DYNAMIC THRESHOLDING GA-BASED ECG FEATURE SELECTION IN CARDIOVASCULAR DISEASE DIAGNOSIS

Hasanain F. Hashim [1]
[1] LR11ES03 SMART Lab, ISG Tunis, Le Bardo, Tunis, Tunisia Université de Tunis, *kut, Waset, Iraq,*
hasaneanduh@gmail.com

Meriam JEMEL and Nadia Ben Azzouna
LR11ES03 SMART Lab, ISG Tunis, Le Bardo, Tunis, Tunisia
Université de Tunis
meriam_jemel@yahoo.fr, nadia.benazzouna@ensi.rnu.tn

***Abstract* - Electrocardiogram (ECG) data are usually used to diagnose cardiovascular disease (CVD) with the help of a revolutionary algorithm. Feature selection is a crucial step in the development of accurate and reliable diagnostic models for CVDs. This research introduces the dynamic threshold genetic algorithm (DTGA) algorithm, a type of genetic algorithm that is used for optimization problems and discusses its use in the context of feature selection. This research reveals the success of DTGA in selecting relevant ECG features that ultimately enhance accuracy and efficiency in the diagnosis of CVD. This work also proves the benefits of employing DTGA in clinical practice, including a reduction in the amount of time spent diagnosing patients and an increase in the precision with which individuals who are at risk of CVD can be identified.**

Index Terms- ECG; Genetic Algorithm; Feature Selection; Dynamic Thresholding.

I. INTRODUCTION

Recent improvements in medical equipment enable ubiquitous patient monitoring. With the ascent of ubiquitous computing, data mining (DM) technology has become increasingly prevalent and may be particularly effective in e-healthcare [1–3]. DM can be considered a specific kind of knowledge discovery method. It can be defined as the study of a volume of data sets to identify potentially important relations and to summarize these relations in a novel and intelligible way. Identifying the most relevant and informative features for a given task improves performance and efficiency in machine learning models. Without carefully considering features, the selection process may include irrelevant or redundant features, which can affect model accuracy (ACC) and generalization negatively. Early cardiovascular disease (CVD) diagnosis is essential for its effective treatment and management. However, early diagnosis can be difficult due to the disease's complexity and the need for specialized knowledge. As a result, interest in the development of algorithms and instruments for electrocardiogram (ECG) DM as a diagnostic aid is increasing. To develop effective ECG DM algorithms, obstacles and limitations, such as variability in ECG signals, lack of standardization in acquisition and processing, and the need for clinical relevance, must be addressed.

Feature selection is a crucial step in developing accurate and reliable diagnostic models for CVDs. With the increasing availability of large datasets and advanced machine learning techniques, identifying the most informative features that can accurately predict the presence of CVDs is crucial. The selection of relevant features can help reduce the dimensionality of the data, improve model ACC, and avoid overfitting. This research explores the importance of feature selection in CVD diagnosis and highlights some of the key methods used in this process.

The use of genetic algorithms (GA) in feature selection is a substantial contribution to the field of ECG classification [8]. By optimizing the selection of relevant features, GAs can effectively reduce the dimensionality of the data and prevent overfitting, ensuring that the model is robust and reliable. As such, this approach has the potential to revolutionize ECG classification by providing clinicians with efficient and accurate diagnostic tools for CVDs. In addition, GA is resilient in selecting the best attribute for the classification procedure, which is the source of the most glaring problems. The neural network (NN) structure was optimized to achieve convergence and efficiency [4].

Selection is an important factor in binary GAs because it determines which individuals are selected for replication or modification. The goal is to select individuals with high fitness while allowing for search space exploration. Binary coding GAs often require a sizable amount of time to converge and can struggle to deviate from local optimal values, which can limit the search process [6] [7]. The utilization of a GA-based approach for ECG feature selection, employing dynamic thresholding, shows promise in the field of CVD detection. The proposed approach employs GAs as a means of selecting the most pertinent ECG characteristics to effectively diagnose cardiovascular ailments. The incorporation of dynamic thresholding in this approach guarantees that the selected characteristics remain unaffected by noise or extraneous signals. The utilization of this technology has demonstrated meaningful promise in enhancing the precision of CVD diagnosis because it possesses the capability to detect crucial characteristics that might be overlooked by conventional diagnostic approaches. The application of a GA-based ECG feature selection technique that utilizes dynamic thresholding enables clinicians to enhance the diagnosis ACC and improve treatment options for those afflicted with cardiovascular ailments [10].

In general, the contributions of this research are listed as follows: (1) This article attempts to accelerate common classification techniques by substituting statistical algorithms with dynamic threshold genetic algorithm (DTGA) in the context of feature selection; and (2) to compare the proposed approach with the classical GA (CGA) approach. Several experiments show that the proposed DTGA for ECG feature selection approach outperforms previous, more involved prototypes in terms of ACC and processing speed.

This remaining sections are organized as follows: Section 2 contains a listing of contemporary pertinent works pertaining to ECG feature selection in CVD diagnosis. Section 3 provides a detailed explanation of the proposed DTGA algorithm. Section 4 encompasses the presentation and analysis of the experimental findings, as well as the subsequent discussion. Section 5 presents the conclusion.

II. RELATED WORKS

In recent years, DM algorithms and upgrades for health care DM have risen in importance within the scientific community as health care companies adopt analytical tools to improve patient outcomes. Feature extraction can be a crucial stage in most DM applications, especially those pertaining to healthcare. Although numerous methodologies have been utilized, their efficacy and usability differ depending on their application. Feature extraction is a crucial step in health care DM techniques for collecting useful information from the underlying dataset. In this literature review, recent advances and existing methods for feature extraction in healthcare DM will be explored [14].

