International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhancing Cardiovascular Disease Diagnosis: A Hybrid Model of Whale Optimization Algorithm with Multilayer Deep Perceptive Classifier

G. Angayarkanni [1*], Dr. M. Rajasenathipathi [2]

Submitted:10/03/2024 **Revised**: 25/04/2024 **Accepted**: 02/05/2024

Abstract: An efficient Hybridization of Whale Optimized MultiLayer Deep Perceptive Classifier (HWO-MLDPC) is proposed to improve the diagnosis accuracy of cardiovascular disease. The proposed technique includes three main stages: preprocessing, feature selection, and classification. First, the data is preprocessed using the Theil-Sen Regressive Discretized Binning method, which smoothes the raw data into a structured format based on median estimation. After preprocessing, the feature selection process uses stochastic bivariate correlation to identify relevant features based on maximal mutual information. Next, classification with the selected relevant features is performed using the Hybridization of Whale Optimized MultiLayer Deep Perceptive Classifier. The proposed MultiLayer Deep Perceptive Classifier comprises several layers. First, the number of selected features is given to the input layer. Then, the input is transferred to the hidden layer, where feature analysis is performed using the Generalized Tversky index similarity. The sigmoid activation function provides the final disease classification results. At the same time, whale optimization updates the weights of inputs with lesser error to achieve accurate classification results with minimum error at the output layer. Based on the classification results, cardiovascular disease can be diagnosed correctly. Experimental evaluation is carried out using different quantitative metrics such as accuracy, precision, recall, F-measure, and time complexity. The analyzed results demonstrate the superior performance of the proposed technique.

Keywords: *Cardiovascular disease, Discretized Binning, Generalized Tversky index similarity, Multilayer Perceptron, Stochastic Bivariate Correlation, Theil-Sen Regression, Whale Optimization Algorithm.*

1. Introduction

Early prediction of cardiovascular disease aids physicians in making more precise decisions about their patients' health statuses. However, identifying cardiac disease based on early-stage indicators poses a notable challenge in medical practice. Consequently, employing machine learning (ML) methods offers a solution to discern symptoms associated with heart disease. While diverse data analytics and mining techniques have been applied for this purpose, the vast volume of data often hinders the rapid enhancement of disease detection accuracy. A conventional deep learning model was developed to identify normal and abnormal cases of heart disease, but due to the massive amount of data, accuracy in heart disease detection is not effectively improved with less dimensionality. However, this work aims to apply a novel deep learning model for dimensionality reduction and heart disease detection by using a feature selection technique. Cardiovascular diseases (CVDs) pose significant challenges in the medical field, including heart diseases in heart patients. Conventional examination methods have been employed to find heart disease, but they are difficult. Due to the non-availability of medical diagnostic devices and medical proficiency, especially in undeveloped countries, diagnosing and treating heart disease is very difficult. However, accurate and timely diagnosis of heart disease is crucial to prevent further damage to the patient. Traditional techniques often lead to imprecise diagnoses and take more time due to human errors. To address these problems, in this work, a Theil-Sen Regressive Stochastic Bivariate Correlation-based Hybridization of Whale Optimization Algorithm with Multilayer Deep Perceptive Classifier (HWOA-MLDPC) is designed. The dataset undergoes preprocessing via the Theil-Sen Regressive Discretized Binning technique. Subsequently, bivariate correlation is employed to identify pertinent features and discard irrelevant ones. Following feature selection, classification is conducted utilizing a multi-layer deep perceptive classifier, analyzing both training patient data and disease testing data. The sigmoid activation function is then utilized to evaluate the similarity outcomes, enabling more accurate classification of disease and normal patient data with reduced time consumption.

2. Related Work

To accurately diagnose patients' risk of cardiovascular illnesses, Ying An et al. [1] developed an end-to-end method known as Deep Risk. To automatically extract the high-quality features, the designed approach was used. Next, a more precise and healthy presentation from Electronic Health Records (EHRs) was achieved. The DeepRisk was utilized to minimize the dimensionality reduction. But, the designed mechanism has not improved the performance of

[1] Research Scholar, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, Tamilnadu, India
Email: g.angayarkanni@gmail.com
[*] (Corresponding Author)
[2] Associate Professor, Department of Computer Technology, Nallamuthu Gounder Mahalingam College, Pollachi, Tamilnadu, India
Email: r.senathipathi@gmail.com

diagnosis of cardiovascular disease with the large volume of data. Awais Mehmood et al., [2] developed Convolutional neural networks (CNN) model for heart disease prediction. The designed model has not enhanced heart disease prediction. Pengpai et al., [3] examined a new multi-modal method for predicting cardiovascular diseases. Nabaouia Louridi et al., [4] discussed Machine learning techniques for identifying a patient's heart condition. However, it failed to consider a recall.

