INTERNATIONAL JOURNAL OF
Telecommunications
AIR DEFENSE COLLEGE

A Machine Learning Model Based on the Archimedes Optimization Algorithm for Heart Disease Prediction

Mohamed S. ElGendy [1], Hossam El-Din Moustafa [2], Hala B. Nafea [2], Warda M. Shaban [1]

[1] Affiliation 1; mohamedelgendy@nilehi.edu.eg and warda_mohammed@nilehi.edu.eg
[2] Affiliation 2; hossam_moustafa@mans.edu.eg and halabahyeldeen@mans.edu.eg
* Correspondence: engmosaad22@gmail.com

Citation: ElGendy, M.; Moustafa, H.; Nafea, H.; Shaban, W.
Title. *Inter. Jour. of Telecommunications,* ***IJT'*2025**,
Vol. 05, Issue 01, pp. 1-21, 2025.

Editor-in-Chief: Youssef Fayed.

Received: 21/11/2024.

Accepted: date 02/01/2025.
Published: date 05/01/2025.

Publisher's Note: The International Journal of Telecommunications, IJT, stays neutral regard-ing jurisdictional claims in published maps and institutional affiliations.

Abstract: Predicting and diagnosing cardiac conditions is a major challenge in medicine because of the many variables involved, including the results of a physical examination and the patient's symptoms and indicators. According to data from the World Health Organization (WHO), heart disease is the leading cause of death worldwide, taking the lives of 18 million people annually. Machine Learning (ML) algorithms are essential to modern medicine, particularly when it comes to using medical databases to diagnose illnesses. In this paper, a novel ML model called Heart Disease Detection Model (HD2M) is presented. The suggested HD2M has four phases: (i) data collecting and preprocessing, which includes converting non-numerical values into numbers, eliminating outliers, and filling in missing values. (ii) feature selection using Archimedes Optimization Algorithm (AOA). Actually, AOA is used to select the most important features. (iii) Patient detection through this phase the selected features are used to fed various ML classifiers. These classifiers are Extreme Gradient Boost (EGB), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). (iv) Evaluate Model and Disease Prediction. Experimental results indicate that HD^{2}M outperforms its rivals in terms of F1-measure, precision, accuracy, and recall.

Keywords: Machine Learning (ML), Archimedes Optimization Algorithm (AOA), Heart disease, Feature Selection.

1. Introduction

The World Health Organization (WHO) reports that cardiac arrest is one of the most prevalent conditions with a high fatality rate. Hypertension affects around one-third of the population, and cardiovascular illness is more common in cities than in rural areas [1,2]. People who are in their forties and fifties are also more likely to acquire heart disease. Numerous elements, including personal and professional habits, genetic predispositions, medical history, and lifestyle choices, are important in the development of heart disease. A sedentary lifestyle, heavy alcohol use, smoking, depression, obesity, hereditary diseases, long-term stress, and pre-existing heart issues are all common risk factors for heart disease. Giving the right care requires an accurate diagnosis and treatment of the patient's problems; a major obstacle for healthcare organizations is addressing cost in order to implement highly preventive measures and provide exceptional and efficient medical care [3].

The integration of cutting-edge technology is necessary to overcome these challenges. This means looking into methods like using data analysis to improve the way healthcare workers make decisions. Healthcare systems may optimize patients' conditions by utilizing these technologies, which eventually creates the foundation for more precise and effective patient care. A sizable medical data warehouse houses patient information and medical records. Medical professionals usually base their conclusions on their experience and training rather than on a data-based understanding. Machine Learning (ML), with its capacity to autonomously adapt and enhance models, is leading the way in this revolution, empowering healthcare practitioners to make well-informed decisions and select the most effective treatments from the onset of a medical condition. ML, with its ability to recognize patterns and extract valuable insights, is crucial for efficient problem-solving in the fast-paced world of today. These advancements enhance prediction accuracy and allow for a deeper understanding of intricate, non-linear interactions within large datasets [4]. Developing a decision support system that can identify cardiac disease from clinical data is a simple and affordable way to apply ML and data mining techniques. This kind of decision-making system can help in early disease diagnosis [5].

Hence, in order to determine the probability of heart disease occurrence in specific populations, it is essential to employ efficient data mining methods, which involve examining the dataset to identify significant patterns and valuable insights. The dataset is a collection of various data relevant to the cause as well as the resulting features for each patient under observation. Furthermore, feature selection techniques are integrated to address the outcomes transcending some inconsequential features in the dataset. Feature selection involves the selection of characteristics based on their relative value to the final feature [6]. These estimation evaluations help build an ML model that can predict cardiac illness in testing data by exploiting the patterns or insights found in the training data.

In this paper, a heart disease prediction model based on ML is introduced called Heart Disease Detection Model (HD^{2}M). The proposed ^{HD2M} is based on The Cleveland Clinic's Cardiovascular Disease Dataset that was obtained from Kaggle. Actually, HD^{2}M is composed of four main parts, which are (i) data collection and preprocessing, which includes several stages; removing outlier items, filling or inputting missing values, and converting non-numerical values to numerical values, (ii) Feature Selection, which uses Archimede Optimization Algorithm (AOA), (iii) patient detection using various ML classifiers such as KNN, SVM, RF, and EGB. (iv) Evaluate the model and disease prediction. The main goal is to enhance the precision of diagnosing heart disease. Utilizing extensive medical databases, ML algorithms play a basic role in the medical industry by forecasting and diagnosing illnesses. An additional motivation is to analyze non-ensemble classification models' performance. Even though ensemble methods like RF and XGBoost have proven effective in many applications, it is still important to determine whether the models employed in this work, along with a novel feature selection method, can reliably predict heart disease to select the most informative features. In fact, improving classification accuracy is possible by identifying and using the most relevant features.

The next parts are organized as follows: Presented in Section 2 are the background and basic ideas. The proposed system and the accompanying literature are investigated in Sections 3–4 at the same time. Section 5 demonstrates the findings from the experiment, and the article will be wrapped up in Section 6.

2. Background and Basic Concepts

The component used in this paper named AOA, are covered more thoroughly in the subsections that follow.

2.1. Archimedes Optimization Algorithm (AOA)

An efficient metaheuristic algorithm that makes use of Archimedes' law is AOA. For difficult numerical optimization problems and engineering design optimization problems, this population-based methodology, AOA [7], can be used. According to this methodology, the individuals, O_j, that are submerged in the population begin at any arbitrary spot in the search area with a lower bound $\boldsymbol{lb}_j$ and an upper bound $\boldsymbol{ub}_j$, and the entire population, N, is considered as the object. Additionally, AOA initiates search operations by using a starting population of objects that have arbitrary accelerations, densities, and volumes. At this point, the random place of each object in the fluid is likewise initialized. AOA employs an iterative process, continuously evaluating the fitness of the initial population until a termination condition is satisfied. During each iteration, AOA alters the volume and density of the object. When an object comes into contact with another object in close proximity, its acceleration is altered accordingly. The updated position of an object is determined by its revised density (den), volume (vol), and acceleration (acc). The precise mathematical representation of the AOA steps is displayed below. Figure 1 shows AOA algorithm flowchart.

2.2. AOA algorithm steps

In this study, the AOA algorithm is presented in mathematical form. Because it handles both the exploration and exploitation phases, AOA is theoretically a global optimization method. The suggested Algorithm 1 shows pseudo-code, which includes parameter updates, population evaluation, and population initialization. The steps of the suggested AOA are broken down mathematically as follows:

Step 1: Initialize AOA parameters

Set each object's position initial using Equation 1:

$$O_j = rand \times (ub_j - lb_j) + lb_j \qquad (1)$$

Where $rand$ indicates a random value between 0 and 1, and $j = 1,2,3,4, \ldots\ldots, N$ and O_j, among a population of N objects, is the jth object. The search space's lower and upper boundaries are denoted by $lb_j$and ub_j, respectively. Next, set the initial den and vol for every jth object by using Equation 2 and Equation 3 respectively:

$$den_j = rand \qquad (2)$$

$$Vol_j = rand \qquad (3)$$

Where den_j, and Vol_j are the density and volume of object j. Rand is a D-dimensional vector that produces random values arbitrarily in the interval [0, 1]. Then, the acceleration of all the objects is measured by using Equation 4:

$$acc_j = rand \times (ub_j - lb_j) + lb_j \qquad (4)$$

Where acc_j represents the acceleration of the jth object. Then, select the object with the highest fitness value from the original population during the subsequent step of evaluation. Assign the vol_{best}, acc_{best}, den_{best}, and x_{best}.

Step 2: update and Adjust volumes and densities.

