Journal of Rare Cardiovascular Diseases
ISSN: 2299-3711 (Print) | e-ISSN: 2300-5505 (Online)
www.jrcd.eu
RESEARCH ARTICLE

An Ensemble Deep Learning and Intelligent Feature Selection Model for the Prediction of Heart Disease Data

N. Vellingiri[1*] and S. Ranjitha Kumari[2]
[1*]PhD Research Scholar, Department of computer science, RVS College of Arts and Science, Sulur, Coimbatore.
[2]Associate Professor, Department of computer science, RVS College of Arts and Science, Sulur, Coimbatore.

*Corresponding Author
N. Vellingiri
(vellingiri.nm@gmail.com)

Article History
Received: 10/07/2025
Revised: 14/08/2025
Accepted: 15/09/2025
Published: 30/09/2025

Abstract: Cardiovascular disease (CVD) and heart diseases (HD) constitute one of the most prevalent and fatal illnesses impacting people's health. HD has increased as a result of the development of bad habits including smoking, overeating, and inactivity. Although angiography is often acknowledged as the most effective method for identifying coronary artery disease, it is costly and has a number of serious side effects. The ability to identify HD early on by recognizing its early signs is one of the most crucial things in the current world. Heart disease's fatal characteristic has given it the moniker "silent killer," as there are seldom any evident warning signs. In order to develop an effective strategy for recognizing the early detection (ED) of HD using fundamental knowledge and symptoms, further research is thus required. To overcome these challenges, our study effort produced an ensemble (DL) deep learning (EDL) approach and efficient feature extraction (FE) method for the categorization of HD. The four primary stages of this study are FE, feature section (FS), preprocessing, and classification. First, pre-processing is done using min-max normalisation (MMN). To increase classifier (ACC) accuracy, clustering is done using the K-Means Clustering (KMC) technique. Principal Component Analysis (PCA) is recommended in this article as an effective FE method. Following that, the modified whale optimisation algorithm (MWOA) approach is used to complete the FS process. To choose the most significant and pertinent attributes from the CVD dataset, it computes the optimal fitness value (FV).The Elman NN (ENN), Generative Adversarial Neural Network (GAN) and a Recurrent NN (RNN) are used in a EDL model to identify HD data. According to the study results, machine learning (ML) techniques are not as effective in predicting HD data as the EDL model that is suggested.

Keywords: Heart disease data, min-max normalization, k-means clustering, Principal Component Analysis (PCA), Generative Adversarial Neural Network (GAN), modified whale optimization algorithm, Elman Neural Network (ENN), and a Recurrent NN (RNN).

INTRODUCTION

The World Health Organisation (WHO) estimates that HD kills 12 million people worldwide each year. CVD is the cause of half of deaths in the US and other wealthy nations [1]. In many developing nations, it is also the major reason for mortality. Overall, it is thought to be the main cause of adult mortality. The phrase "HD" refers to a variety of heart-related conditions. In many nations, including India, HD was the leading cause of mortality. Every 34 seconds, HD kills one person.

There are several types of HD, including coronary HD, cardiomyopathy, and CVD. A vast array of disorders affecting the heart, blood vessels (BV), and circulation of blood throughout the body are included under the general term "CVD." Numerous illnesses, disabilities, and fatalities are caused by CVD [2]. One of the most important and complex tasks in medicine is disease diagnosis. Medical diagnosis is seen as a significant but challenging duty that requires precise and effective execution. It would be really beneficial to automate this system. However, not all doctors are specialists in every subspeciality, and resource personnel are scarce in some areas. Therefore, it would likely be quite advantageous to incorporate all of these factors into a computer-based clinical diagnosis system [3]. Effective computer-based information and/or decision support systems (DSS) can help reduce the cost of clinical testing. A comparative analysis of the different approaches available is necessary for the precise and efficient implementation of automated systems.

The procedure of extracting relevant data from massive datasets through the application of methods like classification, clustering, association, etc., in order to predict or describe the data is known as data mining (DM). DM has been widely used in the healthcare sector, including in identifying effective patient care cost structures, forecasting illness risk factors, and categorising the best treatment approaches [4]. Despite being preventable, the high mortality rates associated with these diseases must be reduced by evaluating a patient's risk and determining their early prognosis. Heart failure (HF), hypertensive HD, coronary HD, cardiomyopathy, and other conditions are examples of common CVD [5].

HD is frequently caused by a high-cholesterol diet, diabetes (DB), smoking, inactivity, high blood pressure (BP), and other conditions. With high levels of ACC, research on CVD using DM has been a continuous endeavour that includes risk score analysis, treatment, and prediction.

The prediction of HD is significantly influenced by computational intelligence (CI). Relationships between patient characteristics and various diseases can be found using concepts utilised in therapy. Many researchers used the FS approach to predict HD in their findings in recent studies [6]. Variable selection or (AS) attributes selection are other names for FS. (DR) dimensionality reduction is not the same as FS. While DR reduces the attribute set by creating new attributes from a given attribute set, FS concentrates on minimising the amount of irrelevant attributes through various strategies, such as attribute subset selection. FS algorithms [7] and techniques can be used with a variety of tools. This Rapid Miner tool is readily accessible via the internet and is open source.

Diagnostic models based on ML that can forecast a patient's likelihood of having a disease based on a collection of risk factors are among the techniques that are frequently incorporated into clinical DSS (CDSS). A cardiovascular HD prediction model is constructed in this facility [8]. Even though ML models have been studied in depth and shown to be highly successful, HD prediction remains a difficult problem with many opportunities for development and investigation. This kind of problem is categorised as a supervised learning (SL) task in ML. To determine the relationship between a set of features and the target class, researchers employ classification algorithms [9]. In this instance, the patient's class is HD, and risk factors for HD include age, cholesterol, and the outcomes of other clinical trials.

Numerous techniques, including the K-Nearest Neighbour Algorithm (KNN), Genetic Algorithm (GA), Decision Trees (DT), and Naive Bayes (NB), are used to determine the severity of diseases [10]. Because HD is so complicated, care must be taken with caution. If this isn't done, it might affect the heart and cause an early death. To identify various metabolic illness kinds, DM and medical research perspectives are employed. Classification-based DM is important for both data exploration and HD prediction. DT have also been applied to the accurate prediction of HD-related events [11]. Various approaches have been used to apply known techniques for DM to extract knowledge for the Predicting HD. Still, ACC of the classifiers declines with increasing dataset complexity. Thus, this research effort generated a EDL algorithm and an effective FS for the categorization of heart condition.

The remaining study is organised as follows: Some of the recent techniques for HD data detection are reviewed in Section 2. The process of the recommended practice is shown in section 3. The findings and discussion are presented in section 4. The conclusion and upcoming tasks are covered in section 5.

LITERATURE REVIEW

Some of the recent technique for the prediction of HD data by ML and DL methods are reviewed in this section.

A DR method and the application of an FS technique to identify HD features have been presented by Gárate-Escamila et al. [12]. The HD UCI ML Repository provided the data used in this investigation. When six alternative ML classifiers were used to validate a 74-feature dataset for HD prediction, the greatest ACC rates (98.7% for Cleveland, 99.0% for Hungarian, and 99.4% for Cleveland-Hungarian (CH) datasets) were obtained by combining Chi-square (CHI), PCA, and Random Forests (RF). The study, conducted by ChiEf Selector, extracted features of anatomical and physiological significance related to heart health including cholesterol, heart rate, chest pain, ST depression features, and heart vessels. The outcomes of the simulation showed that integration of the CHI and PCA improves the performance of most classifiers. Lower outcomes were obtained when PCA was used straight from the raw data; higher dimensionality would be needed to enhance the findings.

To effectively forecast HD, two DM techniques: NB & DT were created by Priyanka et al. [13]. To determine which of the two approaches is superior, it analyses their ACC and efficiency. Patients and medical professionals can both greatly benefit from the use of appropriate technology support in this field.

An innovative approach that utilizes ML approaches to identify important traits was put out by Mohan et al. [14]. The ACC in the prediction of CVD may improve as a result. The prediction model (PM) is introduced using popular classification techniques and a range of feature combinations. With an ACC level of 88.7%, the hybrid RF with a linear model (HRFLM) PM for HD also produces a higher performance level.