Ali A. Samir et al., [5] developed CNN-jSO optimization approach for predicting the heart diseases. But the, heart disease was not accurately predicted. D. Shiny Irene et al., [6] introduced deep belief network and an extreme learning machine (DBNKELM) based on weighted attributes. However, the time complexity of disease prediction was not improved. Anna Karen Garate-Escamila et al., [7] developed different machine-learning classifiers to predict the patient with heart disease. However, the larger dataset was not applied with different feature selection techniques. Zhang et al. [8] developed a Deep Neural Network (DNN) coupled with a feature selection approach for predicting heart disease. Nonetheless, they did not employ efficient optimization techniques in deep learning to achieve improved performance. Pooja Rani et al., [9] investigated hybrid decision support system for early prediction of heart disease. However, the designed system of heart disease diagnosis was not sufficient. In order to create an effective model for predicting cardiovascular illness, Jameel Ahamed et al. [10] investigated machine learning algorithms. However, the minimum error rate was not reached by the accuracy-designated algorithms.

Recurrent Neural Network (RNN) and Logistic Chaos-Based Whale Optimization (LCBWO) were used for the purpose of identifying heart disease data by P. Priyanga et al. [11]. It was not stated, yet, how accurate the classification was for identifying heart disease. The Cross Industry Standard Process for Data Mining (CRISP-DM) technique was studied by Barbara Martin et al. [12] utilizing classifiers; nevertheless, the prediction performance of disease detection was not improved. Cardiovascular disease was classified and diagnosed using Genetic Algorithm (GA)-Linear Discriminant Analysis (LDA) by V. Jothi Prakash and N. K. Karthikeyan [13]. The optimal features were chosen for improved prediction, but the intended strategy was not found. Gradient Descent Optimization was developed by Muhammad Saqib Nawaz et al. [14] to predict cardiovascular disease. However, the enormous datasets were not taken into account. A weight-learning technique was used by Jiang Xie et al. [15] for accurate cardiovascular disease prediction; however, prediction accuracy assessment was not the main focus.

To predict cardiovascular illness, Sudarshan Nandy et al. [16] used a Swarm-Artificial Neural Network (Swarm-ANN) technique, however, they had trouble identifying important elements from high-dimensional datasets. Particularly in situations of type 2 diabetes, cardiovascular disease diagnosis prediction was carried out by W. Dong et al. [17]. Gihun Joo et al. [18] created machine learning methods utilizing big data to forecast the onset of cardiovascular illness; deep feature learning was not included to improve accuracy.

Chunyan Guo et al. [19] introduced a Recursion-Enhanced Random Forest with an Improved Linear Model (RFRF-ILM) to identify heart disease, yet they did not apply deep feature learning methods to improve accuracy. Aqsa Rahim et al. [20] presented a Machine Learning-based Cardiovascular Disease Diagnosis approach for effectively identifying cardiovascular diseases with high precision, though the computational time was relatively higher. Current CVD diagnosis models focus on feature selection and prediction, neglecting input weight optimization for improved classification accuracy. Atimbire SA et al [21] present a novel heart disease prediction approach utilizing the Whale Optimization Algorithm for feature selection. Through extensive dataset analysis and evaluation metrics, the study showcases notable enhancements in model accuracy and performance metrics, underscoring the efficacy of WOA in optimizing predictive modeling for healthcare applications.

Deep learning models struggle to improve accuracy with massive datasets, as seen in DeepRisk [1] and CNN models [2]. A few of the research [7, 14] do not make use of large datasets, while others [17] concentrate on particular patient populations. This restricts how broadly the models may be applied. Moreover, Most of the above studies prioritize either feature selection or prediction accuracy, neglecting the optimization of input weights for classification [5, 8, 13, 14]. This research proposes a novel approach, HWOA-MLDPC that addresses these limitations. It combines a Hybrid Whale Optimization Algorithm (HWOA) with a Multilayer Deep Perceptive Classifier (MLDPC) for simultaneous feature selection, weight optimization, and improved classification accuracy in CVD diagnosis. This comprehensive strategy offers a significant advancement over existing methods.

2.1 Major Contributions of the study:

Achieves superior performance in CVD diagnosis compared to existing methods, with an accuracy of 96.42%, precision of 97.59%, recall of 98.63%, and F-measure of 98.10%. Offers a more comprehensive approach to CVD diagnosis by combining data preprocessing, feature selection, and classification with optimization techniques.

3. Methodology

The suggested HWOA-MLDPC approach comprises three primary steps: data preprocessing, feature selection, and classification, aimed at enhancing disease diagnosis accuracy with extensive

datasets. Initially, the comprehensive dataset is analyzed for disease diagnosis, featuring various attributes and numerous data points. Specifically, the cardiovascular disease dataset encompasses 13 attributes including identification, age, height, weight, gender, blood pressure measurements (ap_hi, ap_lo), cholesterol and glucose levels, smoking and alcohol habits, physical activity, and disease presence indicators.

3.1 Dataset

The experimental assessment of the proposed technique, along with established methods such as DeepRisk [1], CNN models [2], SVM, and RF, is conducted in Python using a cardiovascular disease dataset obtained from www.kaggle.com. The primary objective is to detect the presence or absence of cardiovascular disease in patients with diabetes. The dataset comprises 13 attributes and encompasses 70,000 instances.