By using Equation 5 and Equation 6 respectively and the iteration $t + 1$ the object jth volume and density are updated.

$$vol_j^{t+1} = rand \times \left(vol_{best} - vol_j^m \right) + vol_j^m \qquad (5)$$

$$den_j^{t+1} = rand \times \left(den_{best} - den_j^m \right) + den_j^m \qquad (6)$$

Where vol_j^{t+1} and den_j^{t+1} are the volume and density of object j at iteration $(t + 1)$. The optimal item is characterized by its volume and density discovered thus far are denoted by vol_{best} and den_{best}, whereas the random number rand is uniformly distributed.

Step 3: Transfer Operator and Density Factor

At first, the objects smash into each other. After a while, they try to find a balance. This is accomplished in AOA by means of the transfer operator TF, which is defined in Equation 7, shifting the search function from exploration to exploitation.

$$TF = exp(\frac{t - t_{max}}{t_{max}}) \qquad (7)$$

Where transfer TF steadily rises to 1 with time. t and t_{max} stand for the current iteration and maximum iteration number, respectively. In a similar vein, density reducing factor D helps AOA with both local and global search. With time $(t + 1)$, it gets smaller by utilizing Equation 8:

$$d^{t+1} = exp(\frac{t_{max} - t}{t_{max}}) - (\frac{t}{t_{max}}) \qquad (8)$$

Where d^{t+1} is the density at iteration $(t + 1)$ that gradually lowers, allowing convergence in the previously determined favorable zone. To maintain a healthy equilibrium between exploration and exploitation in AOA, it is important to manage this variable appropriately.

Step 4.1: Exploration Phase

Select a random material (rm) and adjust the object's acceleration for iteration m + 1 by using Equation 9 if TF ≤ 0.5, which indicates that an object collision has occurred.

$$acc_j^{t+1} = \frac{acc_{rm} \times vol_{rm} + den_{rm}}{vol_j^{t+1} \times den_j^{t+1}} \qquad (9)$$

Where acc_j^{t+1} is the acceleration of object j at iteration $(t+1)$. Additionally, den_j, vol_j, and acc_j are the density, volume, and acceleration of object j at the same order. Moreover, the volume, density, and acceleration of a random material are represented by vol_{rm}, den_{rm}, and acc_{rm} at the same order. It is noteworthy to emphasize that exploration is guaranteed for one-third of the iterations if TF ≤ 0.5. Exploration-exploitation behavior will alter when a value other than 0.5 is applied.

Step 4.2: Exploitation phase

If TF > 0.5 indicates that there are no object collisions, by using Equation 10 to adjust the object's acceleration for iteration $(t+1)$.

$$acc_j^{t+1} = \frac{acc_{best} \times vol_{best} + den_{best}}{vol_j^{t+1} \times den_j^{m+1}} \qquad (10)$$

Where acc_{best} is the best object's acceleration.

Step 4.3: Normalize acceleration

Determine the percentage of change by standardizing the acceleration using Equation 11:

$$acc_{j-norm}^{t+1} = l + \frac{acc_j^{t+1} - min(acc)}{max(acc) - min(acc)} \times u \qquad (11)$$

Where $\mathrm{acc}_{\mathrm{j-norm}}^{\mathrm{t+1}}$ is the normalized acceleration at iteration $(t+1)$. u and l are the normalization range and set to 0.9 and 0.1, respectively. Additionally, $max(acc)$ and $min(acc)$ are the maximum and minimum acceleration at the same order.

The percentage of a step that each agent will alter is determined by the acc_{j-norm}^{m+1}. Object j is in the exploration phase if its acceleration value is large and in the exploitation phase otherwise, depending on how far it is from the global optimum. There has been a change from exploring to exploiting in the search process, as shown here. Typically, the acceleration factor is high to begin with and then decreases over time. This helps search agents move from finding the best local solution to finding the best global solution. But keep in mind that some search agents may still need more time than normal to stop exploring. That is why AOA strikes a balance between the two extremes; exploration and exploitation.

Step 5: Update position

By using Equation 12, determine the position of the j^th object for the next iteration $(t+1)$. If TF ≤ 0.5 (exploration phase).

$$x_j^{t+1} = x_j^t + K_1 \times acc_{j-norm}^{t+1} \times rand \times d \times \left(X_{rand} - x_j^t\right) \qquad (12)$$

Where x_j^{t+1} is the j*th* position of the next iteration $(t+1)$ and x_j^t is the position of j^th object in iteration t. Additionally, K_1 is constant equivalent to 2. X_{rand} represents random position. On the other hand, objects adjust their positions if TF > 0.5 (exploitation phase) by using Equation 13.

$$x_j^{t+1} = x_{best}^t + F \times K_2 \times acc_{j-norm}^{t+1} \times rand \times d \times \left(T \times X_{best} - x_j^t\right) \qquad (13)$$

Where x_{best}^t is the best position of j^th object at iteration t and K_2 is constant and equivalent to 6. T is exactly proportional to the transfer operator and rises with time, additionally, Equation 14 is used to define it.

$$T = K_3 \times TF \qquad (14)$$

Where K_3 is constant and T initially takes a specific percentage from the best position and increases with time in the range of $[K_3 * 0.3, 1]$. It begins with a low percentage since this causes the best position and current position to diverge significantly, which raises the random walk's step-size. This percentage steadily rises as the search moves forward, reducing the gap between the ideal position and the current position. As a result, exploration and exploitation are balanced appropriately. The flag denoted by F represents a change in motion direction by using Equation 15.

$$F = \begin{cases} -1 & if\ P > 0.5 \\ +1 & if\ P \leq 0.5 \end{cases} \qquad (15)$$

Where P is a random value between 0 and 1, $P = 2 \times rand - K_4$.

Step 6: Evaluation

Consider the best solution found thus far as you use objective function f to assess each object. Assign the vol_{best}, acc_{best}, den_{best}, and x_{best}.

Algorithm 1. AOA Pseudo code

```
Input
Initialize AOA parameters {Population size (N), Maximum iteration ($t_{max}$) and
Constants {K1, K2, K3, K4}
Output
The best object position
Steps
Start
Generate initial object population with random position using Equation (1), (2),
(3) and (4), respectively.
Choose the initial highest fitness value and assign the $x_{best}(0)$.
Set iteration counter t=1 and object counter j=1
  While $t \leq t_{max}$ do
     For each object j do
          Update volumes and densities using Equations (5) and (6) respectively.
          Update T.F and Density Factor using Equations (7) and (8) respectively.
       If TF ≤ 0.5 then                    (Exploration Phase)
           Update acceleration using Equation (9)
           Update Normalize acceleration using Equation (11)
           Update object position using Equation (12)
       else                                (Exploitation Phase)
           Update acceleration using Equation (10)
           Update Normalize acceleration using Equation (11)
           Update the flag (F) change in motion direction using Equation (15)
           Update object position using Equation (13)
       end if
     end for
Evaluate each object and choose the fitness value and assign the $x_{best}$(m).
Set t=t+1
   end while
optimum solution printing.
End
```

