Received: June 15, 2022. Revised: September 2, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Z-Score Normalized Feature Selection and Iterative African Buffalo Optimization for Effective Heart Disease Prediction

P. Muthulakshmi[1]* **M. Parveen[1]**

[1]*Cauvery College for Women (Autonomous), Affiliated to Bharathidasan University, Trichy*
* Corresponding author's Email: muthulakshmi.cs@cauverycollege.ac.in

Abstract: At present, health prediction in contemporary life set off very much indispensable. Big Data exploration plays a major contribution to predict subsequent status of health and offers outstanding health consequence to people. A lot of research is persisting on predictive analytics utilizing optimized machine learning techniques to disclose healthier decision making. Big Data analytics strengthens exceptional opening to predict future health condition from health criterions and bestow finest outcomes. We used Big Data Predictive Analytic model for heart disease prediction using Z-score Normalized Iterative African Buffalo Optimization (ZN-IABO). It fills the missing values in the database based on Z-score Normalized Data Pre-processing that standardize the scores on same scale by dividing score deviation by standard deviation. Z-score normalization model is suitable for huge data sets especially for big data. Next, Iterative African Buffalo Optimization based Feature Selection process is applied to the pre-processed data that addresses pre-mature convergence. In the simulation, the proposed ZN-IABO method is also compared with numerous well-known algorithms ZN-IABO (without optimization), imperialist competitive algorithm, and FODW. The result illustrates that this proposed algorithm is very competitive compared with the cardiovascular disease dataset for addressing the theoretical issue and superior to solving the real-world issue. The proposed ZN-IABO method (with optimization), achieves enhancement in the heart disease prediction accuracy by 17%, minimization of heart disease prediction time, error rate, and space complexity by 22%, 39%, and 37% as compared to the ZN-IABO (without optimization), imperialist competitive algorithm, FODW, MLBO respectively.

Keywords: Big data, Machine learning, Z-score, Normalized iterative, African buffalo optimization.

1. Introduction

An imperialist competitive algorithm with meta-heuristic approach was introduced in [1] to choose the prominent heart disease features. The designed algorithm presented optimal response for feature selection toward genetic than optimization algorithms. Also, K-nearest neighbor algorithm was employed for classification. However, the designed algorithm failed to perform feature selection method for incomplete and missed data and also the meta-heuristic function involving data classification in medical application was found to be highly sensitive.

Predictive analytics based on Feature Optimization by Discrete Weights (FODW) was carried out in [2] with minimal false alarming training data corpus and optimal feature selection. Here, a new feature selection approach was introduced to perform supervised learning. Also, minimal false alarming concerning prediction of heart disease was made in an efficient manner. Finally, a dynamic n-gram Features Optimization was carried out with the aid of discrete weight of feature correlation. However, dimensionality issues were not handled in a proper manner.

A new 0-1D coupled, personalized hemodynamic model of cardiovascular system (CVS) was designed in [3] to predict pressure waveforms and flow velocities in arteries. The multi-scale CVS model was integrated with Levenberg–Marquardt optimization algorithm for addressing inverse problem depending on measured blood pressure waveforms. Hemodynamic characteristics with brachial arterial pressure waveforms, artery diameters, stroke Vol. s, and flow velocities were computed. However, the error rate was not minimized by designed model.

An intelligent computational predictive system was introduced in [4] for cardiac disease identification and diagnosis. Four different feature selection algorithms were introduced to remove irrelevant and noisy data from extracted feature space. The feature selection algorithm results with classifiers were examined. The accuracy, sensitivity, specificity, AUC, F1-score, MCC, and ROC curve were employed to examine the efficiency and strength of developed model. The prediction accuracy was not improved by designed system.

1.1 Problem definition

Optimization algorithms are generally employed methods in handling optimization issues. Optimization is used for discovering the best possible solution to a problem by considering its constraints. In different fields of science, after facing an optimization issue, first, the variables are identified, and their diverse constraints are considered. For selecting the appropriate features, feature selection plays a crucial role. The existing feature selection method was not performed to choose vital features with less time in medical application. In addition, the conventional pre-processing technique was unable to handle the missing data or values, noisy and inconsistent data with a minimum size of dimensionality. In order to overcome the issue, the objective of this work is to design a heart disease prediction method using feature selection. Feature selection is the pre-processing technique that removes the irrelevant feature not necessary for prediction and also concentrates in the reduction of overall size. Prediction of heart disease can be made by using robust feature selection. In our work, Z-score Normalized Iterative African Buffalo Optimization (ZN-IABO) based feature selection method is proposed.

The contributions of the paper as follows:

- A Log-transformed Z-score Normalized Data pre-processing model for filling the missing values from the cardio vascular dataset is introduced to ensure heart disease prediction accuracy and time significantly.
- The Iterative African Buffalo Optimized Feature Selection is used to select the most pertinent features for predicting heart disease with minimum error and complexity.
- Finally, the method is created to measure the performance of the proposed work in terms of numerous metrics.

The remainder of this manuscript is organized as follows. Related research concerning heart disease prediction is introduced in Section 2. In Section 3, the new Z-score Normalized Iterative African Buffalo Optimization (ZN-IABO) method is described in detail. In Section 4, cardiovascular disease dataset is used in the proposed ZN-IABO method and subjected to experimental analysis. The proposed ZN-IABO method is compared with state-of-the-art methods in order further verify its feasibility and effectiveness and discussed in Section 5. The final section concludes this study with a brief summary in Section 6.

2. Related works

Heart diseases are symbolized as one type of diseases that comprises of heart or vessels also ten percent of overall death in the early twentieth century has contributed due to the prevalence heart diseases. It not only has a greater influence on the human health but also affects the overall economy of the country. In the recent years, large number of data mining mechanisms and machine learning algorithms are designed with the objective of predicting heart disease at the early stage itself.

Z-score was evaluated in [5] for the diagnosis and management of coronary artery dimensions. A hybrid method to predict coronary artery disease using correlation based feature subset in [6] to improve the accuracy involved in prediction. Also as far as heart disease diagnosis is concerned, optimization techniques play a major aspect. An optimization function based on the support vector machine (SVM) in addition to the genetic algorithm (GA) for selecting significant features to get heart disease was proposed in [7]. Despite improvement in accuracy, adaptability issue was not focused. In [8], a feature selection method based on partial least square was proposed with the coordinate descent to select better feature subset, therefore contributing to preferable adaptability.

Among the most serious diseases affecting human is the cardiovascular diseases (CVD) and also these diseases can be mitigated by early diagnosis, therefore a significant mortality rate can be reduced. Risk factor analysis using machine learning models is a promising approach. Separate data collection, data pre-processing and data transformation was applied in [9] to achieve accuracy with minimum error.

A hybrid Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN) was presented in [10] with the objective of ensuring harmonious balance between the training data distribution and XGBoost for significant heart disease prediction. However selecting the features becomes Feature selection becomes noteworthy specifically in the data sets with numerous features involved. It will discard irrelevant features and enhance the accuracy. Critical features were selected on the basis of machine learning models in [11].