The feature selection process has become a crucial component in various practical applications, including but not limited to text processing, face recognition, image retrieval, medical diagnosis, and bioinformatics [19–21]. Numerous endeavors have been undertaken to assess feature selection techniques, which may be categorized into four distinct groups based on the evaluation procedure, namely, filters, wrappers, hybrids, and embedding [22–25]. The classification approach, known as the filter method, encompasses any procedure that selects characteristics without relying on a learning algorithm, such as a standalone preprocessor. The utilization of the filter method for feature selection is an exclusive approach for addressing the feature selection problem without relying on a learning model. However, its implementation necessitates the application of statistical analysis. The wrapper technique utilizes a pre-established learning procedure to assess the efficacy of the filtered datasets. Although wrappers have the potential to enhance performance, their maintenance costs might be substantial, and they may experience failures when burdened with excessive features [26, 27]. To improve feature selection, hybrid techniques take and combine the best parts of filter and wrapper methods. This kind of approach might use filter and wrapper methods to cut down on the number of features and to analyze the remaining ones, respectively. Feature selection is built right into the learning process when embedding techniques are used. This method may work better than filter or wrapper methods because it learns the model parameters and the best set of features simultaneously. In contrast, the use of embedding methods on computers can be difficult and may require considerable data to work correctly. Of course, the dataset and study question at hand will determine which feature selection method is the best.

In their study, the scientists made use of the Heart Disease Data Set (UCI) repository, a comprehensive collection of data pertaining to heart disorders. The inclusion of data extensively expands the search area and introduces greater complexity to the classification models employed. To enhance the classification ACC, feature selection methodologies have been employed in [15] to discover the most essential attributes. This study presents a novel ensemble classification model that incorporates a feature selection strategy. The program utilizes ensemble learning techniques in combination with a GA and biological test values to diagnose heart diseases effectively. Nevertheless, the model that demonstrated the highest level of performance was derived from the integration of the GA and the ensemble learning model. This model attained an ACC rate of 97.57% when applied to the datasets that were considered. The magnitude of this value surpasses those of previously suggested methods.

Using GA and decision tree (DT), Reference [8] suggested a new system for categorizing cardiac arrhythmias. Using the suggested model, two-class and 16-class modes of feature classification were chosen. In the two-class mode, sensitivity and the mean SenSpec value increase, with the best possible ACC of 86.96 and a mean ACC of 85.04. The proposed technique outperformed the competing methods when applied to the 16-class mode, with an ACC parameter of 78.76 [8].

The utilization of DM techniques facilitates the categorization of ECG signals to aid in the diagnosis of heart disorders. To address these obstacles, the optimized discrete kernel vector (ODKV) classifier, accompanied by a noteworthy pre-processing technique, is proposed in Reference [16]. The primary application of adaptive notch filters (ANF) is the mitigation of power line interference in ECG signals. The ODVK classifier is used to mitigate the presence of redundant features and improve the classification ACC of the input ECG data.

Performance metrics, such as sensitivity, specificity, ACC, and mean square error, are calculated to prove the enhancement of the classification technique [16]. Overall, the results suggest that ANN-kNN performed the best among the methods tested, with the lowest values for most of the metrics. SVM-kNN and GB-SVNN also performed relatively well, whereas CNN and ODKV obtained higher values for most of the metrics. However, the relative performance of these methods may vary depending on the specific dataset and problem being addressed.

The approach provided in Reference [17] demonstrates a high level of ACC in classifying ECG data. Utilizing machine learning libraries and implementing the solution in the Scala programming language within the Apache Spark framework achieves this objective. The rhythmic contraction of the heart is the result of coordinated electrical signals produced by several specialized cardiac tissues. The effectiveness of the technique was evaluated using a dataset that consists of 205,146 records. The most challenging element of diagnosing an ECG is managing abnormalities within the ECG signals. The suggested technique is assessed and validated using the MIT-BIH Arrhythmia and MIT-BIH Supraventricular Arrhythmia databases as the baseline datasets. The proposed method achieved an overall ACC of 96.75% and 97.98% for binary identification using the Guangdong Development Bank (GDB) tree algorithm and random forest, respectively. The random forest algorithm demonstrates a precision rate of 98.03% in the context of multiclass classification.

In Reference [18], the utilization of meta-heuristic techniques was proposed to address the issue of redundant and unnecessary information. However, these methods often fail to consider the correlation that exists among a group of selected features. The researchers offer a strategy that incorporates the application of a GA based on community detection. The method functions sequentially, comprising three discrete phases. Initially, the process of computing feature similarities is conducted. Following this process, the characteristics are categorized into clusters. In conclusion, a GA, combining a unique repair operation centered around communities, is utilized to determine feature selection. The efficacy of the proposed methodology was assessed by examining nine standardized classification tasks. They show that their suggested method works better than three feature selection methods that use the particle swarm optimization (PSO), ant colony optimization (ACO), and ant bee colony (ABC) algorithms by conducting a comparison study. Specifically, the ACC of the proposed method is found to be, on average, 0.52% higher than that of the PSO method, 1.20% higher than that of the ACO method, and 1.57% higher than that of the ABC algorithm [18].

CVD is a prominent contributor to global mortality rates, necessitating the timely identification of the condition to facilitate optimal therapeutic interventions. In recent years, machine learning techniques have been increasingly used for CVD diagnosis, with feature selection being a critical step in improving ACC and efficiency. This research proposes a novel approach to a wrapper feature selection method using the DTGA algorithm.

