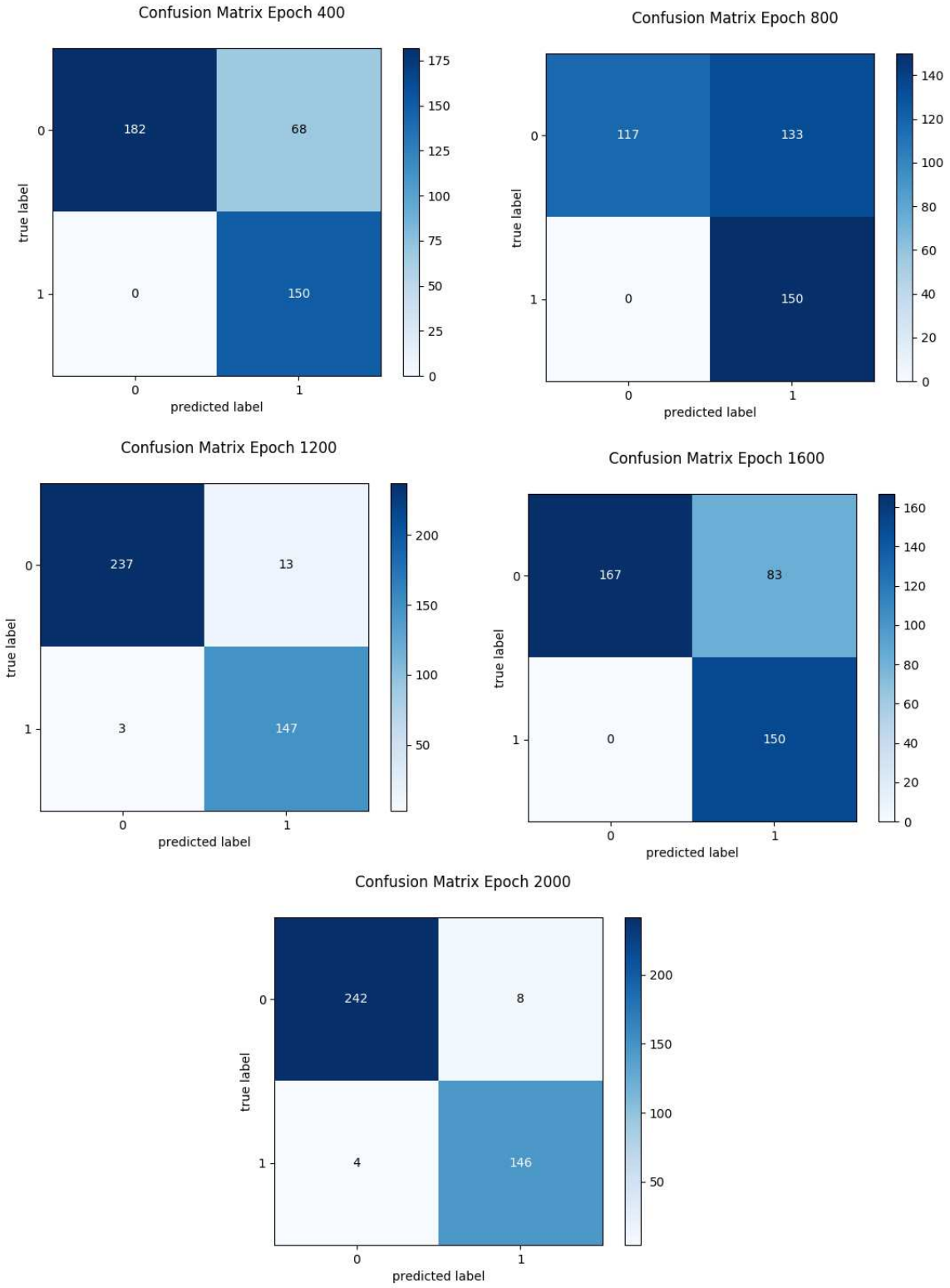


Figure 7. Confusion Matrix from 2000 Epochs of RMSPO-LR without FS.