FS approaches and algorithms were introduced by Bashir et al. [15]. To analyse experiments and show how ACC has improved, author use a number of HD datasets. DT, Logistic Regression (LR), LR SVM, NB, and RF algorithms are employed as FS approaches with Rapid Miner as a tool, and the results are improved by displaying the ACC.

Based on the pertinent features, the ML model for HD prediction was suggested by Sharma et al. [16]. 14 different HD-related parameters are included in the benchmark dataset of UCI HD prediction, and it is used in this investigation. The model has been developed using ML algorithms as RF, Support Vector Machine (SVM), NB, and DT. With the aid of common ML techniques, researchers in the study attempted to identify connections among the various features present in the dataset. The prediction of HD possibilities was then successfully done by this application. The outcome demonstrates that RF provides more ACC for the prediction in less time than other ML approaches. As DSS, medical professionals at their clinic may find this model useful.

To predict risk factors related to HD, Repaka et al. [17] introduced a Smart HD Prediction (HDP) (SHDP) via NB. A standardised format is used to compile the necessary data. The medical profiles are used to predict a patient's risk of HD by retrieving information on blood sugar, sex, age, BP, cholesterol, and other factors.In order to forecast HD, the NB classification uses the gathered features as input. The dataset is divided into two parts: 20% is used for testing, while the remaining 80% is used for training. Dataset collection, application-based user registration and login, classification through NB, prediction, and secure data transfer (SDT) using AES (Advanced Encryption Standard) are the steps in the suggested methodology. A result is then generated. The study uses DM approaches that are used for HDP to elaborate and show various knowledge abstraction methodologies. According to the results, the established diagnostic system helps forecast risk factors for HD in an effective manner.

In order to choose features for disease classification, Abdollahi et al. [18] suggested an ensemble-genetic learning technique utilising wrapper feature reduction (WFR). Results: More accurate diagnosis and lower cost of diagnosis was attained by the development of a medical diagnosis system based on ensemble learning (EL) in HD prediction, when compared to the traditional method. Conclusion: Vascular occlusion and Thallium Scan results were the most important characteristics in a study that used an EL approach to diagnose heart disease (HD). It attains 97.57% ACC rate in differentiating between normal and abnormal patients.

The density-based spatial clustering of applications with noise (DBSCAN) procedure was created by Xie et al. [19]. To effectively predict CVD, the DBSCAN approach is a weight learning methodology that uses dataset density data. The RF algorithm is used in the method to choose significant features. It separates the sample points into three categories and uses weight learning based on density to weight them using various values. Therefore, combining the original features and weight feature may help the created models in learning density data, recognising decision limits more successfully, and perform better. ML models with weight learning with cross-validation showed an ACC of 3 % using the Stroke dataset and >10 % using the UCI dataset, in comparison to conventional ML models.

The FS approach was developed by Reddy et al. [20]. By choosing the key features, these FS methods could be useful strategies to lower the cost of diagnosis. The goal of the project is to predict the classification model and determine which elements are most crucial for HD prediction using the Cleveland and Statlog Project Heart datasets. The ACC of the RF procedure has been demonstrated to be 90–95% in both the categorisation and FS models, based on three different percentage splits. Performance was best when 8 or 6 characteristics were chosen. PM performance was not enhanced by further lowering the number of chosen features.

By utilising pertinent features selected by multiple FS approaches, Spencer et al. [21] created ML strategies. Four popular HD datasets have been evaluated using PCA, ReliefF, Chi squared testing (CST), and symmetrical uncertainty for creating unique feature sets. A variety of classification techniques have been used to create models that are then assessed to determine the optimal feature combinations so as to improve the ACC of HD prediction. Researchers discovered that the benefits of adopting FS for the heart datasets under consideration differed based on the ML technique employed. With an ACC of 85.00% on the tested datasets, the satisfactory outcomes were attained by integrating the BayesNet method with Chi-squared FS.
A rule-based model was created by Mythili et al. [22]. This suggested model helps in examining the ACC of applying rules to the individual SVM, DT, and LR outcomes on the Cleveland HD Database. Thus, an accurate model of HD prediction was obtained.

Proposed Methodology

This research work introduces an EDL approach and efficient FE method for the categorization of heart condition. The four primary stages of this study are FE, FS, preprocessing, and classification. Initially the MMN is used for pre-processing. The KMC technique is used for clustering with the goal of improving classifier ACC. This work suggests using a PCA to efficiently extract features. After that, the FS process is finished using the MWOA. The best relevant and meaningful features are selected from the CVD dataset by computing the optimal FV. The ENN, GAN and a RNN are used in an EDL model to identify HD data.

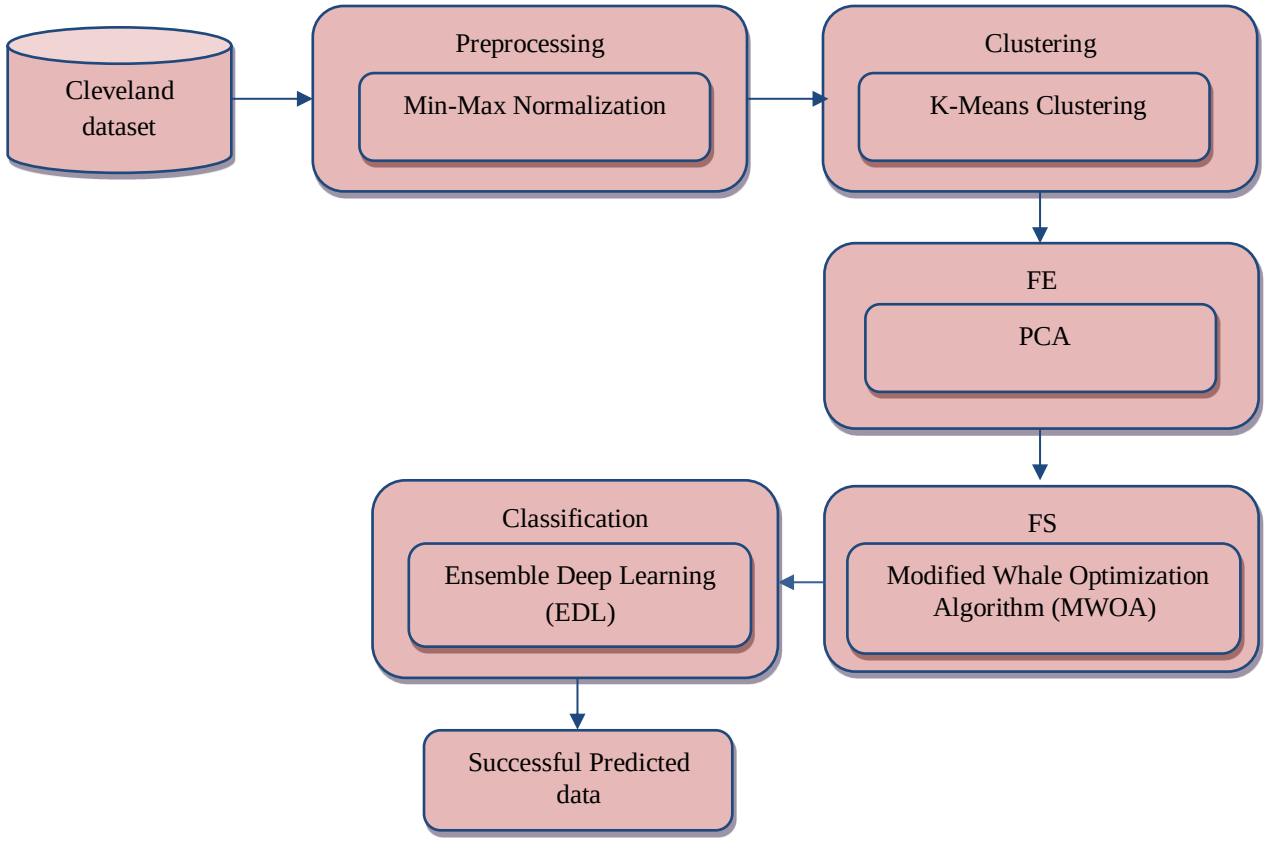


Figure 1. the flowchart of the recommended procedure

3.1. Data normalizations (DN) using MMN

Normalizing the data is necessary to prevent erroneous findings from being produced by variances in the input crime data. This study adopts the MMN paradigm, which involves employing a mathematical function to transform numerical values into a new range. The crime dataset in this proposed study is normalized using often used MMN. The dataset's values are normalized to fall between the minimum and maximum values specified, and each value is then changed using Equation (1)

$$v' = \frac{v - \min_A}{\max_A - \min_A}\left(\text{new\_max}_A - \text{new\_min}_A\right) + \text{new\_min}_A \quad (1)$$

Where,
A –Data Attributes,
Min (A), Max (A) –stands for A's min. and max. absolute values
v'- New values of inputs
v - Old values of inputs
New_ max (A), new_ min (A) imply max and min value ranges (boundaries of required ranges).