III. PROPOSED DTGA-BASED FEATURE SELECTION APPROACH FOR ECG IN CVD DIAGNOSIS

A. *DTGA-based feature selection*

In machine learning and signal processing applications, the use of DTGA is a common method for selecting features [24]. This technique employs GAs to search iteratively for the optimal subset of features that can be used to classify a specific dataset accurately. Utilizing a dynamic thresholding technique to control the selection pressure and prevent premature convergence is the main innovation of the DTGA. Each chromosome in the DTGA represents a subset of features, with each gene indicating whether a particular feature is included or excluded from the subset. One gene is stored and represented using a single bit in DTGA. Each bit can either be set to 1 or 0, depending on the context. The use of bit encoding allows one individual to embody the states instantaneously, forcing the DTGA to be better in terms of diversity compared with the CGA algorithm. As stated in Reference [25], convergence can also be achieved with the bit statement. As a gene approaches 0 or l, the bit chromosome joins into one single state. The fitness of each chromosome is determined by training a classification model using only the chosen features and measuring the model's performance on a validation set. The fitness function is intended to strike a balance between the model's ACC and complexity, with penalties for overfitting and underfitting.

The crossover operation in the DTGA follows a two-point strategy, in which two progenitor chromosomes and two crossover points are selected randomly. The parents switch out the genes in the region between the crossover points, causing the progeny to have two chromosomes. Randomly and with a low probability, a gene in a chromosome flips from 0 to 1 or vice versa during a DTGA mutation. The mutation operator contributes to population diversity and prevents premature convergence. The DTGA approach provides an efficient method for selecting features in classification problems, with the dynamic thresholding technique balancing exploration and exploitation during the generation of new populations.

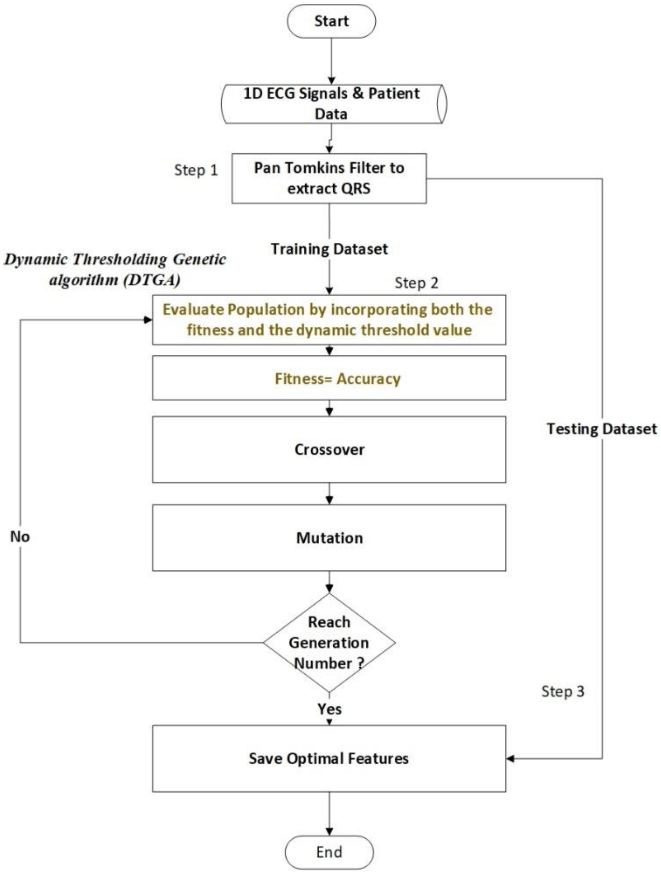
Start
1D ECG Signals & Patient Data
Step 1
Pan Tomkins Filter to extract QRS
Training Dataset
Dynamic Thresholding Genetic algorithm (DTGA)
Step 2
Evaluate Population by incorporating both the fitness and the dynamic threshold value
Fitness= Accuracy
Crossover
Mutation
No
Testing Dataset
Reach Generation Number ?
Step 3
Yes
Save Optimal Features
End

Fig. 1. Flowchart of the proposed approach

In DTGA [24], the binary representation of populations is adopted for minimization problems. The characteristic of the representation is that the bits are utilized to embody the state operator as:

$$q_j^t = [\beta_1^t|\beta_2^t|\dots\dots\dots.|\beta_m^t|] \quad , \tag{1}$$

where q_j^t represents the chromosome of the t-th generation and the j-th individual, and m is the gen β index number. The use of bit encoding allows one individual to embody the states instantaneously, thereby forcing the DTGA to be better in terms of diversity compared with the CGA algorithm. As stated in Reference [31], convergence can also be achieved with the bit statement. β attitudes to 0 or l.

Dynamic thresholding refers to a technique that can be used to enhance the population diversity and exploration space of a CGA. The following steps represent the mechanism of dynamic thresholding in a GA:

1. The dynamic thresholding technique assigns a threshold value to everyone in the population, denoted as T(i), where i represents the index of the individual.
2. The threshold value T(i) is determined by analyzing the performance of the individuals in the population by measuring the fitness value for each individual, typically based on their fitness or correlation coefficient values.
3. If the population is evolving rapidly, as determined by a correlation coefficient, then the threshold value T(i) is increased by one by increasing the genes in the population to adapt to the changing dynamics.
4. Conversely, if the population is stagnating or the fitness value does not change, then the threshold value T(i) is decreased by one by decreasing the genes in population to encourage exploration and avoid premature convergence.
5. Solutions from the population are selected for further evaluation and modification based on their fitness or correlation coefficient values relative to the threshold value T(i). Individuals whose performance exceeds or meets the threshold T(i) are considered potential solutions for further processing.
6. The selected solutions are subject to modification through crossover and mutational genetic operators to introduce diversity and explore different regions of the search space.
7. The algorithm terminates when the GA completes its execution or when a termination criterion is met. The high-quality solutions obtained during the dynamic thresholding process are returned as the final output of the algorithm.

This instance can improve performance, and the GA can provide better solutions [24].