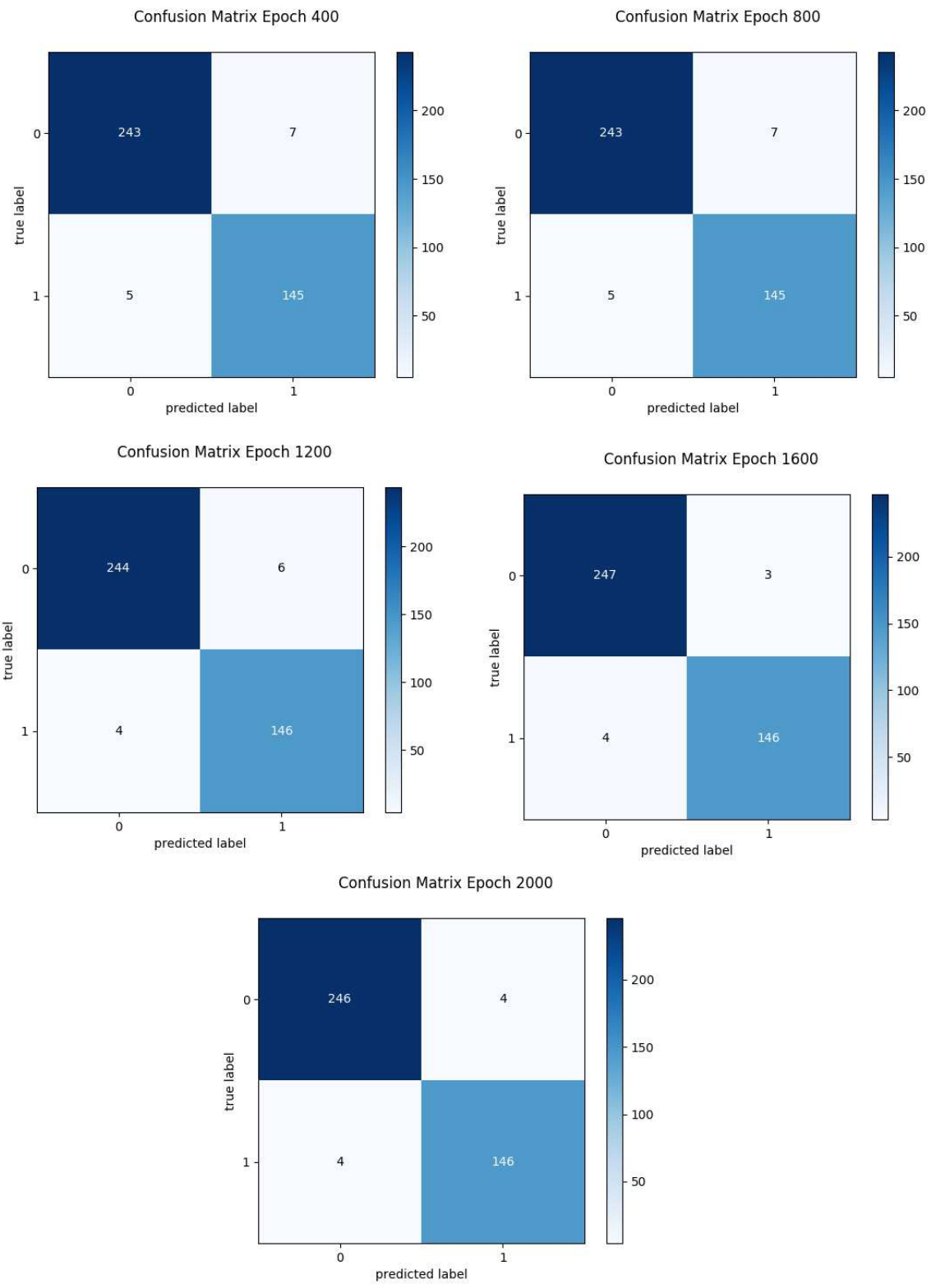


Figure 8. Confusion Matrix from 2000 Epochs of SA-RMSPO-LR.