3.2. KMC

Until an optimal centroid is found, the KMC algorithm calculates centroids and repeats. The number of clusters is assumed to be known. It is also known as the flat clustering algorithm. The number of clusters that the K-means technique has found in the data is represented by the letter "K" [23]. This approach involves allocating data points (DP) to clusters in a method that minimises the totality of the squared distances among the DP and the centroid. It is important to remember that more identical DP occurs within a cluster when there is less diversity in it.

The steps listed below will assist us in comprehending the process of the KMC technique:

- *Phase 1:* It is necessary to first specify the amount of clusters, K, that this procedure must generate.
- *Phase 2:* Then, allocate each K DP to a cluster after selecting them at random. Sort the data according to the DP count.
- *Phase 3:* The next step is to calculate the cluster centroids.
- *Phase 4:* To find the optimal centroid, repeat the following steps. Optimal centroid is the process of allocating DPs to clusters that are constant.

4.1 The first calculation would be the sum of squared distances between the centroids and DP.

- 4.2 The cluster that is closest to the others (the centroid) must now be assigned to each DP.
- 4.3 At last, the DP of each cluster is averaged to determine the cluster centroids.

For the purpose of resolving this issue, the Expectation-Maximization technique is utilized by K-means. The Maximisation step (M-step) determines each cluster's centroid, and the Expectation step (E-step) assigns DP to the cluster that is closest to it. Assigning the DP to the nearest cluster is the E-step. Each cluster's centroid is calculated using the M-step. Feel free to skip the explanation regarding the way that we may resolve it mathematically.

The objective function (OF) is

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \left\| x^i - \mu_k \right\|^2 \quad (2)$$

If DP x^i is a member of cluster k, $w_{ik} = 1$; if not, $w_{ik} = 0$. Additionally, the centroid of the cluster of $x^i$is μ_k.

3.3. DR using PCA

On normalizing data their dimensionalities are reduced. PCA is being used in this effort to reduce dimensionality. A DR method called PCA creates new features by combining the original features in a linear fashion. Instances of the dataset are mapped via PCA from d-dimensional spaces to k-dimensional subspaces where k<d. The Principal Components (PC) are the collection of k newly created dimensions. Each PC aims to maximize variance while avoiding variation that has already been taken into account by all of its previous components [24]. The first component, then, covers the highest variation, while each subsequent component covers a smaller degree of variance. PC can be computed using:

$$PC_i = a_1 X_1 + a_2 X_2 + \cdots + a_d X_d \quad (3)$$

Here, X_j is the original feature "j," PC_i – PC 'i' and the numerical coefficient for X_j is denoted as a_j.

3.3. FS using MWOA

The WOA technique was inspired by humpback whale (WH) hunting strategies and the creation of an equation for those strategies. The WH hunt by circling and consuming its prey, that is primarily small fish, by utilising a bubble-net (BN) technique. WH appear and start to rise to the surface, creating a wide ring of bubbles deep beneath the fish. The bubbles serve as a trap, forcing the fish to the surface. Fish that are ascending to the surface are pursued by WH [25]. Exploitations, circling, and explorations are three steps to the hunting process in theory. The quality of the fishes eaten is determined by the explorations. Encirclement phase: WH finds the fish and surrounds it.

The starting point for the optimal location is originally arbitrary and undetermined. The ideal locations to reach targets are those that have been updated by other AG in response to the random initiations.

Individuals in animal swarms learn from both their neighbours. Personal traits can be greatly improved by this. In order to create WH's adaptable neighbourhoods and strengthen relationships between groups, the adaptive social learning (ASL) technique creates neighbourhoods by identifying social ranking, social impacts, and social networks. In order to maintain computation ACC and enhance population diversity, a unique approach based on neighbourhood updates is developed.

The location of the WH and the surrounding area can be shown using equations (4) and (5).

$$\vec{H} = |\vec{C} x \vec{X} * (t) - \vec{X}(t)| \quad (4)$$

$$\vec{X}(t+1) = \vec{X}(t) - \vec{A} \times \vec{H} \quad (5)$$

Vector coefficients derived from equations (6) and (7) are represented by $\vec{A}$and $\overrightarrow{C}$,the current iterations are denoted as t, positional vectors are provided by $\vec{X}$, and arbitrary solutions started arbitrarily are denoted by $\vec{X}*$.

$$\vec{A} = 2\vec{a} \times \vec{r} - \vec{a} \quad (6)$$

$$\vec{C} = 2x\vec{r} \quad (7)$$

A random value between 0 and 1 is denoted as $\vec{r}$, and the elements of $\vec{a}$ are decreased linearly from 2 to 0 each time during each iteration.

Humpback WH hunt by circling their prey using BN tactics. Fish are among the prey that WH encircle before changing their positions to determine the best way to proceed. The main mathematical component of the WOA is shown in equations (8) and (9).

$$X(t+1) = X^*(t) - A.|C.X^*(t) - X(t)| \; if \; p < 0.5 \quad (8)$$

$$X(t+1)|C.X^*(t) - X(t)|.e^{bl} \cos(2\pi t) + X^*(t) \; if \; p \geq 0.5 \quad (9)$$

In this case, time or iteration indices are denoted as t. Vectors of WH locations are represented by X. T he optimal solutions that were obtained is denoted as X^* . A=2a. (r-a); C=2.r Random vectors having values between 0 and 1 are denoted as r. Then, a represents coefficient vectors that reduce linearly from 2 to 0 on iterations, whereas b represents constant values based on specific paths and dictates the form of logarithmic spirals, where its value in this work is 1. The random number that falls from -1 to 1 is l .

WH have an equal chance of randomly selecting either course through the optimisation stage in Eqs. (10) and (11), since the probabilities are 50% and 50%, respectively. To update the positions of the WH, p, a random value between 0 and 1 is used to alternate between (8) and (9). During the BN stage, vector A's random value is [-1, 1], while during the searching stage, it may be more or less than 1. Equation (10) provides an illustration of the search strategy.

$$X(t+1) = X_{rand} - A.|C.X_{rand} - X(t)| \quad (10)$$

This random search (RS) methodology enables the WOA method to perform a global search (GS). When |A| is greater than 1, this RS gives the search process priority. Creating random solutions is the first step in the WOA searching procedure. These solutions are then continuously updated using the procedure shown in Table 1. Until a predefined max. iteration is achieved, searches continue.

Exploitation stages: (i) surrounding and (ii) spirally are the 2 stages for updating positions. Decreasing $\vec{a}$ linearly from 2 to 0 for every iteration is one method of defining encircling behaviours. Updating the position of spiral: The WH's relation to the fish and its helical movement is established as

$$\vec{X}(t+1) = \vec{D}xe^{b1}xcos(2\pi l) + \vec{X} * (t) \quad (11)$$

In this case, the current distance between the fish and the WH is $\overrightarrow{D'} = |\vec{X} * (t)\_\vec{X}(t)|$. An immovable object that represents the spiral motion of the WH is denoted as b. A random vectors with interval [-1,1] is denoted ad b. Additionally, there is the option to randomise vector values of p in the interval [0,1] or to make deep plunges by circling and creating spirals, as provided by equation 11.

Exploration stage: As part of a global fish exploration task, WH search for fish that are becoming more and more prominent. Values of vectors in the interval [0,1] are implied by $\vec{A}$. Explorations are indicated by 0 , and exploitations by 1. Decisions to switch between exploration and exploitation are based on these values. Additionally, the new location of the WH is shown in equations (12) and (13)

$$\vec{H} = |\vec{C}x\vec{X}_{rand} - \vec{X}| \quad (12)$$

$$\vec{X}(t+1) = \vec{X}_{rand} - \vec{A}x\vec{H} \quad (13)$$

Here, new locations of the WH are chosen at random from the remaining WH by $\vec{X}_{rand}$.