Due to ubiquitous nature of big data, therefore results in the increase in the data dimensionality. With the increase in the feature selection process, one of the prominent tools for feature selection involves the evaluation and feature acquisition process. This is found to be increasingly correct as far as healthcare domain for analyzing heart disease is concerned, where data is accumulated and under-utilized.

Convolution Neural Network was applied in [12] for determining heart disease based on the structured data, therefore improving accuracy significantly. Yet another novel fast conditional mutual information feature selection model was designed in [13] to identify heart disease in an intelligent manner. However, healthcare heart disease big data remains under-utilized due to numerous conflict proceedings between domains of data analytics and healthcare. Feature selection based on Bayesian learning was applied in [14], therefore minimizing false positive and false negative rates considerably.

One of the paramount origins of mortality globally is due to the Coronary Artery Disease (CAD) and is connected with big data or in other words the nature or dimensionality of data involved. Researchers are inspired to appertained Machine Learning (ML) for swift and precise CAD detection. However, the analysis of automated systems purely depends on the feature quality being utilized. A novel hybrid feature selection algorithm was utilized in [15] to concentrate on the accuracy matters.

In [16], a novel method that aims at identifying relevant features by means of machine learning techniques contributing to enhancing the cardiovascular disease prediction accuracy was proposed. Best first search feature selection was applied in [17] for enhancing the disease performance analysis. Cardio disease prediction using rule-based algorithm to select the pertinent features was applied in [18] to concentrate on the prediction accuracy. Artificial Intelligence based smart prediction was proposed in [19] to enhance prediction accuracy.

Stochastic komodo algorithm (SKA) was introduced in [20] to change the Komodo mlipir algorithm (KMA) with minimum redundancy. But, this algorithm was not tested to handle the real-world issue. Fixed-step average and subtraction-based optimizer (FS-ASBO) was developed in [21] for enhancing the ASBO. FS-ASBO of the objective function was more difficult. Mixed Leader Based Optimizer (MLBO) was discussed in [22] for solving the optimization issue. However, the multi-objective of MLBO was not focused. Multi Leader optimizer (MLO) was introduced in [23] for obtaining the quasi-optimal solution.

Three Influential Members Based Optimizer (TIMBO) was examined in [24] for offering appropriate quasi-optimal solutions. Random Selected Leader Based Optimizer (RSLBO) was developed in [25] to optimize the objective function. But, the optimization performance was not enhanced. Squirrel Search Optimizer (SSO) was investigated in [26] with a higher search ability of the algorithm. Puzzle Optimization Algorithm (POA) was introduced in [27] for examining the performance of POA. Ring Toss Game-Based Optimization (RTGBO) algorithm was designed in [28] for discovering the global optimal solution. The algorithm failed to design of multi-objective version of RTGBO.

In related work, several articles are discussed and the issue is explained. The error rate, complexity, time, and accuracy have not been fully explored for heart disease prediction. Because the machine learning techniques use the entire feature selection therefore it is infeasible and inaccurate. Hence, the dimensionality of the dataset needs to minimize for accurate heart disease prediction with lesser complexity. Therefore, in this work, we address the error rate, heart disease prediction accuracy, heart disease prediction time, and space complexity as compared to the existing feature selection model with the aid of Z-score Normalized Iterative African Buffalo Optimization (ZN-IABO).

3. Methodology

Big Data involves a collection of huge volume of data that are said to escalate with the evolution of time. As per the studies analyzed, Big Data analytics is found to be of highly useful in heart disease prediction and also the technologies utilized in Big Data are immensely pivotal to the management for cardiovascular disease. Several factors are found to be influenced that increase the risk of heart disease and to name a few are, smoking habit, body cholesterol level, obesity, high blood pressure, and lack of physical exercise and so on. Data pre-processing and feature selection are performed in our work for efficient heart disease prediction. The data pre-processing is performed in our work to refill the missing values in the input database.

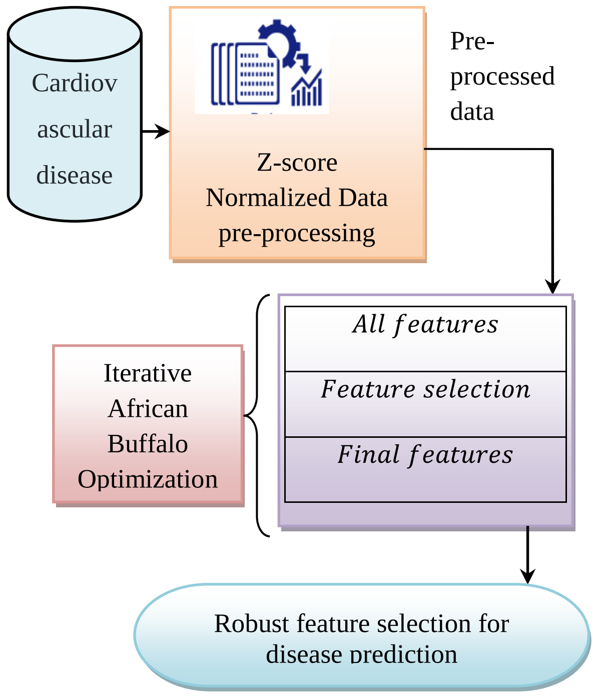

Cardiovascular disease
Pre-processed data
Z-score Normalized Data pre-processing
Iterative African Buffalo Optimization
All features
Feature selection
Final features
Robust feature selection for disease prediction

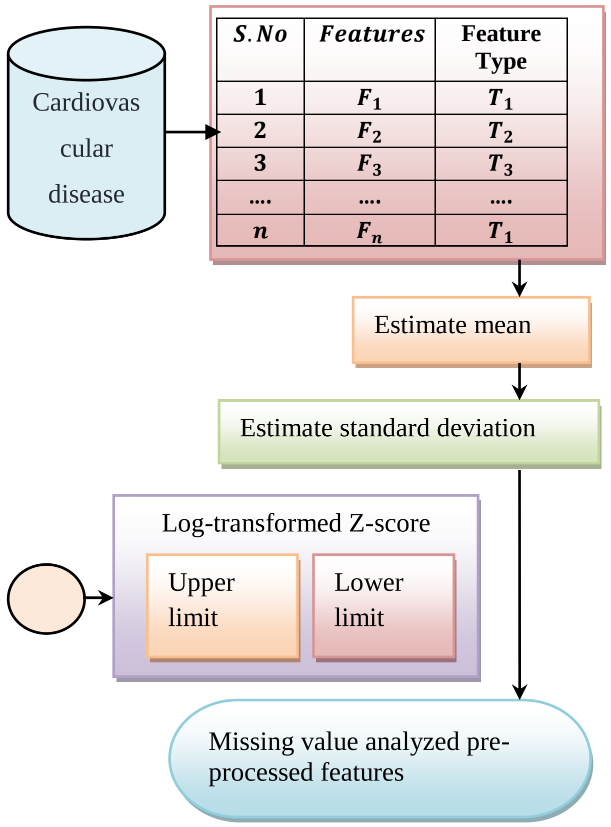
Figure. 1 Block diagram of ZN-IABO method

Followed by which the process of feature selection is carried out to select relevant features from pre-processed data. In this work, a Z-score Normalized Iterative African Buffalo Optimization (ZN-IABO) method is proposed to concentrate more on heart disease prediction time and accuracy. Fig. 1 shows the block diagram of Z-score Normalized Iterative African Buffalo Optimization (ZN-IABO) method.