Algorithm 1 represents the DTGA pseudocode. The DTGA algorithm takes a dataset T, the number of generations t, the number of individuals j, the initial populations Pops, and a dynamic thresholding function DT as inputs. Subsequently, the algorithm encodes Pops using bit_Encoding, evaluates the fitness values of Pops using Fitness_Evaluation, and selects the best individuals from Pops based on their fitness values using Selection_Best. Next, the algorithm updates Pops using Dynamic_Thresholding_Function(pop) and checks if the termination condition is true or false. If it is false, then the algorithm performs crossover and mutation on Pop to generate new individuals and sets Pops to New_Pops. If it is true, then the algorithm returns Pops. The best individual is the one in Pops with the highest fitness value. The Dynamic_Thresholding_Function(Pop) computes a threshold, increases or decreases based on whether the population is evolving rapidly or stagnating, selects solutions based on the threshold, evaluates and modifies them, and returns high-quality solutions.

Algorithm 1: DTGA

```
Inputs: Dataset T, Number of generations t, Number of individuals j, Initial Populations Pops, and Dynamic thresholding DT
Outputs:
1. While t is less than MAX_GENS, do the following:
2. Increment t by 1
3. Encode Pops using bit_Encoding
4. Evaluate the fitness values of Pops using Fitness_Evaluation
5. Select the best individuals from Pops based on their fitness values using Selection_Best
6.Update Pops using Dynamic_Thresholding_Function(pop)
7. If Termination_Condition is False, then do the following:
    a. For each i from 0 to j-1, do the following:
        i. Perform crossover on Pop to generate a new individual New_Pops(i)
        ii. Perform mutation on New_Pops(i)
    b. End for loop
    c. Set Pops to New_Pops
8. End if statement
9. If Termination_Condition is True, then do the following:
    a. Return Pops
10. End if statement
11. End while loop
12. Best individual is the one in Pops with the highest fitness value.
```

Dynamic_Thresholding_Function(Pop)

```
threshold = computeThreshold(Pop)
while geneticAlgorithmIsRunning():
    if populationIsEvolvingRapidly():
        threshold = increaseThreshold(threshold)
    else if populationIsStagnating():
        threshold = decreaseThreshold(threshold)
    selectedSolutions = selectSolutions(pop, threshold)
    evaluateSolutions(selectedSolutions)
    modifySolutions(selectedSolutions)
return highQualitySolutions
```

B. *Approach overview*

DTGA is used in this research to determine the most informative features for detecting and classifying various forms of arrhythmias. This approach can aid in the development of more precise and effective algorithms for diagnosing and treating cardiac conditions. The DTGA methodology involves multiple stages. Fig. 1 represents the steps of the DTGA model for feature selection. Initially, a random population of candidate feature subsets is generated. Then, each subset is evaluated using a fitness function, which uses ACC as the fitness value that gauges its ability to classify the ECG signals in the database accurately.

Important ECG signal characteristics include the QRS complex, P, T, and U waves. The QRS complex depicts the rapid depolarization of the right and left ventricles of the heart. P and T waves represent the depolarization of the atria and the repolarization of the ventricle, respectively. The U wave refers to the repolarization of the interventricular septum [29]. These defining points play a crucial role in defining an ECG signal [30, 31]. The primary procedures in our methodology are presented as follows:

B.1 Data description

The MIT-BIH Arrhythmia Database is used in this research,. It contains ECG recordings of patients with different types of arrhythmias. The MIT-BIH Arrhythmia Database is a group of ECG records from people with different heart problems. This database has ECG signals from two channels and 48 records that last 30 min each. Each record in the database has been marked up by trained cardiologists, who have found and named the different types of rhythms that were present in the recordings. The database lists 23 different types of rhythms. Each record in the MIT-BIH Arrhythmia Database has a time series of voltage readings taken from the two ECG channels. Mostly, these voltage readings are taken at a rate of 360 Hz. In addition to the raw voltage data, each record includes annotations regarding the beginning and ending times of various beats. Most of the time, these notes come in the form of binary labels that indicate whether a certain arrhythmia is happening during a certain time interval. Overall, the MIT-BIH Arrhythmia Database consists of 48 half-hour, two-channel ambulatory ECG recordings from 47 subjects, obtained between 1975 and 1979. The recordings were digitized at 360 samples per second per channel with an 11-bit resolution over a 10-mV range. The dataset includes approximately 110,000 annotations for each beat, and two or more cardiologists independently annotated each record. The database has been freely available since September 1999, with 25 of the 48 complete records and reference annotation files available. The 23 remaining signal files were posted in February 2005.

B.2 Approach description

Step 1: Data Preprocessing

The ECG signals from the database that contains the ECG recordings of patients are preprocessed, and the pertinent features are extracted. Signals are filtered, resampled, and segmented to obtain a set of features that can be used for classification.

Step 2: Population at the Outset

In the first step of the DTGA, a set of features is extracted from the dataset. Each individual in the population corresponds to a subset of these features, represented as a binary string where each bit indicates whether a particular feature is included or excluded from the subset. For example, if the dataset contains 10 features, an individual may be represented as a binary string with a length of 10, where a bit value of 1 indicates that the corresponding feature is included in the subset and a bit value of 0 indicates that it is excluded.

Each subset corresponds to the bits set to 1 in the individual. For example, if an individual is represented as 0101100010, then it corresponds to a subset of features that includes the second, third, fifth, eighth, and ninth features. The size of the population corresponds to the number of considered subsets. A larger population size allows for more diverse subsets to be explored during the GA search process. However, it also increases the computational cost of the algorithm.

To begin the GA search process, a random sample of prospective feature subsets is generated. Each subset is depicted as a binary string, as described above. The size of this initial population can be set based on prior knowledge or experimentation.

Step 3: Fitness Assessment

Each candidate feature subset is evaluated using a fitness function called "accuracy," which measures its ability to classify the ECG signals in the database precisely. The fitness function is obtained by applying a classifier to the data using the considered feature subset.