Table 1. Attribute Description

S.No	Attributes	Description
1	Id	Patient id number
2	Age	Patient age in days
3	Height	Patient height in cm
4	Weight	Patient weight in kg
5	Gender	1-women, 2-men
6	ap_hi	Systolic blood pressure
7	ap_lo	Diastolic blood pressure
8	Cholesterol	Cholesterol 1: normal, 2: above normal, 3: well above normal
9	gluc	Glucose 1: normal,2: above normal 3: well above normal
10	Smoke	Smoking 1: Yes, 0:no
11	alco	Alcohol intake 1: Yes, 0:no
12	active	Physical activity
13	cardio	1 presence , 0 absence

3.2 Training and Testing Procedure

Training data represents a selected subset of the entire dataset, which comprises 70,000 records. The selection of this subset is based on a specified number of patient records for each evaluation (10,000, 20,000, etc.). Training data plays a crucial role in the machine learning process, as it enables the model to learn patterns and relationships within the dataset. By adjusting and optimizing the model's parameters using the training data, we can significantly improve its performance.

Testing data is a carefully selected portion of the remaining dataset that was not used for training. It is used to evaluate the model's performance on previously unseen data. The testing data enables us to assess how well the model generalizes to new instances and provides insights into its accuracy and effectiveness. The selection of appropriate testing data is critical to ensure that our machine learning model is robust and reliable.

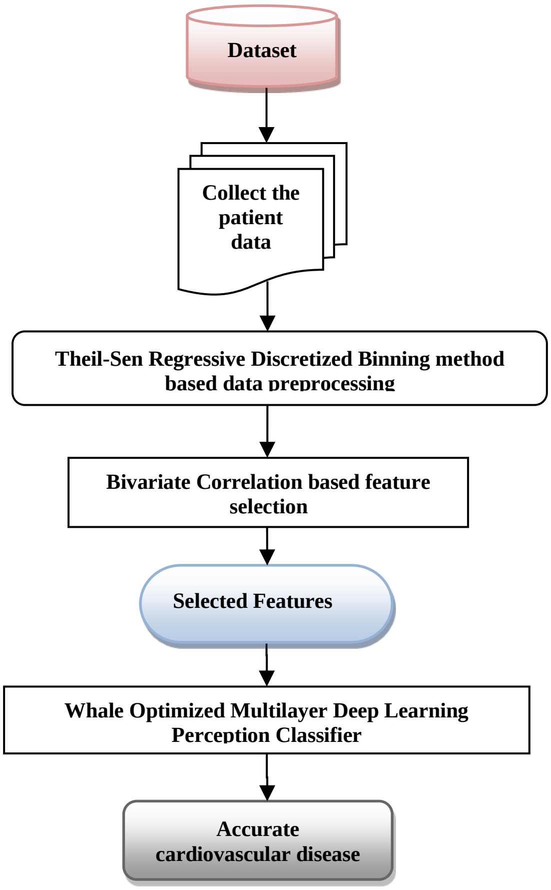
Dataset
Collect the patient data
Theil-Sen Regressive Discretized Binning method based data preprocessing
Bivariate Correlation based feature selection
Selected Features
Whale Optimized Multilayer Deep Learning Perception Classifier
Accurate cardiovascular disease

Fig. 1 Architecture diagram of proposed HWO-MLDPC technique

3.3 Theil-Sen Regressive Discretized Binning Method

Data preprocessing is the processing of transforming the data into a structured format that helps to effectively process the classification. The advantage of Data pre-processing is to improve accuracy and minimize the time complexity. Here Theil-Sen Regressive Discretized Binning method for data preprocessing is used and it is a median-based estimator to smooth the raw data into a structured format by removing the noisy dataset based on the neighborhood of feature values.

Algorithm 1: Theil-Sen Regressive Discretized Binning Method for data preprocessing

Input: Dataset, features $\beta_f = {\beta_f}_1, {\beta_f}_2, \dots, {\beta_f}_n$ and feature values

Output: Preprocessed data

Begin

```
    Number of features $\beta_f = \beta_{f_1}, \beta_{f_2}, \dots, \beta_{f_n}$ taken as input
      For each feature '$\beta_{f_i}$'
        Sorting the feature value in ascending order
        Partition the features value into an equal number of bins
    for each bin
              Find the median value
    end for
         Replace the median value of features in the bins
         Find the minimum and maximum boundary in the bins
         Replace each bin value with neighborhood boundary value
    End for
End
```

3.4 Stochastic Bivariate Correlative Feature Selection

To perform the Dimensionality reduction, the Stochastic Bivariate correlation method is employed to select pertinent features. Stochastic processes entail the consideration of probabilities for various potential outcomes, introducing random variation in one or more inputs (i.e., features) over time. Maximal mutual information serves as a metric for evaluating the mutual dependence between two variables. Bivariate correlation, a statistical tool, assesses the level of mutual dependence between two features. Let us consider the number of features as $\beta_{f_1}, \beta_{f_2}, \dots, \beta_{f_n}$. The correlation between the features is measured as given below,

$$\delta = \frac{n * \sum \beta_{f_i} * \beta_{f_j} - \left(\sum \beta_{f_i}\right)\left(\sum \beta_{f_j}\right)}{\sqrt{\left[n * \sum {\beta_{f_i}}^2 - \left(\sum \beta_{f_i}\right)^2\right]}\sqrt{\left[n * \sum {\beta_{f_j}}^2 - \left(\sum \beta_{f_j}\right)^2\right]}} \quad (1)$$