ASL approach

According to social learning theory, neighbourhood memberships for each WH can improve information sharing between groups, imitation of the current best practices is promoted, and it may also increase a population's ability to deviate from local optimal solutions.

$$G(t) = \{x_1(t), x_2(t), \ldots., x_N(t)\} \quad (14)$$

Here, the population's size is denoted as N. The fitness of individuals is assessed and rated in ascending order to produce a sorted population.

$$G_1(t) = \{x_{(1)}(t), x_{(2)}(t), \ldots., x_{(N)}(t)\} \quad (15)$$

The following represents the social ranks of $x_{(i)}(t)$:

$$I_{(i)}(t) = \frac{R_{(i)}(t)}{N} i = 1,2,\ldots,N \quad (16)$$

When random numbers are implied by $R_{(i)}$. The relationship between an individual and another individual is represented by$I_{(i)}(t)$.

Therefore, group collaboration enables exploration, the technique's exploitation phase mostly concentrates on identifying the optimal solution. Using an adaptable social neighbourhood approach, a novel WH search strategy is created.

Since the WOA exploitation phase only imitates and learns from the actions of the current best solution, the variation of individuals is reduced. The method's rate of convergence is impacted by blindness and inefficient information transmission between groups that occur during the WOA's exploration phase while learning from random individuals. Therefore, it is troublesome that the traditional WOA has a slow convergence speed. To speed up convergence, MWOA has been suggested as a solution to the previously discussed problem.

probability function approach (PFA)

PFA for the penalty function is incorporated into the design along with a new initial population strategy. Equations (17), (18), (19), and (20) show a slight modification of PFA for the penalty function.

$$\Phi(X) = F(X)(1 + penalty) \quad (17)$$

$$penalty = g_{ave}\left(\frac{(g_{max}+g(i))}{(g_{max}-g_{ave})}\right) g(i) \geq g_{ave} \quad (18)$$

$$penalty = g_{ave}\left(\frac{(g_{ave}+g(i))}{(g_{ave}-g_{min})}\right) g(i) < g_{ave} \quad (19)$$

$$penalty = 0 \qquad g(i) = 0 \quad i = 1, \ldots\ldots, m \quad (20)$$

X is the vector for the design variables (DV) in Eqs. (17–20). The OF for minimum volume is *F*(*X*). The total number of constraints is *n*. The maximum, minimum, and average violation values of the generation are denoted by g_{max}, g_{min}, and g_{ave}. A modified OF is U(*X*). The total violation value of the normalized displacement g_{dj} (*x*), is denoted by *g*(*i*). The stress constraints of the *i*th individual is denoted as g_{sr} (*x*).

$$g_{dj} = d_j/d_{uj} - 1 \qquad j = 1 \ldots\ldots m \quad (21)$$

$$g_{sr} = g_r/g_{ar} - 1 \qquad r = 1 \ldots\ldots m \quad (22)$$

Here, the maximum displacement at the *j*th node is denoted by d_{uj}. The permitted stress value in the *r*th element is gar. The count of displacement constraints is represented as *m*. The number of stress constraints is denoted by l. The violations of normalised constraints serve as the foundation for the formulation of the unconstrained optimisation problem. These formulations are missing from the previous version. In order to force the algorithm to remove designs with serious violations and larger OV from the generation, g_{ave} was included as a penalty parameter. This is justified by the fact that the likelihood of survival is not eliminated for designs with minor violations and smaller OV. The following is the MWOA probability function (PF):

$$p_m = 0.5(f_{max} - f)/(f_{max} - f_{ave})\ f \geq f_{ave} \quad (23)$$

$$p_m = (f_{ave} - f)/(f_{ave} - f_{min})\ f < f_{ave} \quad (24)$$

$$p_c = (f_{max} - f')/(f_{max} - f_{ave})\ f' \geq f_{ave} \quad (25)$$

$$p_c = 1.0 f' < f_{ave} \quad (26)$$

Here, an individual's fitness is denoted as *f*. The population's average FV is denoted as f_{ave}. The population's highest FV and minimum FV are denoted by f_{max} and f_{min}. The lower FV of the solutions to be crossed is denoted by *f*0. According to the degree of constraint violation, an individual's DV is arranged for PF. Depending on the individual's FV, different probabilities indicate how many DV will be interrupted. As a result, during the evolutionary process, the PFA can automatically adapt. As a result, no pre-established parameters are required by the algorithm.

3.4. Classification using EDL

Classifiers are used in a decision-making (DM) process known as Majority Voting (MV), which has been executed n times with increasing power each time. Let *C* be a set of *Q* classes and χ be a set of N instances. An algorithm set $S = \{A_1, A_2, A_M\}$ is defined. The M classifiers used in the vote are included in here as well. One of the Q classes has been assigned to each instance of $x \in \chi$. A classifier makes a prediction for each example each time it runs.

Every sample is given a final class that corresponds to the class that the majority of classifiers predicted. The accuracy value of the classifier's prediction *Acc*, determines the weighting of each vote in MV. Consequently, the following is one way to indicate the total number of votes for class c_k:

$$T_k = \sum_{l=1}^{M} Acc\ (A_l) \times F_k(c_l) \quad (27)$$

$$F_k(c_l) = \begin{cases} 1 & c_l = c_k \\ 0 & c_l \neq c_k \end{cases}$$

Here, the classes of *C* are c_l and c_k. The class with the highest total weight is chosen. The classifiers are usually trained on several independent training sets with weights applied in order to determine whether the data is positive or negative.