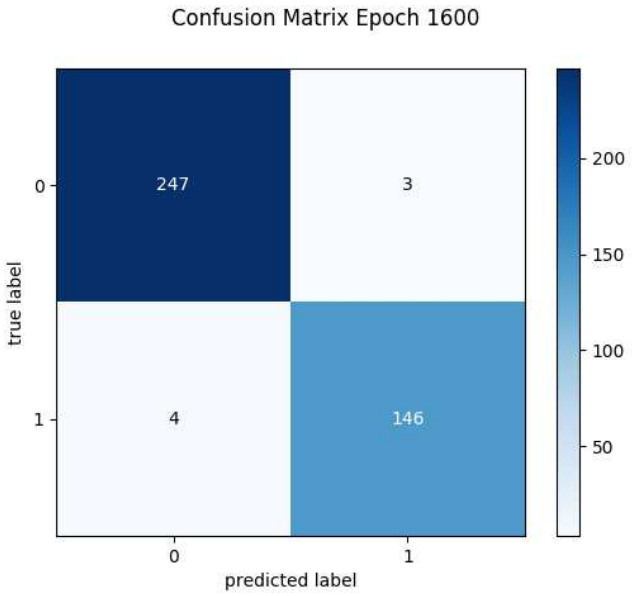
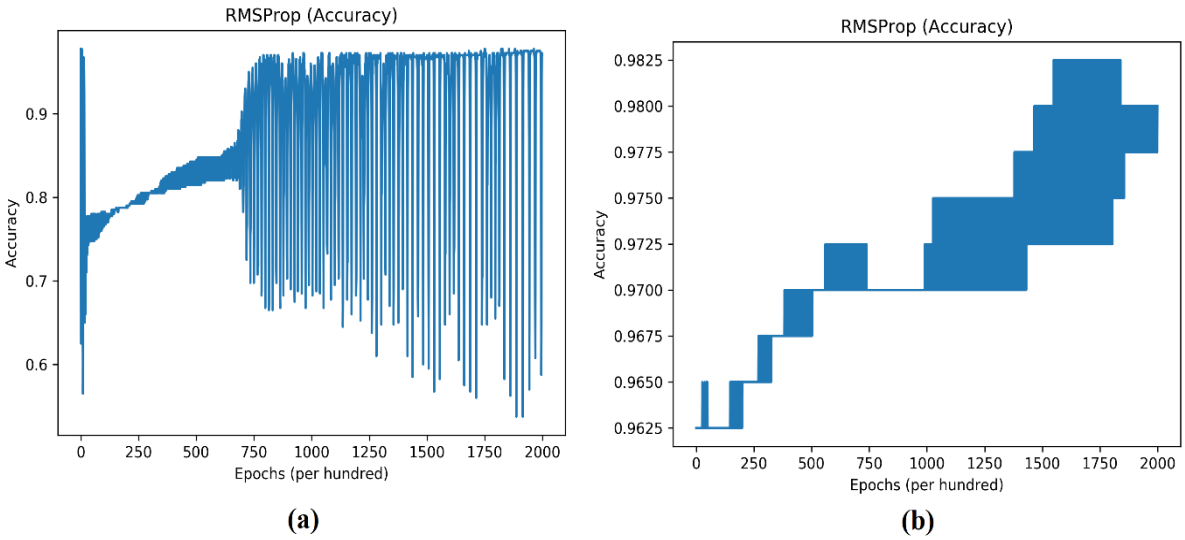


Figure 9. Confusion Matrix of 1600th Iteration.



(a) **(b)**

Figure 10 a) Accuracy Graph RMSPO-LR of 2000 Epochs b) Accuracy Graph SA-RMSPO-LR of 2000 Epochs.

Figure 10 shows the accuracy graph of the RMSPO-LR and SA-RMSPO-LR models under the varying epoch count of 2000. It is depicted that the SA-RMSPO-LR model leads to a maximum accuracy over the RMSPO-LR model. The accuracy of the proposed models begins to rise with an increase in epoch count and remains saturated at 2000 epochs.

Figure 11 depicts the loss graph of the RMSPO-LR and SA-RMSPO-LR models under the varying epoch count of 2000. It is demonstrated that the SA-RMSPO-LR model results in a minimum loss rate over the RMSPO-LR model. The loss rate of the proposed models begins to fall with an increase in epoch count and remains saturated at 2000 epochs. It is noted that the loss graph gets drastically reduced by the inclusion of SA based FS process.