Step 4: Selecting the Best Populations

The GA is used to select the feature subsets with the highest performance from the initial population. This objective is achieved using tournament selection operators, which are a justified choice for GAs because of their ability to maintain diversity, efficiency, robustness, and scalability. Tournament selection promotes population diversity by selecting individuals randomly and permitting occasionally weaker individuals to win, thereby preventing premature convergence and guaranteeing that the search space is exhaustively explored. Each tournament selection iteration only requires a small subset of the population to be evaluated, making it appropriate for problems involving large populations or limited computational resources [18]. Tournament selection is robust to noise and outliers in fitness values because it only assesses the relative fitness of individuals within a tournament and not their absolute fitness [18]. By adjusting the tournament size, tournament selection can be readily adapted to varying problem types and fitness functions. Consequently, it can be tailored to specific problem domains and spaces [26].

Step 5: Crossbreeding and Mutation

New candidate solutions are generated by applying the genetic operators of crossover and mutation to the selected subsets. The crossover used by DTGA consists of two points, implying that two random sites on the parent chromosomes are selected, and the subsequences between them are swapped to create the offspring chromosomes. This form of crossover is frequently used in GAs and has been demonstrated to balance exploration and exploitation of the search space effectively. The mutation used in DTGA is a bit-flip mutation [16]. This means that for each offspring chromosome, each bit (or gene) has a small probability of being flipped from 0 to 1 or vice versa. This type of mutation is simple and commonly used in GAs because it allows for small changes in the offspring chromosomes that can help explore new regions of the search space [18].

Step 6. Steps 3 to 5 are repeated until a satisfactory solution is reached or a stopping criterion is met. In our context, the stopping criterion is reaching the generation number.

Step 7: Complete the Feature List

Dynamic thresholding in CGA improves population diversity and exploration space by assigning a threshold value to each individual based on fitness or correlation coefficient values. High-performance solutions are evaluated and modified using crossover and mutational genetic operators. The algorithm terminates when it completes or a termination criterion is met, resulting in high-quality solutions. The final feature set is obtained by selecting the subset of the previous generation of candidate solutions with the highest performance.

Through ECG DM, DTGA has the potential to revolutionize the diagnosis and management of CVD. Nevertheless, several obstacles, such as the need for large and diverse datasets, standardization in data acquisition and processing, and ensuring clinical relevance in the selection of ECG features, must be overcome. To optimize the performance of DTGA, various parameter settings must be explored, the use of other GAs should be investigated, and additional research is required. With sustained collaboration between researchers and healthcare providers, DTGA can become a valuable tool for identifying and providing personalized treatment to individuals at risk of CVD.

IV. SIMULATION RESULTS AND ANALYSIS

The scale factor procedure is used in our approach to normalize the feature values to ensure that they have equal importance in the classification model.

C. *Evaluation Metrics*

ACC, F-score, or sensitivity true positive rate (TPR), false positive rate, true negative rate (TNR), and false negative rate are types of measurements that are considered when evaluating and comparing the performance of KNN classification systems, as illustrated in Equations (2–6) [25]. Some results are used as true positive (TP), true negative (TN), false positive (FP), and false negative (FN)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (1)$$

$$TPR = \frac{TP}{TP+FN}, \quad (2)$$

$$TNR = \frac{TN}{TN+FP}, \quad (3)$$

$$PPV = \frac{TP}{TP+FP}, \quad (4)$$

$$F\_Score = \frac{2TP}{2TP+FP+FN}. \quad (5)$$

The size and complexity of the dataset, the availability of computational resources, and the desired level of ACC all play a role in the parameter selection in DTGA. Generation number determines the number of iterations in the method, whereas population size determines the evaluation of candidate feature subsets in each generation. The crossover ratio controls the fraction of candidate solutions that mutate during each iteration, whereas the mutation ratio controls the likelihood that a bit in a solution flips during each iteration. Furthermore, the threshold establishes the bare minimum of fitness values. The values of the parameters employed in our trials are listed in Table 1.

TABLE I. DTGA ALGORITHM PARAMETERS FOR EXPERIMENTS

Parameter	*Default Value*
Population size	[5,10,15,…..65]
Generation Number	[5,10]
Crossover Ratio	0.5
Mutation Ratio	0.5
Threshold	18

D. *Simulation setup*

We present experimental evidence supporting the efficacy of the proposed method. The output of the DTGA is compared with that of the suggested approach for KNN classification evaluation. Using an ECG problem, the DTGA algorithm provides a novel approach for the remote diagnosis of CVD. DTGA offers an alternative to conventional research methods by providing different models in which model populations are generated through the evolution of random starting models using a GA.

The framework is made available as a MATLAB library for integration into user-specific programs. The evaluations were performed on a PC equipped with an Intel(R) Zeon(R) CPU E5430@ 2.66GHz (2 processor), 16GB RAM, and Microsoft Windows 10-64 bit. The simulation results show that the proposed method may successfully generate a comprehensive modulation categorization.

a) Several commercial devices employ the accurate real-time ECG pulse detector introduced by Pan and Tompkins' algorithm [16]. This algorithm is initially implemented on a microprocessor and then in C. The objective of this work is to make the same technique more accessible to biomedical engineering researchers by implementing it in the MATLAB environment. This strategy consists of two steps:

b) The first step is preprocessing, which consists of five processes, beginning with a bandpass filter (5–15 Hz), then a deriving filter to highlight the QRS complex, then signal squared, signal averaged (MWI) to remove high frequency noise (0.150 s length), and finally, the filtering options are changed to best match the characteristics of the ECG signal depending on the sampling frequency of your signal.

c) The Decision Rule is the second stage. At this point in the algorithm, the preceding steps have generated an output signal waveform that resembles a pulse. Fiducial mark, thresholding, search back for missed QRS complexes, elimination of multiple detections within refractory period, and T wave discrimination are used to determine whether a pulse corresponds to a QRS complex (rather than a high-sloped T-wave or noise artifact) [24].

d) Images from Figure 2–11 illustrate the output of Pan and Tompkins' real-time ECG pulse detector. Table 2 displays the sampling frequencies utilized by the Pan and Tompkins algorithm.