3.1 Log-transformed Z-score normalized data pre-processing

Data Pre-processing refers to the process of converting the raw data (i.e., raw records of patient data) into a clean data set. In other words, whenever the data is gathered from different types of input features (i.e., objective, examination and subjective from cardiovascular disease dataset), it is acquired in raw format which is not practicable for the analysis. Hence, definite steps are carried out to convert the data into a precise dataset. This is referred to as Data Pre-processing. Most real world dataset consists of missing data or values, noisy and inconsistent data, therefore compromising further processing.

To achieve better results in this work, missing value analysis is made by means of Z-score Normalized Data Pre-processing model. Z-Score Normalized Data Pre-processing in our work is utilized to fill the missing values in the database. Also it is employed for standardizing scores on same scale via dividing score deviation by standard deviation in a database. With this the number of standard deviations corresponding to the given data point from mean is identified. Fig. 2 shows the block diagram of Z-score Normalized Data pre-processing model.



$S.No$	$Features$	Feature Type
1	F_1	T_1
2	F_2	T_2
3	F_3	T_3
....	....	....
n	F_n	T_1

Figure. 2 Block diagram of Z-score normalized data pre-processing model

Consisting of 'n' features '$F = F_1, F_2, \dots, F_n$', where '$n = 13$', with each feature belonging to any one of the three different feature types '$T = T_1, T_2, T_3$ ', namely, objective, examination and subjective respectively. The patient data is then organized as a patient vector matrix 'PVM' of 'm' rows and 'n' columns where 'n' refers to the number of features and 'm' denotes the number of patients.

$$PVM = \begin{bmatrix} P_{11} & P_{12} & P_{13} & \dots & P_{1n} \\ P_{21} & P_{22} & P_{23} & \dots & P_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ P_{m1} & P_{m2} & P_{m3} & \dots & P_{mn} \end{bmatrix} \quad (1)$$

From the above Eq. (1), for example the first row '$\mathrm{P}_{11}, \mathrm{P}_{12}, \mathrm{P}_{13}, \dots, \mathrm{P}_{1n}$' denotes the particular data values of the first patient where 'P_{1i}' represents the value of feature 'i' of the patient number '1'.

$$\text{Mean of features } (\mu) = \begin{bmatrix} P_1F_1 + P_1F_2 + P_1F_3 + \cdots + P_1F_n \\ P_2F_1 + P_2F_2 + P_2F_3 + \cdots + P_2F_n \\ P_3F_1 + P_3F_2 + P_3F_3 + \cdots + P_3F_n \\ \cdots \\ P_mF_1 + P_mF_2 + P_mF_3 + \cdots + P_mF_n \end{bmatrix} \quad (2)$$

From the above Eq. (2), the mean of features ‘μ’ are first obtained for each patient (i.e., in the first row) and in a similar manner, the mean of features is estimated for all the patients with the respective dataset values collected at the moment of medical examination.

$$\text{Standard deviation of features } (\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(F_i - \mu_i)}{n}} \quad (3)$$

From the above Eq. (3), the standard deviation of overall patient vector matrix is estimated by each individual value of features ‘F_i’ and the mean of features ‘μ_i’ to the total number of features ‘n’ respectively. Then, the log-transformed Z-score function for each feature is estimated as given below.

$$Z - score\ (LT) = \frac{\ln(F_{ij}) - \ln(\mu_i)}{\sigma[\ln(F_i)]} \quad (4)$$

From the above Eq. (4), the Z-score log-transformed value ‘$Z - score\ (LT)$’ is obtained based on the logarithmic feature ‘$\ln(F_{ij})$’ and mean value ‘$\ln(\mu_i)$’ to the logarithmic feature standard deviation ‘$\sigma[\ln(F_i)]$’ respectively. Finally, the normalized log-transformed Z-score function is evaluated separately for the lower and upper limit as given below.

$$(Z - score)_L = \frac{(L_{Th} - \mu)}{\sigma} \quad (5)$$

$$(Z - score)_U = \frac{(U_{Th} - \mu)}{\sigma} \quad (6)$$

From the above Eq. (5), the Z-score lower limit ‘$(Z - score)_L$’ is estimated based on the difference ratio of lower threshold ‘L_{Th}’, mean of features ‘μ’ to the standard deviation feature value ‘σ’. This reveals that for a feature to have a length greater that the lower limit, it must be at least Z-score lower limit standard deviations above the mean.

On contrary, from the above Eq. (6), the Z-score upper limit ‘$(Z - score)_U$’ is estimated based on the difference ratio of upper threshold ‘U_{Th}’, mean of features ‘μ’ to the standard deviation feature value ‘σ’. This reveals that for a feature to have a length less than that the upper limit, it must be at least Z-score upper limit standard deviations above the mean. The pseudo code representation of Log-transformed Z-score Normalized Data pre-processing is given below.

Algorithm 1. Log-transformed Z-score normalized data pre-processing

```
Input: Dataset ‘$DS$’, Feature ‘$F = F_1, F_2, \dots, F_n$’, Feature type ‘$T = 1,2,3$’, Feature type 1 ‘$T_1 = O$’, Feature type 2 ‘$T_3 = E$’, Feature type 3 ‘$T_3 = S$’
Output: Precise and computationally efficient pre-processed features
Initialize ‘$n$’
Initialize lower threshold ‘$L_{Th}$’, upper threshold ‘$U_{Th}$’
Begin
    For each Dataset ‘$DS$’ with Feature ‘$F$’ and Feature type ‘$T$’
        Formulate patient vector matrix as in equation (1)
        Estimate mean of features as in equation (2)
        Estimate standard deviation of overall patient vector matrix as in equation (3)
        Evaluate Z-score log-transformed value as in equation (4)
        For each lower threshold
            Estimate normalized log-transformed Z-score function for lower limit as in equation (5)
            Return pre-processed features ‘$(Z - score)_L$’
        For each upper threshold
            Estimate normalized log-transformed Z-score function for upper limit as in equation (6)
            Return pre-processed features ‘$(Z - score)_U$’
        End for
        End for
        Return ‘$PF \to (Z - score)_L \cup (Z - score)_U$’
    End for
End
```

As given in the above algorithm, the objective remains in pre-processing the raw cardiovascular disease dataset with improved heart disease prediction time and accuracy. Here, pre-processing is carried out by filling the missing values by means of Z-score log-transformation. To start with, the mean and standard deviation of each feature for the corresponding patient data are evaluated. Followed by which with the initialized lower threshold and upper threshold, normalized log-transformed Z-score function for lower limit and upper limit is evaluated. Finally, with this, different features of the respective patients are converted to same magnitude and standardization is also performed. The objective of utilizing this algorithm is that the results are not affected by the magnitude of the data due to the reason is that its role remains only in discarding the amount, therefore contributing to heart disease prediction accuracy and time.