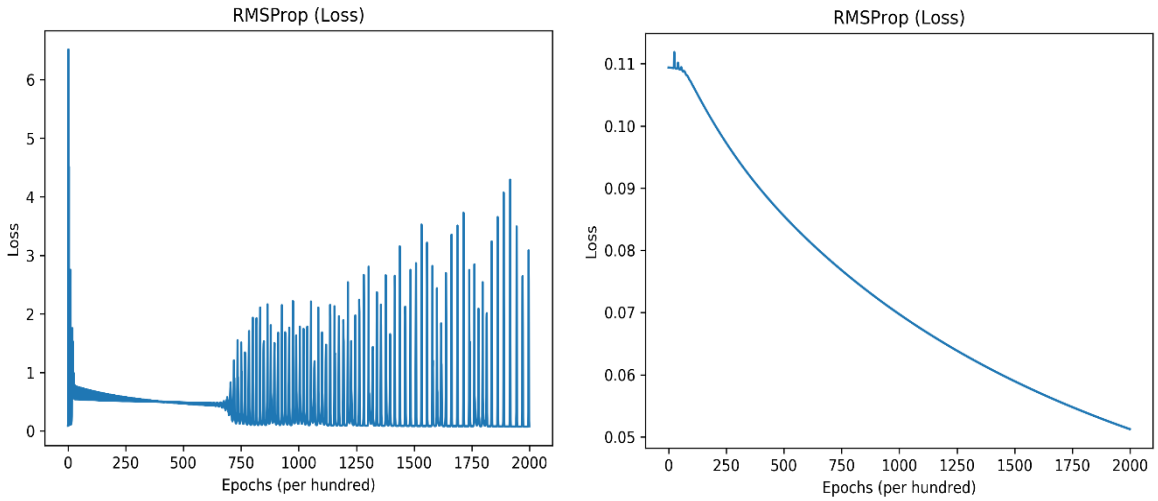


Figure 11. a) Loss Graph RMSPO-LR of 2000 Epochs b) Loss Graph SA-RMSPO-LR of 2000 Epochs.

Table 3. Performance Evaluation of CKD using Proposed Method with various classifiers.

Classifiers	**Performance Measures**				
	Sensitivity	**Specificity**	**Accuracy**	**F-score**	**Kappa**
SA-RMSPO-LR	98.41(%)	97.99(%)	98.25(%)	98.60(%)	96.26(%)
RMSPO-LR	98.37(%)	94.80(%)	97.00(%)	97.58(%)	93.63(%)
FNC	95.68(%)	95.86(%)	95.75(%)	96.63(%)	90.87(%)
D-ACO	96.00(%)	93.33(%)	95.00(%)	96.00(%)	89.33(%)
MLP	92.30(%)	92.86(%)	92.50(%)	94.11(%)	83.78(%)
DT	90.38(%)	89.28(%)	90.00(%)	92.15(%)	78.37(%)
ACO	88.88(%)	84.61(%)	87.50(%)	90.56(%)	72.06(%)
PSO	88.00(%)	80.00(%)	85.00(%)	88.00(%)	68.00(%)
XGBoost	83.00(%)	83.00(%)	83.00(%)	80.00(%)	75.42(%)
LR	83.00(%)	82.00(%)	82.00(%)	79.00(%)	74.60(%)
OlexGA	80.00(%)	66.66(%)	75.00(%)	80.00(%)	46.66(%)

Table 3 shows the results attained by the SA-RMSPO-LR model with existing models with respect to different measures. Figure 12 offered a comparative investigation of the results provided by the SA-RMSPO-LR model in terms of sensitivity and specificity. The figure indicated that the OlexGA is found to the worst performance which has attained a least sensitivity and specificity values of 80% and 66.66% respectively. In addition, it is depicted that the LR reaches to a slightly higher sensitivity and specificity values of 83% and 82% respectively. Besides, it is noted that the XGBoost model has reached to an identical sensitivity and specificity values of 83%. Moreover, it is shown that the PSO algorithm outperformed the earlier models by offering a sensitivity and specificity values of 88% and 80% respectively.

Furthermore, it is observed that the ACO leads to effective results with the sensitivity and specificity values of 88.88% and 84.61%. On continuing with, the DT model results in a slightly manageable performance with the sensitivity and specificity values of 90.38% and 89.28%. In the same way, it is provided that the MLP model reaches to an acceptable classifier results with the sensitivity and specificity values of 92.30% and 92.86%. Simultaneously, the FNC model has shown moderate results with the sensitivity and specificity values of 95.68% and 95.86%. Concurrently, it is depicted that the D-ACO model offered even higher sensitivity and specificity values of 96% and 93.33% respectively. In line with, the RMSPO-LR model has led to a competitive classifier outcome with the sensitivity and specificity values of

98.37% and 94.80%. Generally, the proposed SA-RMSPO-LR models have resulted to a maximum sensitivity and specificity values of 98.41% and 97.99% respectively.