TABLE II. PAN AND TOMPKINS PARAMETERS

Parameter	***Default Value***
Sampling frequency	200

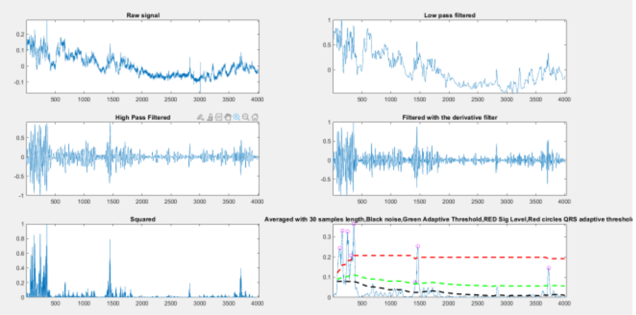

Fig. 2. *ECG Sample 1 Preprocessing outputs.*

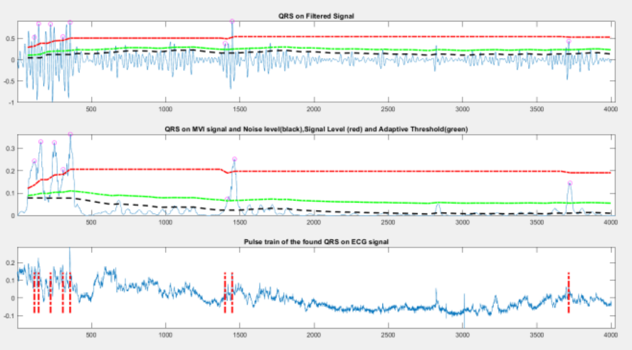

Fig. 3. *ECG Sample 1 Decision Rule outputs.*

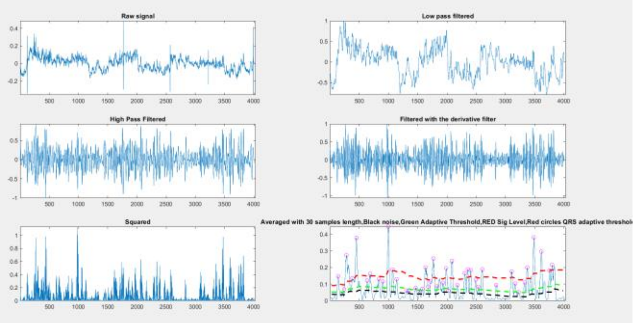

Fig. 4. *ECG Sample 2 Preprocessing outputs*

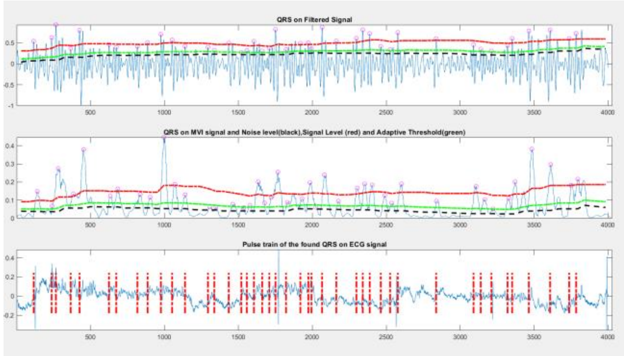

Fig. 5. *ECG Sample 2 Decision Rule outputs.*

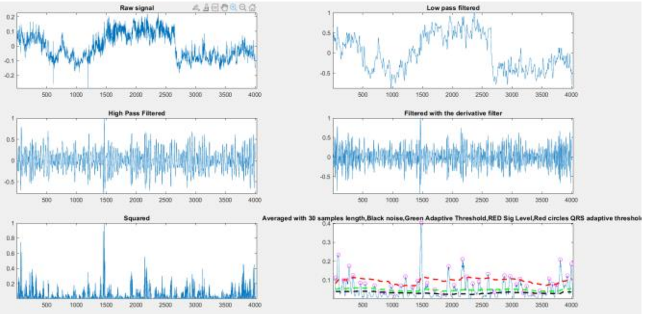

Fig. 6. *ECG Sample 3 Preprocessing outputs.*

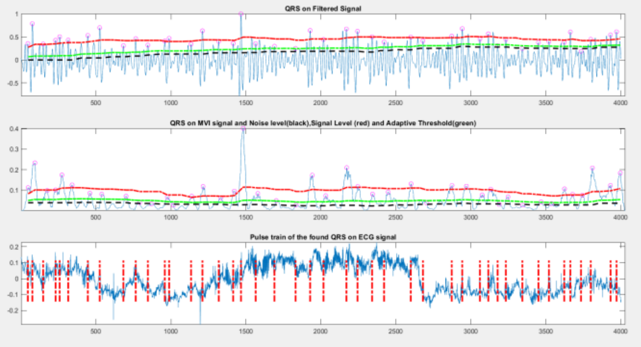

Fig. 7. *ECG Sample 3 Decision Rule outputs*

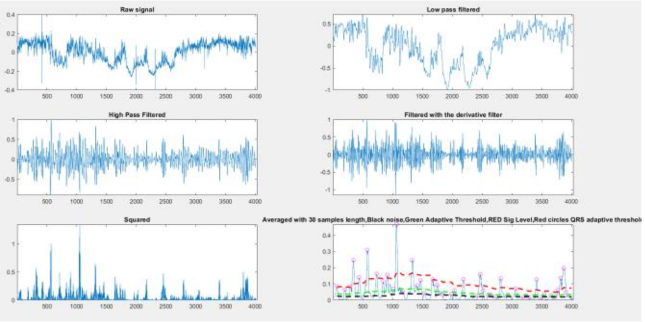

Fig. 8. *ECG Sample 4 Preprocessing outputs.*

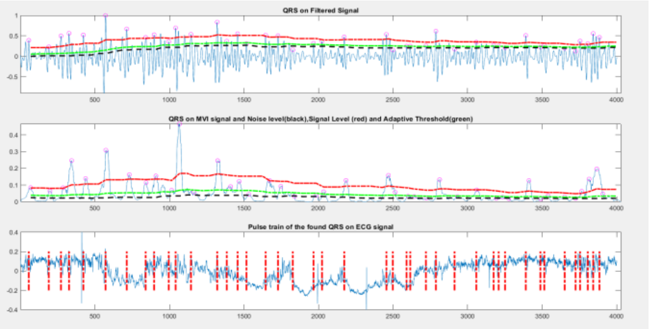

Fig. 9. *ECG Sample 4 Decision Rule outputs*

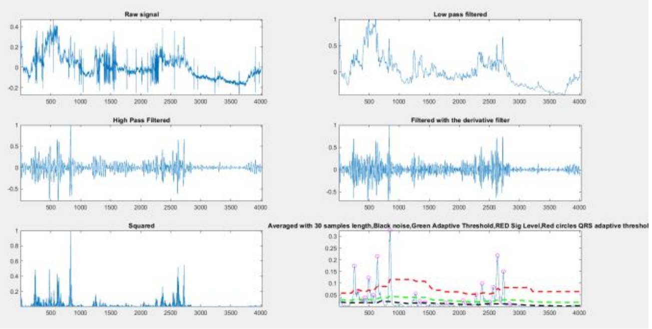

Fig. 10. *ECG Sample 5 Preprocessing outputs.*

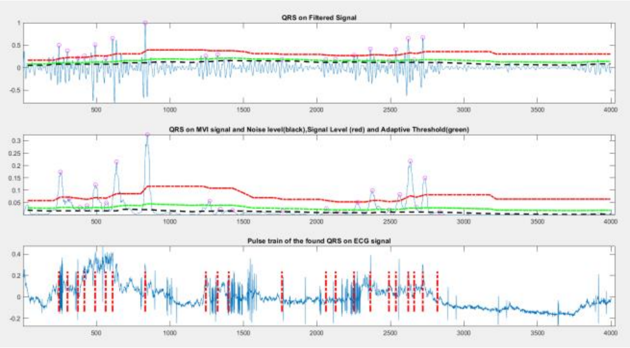

Fig. 11. *ECG Sample 5 Decision Rule outputs*

Table I. PeVRformance of CGA-based ECG DM on Testing 20% from ECG Data Set when Mutation Ratio=0.5, Crossover Ratio =0.5, and Sampling Frequency=200.

GN	*PS*	*Fitness*	*ACC*	*TPR*	*TNR*	*PPV*	*f_score*
5	25	91.837	76.471	0.813	0.684	0.813	0.813
5	50	91.837	76.471	0.813	0.684	0.813	0.813
5	55	91.837	76.471	0.813	0.684	0.813	0.813
5	65	91.837	76.471	0.813	0.684	0.813	0.813
10	5	70.000	76.471	0.813	0.684	0.813	0.813
10	10	91.837	76.471	0.813	0.684	0.813	0.813
10	15	84.000	84.000	0.875	0.778	0.875	0.875
10	35	76.471	84.000	0.875	0.778	0.875	0.875
10	50	91.837	76.471	0.813	0.684	0.813	0.813
10	55	90.000	84.000	0.875	0.778	0.875	0.875

Table 3 shows the performance of CGA-based ECG DM on testing 20% of the ECG dataset. In this case, the mutation and crossover ratios were set to 0.5, and the sampling frequency was 200. The results indicate that the algorithm achieved high fitness scores, with FS_Fitness ranging from 84% to 91%. The ACC of the algorithm ranged from 76.471% to 84%, whereas the TPR, TNR, positive predictive value (PPV), and F-score were consistently ranging from 0.813 to 0.875