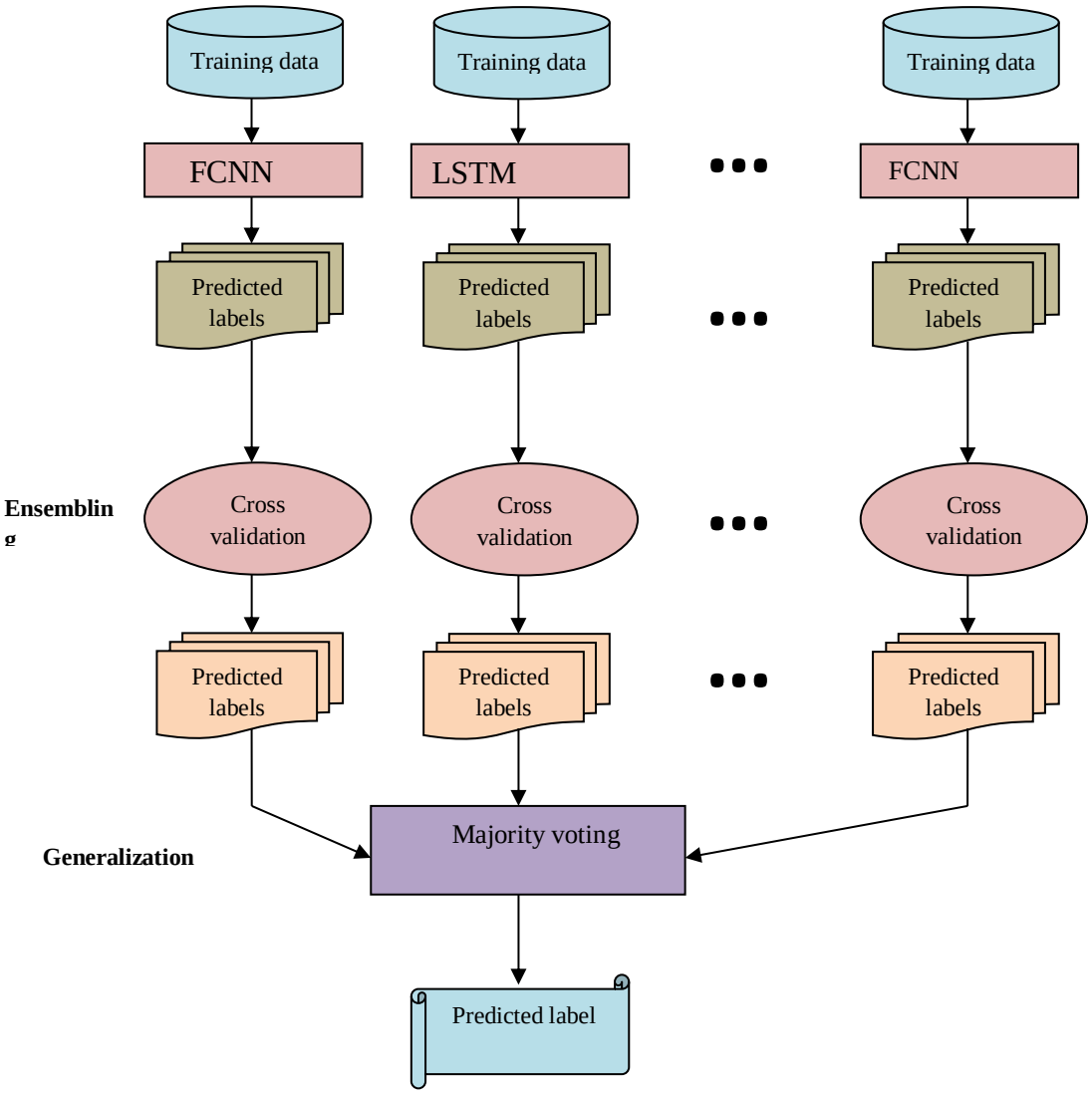


Figure 2: EDL

3.4.1. ENN

The basic configuration of the ENN is displayed in Fig. 3. The Elman network's four primary layers: the input layer (IL), hidden layer (HL), context layer (CL), and output layer (OL). These layers are easily understood. Each pair of adjacent levels is connected by adjustable weights [26]. With extra memory neurons and local feedback, it is typically regarded as a unique type of feed-forward (FF) NN (FFNN). Dynamic system modelling benefits greatly from the Elman network's sensitivity to input data history due to the self-connections of the context nodes. The following is the notation used in this section:

$w1_{ij}$: Node i in the IL and node j in the HL are connected by this weight.
$w2_{ij}$: Node i in the HL and node j in the OL are connected by this weight.
$w3_{ij}$: In the HL, the weight that links context node (Cn) i to node j.

m, n, r : IL, OL, and HL node counts
$u_i(k), y_i(k)$: Here, i=1,2,.....m and j=1,2,.....n .The inputs and outputs of the ENN.
$x_i(k)$: The hidden node (Hn) i output, Here i=l, 2,.....r

$c_i(k)$: The output of the Hn i from the previous time, or the output of Cn i. A unit delay is z^{-1}.

The context unit (CU) is an extra unit that is added for every unit in the HL. In a forward direction, the CU is fully connected (FC) to every hidden unit (HU). This indicates that each CU and each HU have a weight. Recurrent connections exist between the HU and the CU as well. However, as seen in Fig. 3, each HU is only connected to its corresponding CU.

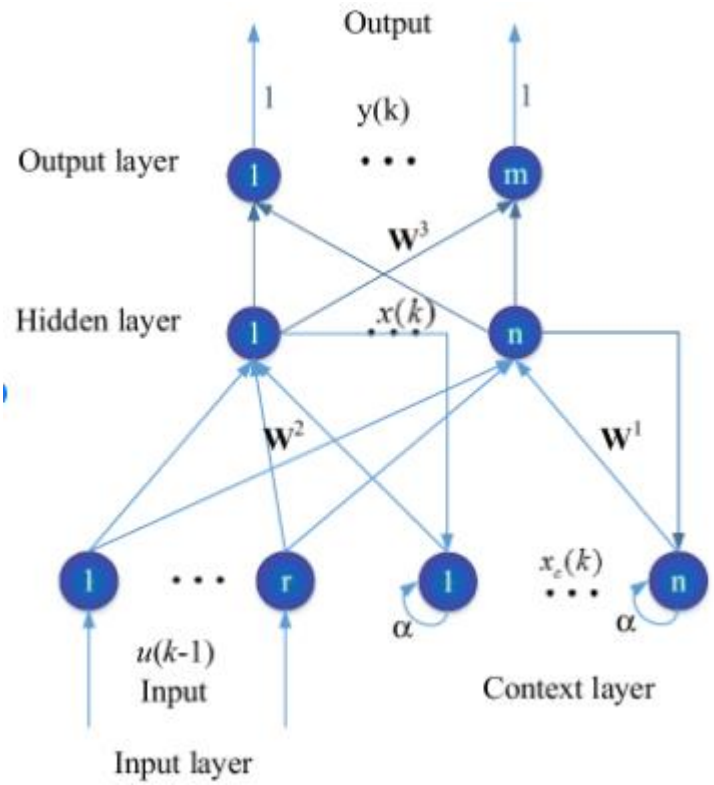


Fig. 3. Configuration of the ENN framaework.

The forward weights are trained using back-propagation (BP). The weights of the recurrent connections are fixed. The CU act similarly to the input units during the forward phase. The same method used for FF networks is employed for computing the values of the output units and the HU. The current values are sent into the relevant CU via the recurrent connections (using a unit delay) following the computation of the HU's outputs. Some initial values for these parameters are fixed at the 1st time step (TS). Then, it will proceed in subsequent TS. The forward weights are modified by the BP.

Target values for the outputs are employed in the backward training phase.
$u(k) \in R^m$ $y(k) \in R^n$, $x(k) \in R^r$ are the network's inputs. The outputs in each layer is expressed as

$$x_j(k) = f(\sum_{i=1}^{m} w2_{i,j} u_i(k) + \sum_{i=1}^{r} w1_{i,j} c_i(k)) \quad (28)$$
$$c_i(k) = x_i(k-1) \quad (29)$$
$$y_j(k) = g(\sum_{i=1}^{r} w3_{i,j} x_i(k) \quad (30)$$

Here, HL and OL linear and nonlinear (NL) output functions are denoted by f(.) and g(.), respectively. Elman networks don't require the state as an input or training signal because its dynamic features are solely supplied by internal connections. The Elman network has this benefit over static FF networks.

3.4.2. Generative Adversial Network (GAN)

Auxiliary Classifier GAN, a type of GAN, was first presented in [27]. The high-quality and high-resolution (HR) images are created by the GAN, and it is the benefits of this network. The generator (*G*) is encouraged to generate diverse and realistic samples through the adversarial training process. The two NNs that comprise the ACGAN model, the discriminator (*D*) and *G*. In a competitive setting, these *D* and *G* are trained simultaneously. Creating synthetic ECG heartbeat (HB) images from random noise vectors conditioned for particular beat types is the main function of the *G*. In contrast, the *D* network plays a dual role as it determines whether the input HB are genuine or synthetic and must predict the class label of those synthetic beats. In terms of its structural composition, the ACGAN model doesn't deviate significantly from previous established models. Nevertheless, it yields exceptional outcomes and demonstrates a notable capacity for stabilizing the training process. The discriminator model is specifically designed for binary real or fake image classification, processing images with dimensions (128, 128, 1). The architecture includes convolutional blocks, starting with a Conv2D layer with 16 filters, a 3 × 3 kernel, and 2 × 2 strides. The following blocks include batch normalisation (BN), dropout (p = 0.5) for regularisation, Leaky ReLU activation (α 0.2), and Conv2D layers with greater filter sizes (32, 64, 128, 256, and 512). These blocks progressively downsample the input image, generating feature maps with reduced spatial dimensions.

The flattened output connects to two Dense layers (DL) for prediction. The first DL outputs the likelihood that the image is real or fake (σ) based on sigmoid activation and a single neurone. Class probabilities are provided by the second DL, which has neurons that correspond to certain classes (S, V, and F) and softmax activation (softmax). The model is built using the Adam optimiser (AO), which has a learning rate of 0.0002 and a β1 value of 0.5. The loss function (LF) reflects the dual goal of the *D* to classify images into classes and ascertain their authenticity by combining binary (CE) cross-entropy (BCE) and sparse categorical CE. Dropout layers contribute to regularization, and batch normalization enhances stability. Conversely, the generator model is designed for conditional image synthesis, taking a latent vector of dimension 100 (*z*) and a categorical class label (*c*) as inputs. The label undergoes embedding (Embedding) and linear transformation (Dense), resulting in a (32, 32, 1) tensor. Concurrently, a DL with ReLU activation processes the latent vector, creating the basis for a 32 x 32 image. The two streams merge and are upsampled through Conv2DTranspose layers with decreasing filter sizes and increasing strides, ultimately generating an image with dimensions (128, 128, 1). Each Conv2DTranspose layer is followed by BN and ReLU activation, and the OL utilizes tanh activation to ensure pixel values fall within the range $[-1, 1]$.