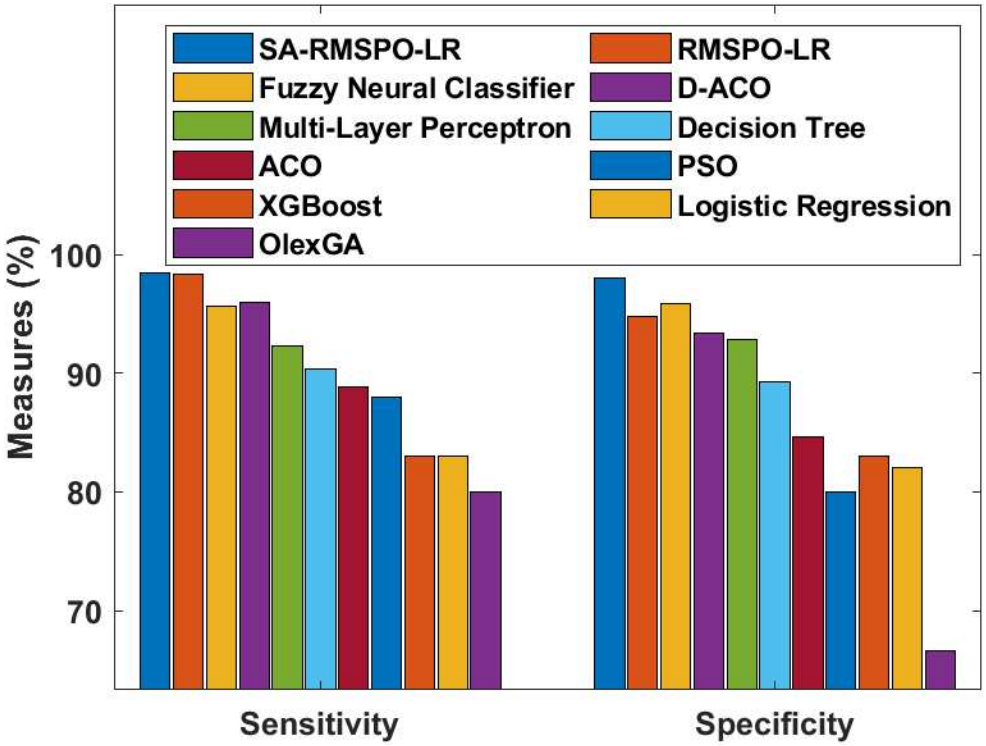


Figure 12. Sensitivity and Specificity analysis.

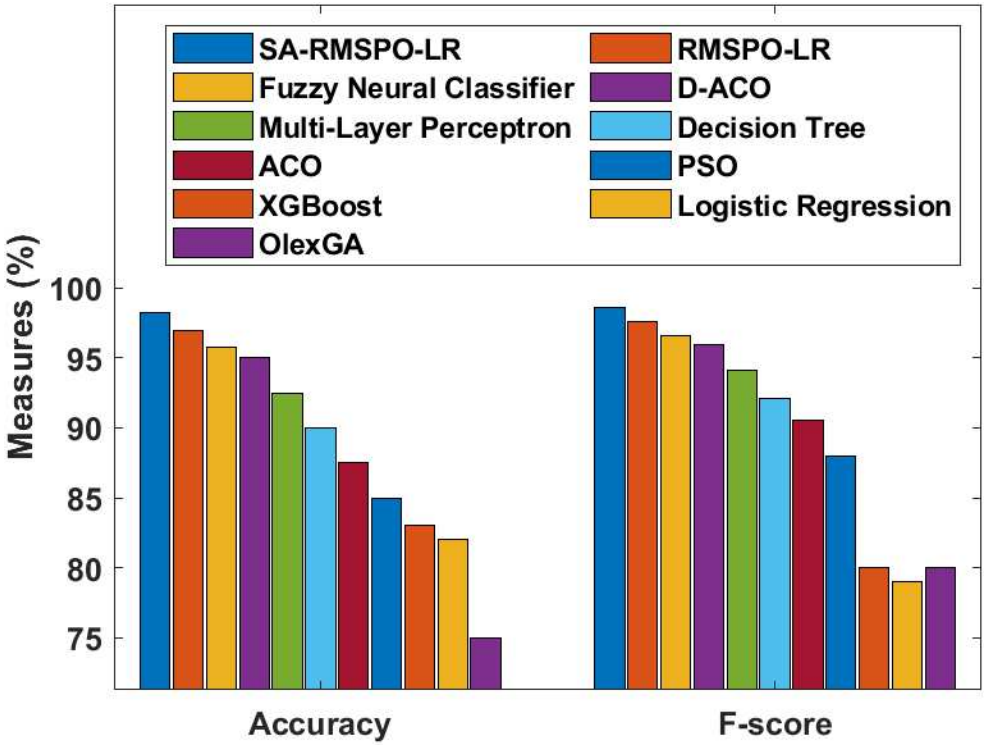


Figure 13. Accuracy and F-score analysis.