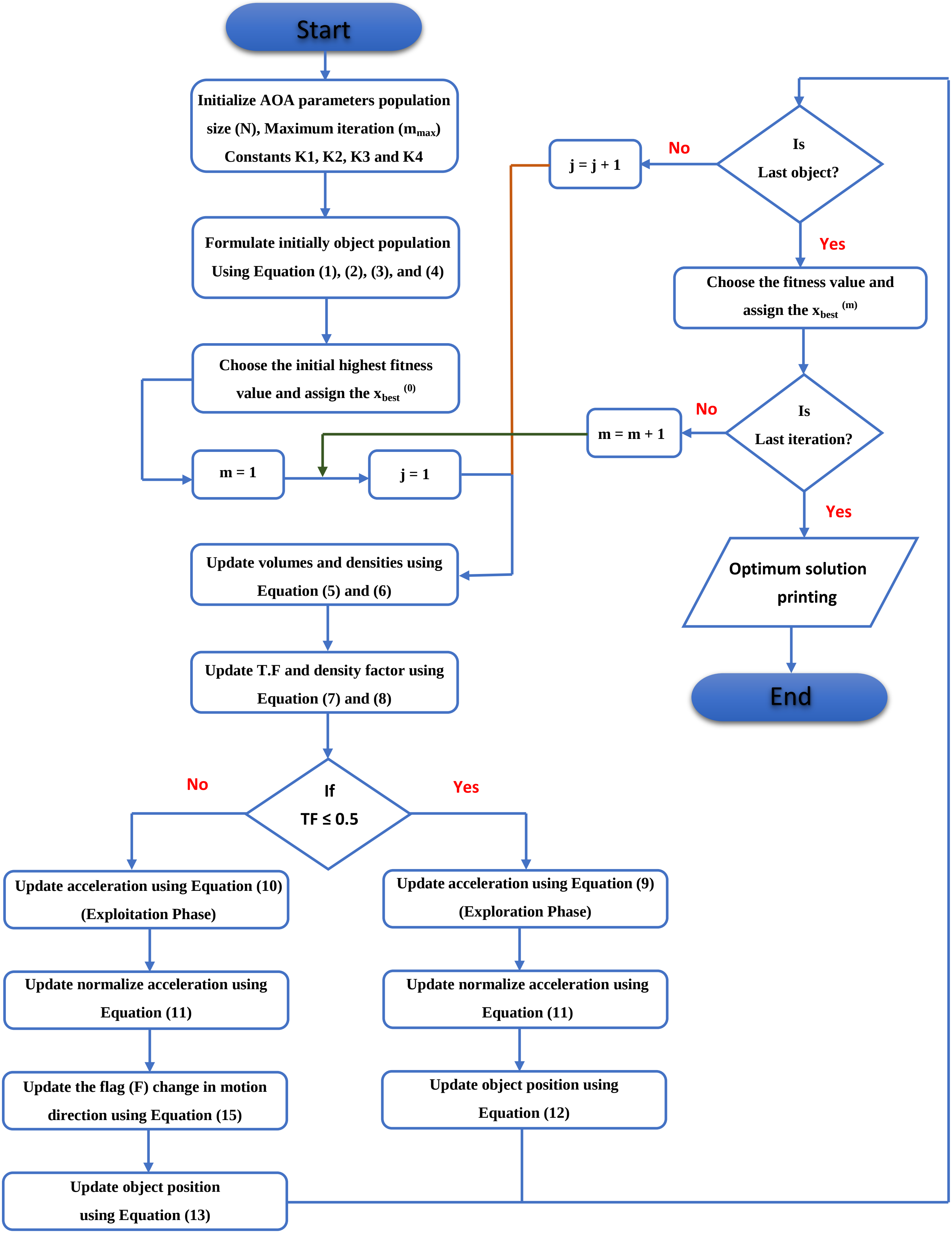

Start
Initialize AOA parameters population size (N), Maximum iteration (m_{max}) Constants K1, K2, K3 and K4
Formulate initially object population Using Equation (1), (2), (3), and (4)
Choose the initial highest fitness value and assign the $x_{best}^{(0)}$
m = 1
j = 1
Update volumes and densities using Equation (5) and (6)
Update T.F and density factor using Equation (7) and (8)
If TF ≤ 0.5
No
Yes
Update acceleration using Equation (10) (Exploitation Phase)
Update normalize acceleration using Equation (11)
Update the flag (F) change in motion direction using Equation (15)
Update object position using Equation (13)
Update acceleration using Equation (9) (Exploration Phase)
Update normalize acceleration using Equation (11)
Update object position using Equation (12)
Is Last object?
No
j = j + 1
Yes
Choose the fitness value and assign the $x_{best}^{(m)}$
Is Last iteration?
No
m = m + 1
Yes
Optimum solution printing
End

Figure 1. AOA algorithm flowchart.

3. Related Work

Data mining refers to the practice of discovering valuable insights within massive databases. Medical data mining uses a variety of classification algorithms to provide clinical support and evaluates these algorithms to see if they can predict patients' cardiac problems. There are a number of ML algorithms used to examine publicly available patient data in the healthcare industry. Specifically, doctors and nurses use this data to make solid diagnoses [8]. Weng et al., for example, came to the conclusion that ML techniques can more accurately identify heart illnesses [9]. Information can be extracted using a variety of techniques, including clustering, regression, association rules, and various classification techniques such as DT, RF, NB, and KNN.

Numerous studies have employed ML and data mining strategies to diagnose cardiac conditions [10]. Reddy et al. used KNN, RF, SVM, NB, and neural networks in conjunction with many feature selection techniques, including the correlation matrix, the learning vector quantization (LVQ) model, and recursive feature elimination (RFE), to categorize cardiac illness as normal or abnormal [6]. According to the findings, RF produced the best outcomes. Recurrent neural networks (RNNs), genetic algorithms, and K-means were utilized by Pillai et al. to prognosticate heart disorders [11]. K-means produced the lowest accuracy, while RNN produced the greatest. The fundamental components of neural networks digital representations of the human brain are interconnected artificial neurons. In ML, they perform a variety of tasks by processing data and drawing conclusions. Time-series analysis jobs benefit greatly from RNN's ability to capture temporal dependencies through the use of sequential information. However, the clustering algorithm's incapacity to handle intricate, non-linear patterns might be partially to blame for some of its reduced accuracy. In addition, the use of genetic algorithms in combination with RNN demonstrates how hybrid techniques can improve predicted accuracy. The application of genetic algorithms aids when optimizing parameters of the model, balancing the shortcomings of RNN and enhancing its strengths.

Raza employed LR, NB, Multilayer Perceptron (MLP), and ensemble learning models to classify cardiac disease [12]. Multiple layers of interconnected nodes (neurons) comprise the structure of ANN, or MLP. With the help of labelled data and backpropagation, the MLP can learn complex patterns and provide accurate predictions by modifying its internal weights to lower prediction errors. MLP is especially useful for complex medical data patterns because of its capacity to capture non-linear correlations. It can automatically extract features due to its hierarchical structure, which enhances its capacity to recognize subtle cardiac illness symptoms. Over time, MLP can increase the accuracy of its predictions by using backpropagation and other training methods. Additionally, the ensemble model's integration of NB and LR in addition to MLP facilitates a thorough examination of heart disease patterns. NB, which is based on probabilistic principles, is highly effective in managing various and incomplete medical information, offering significant insights into potential risk factors. In contrast, LR provides simplicity and interpretability, making the model more transparent for clinical decision-makers. The results indicate that ensemble learning techniques surpass other classifiers in their ability to forecast cardiac disease.

SVM, DT and LR are three algorithms that the authors of [13] propose to use for improved cardiac disease prediction. According to the results, LR outperformed SVM and DT, with an accuracy of 88.52%, and SVM, DT, with a level of accuracy of 83.61% and 80.33%, respectively. Bhatet et al. created a model to diagnose heart disease using an MLP and backpropagation technique. With this established framework, error was reduced, and a maximum accuracy of 80.99% was attained [14]. Abushariah et al. utilized ANN integrated with an Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict cardiac disease. The results indicate that ANN achieved a minimum accuracy of 759.93% and a maximum accuracy of 87.04% [15]. The ANFIS hybrid model integrates the flexibility of neural networks with the comprehensibility of fuzzy logic systems. ANFIS utilizes fuzzy inference systems to establish intricate relationships between input and output data. ANFIS utilizes a hybrid learning technique to adjust the fuzzy rule parameters in accordance with variations in data patterns. When combined with neural networks, fuzzy logic demonstrates superior performance in managing non-linear relationships and uncertainty. ANFIS has shown promise in tackling challenging diagnostic tasks in the medical domain by adeptly discerning and replicating intricate relationships within medical datasets. Moreover, due to its adaptability, ANFIS is particularly advantageous in healthcare settings where data may exhibit dynamic and evolving patterns. The system's ability to adapt to different levels of complexity in medical datasets ensures reliable performance in real-world situations, providing a solid foundation for difficult diagnostic tasks.

Hasan et al. [16] employed MLP with backpropagation and SVM algorithms for the prediction of cardiac disease. The results of the study suggest that MLP has the capacity to achieve a level of accuracy as high as 98%. Sapra et al. examined six ML methodologies, specifically LR, deep learning (DL), DT, RF, SVM, and ensemble learning (gradient boosted tree), for the purpose of diagnosing cardiac disorders. This investigation was conducted utilizing two datasets [17]. Compared to the other approaches, the gradient-boosted tree demonstrated the highest accuracy of 84%. Chen et al. utilized ANN to predict cardiac disease by incorporating a variety of factors [18].

Dutta et al. created an efficient convolutional neural network (CNN) architecture [19]. The results indicate that their CNN model achieved a classification accuracy of 77% in identifying cases of coronary heart disease in the testing data, which represented 85.70% of the entire dataset. Shorewala's study demonstrated that a stacked model consisting of KNN, RF, and SVM achieved the highest performance, with an accuracy of 75.1%. Additionally, the study found that using bagged models improved accuracy by 1.96% [20]. Chang et al. utilized RF method to create an AI model in order to forecast cardiac illness prognosis [21]. When the Python-based model was used with the training set of data, its accuracy was close to 83%.