3.2 Iterative african buffalo optimization based feature selection

In the recent few years, the extensive magnitude of data is accessible for all aspects of domain that necessitates to be deliberately and efficiently mined. Feature selection plays an extensive part in selecting relevant information from pinnacles of data utilizing a negligible subset of features. Feature selection is one of the demanding optimization issues that aid in selecting optimal features from pre-processed cardiovascular dataset so that preferable predictive rate can be attained. In this work, an Iterative African Buffalo Optimization based Feature Selection model is proposed that eliminates the insignificant and unnecessary features for better heart disease prediction.

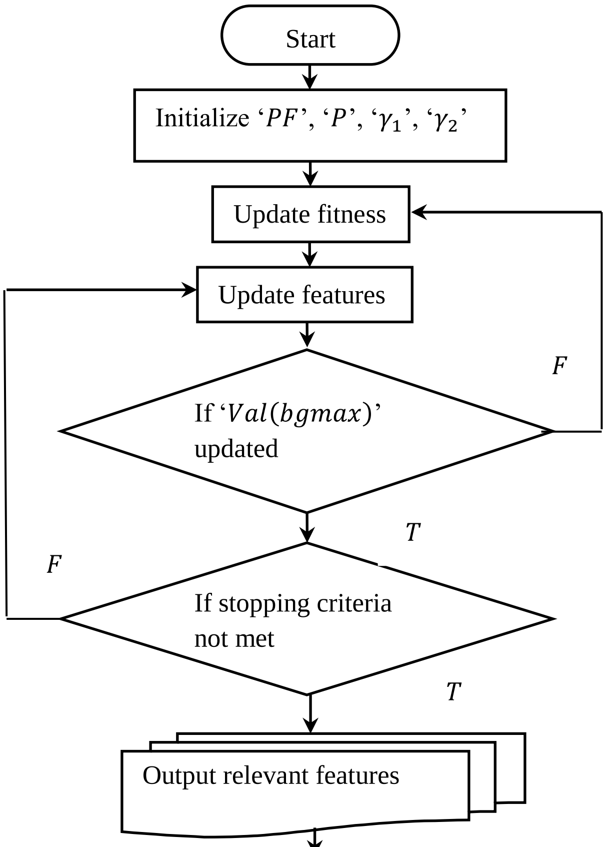

Start
Initialize 'PF', 'P', 'γ_1', 'γ_2'
Update fitness
Update features
If '$Val(bgmax)$' updated
F
T
If stopping criteria not met
F
T
Output relevant features
Stop

Figure. 3 IABO feature selection flow chart

African Buffalo Optimization (ABO) is a user-friendly and efficient algorithm to reveal the exceptional capacity in exploitation and exploration of search space. ABO addresses the pre-mature convergence or stagnation issues through each buffalo (i.e., feature) location is regularly updated in relation to particular buffalo's best previous location and present location of best buffalo in the herd. By this way, relevant features are selected for heart disease prediction with minimal error rate.

As shown in the above Fig. 3, to start with buffalos (i.e., patients considered for simulation) are selected in a random manner to nodes at the solution space (i.e., records of patients' data).

$$m.i + 1 = m.i + \gamma_1(bgmax - w.i) + \gamma_2(bpmax.i - w.i) \quad (7)$$

From the above Eq. (7), the buffalos (i.e., patient's) fitness value is updated based on the exploration '$w.i$', exploitation '$m.i$' iterations respectively, learning parameters 'γ_1', 'γ_2' and the overall buffalos (patients') best fitness '$bgmax$', the individual buffalos (patients') best found relevant features '$bpmax.i$'. In the conventional African Buffalo Optimization (ABO), the discrete updating process is carried out in consonance with the best value of each patient created hitherto '$bpmax$' and the best value acquired by all patients' '$bgmax$'.

Under this discrete learning process, each patient learns from its own finest experience and the finest experience of all patients'. Nevertheless, patient may favour each other after a number of iterations in the recently evolution process, that in turn results in the loss of diversity in the population, hence is susceptible to premature convergence. To address this issue, the conventional ABO is revised by including a novel factor to (7) that denotes the iterative arbitrary learning factor (IALF). The revised discrete learning factor is represented as given below.

$$m.i + 1 = m.i + \gamma_1(bgmax - w.i) + \gamma_2(bpmax.i - w.i) + \gamma_3(F_i - w.i) \quad (8)$$

From the above Eq. (8), by introducing iterative arbitrary learning factor '$\gamma_3(F_i - w.i)$', with 'γ_3' symbolizing the learning parameter arbitrarily selected in the current overall records of patients' data, the revised discrete learning factor is arrived at, therefore addressing premature convergence.

Algorithm 2. Iterative african buffalo optimized feature selection

Input: Dataset ‘DS’, Feature ‘$F = F_1, F_2, \ldots, F_n$’, Feature type ‘$T = 1,2,3$’, Feature type 1 ‘$T_1 = O$’, Feature type 2 ‘$T_3 = E$’, Feature type 3 ‘$T_3 = S$’

Output: Robust feature selection-based disease prediction

```
1: Initialize pre-processed features ‘$PF$’, patients ‘$P = P_1, P_2, \ldots, P_n$’, learning parameters ‘$\gamma_1$’, ‘$\gamma_2$’, ‘$\gamma_3$’
2:     Begin
3:         For each Dataset ‘$DS$’ with Feature ‘$F$’ and Feature type ‘$T$’ and patients ‘$P$’ //patient initialization
4:             Update patient fitness value as in equation (7)
5:             Estimate revised discrete learning factor as in equation (8)
6:             Update feature of each patient as in equation (9)
7:             If ‘$val(bgmax)\ is\ updated$’
8:                 If stopping criteria not met go to step 6
9:                 else
10:                    Output relevant features
11:                End if
12:            Go to step 6
13:        End if
14: End for
15: End
```

Followed by which the feature ‘F’ of patient ‘i’ is updated as given below.

$$w.i + 1 = \left[\frac{w.i+m.i}{\pm 0.5}\right] \quad (9)$$

From the above Eq. (8), the feature of each patient ‘i’ are updated for identifying relevant feature for disease prediction based on the exploitation ‘$m.i$’ and exploration ‘$w.i$’’ factors. The pseudo code representation of Iterative African Buffalo Optimized Feature Selection is given below.