Table 4 shows the performance of DTGA-based ECG DM on testing 20% of the ECG dataset. The mutation and crossover ratios were set to 0.5, the threshold was set to 18, and the sampling frequency was 200. The results indicate that the algorithm achieved high fitness scores, with FS_Fitness ranging from 80% to 96%. The ACC of the algorithm ranged from 69.231% to 91.837%, whereas the TPR, TNR, PPV, and F-score were consistently high across all experiments. Notably, the algorithm achieved a maximum ACC of 96% in Experiment 4 with a population size of 65 and generation number of 5. DTGA is a potent feature selection approach for the MIT-BIH Arrhythmia Database and other comparable datasets. By identifying the most informative features, this method can aid in the development of more accurate and efficient algorithms for detecting and classifying arrhythmias based on ECG data.

Table III. Performance of DTGA based ECG DM on Testing 20% from ECG Data Set when Mutation Ratio=0.5, Crossover Ratio =0.5, Threshold =18 and Sampling Frequency=200.

GN	*PS*	*Fitness*	*ACC*	*TPR*	*TNR*	*PPV*	*f_score*
5	25	91.837	76.471	0.813	0.684	0.813	0.813
5	50	91.837	91.837	0.938	0.882	0.938	0.938
5	55	80.000	84.000	0.875	0.778	0.875	0.875
5	65	96.000	91.837	0.938	0.882	0.938	0.938
10	5	76.471	69.231	0.750	0.600	0.750	0.750
10	10	95.000	91.837	0.938	0.882	0.938	0.938
10	15	91.837	84.000	0.875	0.778	0.875	0.875
10	35	91.837	76.471	0.813	0.684	0.813	0.813
10	50	94.000	91.837	0.938	0.882	0.938	0.938
10	55	91.837	84.000	0.875	0.778	0.875	0.875

Table 5 presents a comparison between the performance of CGA and DTGA algorithms in ECG DM. The population size for CGA is 15, while DTGA has a population size of 10, with both algorithms running for 10 generations.