4. The proposed system

Via this section, the proposed Heart Disease Detection Model (HD^{2}M) will be discussed. Figure 2 shows the workflow of the proposed system. As shown in Figure 2, the proposed model is divided into four main parts (i) data collection and preprocessing, (ii) feature selection, (iii) patient detection, and (iv) Evaluate the model and disease prediction. In the following subsection, each part will be briefly explained.

4.1. Data Collection and Preprocessing

Preparing and gathering data make up the first part of the proposed HD^{2}M. At this point, the proposed model's subsequent stages rely on the data that was prepared at this stage. The Heart Disease Dataset from Cleveland Clinic obtained through Kaggle serves as the foundation for the proposed HD^{2}M [22]. Following that, these data are preprocessed using preprocessing stage, which includes removing outlier items, filling or inputting missing values, and converting non-numerical values to numerical values.

4.2. Normalization

A crucial phase in the data preprocessing process is normalization. To reduce the negative effects of individual data samples, normalization primarily aims to confine preprocessed data within a range [23]. A helpful method for fitting data into a specific range is normalization. Gradient descent is accelerated and improved via normalization. In this study, min-max normalization is utilized to fit data to particular ranges. By essentially transforming the original data linearly, this is accomplished. This formula can be used to calculate normalized data. To calculate normalized data using Equation 16[23].

$$F_{norm} = \frac{f - f_{min}}{f_{max} - f_{min}} \qquad (16)$$

The lower value is denoted by f_{min} and the greater value by f_{max} in the case where f donated the previous value. Next, a value between $[f_{min}, f_{max}]$ is assigned to the normalised value F_norm, where $F_{norm} \in [0,1]$.

4.3. Remove Outlier

In this paper, Interquartile Range (IQR) method is used to check if there is an outlier items. The variation between values in a data set ranked in the 25% and 75% is known as the interquartile range. The values are designated as Q1 and Q3, in that order. IQR using Equation 17 - 19.

$$Tmin = Q1 - C * IQR \tag{17}$$
$$Tmax = Q3 + C * IQR \tag{18}$$
$$IQR = Q3 - Q1 \tag{19}$$

Where C is the user-specified value, which is typically 1.5 [24]. Regretfully, when $Q3$ falls inside the outliers, IQR is likewise impacted by them. Figure 3 shows the outlier items. As shown in Figure 3, the used dataset in this paper [22] contains outlier items that can be removed using Genetic Algorithm (GA) [31].

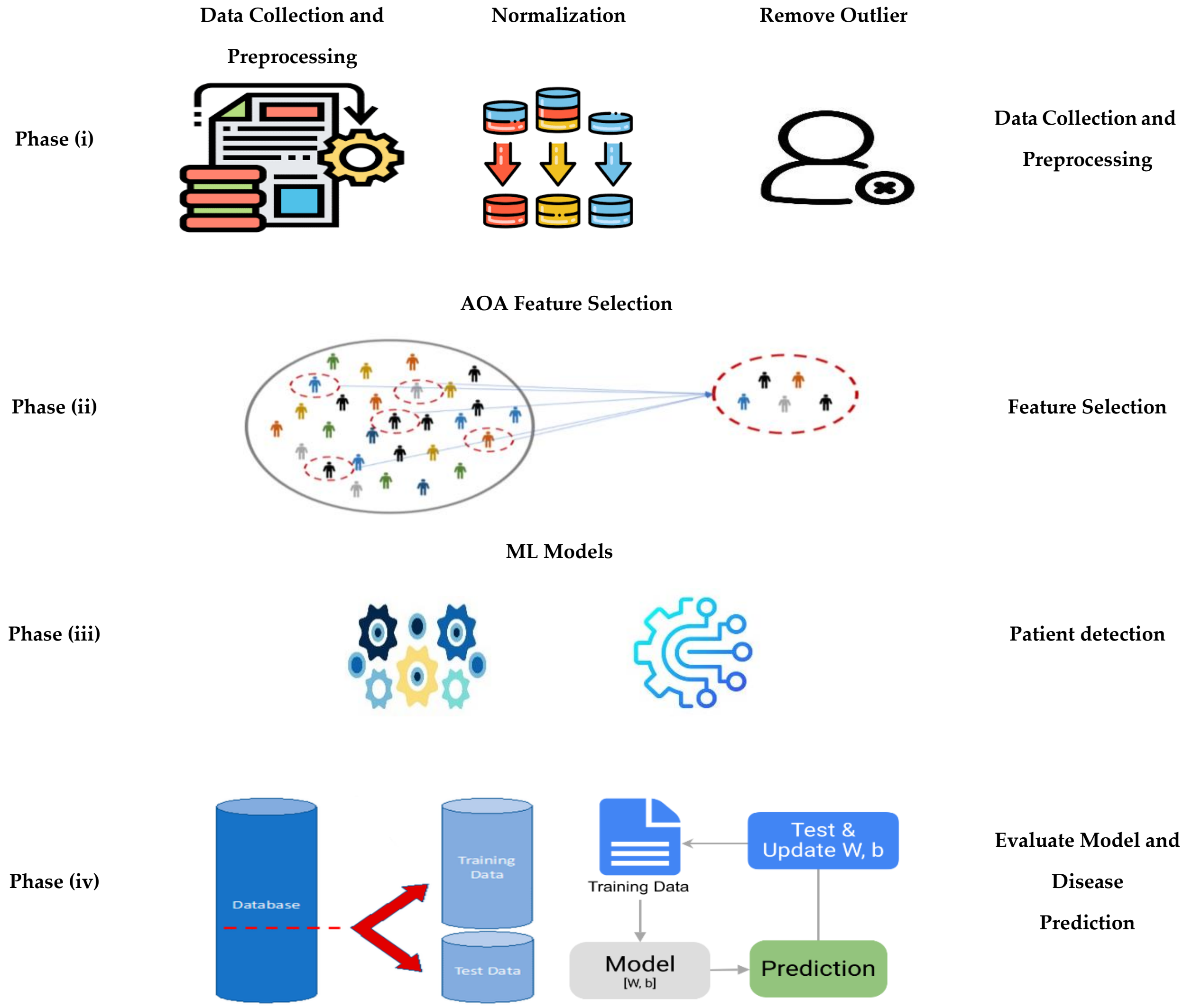


Figure 2. workflow of the proposed system.

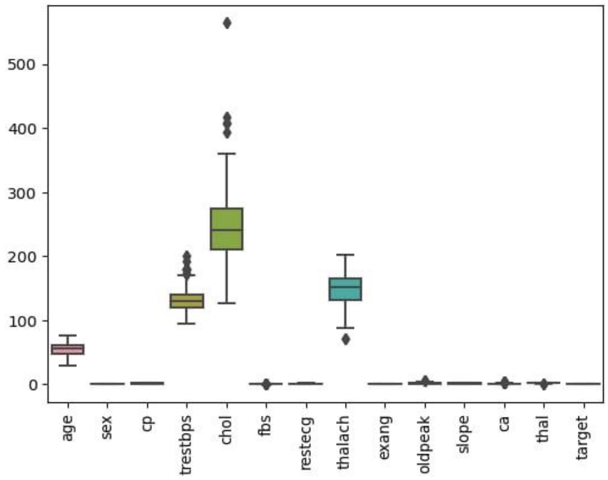


Figure 3. *Boxplot of the used dataset.*

4.4. Feature Selection

Feature selection methods are valuable auxiliary strategies in ML procedures to generate a strong and reliable model. Feature selection methods assess the predictive efficacy of each feature in relation to the target variable and suggest new subsets of features. The aims of feature selection techniques are to refine the model, remove noise, avert overfitting, enhance the efficiency of ML algorithms, and elevate performance results [29].

Feature selection techniques can be classified into three categories: filter-based, wrapper-based, and embedding methods. This classification is predicated on the integration of the selection method within the model-building process [29,30]. Filter-based approaches choose variables independently of the model, resulting in the inclusion of variables that lack correlation with the classifier. Filter-based techniques utilize correlations to effectively detect redundant features. These strategies are effective in minimizing computing time and mitigating the issue of overfitting. The wrapper-based method facilitates the detection of variable interactions, although it leads to increased processing duration and a greater risk of overfitting. The embedded method integrates the benefits of both techniques by concurrently performing feature selection and classification [30].