Where, δ denotes a correlation coefficient, 'n' symbolizes a number of features. $\sum \beta_{f_i} * \sum \beta_{f_j}$ denotes a sum of the product of paired score of two features. $\sum {\beta_{f_i}}^2$ represents a squared score of β_{f_i} and $\sum {\beta_{f_j}}^2$ represents a squared score of β_{f_j}. The Maximal mutual information quantifies the probability of the features to be selected based on correlation value.

$$P\ (\delta\ |y) = \max\left(\frac{P(\delta, y)}{P(\delta)P(y)}\right) \quad (2)$$

$$P\ (\delta\ |y) = \begin{cases} 1 & ;\ select\ the\ features \\ 0 & ;\ otherwise \end{cases} \quad (3)$$

By using (3), $P\ (\delta\ |y)$ denotes a probability of the features to be selected, max denotes the maximum probability of the features to be, δ denotes correlation outcomes, y denotes outcomes of feature selection.

Algorithm 2: Bivariate Correlative Feature Selection
Input: Preprocessed data **Output:** select significant features

```
Begin
    Number of features $\beta_f = \beta_{f_1}, \beta_{f_2}, \dots, \beta_{f_n}$ taken as input
    for each feature $\boldsymbol{\beta_{f_i}} \in \boldsymbol{\beta_f}$
    Measure the Bivariate correlation '$\delta$'
       Apply Maximal mutual information
    Measure the maximum probability of the features to be selected ' $P(\delta|y)$'
    if $(P(\delta|y) = 1)$ then
    Feature is said to be relevant
    Select the relevant features
    else
    Feature is said to be irrelevant
    Remove the irrelevant features
     End if
     end for
End
```

3.5 Hybrid Whale Optimization Algorithm with Multilayer Deep Perceptive Classifier for disease diagnosis

Ultimately, disease classification is conducted utilizing the Hybrid Whale Optimization Algorithm in conjunction with the Multilayer Deep Perceptive Classifier, incorporating the chosen features. The Multilayer Deep Perceptive Classifier, a fully connected feedforward artificial neural network, produces a series of outputs based on a set of selected input features, featuring three primary layers: input, hidden, and output. Each artificial neuron within the network receives input from a specified number of features originating from the input layer or outputs from neurons in preceding layers. The connections between nodes or neurons are referred to as synapses. In the input layer, the selected input features are received, with each feature assigned a weight denoted as '$\alpha_1, \alpha_2, \dots, \alpha_n$' and then combined with a bias term 'g'.

$$Q(t) = \left[\sum_{i=1}^{n} \beta_{f_i}(t) * \alpha_i\right] + g \quad (4)$$

From equation (4), the activity of neuron at input layer '$Q(t)$' specifies that the weighted 'α_i' sum of the input features '$\beta_{f_i}(t)$' and add to the bias function 'g' that stored the value is '1'. Subsequently, the input is conveyed to the initial hidden layer, followed by a series of hidden layers situated between the input and output layers. These hidden layers are composed of small individual units referred to as neurons or nodes. The operation of the artificial neuron within the hidden layer is illustrated in Figure 2.

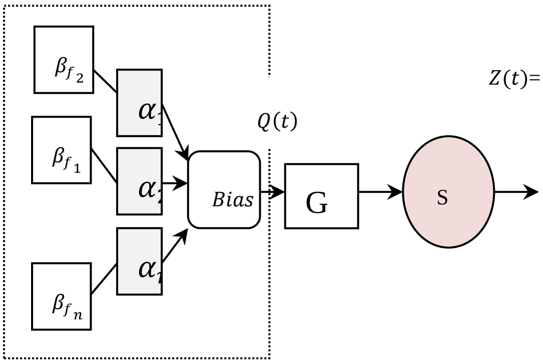

Fig. 2 Flow process of artificial neuron

As shown in Figure 2, an artificial neuron receives the weighed sum of features with bias as input '$Q(t)$'. Then the Generalized Tversky index similarity is used to measure the relationship between testing and training data. Here, the Generalized represents the involving many different data. The similarity measure between the training and testing disease data are estimated.

$$B = \frac{[\beta_{f_t} \cap \beta_{f_{dt}}]}{u\left(\beta_{f_t} \Delta \beta_{f_{dt}}\right) + v\left(\beta_{f_t} \cap \beta_{f_{dt}}\right)} \qquad (5)$$

Where 'B' specifies a similarity coefficient, β_{f_t} signifies the training data, $\beta_{f_{dt}}$shows the testing disease data, $\beta_{f_t} \cap \beta_{f_{dt}}$ designates a mutual dependence between the two data, $\beta_{f_t} \Delta \beta_{f_{dt}}$ indicates a variance between the two data. Then the similarity coefficient output is given to the Sigmoid Activation Function (SAF) for producing the final classification results. The advantage of the activation function is to help the network for learning the complex testing and training patterns in the data. The sigmoid activation function provides the best normalized functions output with 1 and 0, it makes an accurate disease prediction. Hence the proposed deep learning classifier uses the sigmoid activation function for accurate disease prediction.

$$F = \frac{1}{1 + \exp(-B)} \qquad (6)$$

From (6), Where, F denotes a sigmoid activation function, 'B' indicates the similarity coefficient results.