Figure 13 provides a comparative analysis of the results attained by SA-RMSPO-LR method with respect to accuracy and F-score. The figure shows that the OlexGA performs a worst function that has reached a lower accuracy and F-score values of 75% and 80% correspondingly. Also, it is illustrated that the LR accomplish with little higher accuracy and F-score values of 82% and 82%. On the other hand, it is pointed that the XGBoost approach has reached to similar accuracy and F-score values of 83% and 80% respectively. Furthermore, it is given that the PSO technique outperformed than the previous models by attaining the accuracy and F-score of 85% and 88% respectively. Moreover, it is noted that the ACO results in a productive outcome with the accuracy and F-score measures of 87.50% and 90.56% respectively. Similarly, the DT scheme tends to provide a slightly reasonable function with the accuracy and F-score values of 90% and 92.15% correspondingly. Likewise, it is given that the MLP approach accomplish to a manageable classifier outcome with the accuracy and F-score values of 92.50% and 94.11% respectively. At the same time, the FNC model has shown gradual results with the accuracy and F-score values of 95.5% and 96.63%. Simultaneously, it is illustrated that the D-ACO approach provides even better accuracy and F-score values of 95% and 96.63%. Likewise, the RMSPO-LR model resulted to a competitive classification outcome with the accuracy and F-score values of 97% and 97.58% respectively. Finally, the projected SA-RMSPO-LR model provides an optimal accuracy and F-score values of 98.25% and 98.60% correspondingly.

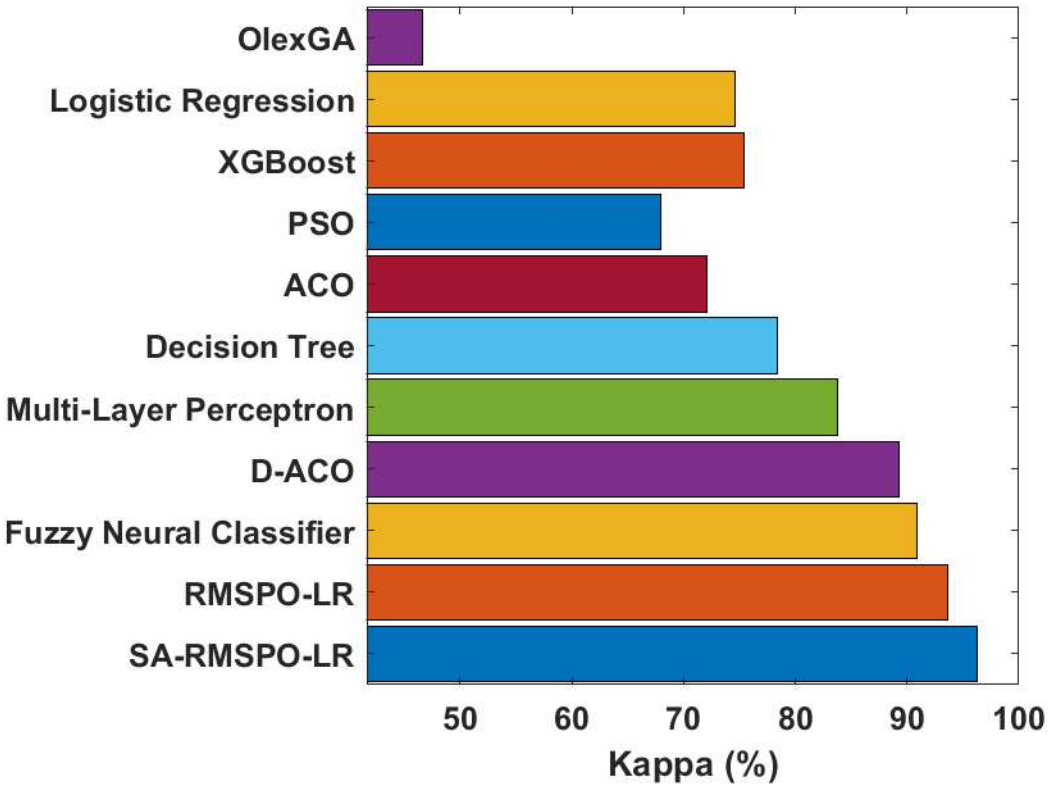


Figure 14. Performance of the Kappa Analysis.

Figure 14 shows a relative examination of the results achieved by the SA-RMSPO-LR technique by means of Kappa. The figure implied that the OlexGA is found to be a poor performance that has reached lower Kappa value of 46.66%. Additionally, it is illustrated that the LR attains to a slightly better Kappa value of 74.60%. On the other hand, it is evident that the XGBoost model has accomplished to a similar Kappa value of 75.42%. Also, it is pointed that the PSO model outperformed the existing frameworks by giving Kappa value of 68%. Moreover, it is monitored that the ACO tends to provide efficient results with the Kappa value of 72.06%. In line with this, the DT model reached slightly appreciable results with the Kappa value of 78.37%. Similarly, it is given that the MLP model attains an acceptable classifier results with the Kappa value of 83.78%. At the same time, the FNC model has shown a gradual outcome with the Kappa

value of 90.87%. Simultaneously, it is evident that the D-ACO approach offered even better Kappa value of 89.33%. The RMSPO-LR model tends to provide competitive classification outcome with the Kappa value of 93.63%. Consequently, the proposed SA-RMSPO-LR system has provided the results with higher Kappa value of 92.26%. By looking into the above mentioned tables and figures, it is ensured that the RMSPO-LR model is found to be an appropriate tool for CKD diagnosis and can be implemented in real time environment.