Consequently, a key component of ML-based prediction models is feature selection, which identifies the collection of variables or features that are most relevant and affect forecast accuracy. Among the several feature selection techniques, AOA has gained popularity, considering that it can manage highly dimensional datasets and identify statistically significant features [13]. Fig. 1 shows AOA algorithm flowchart. In this study, AOA feature selection selects the top 6 features, which include age, serum cholesterol in mg/dl (CHOL), exercise electrocardiography (THALACH), exercise angina (EXANG), OLDPEAK (means ST depression) and SLOPE (ST segment's peak slope during exercise).

4.5. Patient Detetction

The next part of the proposed HD^{2}M is patient detection. Hence, in this study, multiple classifiers which are, KNN, SVM, RF and EGB are used and the final decision is based on the majority voting to detect the infected patients.

4.5.1. K-Nearest Neighbors (KNN)

Among the most strongly popular ML techniques for classification, regression, and pattern recognition is KNN, a straightforward model. Classification issues are addressed by this approach. To locate nearby data points, KNN calculates their distance using the standard formula for geometric distance. This method tackles problems with regression and classification. If you want to find the new case in the same category as all the others, you can do it by finding all the similar existing feature cases and surrounding them with the value of k, where k is a constant that you define. Careful consideration of the K value is required to avoid overfitting the system if K is too small. Low performance with large training datasets is one of the downsides [25].

4.5.2. Support Vector Machine (SVM)

SVM have shown effectiveness in an extensive range of categorization tasks. The objective is to determine the most favorable hyperplane among the classes by tallying the number of points on the boundary of the class descriptors. The margin refers to the distance or gap that exists between several classifications or categories. Enhanced precision in classification can be attained by widening the margin. The data points in close proximity to the border are considered as the support vectors. Classification and regression issues are both addressed by SVM. This approach works well for issues involving nonlinear datasets as well as linear ones. The SVM algorithm uses a number of different kernel types for prediction models, such as sigmoid, linear radial basis function (RBF), and polynomial. When categorizing data points grouped into two separate sets, SVM finds the optimal hyperplane by operating on a high-dimensional space for features. For both smaller and larger datasets that are insurmountable, it is effective [25].

4.5.3. Random Forest (RF)

An ensemble model known as the Random Forest (RF) can also be applied as a kind of nearest neighbor predictor. The fundamental tenet of ensemble techniques is that several models combined will result in a strong model. RF, as a composite term, comes with a built-in ML decision tree. This algorithm starts by taking an input at the top and then divides the data into progressively smaller sets as it traverses the tree. In order to build upon this idea, the random forest combines trees with an ensemble concept. The ability to deal with missing data and balanced, concise running times are two benefits of random forest classifiers [26]. Every generated subtree in the random forest receives the new dataset, often known as the testing data. The dataset's class can be determined by any decision subtree in the forest; the model will select the best class by majority vote. Figure 4 shows RF technique [27].

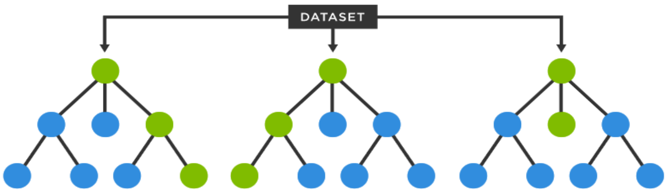


Figure 4. RF technique.

4.5.4. Extreme Gradient Boost (EGB)

The fundamental tenet of optimizing algorithms is that weak classifiers can be combined to create a stronger classifier that is more accurate than the base classifiers from which it was originally constructed. The term for this method of combining strategies is the ensemble approach. The accuracy of ensemble approaches, such as EGB and gradient boosting, is higher than that of non-ensemble methods. Strong classifiers are the outcome of ensemble techniques that combine the benefits of stacking, boosting, and bootstrapping [28].

4.6. Evaluate model and disease prediction

The final stage is evaluating the model and predicting the disease. In the event that the patient has cardiac disease, the results of the prediction show it (0 for "no disease" and 1 for "disease"). While the F1-score, recall, accuracy, and precision will form the basis of the model's evaluation. To apply the proposed model, the dataset was split into 80% to be used as training and 20% for testing. Figure 5 shows the proposed system steps to evaluate the model.

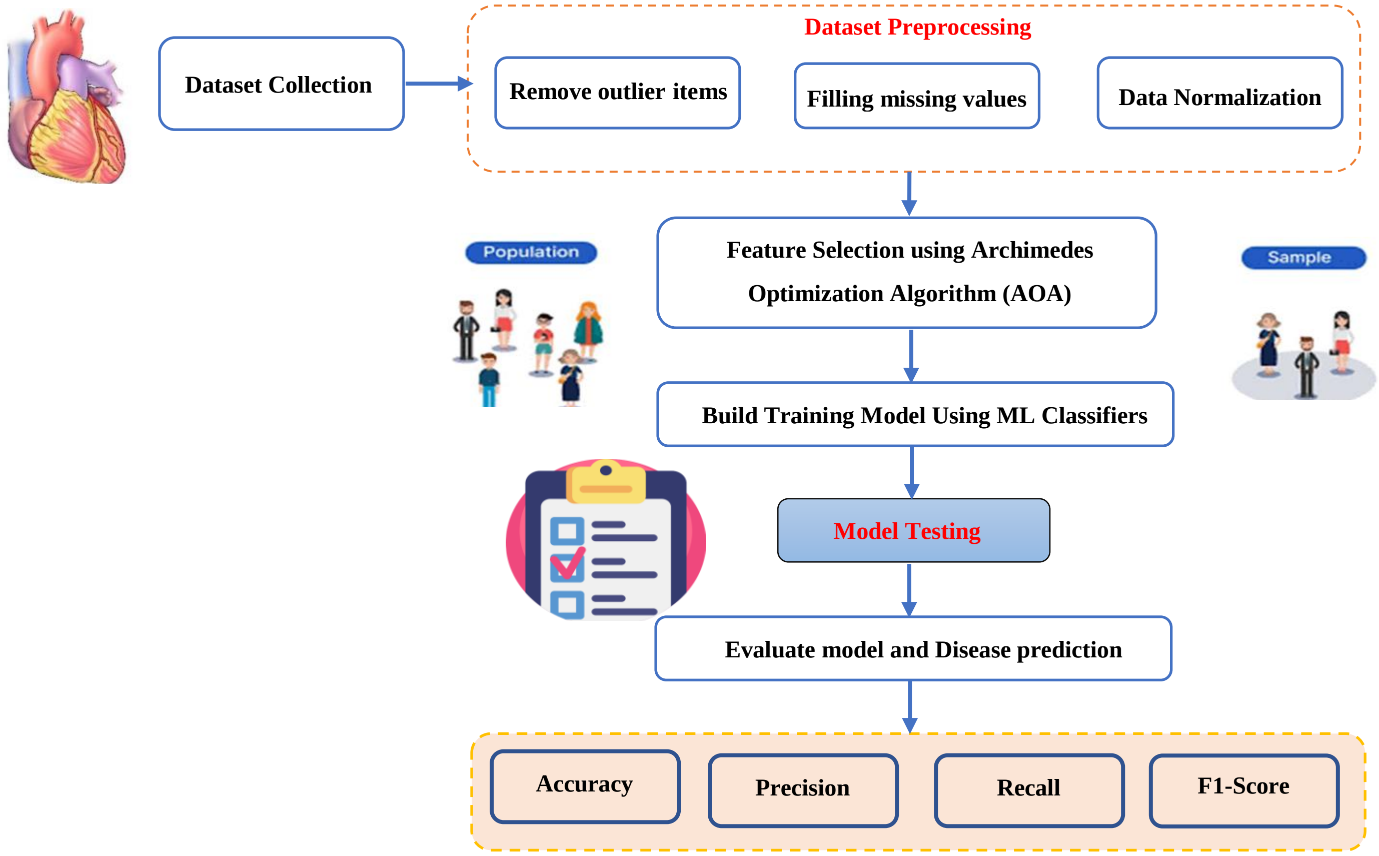


Figure 5. proposed system steps.

5. Experimental results

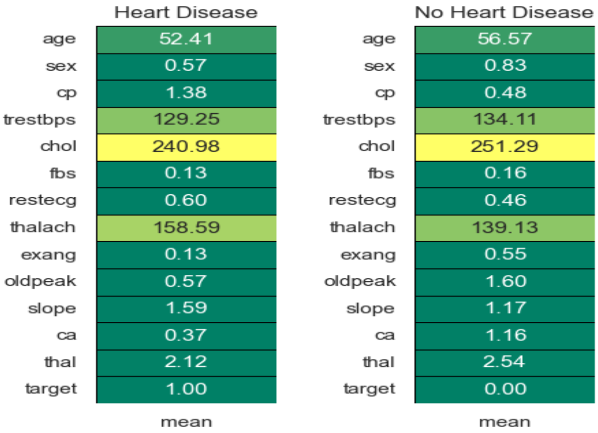
Via this section of the proposed HD2M will be evaluated.