$$F = \begin{cases} 1 \ ; Disease\ diagnosed \\ 0\ ; \qquad otherwise \end{cases} \qquad (7)$$

The sigmoid activation function provides '1' indicates that the disease is correctly diagnosed based on the similarity between testing and training data. Based on the activation function results, the disease is correctly diagnosed. During the learning that occurs in the perceptron, connection weights are updated after each part of data gets processed and measure the error rate. The error rate of data classification is measured as follows,

$$e = \frac{1}{2}\left(Y - Z(t)\right)^2 \qquad (8)$$

Where the error rate 'e' is measured as a squared difference between the actual classification results 'Y' and output produced by the perceptron '$Z(t)$'.The weight is updated as follows,'

$$\nabla \alpha_{t+1} = \alpha_t * \eta \left(\frac{\partial e}{\partial \alpha_t}\right) \qquad (9)$$

Where, $\nabla \alpha_{t+1}$ denotes an updated weight, α_t denotes a current weight, η denotes a learning rate $(\eta < 1)$. A higher learning rate enables the model to learn more rapidly compared to a lower value, '$\frac{\partial e}{\partial \alpha_t}$' denotes a partial derivative of the error 'e' with respect to current weight 'α_t'.

In order to minimize the error, the attributes of the perceptron network called optimal weights are identified by using whale optimization technique. The whale optimization is used to solve optimization problems by minimizing the error rate. Whale optimization is a nature-inspired meta-heuristic algorithm that emulates the hunting behavior of humpback whales. In this algorithm, whales search for prey by generating various bubbles along a path. Here, different weights are considered as whales, and error rate is considered as prey. As a result of optimization, the best solution (i.e. optimal weight) is identified for minimizing the error rate. First, the numbers of whales (i.e. weight values) are initialized. The fitness of the whale is calculated for minimizing the error rate.

$$f = \arg\min e \qquad (10)$$

From (10) f fitness function, arg min denotes an argument of the minimum function, 'e' denotes an error. Based on the fitness, the current best weight is identified to minimize the error of classification results. After that, three different behaviors are carried out such as encircling prey, bubble-net feeding method, and searching the prey. Followed by, an optimal weight value is selected. In encircling prey behavior, the whale determines the location of prey and surrounds them. Since the location of prey in the search space was not known previously. Therefore, the proposed optimization algorithm considers the current best solution (i.e. weight value) is the best optimal. Therefore, the whale position is updated as follows,

$$P_w(i+1) = P_{best}(i) - A.B \quad (11)$$

$$B = \left|\varphi.P_{best}(i) - P_p(i)\right| \quad (12)$$

From the above equation, $P_w(i+1)$ denotes an updated position of whale, $P_{best}(i)$ denotes a current best position of the whale, $P_p(i)$ denotes a position vector of the prey, $A\ and\ B$ represents a coefficient vector. Therefore, the coefficient vector is expressed as follows,

$$A = (2k - 1)r \quad (13)$$

$$\varphi = 2k \quad (14)$$

From (3), ‘r’ is linearly reduced from 2 to 0 over the way of iterations, and ‘k’ indicates a random vector [0, 1]. The bubble-net behavior of whales is executed based on shrinking encircling approach and spiral updating position. The Shrinking encircling mechanism is achieved by reducing the value of r from 2 to 0 over the course of iterations. Then the spiral updating position is executed as follows,

$$P_w(i+1) = D' e^{mn} \cos(2\pi q) + P_{best}(i) \quad (15)$$

$$D = |P_{best}(i) - P_w(i)| \quad (16)$$

Where, $P_w(i+1)$ denotes an updated position of the whale, D denotes an updated distance among the whale current position ‘$P_w(i)$’ and best solution ‘$P_{best}(i)$’, ‘m’ is a constant [0, 1] used to describe the structure of the logarithmic curve, Exponential function ‘e’ is the base of natural logarithms, ‘n’ is the random number ranges are [-1, 1].

Finally, searching the prey behavior is randomly executed according to the position.

$$P_w(i+1) = P_{rand}(i) - A.B \quad (17)$$

$$B = |\varphi . P_{rand}(i) - P_w(i)| \quad (18)$$

From (17) (18), $P_{\ rand}(i)$ denotes a random position vector of a whale, $P_w(i)$ current position of the whale and φ is calculated using (17) (18). Repeat the process till the maximum iteration is achieved. Finally, the best solution (i.e. weight) is obtained to minimize the error of disease classification. Finally, the results are transferred into the output layer of the multilayer deep perceptive classifier.