As given in the above algorithm with the pre-processed features as input, the objective remains in obtaining the pertinent or relevant feature for heart disease prediction with minimum error rate and complexity. In this work new administered feature selection model based on Iterative African Buffalo Optimization is selected based on iterative arbitrary learning factor. The feature selection for disease prediction inhibits the process of choosing a subset of pertinent or relevant features. As Iterative African Buffalo Optimization is with minimizing error rate and space complexity achievement so this work suggests Iterative African Buffalo Optimization as a significant feature selection model for predicting heart disease.

4. Experimental settings

In this section, the experiments have been carried out on the proposed method Z-score Normalized Iterative African Buffalo Optimization (ZN-IABO) and compared with the other two methods, imperialist competitive algorithm [1], FODW [2], and state-of-the-art optimizer such as MLBO [22] The performance analysis has adopted statistical assessment metrics to scale the decision disease prediction accuracy of the proposed method and the other state-of-the-art methods, optimizers and performed in Python.

The metrics used in this regard are heart disease prediction accuracy, heart disease prediction time, error rate, space complexity for different numbers of patients and features. The results obtained for these metrics from the experimental study performed on the proposed method ZN-IABO and the state-of-the-art methods imperialist competitive algorithm [1], FODW [2], MLBO [22] using cardiovascular disease dataset [29]. The cardiovascular disease dataset [29] used in our proposed work is an open-source dataset found on Kaggle.

The dataset comprises of 70,000 patient records split into records with cardiovascular disease and records without cardiovascular disease. Eleven features are present in the dataset of which, four belongs to the demographic, four features are belongs to the examination results and three features of social history.

The features given in the above table are found to be either numerical, whereas some other features are assigned categorical codes, and others are binary

Table 1. Details of cardiovascular disease dataset

S. No	Features(or) Attributes	Categorization
1	Age	Demographic
2	Height	Demographic
3	Weight	Demographic
4	Gender	Demographic
5	Systolic blood pressure	Examination
6	Diastolic blood pressure	Examination
7	Cholesterol	Examination
8	Glucose	Examination
9	Smoking	Social history
10	Alcohol intake	Social history
11	Physical activity	Social history

values. Also the classes are found to be balanced, however the presence of larger number of female patients are found than when compared to the male patients.

5. Discussion and theoretical explanation

In this section, numerous findings related to this work will be discussed. These findings are obtained from the simulation result. First, in general, the proposed method successfully becomes a good metaheuristic algorithm. It overcomes the challenge of finding the near-optimal or acceptable solution within the given iteration. In this section, the experimental evaluation of Z-score Normalized Iterative African Buffalo Optimization (ZN-IABO) is carried out using on factors such as heart disease prediction accuracy, heart disease prediction time, space complexity and error rate with respect to number of patient data. Also in depth discussion with state-of-the-art methods, with the aid of graph and tabulation are provided in detail.

Data Pre-processing and feature selection are two very important methods to measure the heart disease prediction algorithms and handle optimization issues. An algorithm must scan the search space accurately in the initial iterations. During the iterations of the algorithm and after a proper search, the algorithm must reach the appropriate feature selected for heart disease prediction. Iterative African Buffalo Optimization is used to analyze and compares the exploitation power index in optimization algorithms.

Therefore, it can be stated that the proposed ZN-IABO (with optimization) method has been able to provide more appropriate solutions by maintaining exploration and exploitation and is much more competitive than the other three optimization algorithms to implement in solving optimization problems.

5.1 Performance analysis of heart disease prediction accuracy

First, the efficiency of the method in terms of accuracy is estimated. In other words heart disease prediction accuracy refers to the prediction accuracy made with respect to the diseased patient. The heart disease prediction accuracy is mathematically formulated as given below.

$$HDP_{acc} = \frac{P_{AP}}{P_D} * 100 \qquad (10)$$

From the above Eq. (10), heart disease prediction accuracy 'HDP_{acc}' is measured on the basis of the patients accurately predicted with the disease 'P_{AP}' to the actual patients suffering from the disease 'P_D'. It is measured in terms of percentage (%).

The first finding is that this proposed ZN-IABO (with optimization) method is competitive enough compared with the state-of-art algorithms ZN-IABO (without optimization), imperialist competitive algorithm, FODW, and MLBO. This competitiveness gives the prospect that this proposed algorithm is promising to be used in solving real-world optimization issues. Compared with the above method, this proposed algorithm is also competitive enough. Table 2 given below compares the efficacy of the methods in terms of heart disease prediction accuracy.

Table 2. Comparison of heart disease prediction accuracy with the proposed ZN-IABO method (both with and without optimization), imperialist competitive algorithm [1], FODW [2], and MLBO [22]

Patients	Heart disease prediction accuracy (%)				
	ZN-IABO (with optimization)	ZN-IABO (without optimization)	Imperialist competitive algorithm	FODW	MLBO
500	95.91	85.71	93.77	91.83	80.45
1000	94.25	84.15	91.45	89.15	78.25
1500	93.15	82.55	90.25	87.25	76.65
2000	92.25	81	88.15	84.25	75.85
2500	91.15	78.15	86.35	82.15	74.15
3000	90.35	75.25	85.15	80.15	71.35
3500	89.85	73	83.25	78.35	69.65
4000	89.25	70.15	82	76.25	66
4500	89	68.25	81.85	74.25	64.15
5000	87	67	78	70	62

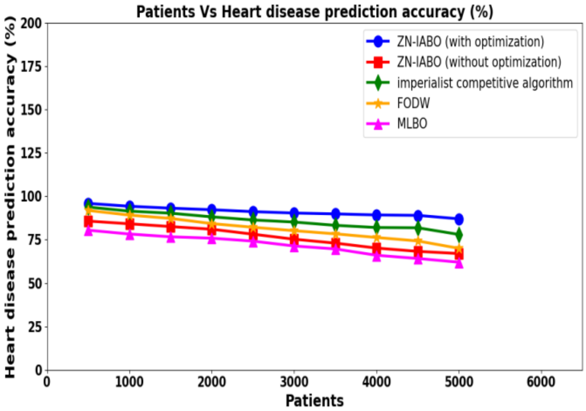


Figure. 4 Graphical representation of heart disease prediction accuracy

Fig. 4 given above illustrates the four different curves with blue curve representing the proposed with optimization, red curve representing the proposed without optimization, green curve [1], orange curve denoting the [2], and finally pink curve indicating the [22] respectively. From the figure a decreasing trend is observed for all the three methods. However, the red curve representing the proposed without optimization is found to be comparatively lesser in performance than the proposed method ZN-IABO (with optimization), [1, 2, 22]. Also, with simulations conducted over 500 patients, and 245 patients' actually detected with the disease and 235 patients' accurately predicted using ZN-IABO (with optimization), 210 patients' accurately predicted using ZN-IABO (without optimization), 230 patients' accurately predicted using [1], 225 patients' accurately predicted using [2], and 402 patients' accurately predicted using [22] the overall heart disease prediction accuracy were found to be 95.91%, 85.71%, 93.77%, 91.83% and 80.45% respectively. From the results the accuracy rate was observed to be higher using ZN-IABO (with optimization) than compare to [1, 2, 22] and ZN-IABO (without optimization). The reason behind the improvement was due to the conversion of different features of respective patients to same magnitude and followed by it standardization was also done. Moreover, with this algorithm the results were not affected by magnitude of data, therefore contributing to heart disease prediction accuracy using ZN-IABO (with optimization) by 20% compared to ZN-IABO (without optimization), 6% compared to [1], 13% compared to [2], and 28% compared to [22].