5.1. Dataset Description

Dataset represents a collection of Heart Disease Dataset from Cleveland Clinic gathered through Kaggle [22], the number of collected case is 1025 and there are 14 attributes in the dataset. Dataset include clinical variables such as ages, sex (1 for "male", 0 for "female"), and some heart diseases such as cp stands for chest pain, resting blood pressure (trestbps) and regular test data such as serum cholesterol (chol) in mg/dl, fasting blood sugar (fbs), resting electrocardiographic diagram (restecg), exercise electrocardiography (thalach), exercise angina (exang), oldpeak (means st depression) and slope (st segment's peak slope during exercise). there are two possible outcomes referred to target, the target results indicate if the patient has cardiac disease whether or not (0 for "no disease" while 1 for "disease"). Table 1 shows dataset distribution. Figure 6 shows mean value of heart disease and non-disease and Figure 7 - 10 shows dataset description.

Table 1. Dataset distribution.

Criteria	Dataset overview									
Total number of patients	Male					Female				
	713					312				
Sick cases	Disease					Non-Disease				
	526					499				
Heart Disease patients	29-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-77
	10	32	95	62	119	87	54	47	17	3



	Heart Disease		No Heart Disease
age	52.41	age	56.57
sex	0.57	sex	0.83
cp	1.38	cp	0.48
trestbps	129.25	trestbps	134.11
chol	240.98	chol	251.29
fbs	0.13	fbs	0.16
restecg	0.60	restecg	0.46
thalach	158.59	thalach	139.13
exang	0.13	exang	0.55
oldpeak	0.57	oldpeak	1.60
slope	1.59	slope	1.17
ca	0.37	ca	1.16
thal	2.12	thal	2.54
target	1.00	target	0.00
	mean		mean

Figure 6. Mean Value of Heart Disease and Non-Disease.

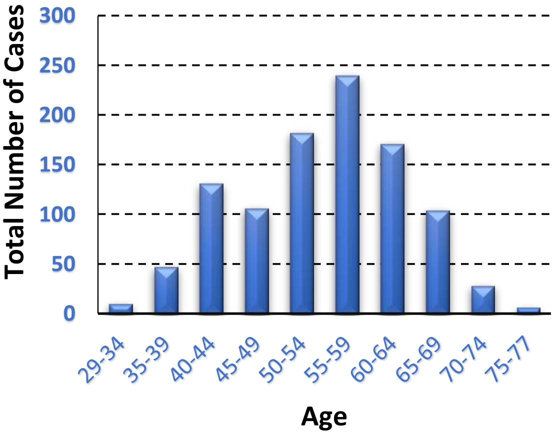


Figure 7. Total Number of Patients According to Age.

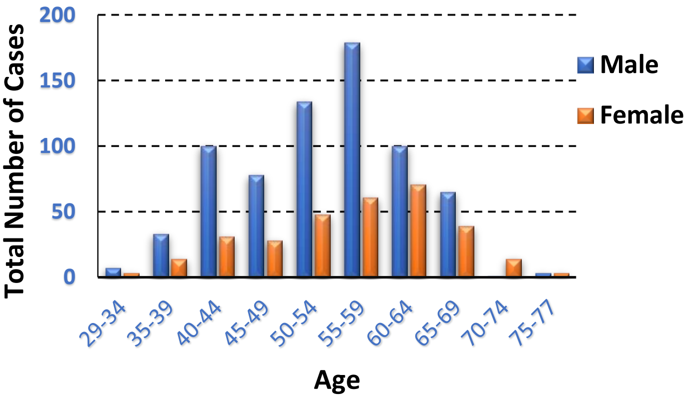


Figure 8. Total Number of Patients According to Age and Sex.

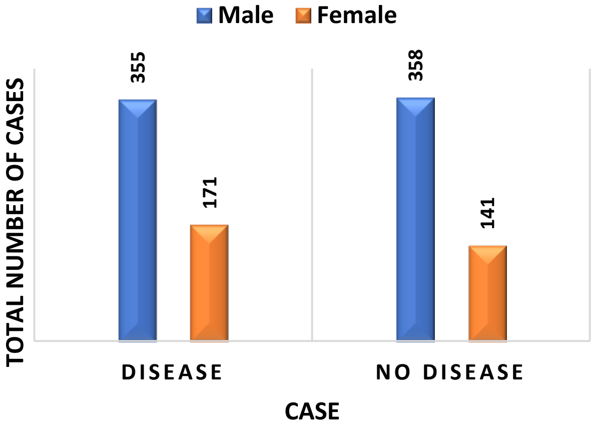


Figure 9. Disease and Non-Disease Distribution.

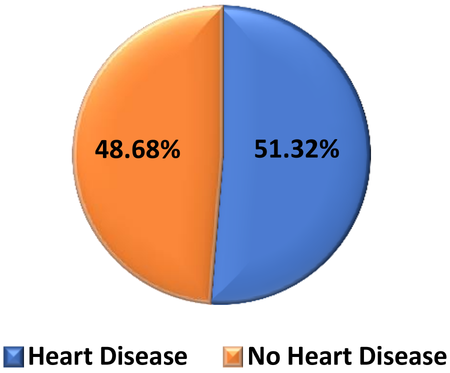


Figure 10. Heart Disease Percentage.

5.2. Evaluation Metrics

Assessing the level performance of the prediction model on testing data, which contains unfamiliar variables, is crucial once it has been trained on the training dataset. Evaluation metrics refer to performance metrics that help assess the suitability of a model for real-world applications. Confusion matrices, often referred to as matrices of confusion, offer a comprehensive perspective on how well a categorization model performs. They display the total count of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as shown in Table 2. A confusion matrix is employed to compute performance metrics such as F1 measure, recall/sensitivity, accuracy, and precision [29-31]. The assessing metrics used include recall, accuracy, precision, and F1-score [28]. The metrics are computed using the subsequent formulas:

- Accuracy of the model's predictions is measured using Equation (20).

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{20}$$

- Precision is a metric of the accuracy of positive predictions, calculated by splitting the number of correctly predicted positive instances by the total number of projected positive instances, as shown in Eq. (21).

$$precision = \frac{TP}{TP+FP} \tag{21}$$

- Recall is the percentage of true positive occurrences out of all positive predictions which can be calculated using Eq. (22).

$$recall = \frac{TP}{TP+FN} \tag{22}$$

- F1 measure is a reproduction statistic that strikes a balance between recall and precision; it provides an accurate assessing of the model's correctness by summing recall and precision into a single number as presented in Eq. (23).

$$F1\ score = \frac{2*(precision*recall)}{(precision+recall)} \tag{23}$$

Figure 11 shows accuracy of ML classifiers, Figure 12 shows the target correlation, Figure 13 – 16 show confusion matrix of each algorithm, and Figure 17 shows the heatmap correlation of all features. Table 3 shows the performance of ML algorithms.

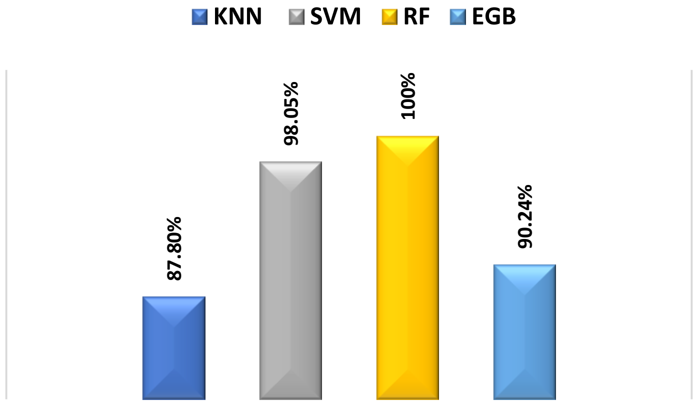


Figure 11. ML classifiers accuracy.

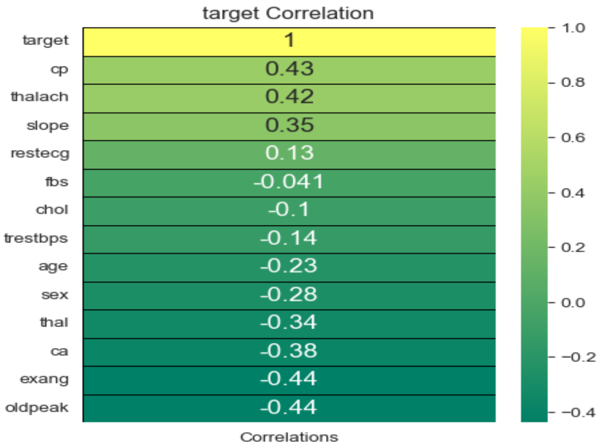


Figure 12. Target Correlation.