1: **Input:**
• Real samples *X*real
• Noise samples *Z*gen from prior *Pg*(*Z*)
• Data distribution *P*data(*X*)
• Number of epochs, *n*epochs
• Number of steps per epoch, *e*steps
• Batch size, *n*batch
• Latent dimension, *latent_dim*
2: **Output:** Trained ACGAN
3: **procedure** TrainACGAN
4: Initialize *D* and generator *G* parameters.
5: **for** *e* in 1 to *n*epochs **do**
6: **for** *s* in 1 to *e*steps **do**
7: Sample *M* noise samples *Zk* from *Pg*(*Z*).
8: Sample *M* examples *Xk* from *P*data(*X*).
9: Update *D* by ascending its
stochastic gradient.
10: **end for**
11: **for** *s* in 1 to *e*steps **do**
12: Sample *M* noise samples *Zk* from *Pg*(*Z*).
13: Update *G* by descending its
stochastic gradient (SG).
14: **end for**
15: **end for**
16: **end procedure**
17: **Return:** Trained ACGAN

Algorithm 1. ACGAN process

Algorithm 1 shows the training procedure of the ACGAN framework, involving alternating steps for updating the *D* and the *G*. The mathematical expressions for updating the D and G involve binary (CEL) cross-entropy loss, providing a rigorous framework for the dual objectives of realism and class conditioning in image synthesis. In each training iteration, real samples *X*real and noise samples *Z*gen are selected. The *D* is then updated by ascending its SG, calculated as a combination of the binary CEL for real samples and the generated samples. Subsequently, the *G* is updated by descending its SG, aiming to minimize the binary CEL by fooling the *D*. As the *D* learns to differentiate between generated and real samples, this process iterates over several epochs (E) and steps, modifying the model parameters to improve the *G*'s capacity to generate realistic images. For real samples, $\mathcal{L}_{real}$ calculates the discrepancy between the actual labels and predicted probabilities. $\mathcal{L}_{fake}$ measures the *D*'s capacity to discriminate between generated and real samples. The Lreal, and Lfake are the two primary loss components that are involved in updating the *D*.
The discriminator loss L*D* is given by the sum of these components:

$$\mathcal{L}_D = \mathcal{L}_{real} + \mathcal{L}_{fake} \quad (31)$$

Where

$$\mathcal{L}_{real} = -\frac{1}{2}\mathbb{E}_{x,c}\left[logD_{(x|c)}\right], \quad (32)$$

$\mathcal{L}_{fake} = -\frac{1}{2}\mathbb{E}_{z,c}\left[\log(1 - D_{(G(Z|C))})\right]$ (33)

Similarly, the update of the generator involves two main loss components: Ladv, focusing on generating samples that resemble real ECG signals, and Laux, encouraging the model to produce synthetic samples with accurate class labels. The generator loss L*G* is given by the sum of these components:

$\mathcal{L}_G = \mathcal{L}_{adv} + \mathcal{L}_{aux}$ (34)

$\mathcal{L}_{adv} = -\frac{1}{2}\mathbb{E}_{z,c}\left[logD(G_{(Z|C)})\right]$ (35)

$\mathcal{L}_{aux} = -\mathbb{E}_{z,c}\left[logsoftmax(D_{label}(G_{(Z|C)}))\right]$ (36)

The algorithm 1 converges over training epochs, optimizing the ACGAN for conditional data classification.

3.4.3 RNN

One type of supervised ML model is called an RNN, which is an artificial neuron with many feedback loops. Figure 4 illustrates The loops that repeat over time or in a series are known as feedback loops. An input-target pair training dataset is necessary for supervised RNN training [28]. By optimizing the network's weights, the goal is to minimize the variation (i.e., the loss value) among the outcome and goal pairings.

A. Model Architecture

A natural RNN has three coatings, as shown in Figure 4: suggestion, unseen repeating, and quantity. The coating has N recommended holes. A set of headers spanning {..., xt−1, xt, xt+1...}, where xt = (x1, x2,..., xN). These are the parameters for this tier. The weight source W_{IH} identifies the links between recommendation and unseen portions of a fully associated RNN's hidden coating. As shown in Fig. 1b, the layer that is hidden consists of M units that are concealed, ht = (h1, h2..., hM), which are linked together over time via recurrent connections. HU can be initialized with minor non-zero components to improve the system's overall eficiency and stability. The HL defines the system's "memory," or state space, as

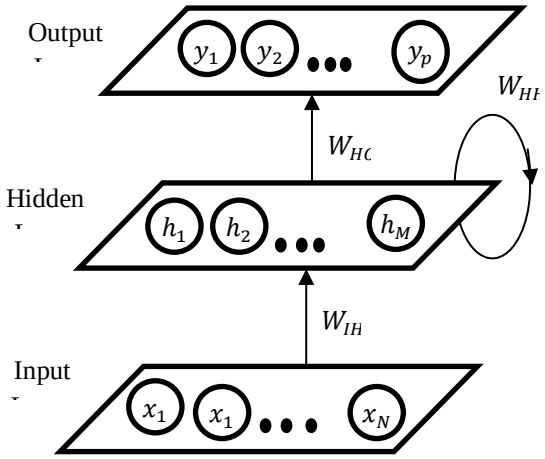
$h_t = f_H(o_t)$ (37)

$o_t = W_{IH}x_t + W_{HH}h_{t-1} + b_h$ (38)

b_h is the HU' bias vector, while fH(·) represents the HL's (AF). The weighted link between the OL and the HU is denoted as W_{HO}. The OL calculates the p units, yt = y1..., yP, as

$y_t = f_0(W_{HO}h_t + b_o)$ (39)

Here, the OL's bias vector is *bo* and the (AF) activation functions are represented by fO(·). The aforementioned stages are then repeated throughout time t = (1,..., T) Because the target and input historical pairs occur sequentially. Equations (1) and (3) exhibit the nonlinear state equations that make up an RNN over time. At each TS, the hidden states create an OL prediction based on the input data. A collection of values, unaffected by external effects, contains every unique piece of data required to reconstruct the network's prior states over a range of timesteps. A collection of values reflects an RNN's hidden state. Accurate predictions may be made at the OL and the future behavior of the network defined by this combined knowledge. Every unit of an RNN employs a basic nonlinear AF. However, if properly taught using TS, even a basic structure may simulate rich dynamics.