5.2 Performance analysis of heart disease prediction time

The second paramount parameter of significance is the heart disease prediction time. Early the prediction is made treatment can be given at the initial stage and therefore reducing the mortality rate. Hence, heart disease prediction time plays a major role in controlling the death rate. The heart disease prediction time is mathematically estimated as given below.

$$HDP_{time} = \sum_{i=1}^{n} P_i * Time\ [FS] \qquad (11)$$

From the above Eq. (11), heart disease prediction time 'HDP_{time}' is measured based on the sample of patients' involved for simulation 'P_i' and the time consumed in obtaining the features or selecting the features '$Time\ [FS]$'. It is measured in terms of milliseconds (ms).

The second finding is that the heart disease prediction time of this proposed method is fast enough. As indicated in Table 3, a heart disease prediction time can be achieved in ZN-IABO (with optimization) than the existing methods. Although in general, this proposed ZN-IABO (with optimization) is better than other methods, its performance is good in solving optimization functions. Besides, better results can be simply achieved by expanding ten iterations. Table 3 given above compares the efficacy of the methods in terms of heart disease prediction time.

Fig. 5 given above illustrates the graphical representation of heart disease prediction time with respect to 5000 different numbers of patients

Table 3. Comparison of heart disease prediction time with proposed ZN-IABO method (both with and without optimization), imperialist competitive algorithm [1], FODW [2], and MLBO [22]

Patients	Heart disease prediction time (ms)				
	ZN-IABO (with optimization)	ZN-IABO (without optimization)	imperialist competitive algorithm	FODW	MLBO
500	567.5	1007.5	580	597.5	1075.25
1000	635.25	1125	735.85	815.55	1185
1500	815.35	1185	1025.45	1085.25	1255.15
2000	955.15	1235	1085.95	1155.45	1295.15
2500	1025.25	1315	1135.45	1175.35	1375
3000	1085.35	1455	1185.25	1235.15	1515.45
3500	1125.45	1485	1205.35	1315.45	1585
4000	1155.75	1535	1225.45	1455.25	1655.35
4500	1205.25	1595	1325.55	1535.85	1685
5000	1315.45	1835	1425.55	1615.25	1915.45

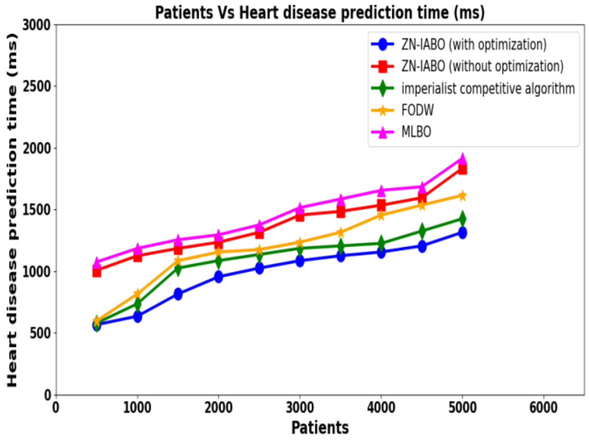


Figure. 5 Graphical representation of heart disease prediction time

involved in the simulation. From the figure increasing the number of patients causes an increase in the heart disease prediction time. In other words, with the increase in the patients' considered for simulation, the features involved in prediction increases and therefore causes an increase in the heart disease prediction time. However, with simulations conducted for 500 patients', the heart disease prediction for single patient was observed to be 1.135ms using ZN-IABO (with optimization), 2.015ms using ZN-IABO (without optimization), 1.160ms using [1], 1.195ms using [2] and 2.1505 ms using [22]. From these results, the overall prediction time involved in heart disease was found to be 567.5ms using ZN-IABO (with optimization), 1007.5ms using ZN-IABO (without optimization), 580ms [1], 597.5ms [2] and]. 1075.25 ms [22With these results, the time involved in predicting heart disease using ZN-IABO (with optimization) was found to be comparatively better than all the other three methods. The prediction time improvement was found owing to the application of Log-transformed Z-score Normalized Data pre-processing algorithm. By applying this algorithm, Z-score log-transformation was applied via normalized log-transformed Z-score function separately for lower limit and upper limit. Using these values only, the Z-score values were obtained and performing computationally efficient pre-processing, therefore reducing the heart disease prediction time using ZN-IABO (with optimization) by 29% compared to ZN-IABO (without optimization), 10% compared to [1], 17% compared to [2] and 33% compared to [22].

5.3 Performance analysis of error rate

While predicting heart disease, a significant amount of error is said to occur by wrongly predicting the patient not having the disease with the presence of the disease. The error rate in our work is mathematically expressed as given below.

$$\mathrm{Err} = \frac{\mathrm{P_{WP}}}{\mathrm{P_D}} * 100 \tag{12}$$

From the above Eq. (12), error rate 'Err' is measured on the basis of the number of patients wrongly predicted with the heart disease though not actually having the disease '$\mathrm{P_{WP}}$' to the patients actually detected with heart disease '$\mathrm{P_D}$'. It is measured in terms of percentage (%).Table 4 given below compares the significance of the methods in terms of error rate.

Table 4. Comparison of error rate with proposed ZN-IABO method (both with and without optimization), imperialist competitive algorithm [1], FODW [2], and MLBO [22]

Patients	Error rate (%)				
	ZN-IABO (with optimization)	ZN-IABO (without optimization)	imperialist competitive algorithm	FODW	MLBO
500	4.08	14.28	6.12	8.16	16.55
1000	5.15	15.15	7.35	9.15	17
1500	6.35	15.85	8.15	9.85	17.45
2000	7.15	16	9.05	10.35	17.65
2500	8.25	16.35	9.35	10.85	17.85
3000	8.85	16.85	9.85	11	18.15
3500	9	17	10	11.35	18.45
4000	9.15	17.15	10.15	11.55	18.75
4500	9.35	17.35	10.25	11.75	19.25
5000	9.55	17.55	10.4	12	19.55