Algorithm 3: Hybridization of Whale Optimized MultiLayer Deep Learning Perceptive Classifier

Input: Selected relevant features $\beta_f = \beta_{f_1}, \beta_{f_2}, \dots, \beta_{f_n}$ and data $\beta_d = \beta_{d_1}, \beta_{d_2}, \dots, \beta_{d_n}$

Output: Increase the disease diagnosis accuracy

```
Begin
    Number of selected features $\beta_f = \beta_{f_1}, \beta_{f_2}, \dots, \beta_{f_n}$ taken at the input layer
      For each feature $\beta_{f_i}$
        Assign weight ‘$\alpha_i$’ and add bias ‘$g$’
        Obtain the neuron activity at input layer ‘$Q(t)$’
    end for
    For each training data with testing disease data –[hidden layer]
        Perform the generalized Tversky index similarity measure ‘$B$’
        Apply sigmoid activation function ‘$F$’
        If ($F = +1$) then
            Correctly diagnosed as disease
        else
            Correctly diagnosed as normal
        End if
      For each results
        Measure the error rate ‘$e$’
            Update the weight ‘$\nabla\alpha_{t+1}$’
        Find minimum error by identifying optimal weight
           Initialize the whale's populations (i.e. weights)
        for each whale
            Calculate the fitness ‘$f$’ using (10)
            Find current best ‘$P_{best}$’ whale
        while( t < maximum number of iterations )
        if (P<0.5)
         if (|$A$|<1) then
        Update the position to select the optimal whale using (11)
        else
        Select a random position of whale $P_{rand}(i)$
        Update the position of current best solution using (17)
        else
        Update the position of current best solution (15)
        end if
         end if
        Obtain the best solution
        end for
         end for
        t= t+1
         end while
        Return (global best optimal solution )
          Obtain the final classification results with minimum error at the output layer
End
```

Algorithm 3 outlines the process of diagnosing cardiovascular diseases accurately and quickly. The MultiLayer Deep Learning Perceptive classifier analyzes input using several layers. The whale optimization is used to minimize the error and find the optimal weight value. The globally best weight value is determined to minimize the error rate of disease classification. The results are displayed at the output layer, increasing accuracy and minimizing false positives.

4. Results and Discussion

In this section, performance results and discussion of the HWOA-MLDPC is compared with DeepRisk [1], CNN models [2], and two state-of-the-art methods namely Support Vector Machine (SVM), Random Forest (RF). A comprehensive analysis is conducted using various metrics including accuracy, precision, recall, F-measure, and time complexity. The performance of the HWOA-MLDPC method across these metrics is discussed, supported by both tabular and graphical representations. Statistical analysis is employed to assess the utility and efficacy of larger trials, demonstrating the quality of results from a meta-analysis of the proposed technique.

4.1. Performance Analysis of Accuracy

It is defined as the number of patient data accurately diagnosed as cardiovascular disease presence or absence to the total number of patient data. Therefore, the overall accuracy rate is measured as follows.

$$A_{cd} = \left[\frac{t_{tr}+t_{ne}}{t_{tr}+t_{ne}+f_{pv}+f_{nv}}\right] * 100 \qquad (19)$$

Where, A_{cd}indicates cardiovascular diagnosing accuracy, $t_{tr}\ denotes\ a\ true\ positive, t_{ne}$ indicates a true negative, f_{pv}represents a false positive, f_{nv} denotes a false negative. Therefore the accuracy is measured in terms of percentage (%).

Table 2 Accuracy versus Number of patient data

Number of patient data	**Accuracy (%)**				
	DeepRisk	CNN model	SVM	RF	HWOA-MLDPC
10000	91.84	87.35	82.34	84.5	96.5
20000	86.48	85.38	80.89	83.6	95.25
30000	91.55	89.6	85.3	87.3	94.33
40000	90.75	90.05	85.33	87.66	94.5
50000	90.62	89.5	85.12	86.8	95
60000	90.11	88	86.24	87.98	95.12
70000	90.36	89.9	87.06	88.11	96.42

The average of comparison results proves that the accuracy of the HWOA-MLDPC technique is significantly increased by 6%, 8%, 13%, and 10% when compared to the existing DeepRisk, CNN model, SVM, and RF.

4.2 Performance Analysis of Precision:

It is measured based on true positives and the sum of true positives and false positives. The Precision is formulated as given below,

$$Pr = \left[\frac{t_{tr}}{\beta_{P_t}+f_{pv}}\right] * 100 \qquad (20)$$

Where, Pr denotes Precision,$t_{tr}\ denotes\ a\ true\ positive$,$f_{pv}$ represents a false positive. The Precision is measured in percentage (%).

Table 3 Precision versus Number of patient data

Number of patient data	**Precision (%)**				
	DeepRisk	CNN model	SVM	RF	HWOA-MLDPC
10000	93.5	90.55	86.55	88.65	98.33
20000	92.65	90.78	86.78	88.41	96.91
30000	93.74	91.65	90.44	92.15	96.08
40000	93.65	92.68	90.87	91.63	96.23
50000	93.87	92.85	90.85	91.32	96.84
60000	92.89	93.66	91.45	92.55	96.43
70000	93.66	92.41	90.55	91.44	97.59

The average of ten comparisons results confirms that the precision is significantly increased by 4%, 5%, 8%, and 7% than the DeepRisk, CNN model, and state-of-the-art methods.

4.3 Performance Analysis of Recall:

It is measured based on a truly positive and false negative. The recall is measured using the given formula,

$$Rl = \left[\frac{t_{tr}}{t_{tr}+f_{nv}}\right] * 100 \qquad (21)$$

Where Rl denotes recall, t_{tr} represents the true positive, f_{nv} denotes the false negative. The recall is measured in percentage (%).