(a) Folded

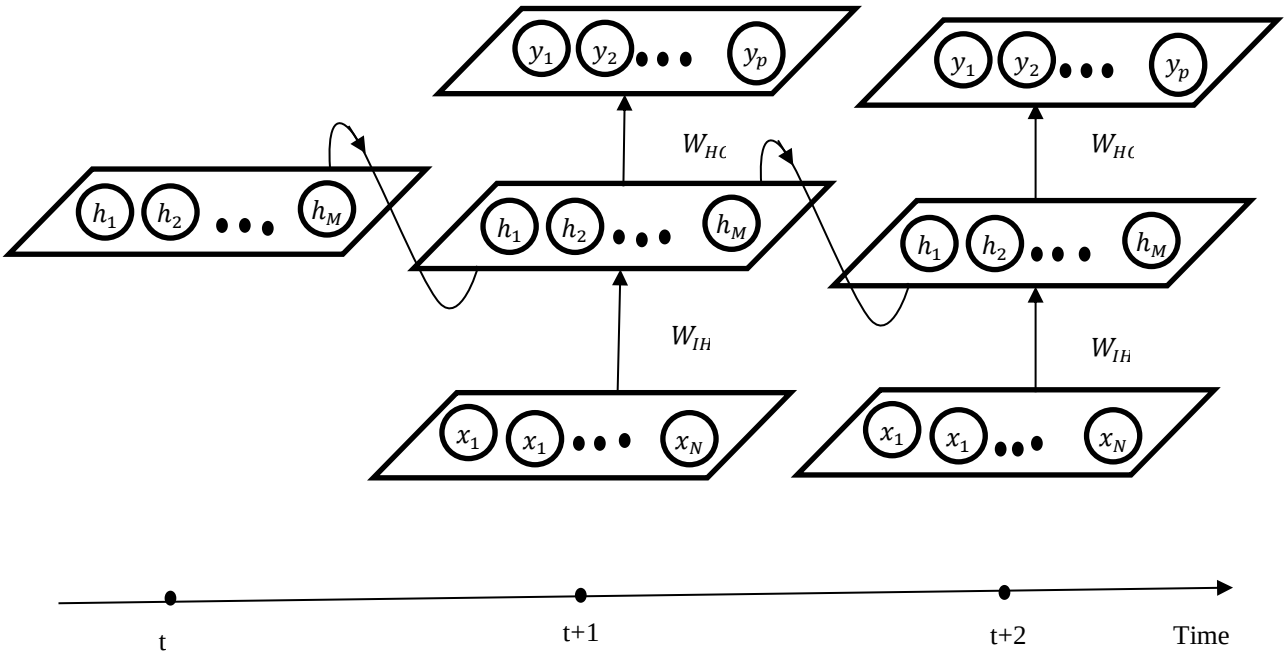


Figure 4. The RNN's Structure

B. Activation Function

In the case of linear networks, a number of linear hidden layers function as one. The ability to define nonlinear boundaries makes nonlinear functions more potent than linear functions. An RNN's hidden levels can include discontinuities, which allow for the identification of input-target links. A selection of the most often utilized stimulation functions are displayed in Find 3. Though not as well-known as the remaining functions of activation, the "sigmoid," "tanh," and ReLU have garnered attention lately. The "sigmoid calculation" is a widely used approach that entails entering a real integer into the range [0, 1]. Utilizing the Cross Entropy functions for loss in the OL, the stimulation function is widely used to train classification models. The following is the definition of the words "sigmoid" and "tanh" activation functions:

$$\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1} \quad (40)$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (41)$$

in that order. The "sigmoid" function of activation and the "tanh" stimulation function are, as it turns out, scaled.

$$\sigma(x) = \frac{\tanh\left(\frac{x}{2}\right)+1}{2} \quad (42)$$

For inputs that have positive values, the widely used ReLU activation function is open-ended and is described as

$$y(x) = \max(x, 0) \quad (43)$$

Most of the time, the best AF depends on the problem and data type. For networks with an output in the range [0, 1], for instance, "sigmoid" is appropriate. The neuron is rapidly saturated by the "tanh" and "sigmoid" AF, which have the ability to dissolve gradients. Ignoring that might lead to unstable dynamics in the weights' gradient updates due to the non-zero centered output of the "sigmoid ". Stochastic gradient descent (SGD) produces quicker convergence and sparser gradients than "sigmoid" or "tanh" activation functions. ReLU may be implemented with a border activation value of zero, making it computationally inexpensive. The neuron may stay dormant during training as the weight matrix expands since ReLU is not robust to strong gradient flows.

C. Loss Function

By comparing the corresponding target (zt) with the outcome (yt), that is known as

$$\mathcal{L}(y, z) = \sum_{t=1}^{T} \mathcal{L}_t(y_t, z_t) \quad (44)$$

It is the total of all losses for every timestep. The particular problem must be considered while selecting a loss function. Popular loss functions for real-value forecasting include the Euclidean and Hamming distances, and for classification issues, CE over the probability distribution of the outputs.

RESULTS AND DISCUSSION

The test outcomes of the recommended framework are presented in detail in this section. The suggested idea is implemented using Mat Lab. The new EDL-MWOA model is compared to the current HFCNN, CNN, and Swarm-ANN models for the Cleveland database from the UCI data repository in terms of precision (P), recall (R), ACC, and f-measure. https://archive.ics.uci.edu/ml/machine-learning-databases/heartdisease/is the link to the Cleveland database. The term "multivariate" describes this kind of dataset, which are numerical data analyses that provide or incorporate wide ranges of mathematical or statistical variables. Fourteen characteristics make up this composite: BP at rest, serum cholesterol, blood sugar while fasting, maximal heart rate reached, angina brought on by exercise, old peak — ST depressions due to activities in comparison to rest —slopes of peak exercise ST segments, and counts of main vessels, and thalassemia. The Cleveland database has 920 occurrences.

P is expressed as the proportion of all expected positive annotations to correctly acquired optimistic findings.

$$P = TP/TP+FP \quad (45)$$

The ratio of accurately recognized positive findings compared to every finding within the actual group is referred to as sensitivity.

$$R = TP/TP+FN \quad (46)$$

The ACC and R weighted average is represented by the F1 score. It can support both FP and FN outcomes.

$$F1\ measure = 2*(R * P)/ (R + P) \quad (47)$$

In terms of positives and negatives, ACC is computed:

$$ACC = (TP+FP)/(TP+TN+FP+FN) \quad (48)$$

TP, FP, TN, and FN stand for True Positive, False Positive, True Negative, and False Negative, respectively.

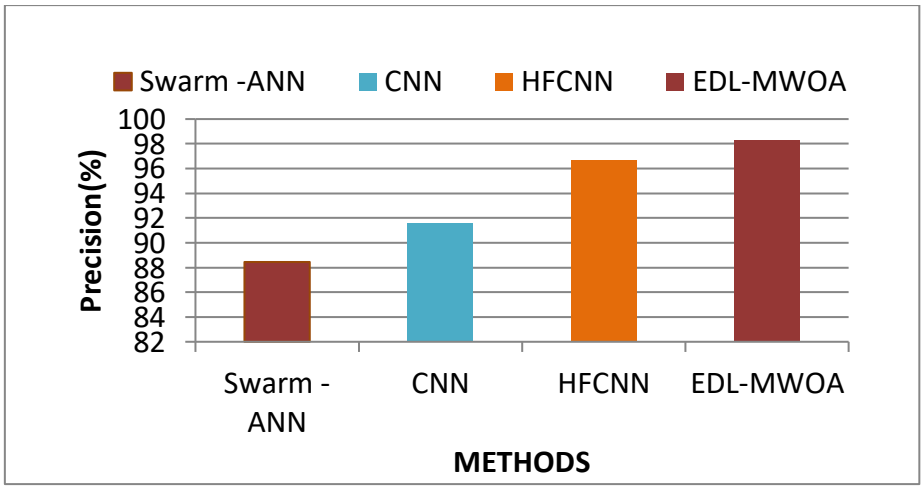


Figure 5. Comparison of the Suggested and Current Approaches by P

The results show that, in the event that adequate data is provided to the classifier model the suggested model may correctly predict HD. By employing improved FE approaches, the suggested model performs better. These outcomes also show that the suggested model is capable of outperforming other models. Figure 5 shows that the CNN technique has 91.57%, the Swarm-ANN technique has 88.45%, the HF-CNN technique has 96.7%, and the suggested EDL-MWOA method has an ACC rate of 98.26%.

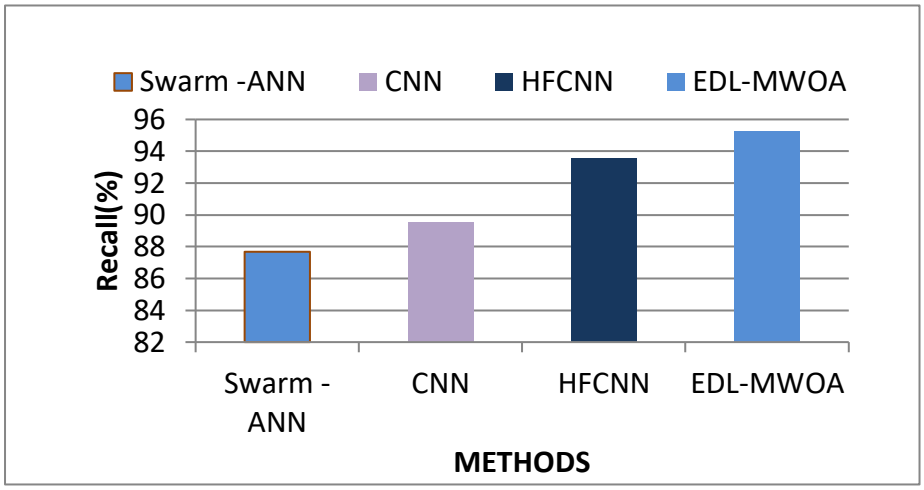


Figure 6. Comparison of the suggested and current approaches by R.

The suggested EDL-MWOA performs better in recall than earlier methods, as shown in Figure 6. With a recall of 95.27%, the EDL-MWOA approach is more effective than the Swarm-ANN method (89.54%), CNN method (87.68%), and HFCNN methodology (93.57%). Recall increases rapidly in the beginning of training, and using the recommended MWOAbased FS, it is evident that this reduces the distance between the points, stabilizing the output.