Fig. 6 given below shows the error rate for varying numbers of patients with three types of input features, namely, objective, examination and subjective. From the figure it is inferred that the error rate is found to be directly proportional to the number of patients involved in simulation. In other words, increasing the number of patients involved in simulation purpose increases the frequency of results of medical examination and therefore a steep rise in the error rate also. However, with simulations done using 500 numbers of patients, actually detected with disease being 245 and wrongly predicted by applying ZN-IABO (with optimization) (10), ZN-IABO (without optimization) (35), (15) by applying [1], (20) by applying [2], (40) by applying [22] the overall error rate was found to be 4.08%, 14.28%, 6.12%, 8.16% and 16.55% respectively. From these results the error rate using ZN-IABO (with optimization) was comparatively better over the other three methods. The reason behind the minimization of error rate was owing to the application of Iterative African Buffalo Optimized Feature Selection algorithm. By applying this algorithm, a discrete learning process is applied instead of the conventional fitness function. Here, each patient obtains the features from its' own finest experience and the finest experience of all patients'. Moreover, after a number of iterations, patient favor is not complied, therefore eliminating the premature convergence and minimizing the error rate using ZN-IABO (with optimization) by 53% compared to applying ZN-IABO (without optimization), 17% compared to [1], 29% compared to [2] and 58% compared to [22] respectively.

Received: June 15, 2022. Revised: September 2, 2022.

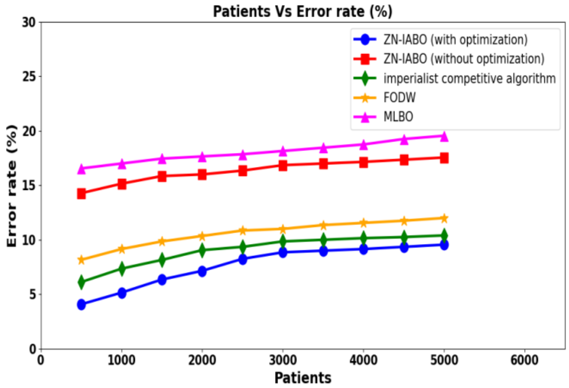


Figure. 6 Graphical representation of error rate

5.4 Performance analysis space complexity

Finally, space complexity involving heart disease prediction is analyzed. This is mathematically expressed as given below.

$$SC = \sum_{i=1}^{n} P_i * Mem\left(FS(Disease\ Prediction)\right) \quad (13)$$

From Eq. (13), the space complexity 'SC' here refers to the memory consumed in the feature selection process for disease prediction '$Mem\left(FS(Disease\ Prediction)\right)$' and the patients considered for simulation 'P_i'. Table 5 given below compares the significance of the methods in terms of error rate.

Table 5 also shows that the proposed ZN-IABO (with optimization) method is competitive enough compared with other algorithms. It outperforms all state-of-art algorithms in solving optimization functions.

Table 5. Comparison of space complexity with proposed ZN-IABO method (both with and without optimization), imperialist competitive algorithm [1], FODW [2], and MLBO [22]

Patients	Space complexity (KB)				
	ZN-IABO (with optimization)	ZN-IABO (without optimization)	imperialist competitive algorithm	FODW	MLBO
500	500	2000	1000	1500	2120
1000	850	2025	1035	1585	2225
1500	975	2085	1085	1635	2285
2000	1025	2155	1155	1725	2335
2500	1135	2235	1325	1935	2375
3000	1285	2285	1455	2015	2425
3500	1345	2315	1535	2035	2485
4000	1525	2455	1715	2085	2585
4500	1915	2535	2085	2215	2675
5000	2035	2725	2235	2355	2855

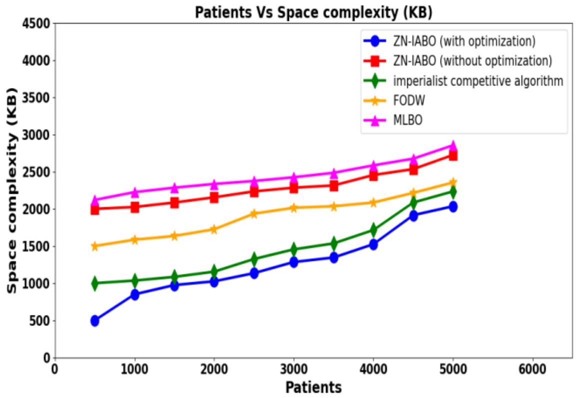


Figure. 7 Graphical representation of space complexity

Finally, Fig. 7 given above illustrates the space complexity involved in analyzing the heart disease prediction. Space complexity also increases with the increase in the number of patients involved in simulation. This space complexity refers to a significant amount of memory incurred during the entire prediction process. So this metric is an overall analysis of the entire prediction process. Also a considerable improvement or minimization of space complexity was involved by applying ZN-IABO (with optimization) upon comparison with ZN-IABO (without optimization), [1, 2, 22]. This was due to the reason that though optimization patterns were utilized in all the four methods, i.e., ZN-IABO (with optimization), [1, 2, 22], by applying Iterative African Buffalo Optimization in our work, subset of pertinent features for heart disease prediction was selected by discarding the irrelevant features using iterative arbitrary learning factor (IALF), therefore reducing the space complexity involved in ZN-IABO (with optimization) by 45% compared to ZN-IABO (without optimization), 16% compared to [1], 36% compared to [2] and 49% compared to [22].

6. Conclusion

Many optimization problems in different sciences should be optimized and solved using appropriate methods. African Buffalo Optimization algorithms are one of the most widely used methods in this field, which provide appropriate solutions to the problem based on random search in the problem-solving space. The main idea of the proposed optimization method is to predict heart disease. In this paper, in order to achieve minimal error rate, space complexity, and heart disease prediction time with maximum heart disease prediction accuracy, the proposed method is designed. First, Log-transformed Z-score was applied to three different types of input features for analyzing the missing value which has been estimated by the proposal built based on the separate upper and lower limit score. Next, the actual relevant feature selection for heart disease prediction was conducted to obtain the relevant feature by means of log transformation. Moreover, the proposed method successfully achieves a globally optimal solution in three functions.

The performance of the proposed method ZN-IABO (with optimization) has been scaled by comparing the other contemporary methods, ZN-IABO (without optimization), imperialist competitive algorithm, FODW, and MLBO through the performance metrics heart disease prediction accuracy, time, error rate, and complexity. The examinations of the execution analysis reveal the notable superiority of the proposed method ZN-IABO (with optimization) compared to the state-of-the-art methods like ZN-IABO (without optimization), imperialist competitive algorithm, FODW, and MLBO.

The authors suggest some ideas and perspectives for future studies. This work is an early improved version of the ABO. It means that the other improvements are still possible and promising. Furthermore, studies that implement this proposed method to be used to solve real-world optimization issues are needed to give a more comprehensive evaluation of this algorithm.

Conflicts of Interest

The authors declare have no conflict of interest.

Author Contributions

The contributions of authors are as follows:

P. Muthulakshmi; Conceptualization, Methodology, software, validation, formal analysis, investigation, data curation and writing-original paper draft.

Dr. M. Parveen: Validation, supervision and project administration.