Table 4 Recall versus Number of patient data

Number of patient data	**Recall (%)**				
	DeepRisk	CNN model	SVM	RF	HWOA-MLDPC
10000	94.65	93.86	90.66	91.23	97.79
20000	92.68	91.87	88.65	90.87	97.73
30000	95.82	94.58	91.74	92.89	97.82
40000	94.33	93.75	91.89	92.15	97.81
50000	94.97	93.26	90.76	92.78	97.87
60000	94.21	93.12	91.74	92.86	98.38
70000	94.56	92.89	90.85	91.45	98.63

The average of ten results indicates that the recall of the HWOA-MLDPC technique is considerably increased by 4%, 5%, 8%, and 6% when compared to conventional methods.

4.4 Performance Analysis of F-measure:

It is measured as the mean of precision as well as recall. It is measured as follows,

$$f - m = \left[2 * \frac{Pr * Rl}{Pr + Rl}\right] * 100 \qquad (21)$$

Where$f - m$ denotes an F-measure, Pr denotes precision, 'Rl denotes a recall. F-measure is measured in terms of percentage (%).

Table 5 F-measure versus Number of patient data

Number of patient data	F-measure (%)				
	DeepRisk	CNN model	SVM	RF	HWOA-MLDPC
10000	94.07	92.17	88.55	89.92	98.05
20000	92.66	91.32	87.70	89.62	97.31
30000	94.76	93.09	91.08	92.51	96.94
40000	93.98	93.21	91.37	91.88	97.01
50000	94.41	93.05	90.80	92.04	97.35
60000	93.54	93.38	91.59	92.70	97.39
70000	94.10	92.64	90.69	91.44	98.10

The average of the comparison result of F-measure is found to be considerably increased by 4%, 5%, 8%, and 7% as compared to the existing methods respectively.

4.5 Performance Analysis of time complexity:

It is measured as the amount of time consumed by the algorithm to diagnose cardiovascular disease. The time complexity is mathematically calculated as follows,

$$Tc = n * time\ (Dpd) \qquad (23)$$

From (22),Tc denotes a time complexity, n represents a number of patient data, $time\ (Dpd)$ denote a time for diagnosing the single-patient data. Therefore, the time complexity is measured in terms of milliseconds (ms).

Table 6 Time complexity versus Number of patient data

Number of patient data	Time Complexity (ms)				
	DeepRisk	CNN model	SVM	RF	HWOA-MLDPC
10000	58	64	74	68	46
20000	63.5	69	80.6	75	52
30000	68.8	76	87.21	83.8	60
40000	75.6	83.5	95.65	90.5	66
50000	89	94	87.5	98.4	72
60000	90.3	97	92.4	108.3	79.8
70000	94.65	109.8	105.35	113.6	85.4

The average of ten comparison results of the HWOA-MLDPC technique is considerably minimized by 15% and 23%, 27%, and 28% when compared to existing methods.

5. Conclusion

This study focuses on cardiovascular disease prediction through the application of the HWOA-MLDPC technique. Initially, the proposed method conducts data preprocessing using the Theil-Sen Regressive Discretized Binning approach. Subsequently, feature selection employs stochastic bivariate correlation fused with maximal mutual information to identify pertinent features while eliminating irrelevant ones. Ultimately, classification is executed using a Hybridization of Whale-Optimized Multi-Layer Deep Learning Perceptive Classifier, enabling the identification of cardiovascular disease by analyzing the similarity between training and testing disease data with minimal error. An extensive experimental evaluation is conducted utilizing a large dataset of cardiovascular disease cases and the quantitative data supporting this claim includes superior performance metrics such as an accuracy of 96.42%, precision of 97.59%, recall of 98.63%, and F-measure of 98.10%. This outperforms existing techniques and showcases the effectiveness of the proposed approach in enhancing the accuracy of cardiovascular disease diagnosis. Strengths of this research include the innovative combination of optimization techniques, feature selection methods, and advanced classification algorithms within the HWOA-MLDPC model, leading to improved accuracy in disease diagnosis. The model's ability to achieve high precision and recall rates demonstrates its effectiveness in accurately identifying cardiovascular diseases. However, a potential weakness could be the complexity of the model, which may require computational resources and expertise to implement effectively. To further enhance the research, future improvements could focus on optimizing the computational efficiency of the proposed method, potentially by exploring parallel processing techniques or optimizing the algorithm for scalability.

References

[1] An Y, Huang N, Chen X, Wu F, Wang J. High-risk prediction of cardiovascular diseases via attention-based deep neural networks. IEEE/ACM transactions on computational biology and bioinformatics. 2019 Aug 14; 18 (3):1093-105. doi: https://doi.org/10.1109/TCBB.2019.2935059

[2] Mehmood A, Iqbal M, Mehmood Z, Irtaza A, Nawaz M, Nazir T, Masood M. Prediction of heart disease using deep convolutional neural networks. Arabian Journal for Science and Engineering. 2021 Apr; 46 (4):3409-22. doi: http://dx.doi.org/10.1007/s13369-020-05105-1.