5. Conclusion

This paper has addressed an optimal IoT and cloud based decision support system for CKD using a SA-RMSPO-LR model. Initially, the data collection process takes place, which collects the patient's data through medical gadgets. Then, preprocessing is carried out to transform the collected data for further processing. Afterwards, SA-FS process gets executed and has chosen a subset of features, which are provided to the RMSPO-LR based classifier. The proposed classifier effectively classified the existence of CKD. Detailed experimental analysis takes place and the results are ensured using the benchmark CKD dataset. The simulation results are examined under varying number of epochs. The experimental results indicated that the proposed SA-RMSPO-LR model leads to effective CKD classification with the maximum sensitivity of 98.41%, specificity of 97.99%, accuracy of 98.25%, F-score of 98.60% and kappa value of 96.26%.The attained results clearly portrayed the enhanced classification performance over the compared methods. As a part of our future work, the performance of the SA-RMSPO-LR CKD diagnosis model can be improved by the inclusion of clustering techniques.

DECLARATIONS

FUNDING

This article has been written with the financial support of RUSA–Phase 2.0 grant sanctioned vide Letter No. F. 24-51/2014-U, Policy (TNMulti-Gen), Dept. of Edn. Govt. of India, Dt. 09.10.2018.

CONFLICTS OF INTEREST

The authors of the present research have no conflict of interest.

AVAILABILITY OF DATA AND MATERIAL

We thank the UCI Machine Learning Repository for publishing attribute information of own and original chronic kidney disease dataset: L.Jerlin Rubini, P.Eswaran & Dr.P.Soundarapandian, M.D., D.M (Senior Consultant Nephrologist).
https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.

ACKNOWLEDGEMENT

We thank Dr.P.Soundarapandian, M.D., D.M (Senior Consultant Nephrologist) and Apollo Hospital, Karaikudi, India for their support and facility at the time of CKD data collection and preliminary study.

AUTHOR CONTRIBUTION

Pramila Arulanthu and Eswaran Perumal conceived the presented idea. Pramila Arulanthu contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. Dr.P.Eswaran investigates and supervised the findings of this work.

References

[1] L. Atzori, A. Iera, G. Morabito, "The internet of things: A survey", Computer Networks, vol. 54, Issue 15, page(s): 2787–2805, 2010. https://doi.org/10.1016/j.comnet.2010.05.010.

[2] B. Xu, L. Xu, H. Cai, L. Jiang, Y. Luo, Y. Gu, "The design of an m-health monitoring system based on a cloud computing platform", Enterprise Information System, vol. 11, No. 1, page(s): 17–36, 2017. https://doi.org/10.1080/17517575.2015.1053416.

[3] Diamantidis, C.J., Becker, S., "Health information technology (IT) to improve the care of patients with chronic kidney disease (CKD)", BMC Nephrology, 15, 7, 2014. https://doi.org/10.1186/1471-2369-15-7.

[4] M.M. Baig, H. Gholamhosseini, "Smart health monitoring systems: An overview of design and modeling, J. Med. Syst., 37 9898, 2013. https://doi.org/10.1007/s10916-012-9898-z.

[5] Thibaud, M., Chi, H., Zhou, W. and Piramuthu, S., "Internet of Things (IoT) in high-risk Environment, Health and Safety (EHS) industries: A comprehensive review", Decision Support Systems, vol. 108, pp.79-95, 2018. https://doi.org/10.1016/j.dss.2018.02.005.

[6] Chiu, R.K., Chen, R.Y., Wang, S.A. and Jian, S.J., "Intelligent systems on the cloud for the early detection of chronic kidney disease", Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, vol. 5, pp. 1737-1742. 2012. DOI:10.1109/ICMLC.2012.6359637.

[7] N.M. Nasrabadi, "Pattern recognition and machine learning", J. Electronic Imaging, 16 (4), 049901, 2007. https://doi.org/10.1117/1.2819119.

[8] J. Schmidhuber, "Deep learning in neural networks: An overview", Neural Networks, vol. 61, pp. 85–117, 2015. Doi:10.1016/j.neunet.2014.09.003.

[9] Miotto, R., Li, L., Kidd, B.A. and Dudley, J.T., "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records", Scientific reports, 6(1), 26094, 2016. https://doi.org/10.1038/srep26094.