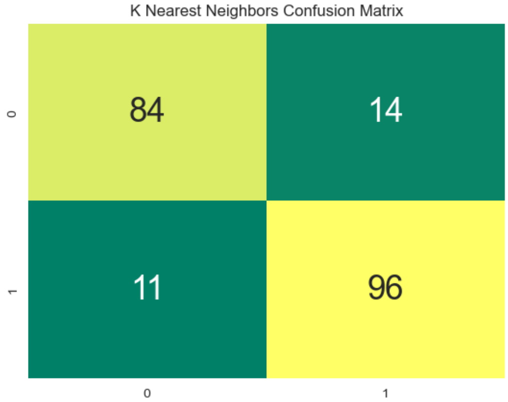


Figure 13. KNN confusion matrix.

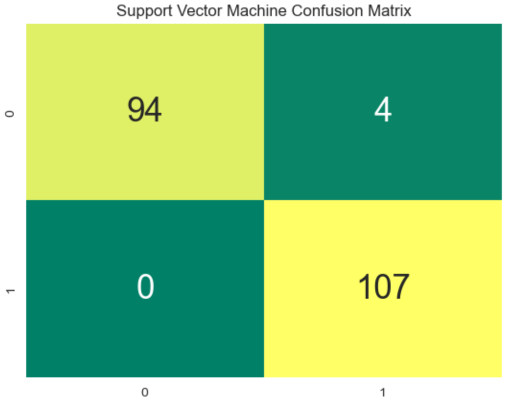


Figure 14. SVM confusion matrix.

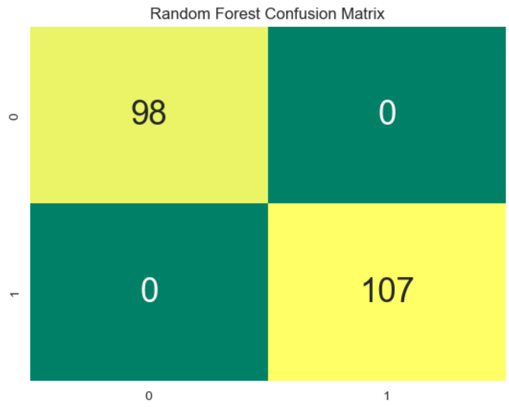


Figure 15. RF confusion matrix.

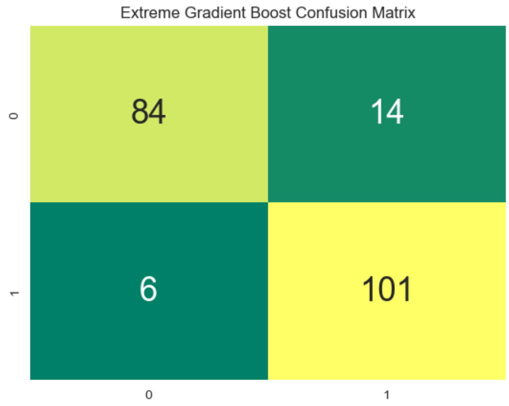


Figure 16. EGB confusion matrix.

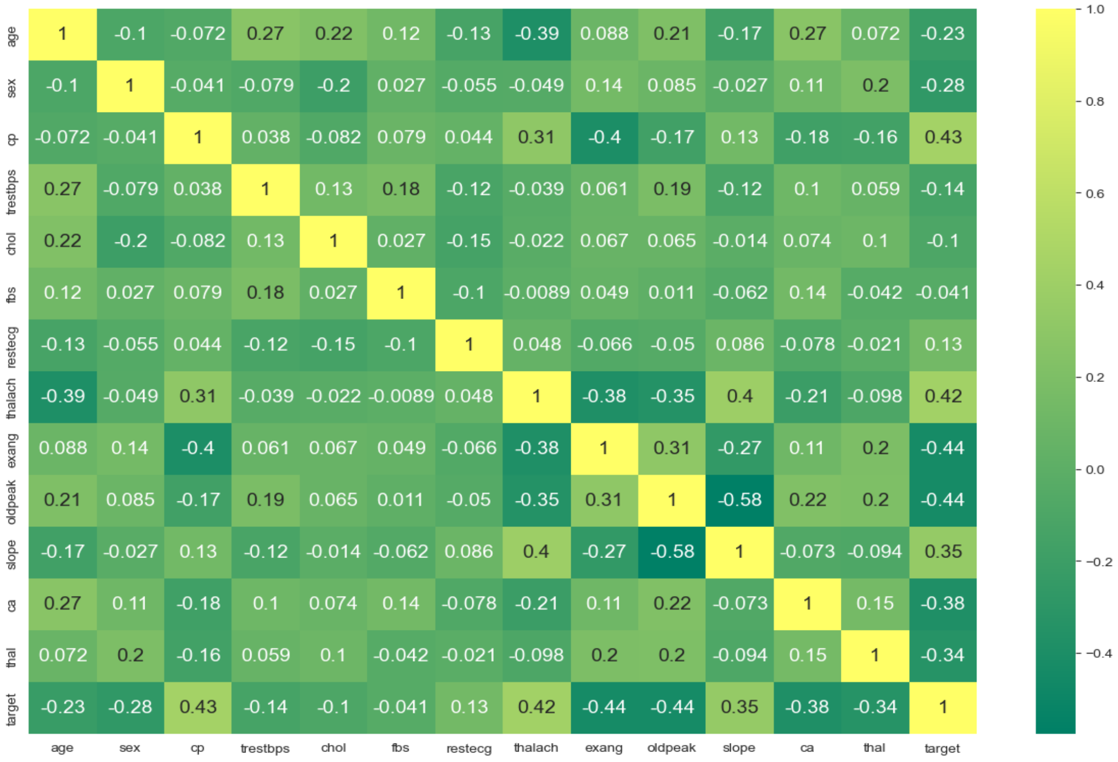


Figure 17. Heatmap Correlation of all features.

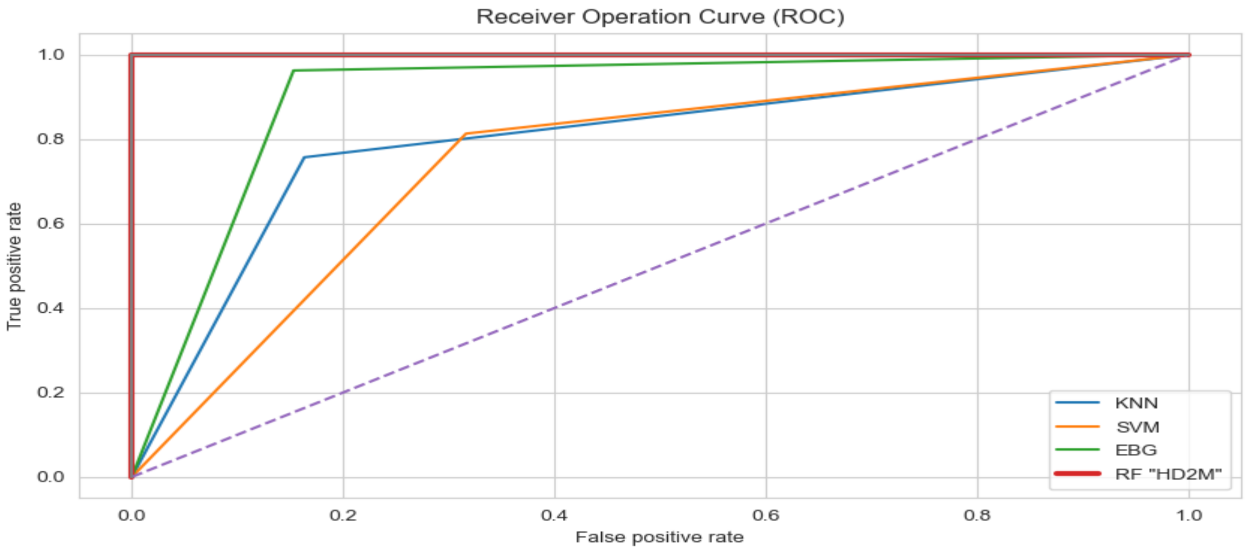


Figure 18. Receiver Operation Curve (ROC).