Acknowledgments

None.

References

[1] J. N. Khiarak, M. F. Derakhshi, K. Behrouzi, S. Mazaheri, Y. Z. Harghalani, and R. M. Tayebi, “New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection”, *Health and Technology, Springer*, Vol. 10, pp. 667-678, 2020.

[2] F. A. M. A. Yarimi, N. M. A. Munassar, M. H. M. Bamashmos, and M. Y. S. Ali, “Feature optimization by discrete weights for heart disease prediction using supervised learning”, *Soft Computing, Springer,* Vol. 25, pp. 1821-1831, 2021

[3] X. Zhang, D. Wu, F. Miao, H. Liu, and Y. Li, “Personalized Hemodynamic Modeling of the Human Cardiovascular System: A Reduced-Order Computing Model”, *IEEE Transactions on Biomedical Engineering*, Vol. 67, No. 10, pp. 2754-2764, 2020

[4] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, “Early and accurate detection and diagnosis of heart disease using intelligent computational model”, *Scientific Reports, Springer*, Vol. 10, No. 19747, pp. 1-17, 2020.

[5] D. L. Robinson, A. L. Ware, M. C. Sauer, R. V. Williams, Z. Ou, A. P. Presson, L. Y. Tani, L. L. Minich, and D. T. Truong, “Implications of Changing *Z*-Score Models for Coronary Artery Dimensions in Kawasaki Disease”, *Pediatric Cardiology, Springer*, Vol. 42, No. 2, pp. 432-441, 2021.

[6] L. Verma, S. Srivastava, and P. C. Negi, “A Hybrid Data MiningModel to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data”, *Journal of Medical Systems, Springer*, Vol. 40, No. 7, 2016.

[7] C. B. Gokulnath and S. P. Shantharajah, “An optimized feature selection based on genetic approach and support vector machine for heart disease”, *Cluster Computing,* Vol. 22, pp. 14777-147, 2018

[8] C. Huang, J. Du, B. Nie, R. Yu, W. Xiong, and Q. Zeng, “Feature Selection Method Based on Partial Least Squares and Analysis of Traditional Chinese Medicine Data”, *Computational and Mathematical Methods in Medicine*, Vol. 2019, pp. 1-11, 2019.

[9] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. J. M. Sharmat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. D. Boer, “Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques”, *IEEE Access*, Vol. 9, pp. 19304-19326, 2021.

[10] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, “HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System”, *IEEE Access*, Vol. 8, pp. 133034-133050, 2020.

[11] R. Chen, C. Dewi, S. Huang, and R. E. Caraka, “Selecting critical features for data classification based on machine learning methods”, *Journal of Big Data, Springer*, Vol. 7, No. 52, pp. 1-26, 2020.

[12] V. V. Shankar, V. Kumar, U. Devagade, V. Karanth, and K. Rohitaksha, "Heart Disease Prediction Using CNN Algorithm", *Computer Science, Springer Nature*, Vol. 1, No. 170, 2020.

[13] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare", *IEEE Access*, Vol. 8, pp. 107562-107582, 2020.

[14] O. Golgstein, M. Kachuee, K. Karkkainen, and M. Sarrafzadeh, "Target-Focused Feature Selection Using Uncertainty Measurements in Healthcare Data", *ACM Transactions on Computing for Healthcar*e, Vol. 1, No. 3, pp. 1-17, 2020.

[15] E. Nasarian, M. Abdar, M. A. Fahami, R. Alizadehsani, S. Hussain, M. E. Basiri, M. Z. Moghadam, X. Zhou, P. Pławiak, U. R. Acharya, R. Tan, and N. Sarrafzadega, "Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach", *Pattern Recognition Letters, Elsevier*, Vol. 133, pp. 33-40, 2020.

[16] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", *IEEE Access*, Vol. 7, pp. 81542-81554, 2019.

[17] Sourabh, V. Mansotra, P. Kour, and S. Kumar, "Voting-Boosting: A novel machine learning ensemble for the prediction of Infants' Data", *Indian Journal of Science and Technology*, Vol. 12, No. 22, pp. 2189-2202, 2020.

[18] P. G. Shynu, V. G. Menon, R. L. Kumar, S. Kadry, and Y. Nam, "Blockchain-Based Secure Healthcare Application for Diabetic-Cardio Disease Prediction in Fog Computing", *IEEE Access*, Vol. 9, pp. 45706-45720, 2021.

[19] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes", *The Journal of Supercomputing, Springer*, Vol. 77, pp. 5198-5219, 2020.

[20] P. D. Kusum and M. Kallista, "Stochastic Komodo Algorithm", *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 4, pp. 1-11, 2022, doi: 10.22266/ijies2022.0831.15.

[21] P. D. Kusuma and A. Dinimaharawati, "Fixed Step Average and Subtraction Based Optimizer", *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 4, pp. 1-13, 2022, doi: 10.22266/ijies2022.0831.31.

[22] F. A. Zeidabadi, S. A. Doumari, M. Dehghani, and O. P. Malik, "MLBO: Mixed Leader Based Optimizer for Solving Optimization Problems", *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 4, pp. 1-9, 2021, doi: 10.22266/ijies2021.0831.41.

[23] M. Dehghani, Z. Montazeri, A, Dehghani, R. A. R. Mendoza, H. Samet, J. M. Guerreroo, and G. Dhiman, "MLO: Multi Leader Optimizer", *International Journal of Intelligent Engineering and Systems*, Vol. 13, No. 6, pp. 1-10, 2020, doi: 10.22266/ijies2020.1231.32.

[24] F. A. Zeidabadi, M. Dehghani, and O. P. Malik, "TIMBO: Three Influential Members Based Optimizer", *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 5, pp. 1-8, 2021, doi: 10.22266/ijies2021.1031.12.

[25] F. A. Zeidabadi, M. Dehghani, and O. P. Malik, "RSLBO: Random Selected Leader Based Optimizer", *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 5, pp. 1-10, 2021, doi: 10.22266/ijies2021.1031.46.

[26] M. Suman, V. P. Sakthivel, and P. D. Sathya, "Squirrel Search Optimizer: Nature Inspired Metaheuristic Strategy for Solving Disparate Economic Dispatch Problems", *International Journal of Intelligent Engineering and Systems*, Vol. 13, No. 5, pp. 1-11, 2020, doi: 10.22266/ijies2020.1031.11.

[27] F. A. Zeidabadi and O. P. Malik, "POA: Puzzle Optimization Algorithm", *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 1, pp. 1-10, 2022, doi: 10.22266/ijies2022.0228.25.

[28] S. A. Doumari, H. Givi, M. Dehghani, and O. P. Malik, "Ring Toss Game-Based Optimization Algorithm for Solving Various Optimization Problems", *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 3, pp. 1-10, 2021, doi: 10.22266/ijies2021.0630.46.

[29] https://www.kaggle.com/sulianova/cardiovascular-disease-dataset