[10] Pramila Arulanthu and Eswaran Perumal, "Risk Factor Identification, Classification and Prediction Summary of Chronic Kidney Disease", Recent Advances in Computer Science and Communications (2020) 13: 1. https://doi.org/10.2174/2666255813666200101100424.

[11] Elhoseny, M., Shankar, K. and Uthayakumar, J., "Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease", Scientific reports, 9 (1), 9583, 2019. https://doi.org/10.1038/s41598-019-46074-2.

[12] Gagnebin, Y., Julien, B., Belén, P. and Serge, R., "Metabolomics in chronic kidney disease: Strategies for extended metabolome coverage", Journal of pharmaceutical and biomedical analysis, vol. 161, pp. 313-325. 2018. https://doi.org/10.1016/j.jpba.2018.08.046.

[13] Chatterjee, S., Banerjee, S., Basu, P., Debnath, M., & Sen, S., "Cuckoo search coupled artificial neural network in detection of chronic kidney disease", Proceedings of 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech). 1-4 (2017). DOI: 10.1109/IEMENTECH.2017.8077016.

[14] Chatterjee, S., Dzitac, S., Sen, S., Rohatinovici, N. C., Dey, N., Ashour, A. S., &Balas, V. E., "Hybrid modified cuckoo search-neural network in chronic kidney disease classification", Proceedings of 14th International Conference on Engineering of Modern Electric Systems (EMES). 164-167, 2017. DOI: 10.1109/EMES.2017.7980405

[15] Chen, Z., Zhang, Z., Zhu, R., Xiang, Y., & Harrington, P. B., "Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers", Chemometrics and Intelligent Laboratory Systems, Vol. 153, pp. 140-145, 2016. https://doi.org/10.1016/j.chemolab.2016.03.004

[16] Arasu, S. D., & Thirumalaiselvi, R., "A novel imputation method for effective prediction of coronary Kidney disease", Proceedings of 2nd International Conference on Computing and Communications Technologies (ICCCT), pp. 127-136, 2017. DOI:10.1109/ICCCT2.2017.7972256

[17] Tan, K. C., Teoh, E. J., Yu, Q., & Goh, K. C., "A hybrid evolutionary algorithm for attribute selection in data mining", Expert Systems with Applications. vol. *36*, issue 4, pp. 8616-8630, 2009. https://doi.org/10.1016/j.eswa.2008.10.013

[18] Chetty, N., Vaisla, K. S., & Sudarsan, S. D., "Role of attributes selection in classification of Chronic Kidney Disease patients", Proceedings of International Conference on Computing, Communication and Security (ICCCS). 1-6 (2015). DOI: 10.1109/CCCS.2015.7374193.

[19] P. Arulanthu and E. Perumal, "Predicting the Chronic Kidney Disease using Various Classifiers," 4th IEEE International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), pp. 70-75, 2019. DOI: 10.1109/ICEECCOT46775.2019.9114653.

[20] Wibawa, M. S., Maysanjaya, I. M. D., & Putra, I. M. A. W., "Boosted classifier and features selection for enhancing chronic kidney disease diagnose", Proceedings of 5th International Conference on Cyber and IT Service Management (CITSM), pp. 1-6, 2017. DOI: 10.1109/CITSM.2017.8089245.

[21] Polat, H., Mehr, H. D., & Cetin, A., "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods", Journal of medical systems. 41(4), 55, 2017. DOI: 10.1007/s10916-017-0703-x.

[22] Pramila Arulanthu and Eswaran Perumal, "An intelligent IoT with cloud centric medical decision support system for chronic kidney disease prediction", International Journal of Imaging Systems and Technology, March 2020, 1-13. https://doi.org/10.1002/ima.22424.

[23] Lakshmanaprabu, S.K., Mohanty, S.N., Sheeba Rani, S., Krishnamoorthy, S., J. Uthayakumar, and K.Shankar, "Online clinical decision support system using optimal deep neural networks. Applied Soft Computing, vol. 81, 105487, 2019. https://doi.org/10.1016/j.asoc.2019.105487

[24] Jongbo, O.A., Adetunmbi, A.O., Ogunrinde, R.B. and Badeji-Ajisafe, B., "Development of an Ensemble Approach to Chronic Kidney Disease Diagnosis", Scientific African, vol. 8, p.e00456, 2020. https://doi.org/10.1016/j.sciaf.2020.e00456

[25] L.Jerlin Rubini, P.Eswaran & Dr.P.Soundarapandian, M.D., D.M (Senior Consultant Nephrologist) (2015) UCI Chronic Kidney Disease Dataset, School of Information and Computer Sciences, University of California, Irvine, CA, USA, https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.