Table 2. Confusion matrix.

		Actual Values	
		1	0
Predicted	1	TP	FP
Values	0	FN	TN

Table 3. performance of ML Algorithms.

ML Classifiers	Accuracy %	precision	recall	F1 score
KNN	87.8 %	87%	90%	88%
SVM	98.048 %	96%	95.5%	95.7%
EGB	90.24 %	88%	94%	91%
HD2M	100 %	98%	96%	97%

5.3. Testing whole system

In this section, the proposed system will be evaluated, keeping all phases working together. As shown in Table 4 the proposed HD2M performs well as it provides 100% accuracy compared to KNN with an accuracy 87.8%, SVM with an accuracy 98.048%, and EGB with an accuracy 90.24%. Figure 18 also displays the Receiver Operation Curve (ROC) of used ML classifiers. Additionally, HD^{2}M provides 98% precision, 96% recall, and 97% f1 score. The reason is that HD2M depended on the most important features selected by AOA.

5.4. Comparison with that state of the art

In this section, we have conducted a comparison between our suggested approach and other recently employed classification techniques in order to showcase its effectiveness. The results are displayed in Table 4. Table 4 demonstrates that the proposed HD2M is highly proficient in identifying individuals with heart disease. According to the data presented in Table 4, the proposed HD2M performs better than other competitors when evaluated using various metrics.

Table 4. Comparison of related works.

Reference	ML Classifier	Accuracy
[13]	LR	88.52 %
	SVM	83.61 %
	DT	80.33 %
[14]	MLP with backpropagation technique	80.99 %
[16]	MLP with backpropagation and SVM	98 %
[17]	DL, RF, DT, LR, SVM, Gradient boosting	84 %
[18]	ANN	80 %
[19]	CNN	85.70 %
[20]	Stacked model involving KNN, RF and SVM	75.1 %
[21]	RF algorithm	83 %
HD^{2}M	Various ML	100 %

6. Conclusions

Heart disease identification is difficult in clinical settings, and the death rate is high as a result, according to WHO statistics. Therefore, a new ML-based model is suggested in this research called Heart Disease Detection Model (HD^{2}M). The proposed model consists of four main parts. These parts are, (i) data collection and preprocessing, which includes removing outlier items, filling or inputting missing values and converting non-numerical values to numerical values, (ii) feature selection, which use a new methodology named Archimedes Optimization Algorithm (AOA), (iii) patient detection using multiple classifiers which are, KNN, SVM, RF and EGB. (iv) Evaluate model and disease prediction. At first, the preprocessed dataset is fed to the feature selection part to select the most effective features using AOA. Then, these features are fed to the patient detection part. Experimental results proved that the proposed HD2M is effective at identifying people with heart disease or not, according to the previous results. It provides 100% accuracy, 98% precision, 96% recall, and 97% f1-score.

References

1. SECKELER, M. D. & HOKE, T. R. 2011. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. *Clin Epidemiol,* 3**,** 67-84.
2. WHO. *World Health Organization,* *https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1* [Online]. [Accessed].
3. GAZIANO, T. A., BITTON, A., ANAND, S., ABRAHAMS-GESSEL, S. & MURPHY, A. 2010. Growing epidemic of coronary heart disease in low- and middle-income countries. *Curr Probl Cardiol,* 35**,** 72-115.
4. WENG, S. F., REPS, J., KAI, J., GARIBALDI, J. M. & QURESHI, N. 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One,* 12**,** e0174944.
5. RANI, P. 2021. A decision support system for heart disease prediction based upon machine. *Journal of Reliable Intelligent Environments,* 7**,** 263-275.
6. REDDY, N. S. C. 2019. Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction. *International Journal of Innovative Computing,* 9.
7. HASHIM, F. A. 2021. Archimedes optimization algorithm: a new metaheuristic algorithm for solving optimization problems. *Applied intelligence,* 51**,** 1531-1551.
8. SHAH, D., SAMIR PATEL, AND SANTOSH KUMAR BHARTI. 2020. Heart disease prediction using machine learning techniques. *SN Computer Science,* 1**,** 345.
9. PATEL, J., DR TEJALUPADHYAY, AND SAMIR PATEL. 2015. Heart disease prediction using machine learning and data mining technique. 7**,** 129-137.
10. PHAM, V., HEMPTINNE, Q., GRINDA, J. M., DUBOC, D., VARENNE, O. & PICARD, F. 2020. Giant coronary aneurysms, from diagnosis to treatment: A literature review. *Arch Cardiovasc Dis,* 113**,** 59-69.
11. BAVISKAR, V., VERMA, M., & CHATTERJEE, P. 2020. A model for heart disease prediction using feature selection with deep learning. *International Advanced Computing Conference***,** 151-168.
12. RAZA, K. 2019. Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. *U-Healthcare Monitoring Systems***,** 179-196.
13. MANIKANDAN, G., PRAGADEESH, B., MANOJKUMAR, V., KARTHIKEYAN, A. L., MANIKANDAN, R., & GANDOMI, A. H. 2024. Classification models combined with Boruta feature selection for heart disease prediction. *Informatics in Medicine Unlocked,* 44**,** 101442.
14. BHAT, R., CHAWANDE, S., & CHADDA, S. 2019. Prediction of test for heart disease diagnosis using artificial neural network. *Indian Journal of Applied Research,* 9.
15. ABUSHARIAH, M. A., ALQUDAH, A. A., ADWAN, O. Y., & YOUSEF, R. M. 2014. Automatic heart disease diagnosis system based on artificial neural network (ANN) and adaptive neuro-fuzzy inference systems (ANFIS) approaches. *Journal of software engineering and applications* 7**,** 1055-1064.
16. HASAN, T. T., JASIM, M. H., & HASHIM, I. A. 2017. Heart disease diagnosis system based on multi-layer perceptron neural network and support vector machine. *International Journal of Current Engineering and Technology,* 77**,** 2277-4106.
17. SAPRA, L., SANDHU, J. K., & GOYAL, N. 2021. Intelligent method for detection of coronary artery disease with ensemble approach. *Advances in Communication and Computational Technology: Select Proceedings of ICACCT 2019***,** 1033-1042.
18. CHEN, A. H., HUANG, S. Y., HONG, P. S., CHENG, C. H., & LIN, E. J. 2011. HDPS: Heart disease prediction system. *computing in Cardiology***,** 557-560.
19. DUTTA, A., BATABYAL, T., BASU, M., & ACTON, S. T. 2020. An efficient convolutional neural network for coronary heart disease prediction. *Expert Systems with Applications,* 159**,** 113408.
20. SHOREWALA, V. 2021. Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked,* 26**,** 100655.
21. CHANG, V., BHAVANI, V. R., XU, A. Q., & HOSSAIN, M. A. 2022. An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics,* 2**,** 100016.
22. KAGGLE. *https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset* [Online]. [Accessed].

23. SHABAN, W. M., ASHRAF, E., & SLAMA, A. E. 2024. SMP-DL: a novel stock market prediction approach based on deep learning for effective trend forecasting. *Neural Computing and Applications,* 36**,** 1849-1873.
24. YANG, J., RAHARDJA, S., & FRÄNTI, P. 2019. Outlier detection: how to threshold outlier scores?. *information processing and cloud computing***,** 1-6.
25. IBRAHIM, I., & ABDULAZEEZ, A. 2021. The role of machine learning algorithms for diagnosing diseases. . *Journal of Applied Science and Technology Trends,* 2**,** 10-1.
26. LAKSHMI, S. V., MEENA, M. K., & KIRUTHIKA, N. S. 2019. Diagnosis of chronic kidney disease using random forest algorithms. *International Journal of Research in Engineering, Science and Management,* 2**,** 559-562.
27. KIRASICH, K., SMITH, T., & SADLER, B. 2018. Random forest vs logistic regression: binary classification for heterogeneous datasets. . *SMU Data Science Review,* 1**,** 9.
28. JINNY, S. V., & MATE, Y. V. 2021. Early prediction model for coronary heart disease using genetic algorithms, hyper-parameter optimization and machine learning techniques. *Health and Technology,* 11**,** 63-73.
29. SHABAN, W. M. 2024. Early diagnosis of liver disease using improved binary butterfly optimization and machine learning algorithms. *Multimedia Tools and Applications,* 83**,** 30867-30895.
30. SHABAN, W. M. 2024. Detection and classification of photovoltaic module defects based on artificial intelligence. *Neural Computing and Applications***,** 1-28.
31. SHABAN, W. M. 2023. Insight into breast cancer detection: new hybrid feature selection method. *Neural Computing and Applications,* 35**,** 6831-685.