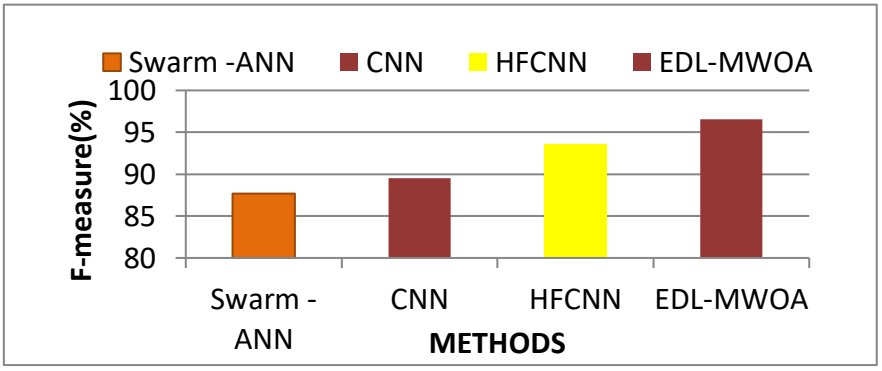


Figure 7. F-meaure review between the recommended and current approaches

As seen in Figure 7, the suggested EDL-MWOA outperforms previous approaches in terms of F-measure. The suggested system's performance was further enhanced by combining the feature selection approach with the MWOA methodology. After selecting the characteristics, used both general and particular feature selection methods to assess their relevance.

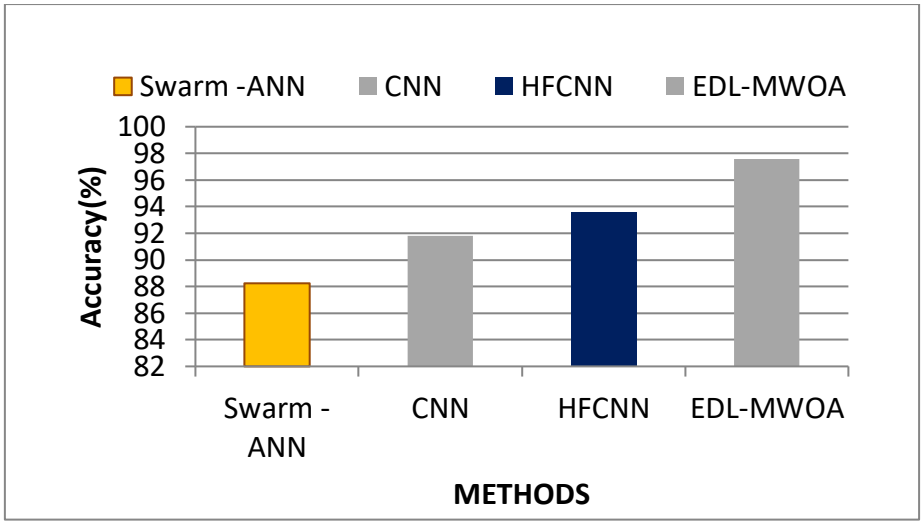


Figure 8. Evaluation of the Suggested and Current Methods' Accuracy

The suggested EDL-MWOA model has the best accuracy (97.58%), as seen in Figure 8. This classifier has better accuracy than others. The obtained findings demonstrate that the suggested model may reliably predict cardiac disease if sufficient information is supplied to the classifier model. These findings also show that the model that was suggested outperforms existing models in data categorization with more characteristics. All baseline models' accuracy improved significantly after utilizing the suggested feature extraction and selection technique.

CONCLUSION

The healthcare industry collects vast amounts of data, but unfortunately, it is not "mined" to find insights that could help in making choices. Unexplored patterns and linkages are frequently discovered. This can be fixed with the use of sophisticated DM strategies. This research work introduces an EDL approach and efficient FE method for the categorization of heart condition. The four primary stages of this study are FE, FS, preprocessing, and classification. Initially the MMN is used for pre-processing. The KMC technique is used for clustering with the goal of improving classifier ACC. This work suggests using PCA to efficiently extract features. After that, the FS process is finished using the MWOA. In order to choose the most pertinent and significant characteristics from the CVD dataset, it computes the optimal FV. The ENN, GAN and RNN are used in an EDL model to identify HD data. The obtained findings demonstrate that the recommended framework may reliably predict HD if sufficient information is supplied to the classifier model. These findings also show that the model that was suggested outperforms existing models in data categorization with more characteristics. All baseline models' ACC improved significantly after utilizing the suggested FE and FS technique. Further this research work concentrate on the EEG signals based prediction.

REFERENCES

1. Katarya, R., & Meena, S. K. (2021). Machine learning techniques for heart disease prediction: a comparative study and analysis. *Health and Technology*, *11*, 87-97.
2. Kumar, M. N., Koushik, K. V. S., & Deepak, K. (2018). Prediction of heart diseases using data mining and machine learning algorithms and tools. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *3*(3), 887-898.
3. Rani, P., Kumar, R., Ahmed, N. M. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, *7*(3), 263-275.
4. Maheswari, S., &Pitchai, R. (2019). Heart disease prediction system using decision tree and naive Bayes algorithm. *Current Medical Imaging*, *15*(8), 712-717.
5. Pattekari, S. A., & Parveen, A. (2012). Prediction system for heart disease using Naïve Bayes. *International journal of advanced computer and mathematical sciences*, *3*(3), 290-294.
6. Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS international conference on computer systems and applications* (pp. 108-115). IEEE.
7. Yahaya, L., Oye, N. D., & Garba, E. J. (2020). A comprehensive review on heart disease prediction using data mining and machine learning techniques. *American Journal of Artificial Intelligence*, *4*(1), 20-29.
8. Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1-4.
9. Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. *Procedia Computer Science*, *85*, 962-969.

10. Aljanabi, M., Qutqut, M. H., & Hijjawi, M. (2018). Machine learning classification techniques for heart disease prediction: a review. *International Journal of Engineering & Technology*, *7*(4), 5373-5379.
11. Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018, March). Prediction of cardiovascular disease using machine learning algorithms. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* (pp. 1-7). IEEE.
12. Gárate-Escamila, A. K., El Hassani, A. H., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, *19*, 100330.
13. Priyanka, N., & Kumar, P. R. (2017, April). Usage of data mining techniques in predicting the heart diseases—Naïve Bayes & decision tree. In *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)* (pp. 1-7). IEEE.
14. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.
15. Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving heart disease prediction using feature selection approaches. In *2019 16th international bhurban conference on applied sciences and technology (IBCAST)* (pp. 619-623). IEEE.
16. Sharma, V., Yadav, S., & Gupta, M. (2020, December). Heart disease prediction using machine learning techniques. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 177-181). IEEE..
17. Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019, April). Design and implementing heart disease prediction using naives Bayesian. In *2019 3rd International conference on trends in electronics and informatics (ICOEI)* (pp. 292-297). IEEE.
18. Abdollahi, J., & Nouri-Moghaddam, B. (2021). Feature selection for medical diagnosis: Evaluation for using a hybrid Stacked-Genetic approach in the diagnosis of heart disease. *arXiv preprint arXiv:2103.08175*.
19. Xie, J., Wu, R., Wang, H., Chen, H., Xu, X., Kong, Y., & Zhang, W. (2021). Prediction of cardiovascular diseases using weight learning based on density information. *Neurocomputing*, *452*, 566-575.
20. Reddy, N. S. C., Nee, S. S., Min, L. Z., & Ying, C. X. (2019). Classification and feature selection approaches by machine learning techniques: Heart disease prediction. *International Journal of Innovative Computing*, *9*(1).
21. Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. *Digital health*, *6*, 2055207620914777.
22. Mythili, T., Mukherji, D., Padalia, N., & Naidu, A. (2013). A heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL). *International Journal of Computer Applications*, *68*(16).
23. Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, 80716-80727.
24. Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100.
25. Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. *Advances in engineering software*, *95*, 51-67.
26. Ren, G., Cao, Y., Wen, S., Huang, T., & Zeng, Z. (2018). A modified Elman neural network with a new learning rate scheme. *Neurocomputing*, *286*, 11-18.
27. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.
28. Salehinejad, H., Sankar, S., Barfett, J., Colak, E., & Valaee, S. (2017). Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*.