[3] Li P, Hu Y, Liu ZP. Prediction of cardiovascular diseases by integrating multi-modal features with

machine learning methods. Biomedical Signal Processing and Control. 2021 Apr 1;66:102474. doi: https://doi.org/10.1016/j.bspc.2021.102474

[4] NabaouiaLouridi, Samira Douzi&Bouabid El Ouahidi. Machine learning-based identification of patients with a cardiovascular defect. Journal of Big Data. Springer, Volume 8, 2021, Pages 1-15. doi: https://doi.org/10.1186/s40537-021-00524-9

[5] Samir AA, Rashwan AR, Sallam KM, Chakrabortty RK, Ryan MJ, Abohany AA. Evolutionary algorithm-based convolutional neural network for predicting heart diseases. Computers & Industrial Engineering. 2021 Nov 1;161:107651. doi: https://doi.org/10.1016/j.cie.2021.107651

[6] Irene DS, Sethukarasi T, Vadivelan N. Heart disease prediction using hybrid fuzzy K-medoids attribute weighting method with DBN-KELM based regression model. Medical Hypotheses. 2020 Oct 1;143:110072. doi: https://doi.org/10.1016/j.mehy.2020.110072

[7] Gárate-Escamila AK, El Hassani AH, Andrès E. Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked. 2020 Jan 1;19:100330. doi: https://doi.org/10.1016/j.imu.2020.100330

[8] Zhang D, Chen Y, Chen Y, Ye S, Cai W, Jiang J, Xu Y, Zheng G, Chen M. Heart disease prediction based on the embedded feature selection method and deep neural network. Journal of healthcare engineering. 2021 Sep 29;2021:1-9. doi: https://doi.org/10.1155%2F2021%2F6260022

[9] Rani P, Kumar R, Ahmed NM, Jain A. A decision support system for heart disease prediction based upon machine learning. Journal of Reliable Intelligent Environments. 2021 Sep;7(3):263-75. doi: https://link.springer.com/article/10.1007/s40860-021-00133-6

[10] Ahamed J, Manan Koli A, Ahmad K, Jamal A, Gupta BB. CDPS-IoT: cardiovascular disease prediction system based on IoT using machine learning. doi: http://dx.doi.org/10.9781/ijimai.2021.09.002

[11] Priyanga P, Pattankar VV, Sridevi S. A hybrid recurrent neural network-logistic chaos-based whale optimization framework for heart disease prediction with electronic health records. Computational Intelligence. 2021 Feb;37(1):315-43. doi: https://doi.org/10.1111/coin.12405

[12] Martins B, Ferreira D, Neto C, Abelha A, Machado J. Data mining for cardiovascular disease prediction. Journal of medical systems. 2021 Jan;45:1-8. doi: https://doi.org/10.1007/s10916-020-01682-8

[13] Jothi Prakash V, Karthikeyan NK. Enhanced evolutionary feature selection and ensemble method for cardiovascular disease prediction. Interdisciplinary Sciences: Computational Life Sciences. 2021 Sep;13(3):389-412. doi: https://doi.org/10.1007/s12539-021-00430-x

[14] Nawaz MS, Shoaib B, Ashraf MA. Intelligent cardiovascular disease prediction empowered with gradient descent optimization. Heliyon. 2021 May 1;7(5). doi: https://doi.org/10.1016/j.heliyon.2021.e06948

[15] Xie J, Wu R, Wang H, Chen H, Xu X, Kong Y, Zhang W. Prediction of cardiovascular diseases using weight learning based on density information. Neurocomputing. 2021 Sep 10;452:566-75. doi: https://doi.org/10.1016/j.neucom.2020.10.114

[16] Nandy S, Adhikari M, Balasubramanian V, Menon VG, Li X, Zakarya M. An intelligent heart disease prediction system based on swarm-artificial neural network. Neural Computing and Applications. 2023 Jul;35(20):14723-37. doi: https://doi.org/10.1007/s00521-021-06124-1

[17] Dong W, Wan EY, Bedford LE, Wu T, Wong CK, Tang EH, Lam CL. Prediction models for the risk of cardiovascular diseases in Chinese patients with type 2 diabetes mellitus: A systematic review. Public Health. 2020 Sep 1;186:144-56. doi: https://doi.org/10.1016/j.puhe.2020.06.020

[18] Joo G, Song Y, Im H, Park J. Clinical implication of machine learning in predicting the occurrence of cardiovascular disease using big data (Nationwide Cohort Data in Korea). IEEE Access. 2020 Sep 3;8:157643-53. doi: https://doi.org/10.1109/ACCESS.2020.3015757

[19] Guo C, Zhang J, Liu Y, Xie Y, Han Z, Yu J. Recursion enhanced random forest with an improved linear model (RERF-ILM) for heart disease detection on the internet of medical things platform. Ieee Access. 2020 Mar 16;8:59247-56. doi: https://doi.org/10.1109/ACCESS.2020.2981159

[20] Rahim A, Rasheed Y, Azam F, Anwar MW, Rahim MA, Muzaffar AW. An integrated machine learning framework for effective prediction of cardiovascular diseases. IEEE Access. 2021 Jul 20;9:106575-88. doi: https://doi.org/10.1109/ACCESS.2021.3098688

[21] Atimbire, S.A., Appati, J.K. & Owusu, E. Empirical exploration of whale optimization algorithm for heart disease prediction. Sci Rep 14, 4530 (2024). Doi: https://doi.org/10.1038/s41598-024-54990-1