The results show that DTGA outperformed CGA, achieving a higher best fitness score of 96 compared with CGA's 91.837. The ACC of DTGA was also higher at 91.837%, while CGA had an ACC of 84%. Despite having a smaller population size, DTGA achieved better results than CGA, indicating that DTGA is a more efficient algorithm for ECG DM. The total populations for CGA and DTGA were 150 and 100, respectively, and the time taken for DTGA was shorter at 41 s compared with CGA's 61.5 s. These findings suggest that DTGA is a promising approach for ECG DM, with the potential to improve the ACC and efficiency of ECG analysis.

TABLE V. COMPARATIVE RESULTS BETWEEN CGA AND DTGA

	CGA	***DTGA***
Population size	15	10
Generation Numbers	10	10
Best fitness	91.837	95
ACC	84.000	91.837
Total Populations	150	100
Time Complexity	61.5	41

In general, a dynamic thresholding principle within GA contributes to a plurality of populations rather than a classical GA. This variety helps to achieve optimum solutions for the right fitness functions. Moreover, in binary chromosomes, the assortment of all possible binary states offers a wide variety over classical representation. To converge the chromosome individuals toward optimal solutions, the dynamic thresholding is implemented. Table 6 presents the results of five different studies that use various machine learning techniques, including GA, DT, and ensemble classification models, to classify ECG signals and diagnose CVDs. The datasets used in these studies include the Heart Disease Data Set-UCI repository, Mobile HEALTH (MHEALTH) database, MIT-BIH Arrhythmia, and MIT-BIH Supraventricular Arrhythmia databases, and nine standardized classification tasks. The results show that the proposed methods achieve high ACC rates that range from 78.76% to 97.57%, with some methods outperforming others. For instance, the ODKV classifier had an ACC rate higher than 98%, while the PSO-based, ACO-based, and ABC-based methods had lower classification accuracies compared with the GA-based method that used community detection. Furthermore, the DTGA algorithm outperformed the CGA algorithm in terms of best fitness score and ACC, achieving a score of 96 and an ACC of 91.837%, respectively. Overall, these studies demonstrate the effectiveness of machine learning techniques in diagnosing CVDs and classifying ECG signals.

TABLE VI. COMPARISON BETWEEN THE PROPOSED METHOD AND THE METHOD USED IN PREVIOUS STUDIES

Study	Focus –Methodologies - Dataset	Results
[15]	• Ensemble Classification Model for Heart Disease Diagnosis • Feature selection strategy, GA, Ensemble Classification, and biological test values • Heart Disease Data Set-UCI repository	ACC rate of 97.57%
[8]	• Categorizing Cardiac Arrhythmias • GA and DT • Data Set: N/A	Two-class mode: ACC of 86.96%; 16-class mode: ACC of 78.76%
[16]	• Classification of ECG Signals with ODKV Classifier • ODKV classifier, ANF pre-processing technique, ANN-kNN, SVM-kNN, GB-SVNN, CNN • MHEALTH database	ACC of ANN-kNN is 80%, SVM-kNN is 83% and GB-SVNN is 96.9%. CNN is 93% and ODKV had ACC higher than 98%
[17]	• ECG Classification with Machine Learning Libraries and Apache Spark • GDB Tree algorithm, Random Forest • MIT-BIH Arrhythmia and MIT-BIH Supraventricular Arrhythmia databases	Overall ACC of 96.75% and 97.98% for binary identification using GDB Tree algorithm and Random Forest, respectively. Random Forest algorithm had a precision rate of 98.03% in multiclass classification.
[18]	• Feature Selection with GA Based on Community Detection • GA based on community detection, PSO, ACO, ABC algorithms • Nine standardized classification Datasets	proposed method obtained a 93.99% classification ACC. In contrast, for PSO-based, ACO-based, and ABC-based methods, these values were reported 92.54%, 91.81%, and 90.35%, correspondingly
Proposed GA and DTGA	• ECG Feature Selection in CVD Diagnosis • GA and DTGA • MIT-BIH Arrhythmia Database	The results show that DTGA outperformed CGA, achieving a higher best fitness score of 96 compared to CGA's 91.837 The ACC of DTGA was also higher at 91.837%, while CGA had an ACC of 84%.

CONCLUSION

For this reason, the DTGA is a good feature selection method using ECG data to diagnose heart disease. Our research reveals that DTGA is successful in selecting relevant ECG features, which ultimately improves the ACC and efficiency in the diagnosis of CVD. The possible benefits of utilizing DTGA in clinical practice include a reduction in the amount of time spent diagnosing patients and an increase in the level of ACC achieved when doing so. To improve the detection and treatment of cardiovascular illness, the application of more sophisticated computational algorithms, such as DTGA, offers a great deal of potential. The clinical value of DTGA should be investigated further in future studies, preferably in larger patient groups and throughout a variety of healthcare settings. Numerous metrics are used to evaluate the DM process. The results indicate that DTGA is a more efficient algorithm for ECG DM compared with CGA. DTGA outperformed CGA in terms of best fitness score, ACC, and time taken. The findings suggest that DTGA is a promising approach for ECG DM and has the potential to improve the ACC and efficiency of ECG analysis. Even though the use of DTGA algorithm for ECG DM is promising, several limitations must be considered. To address the study's limitations, researchers could alter the scale and diversity of the dataset used for scalability testing. This would help determine whether the algorithm’s performance remains stable as data volume and complexity grow. Evaluating sensitivity and specificity in addition to ACC and efficiency could provide a more nuanced comprehension of the algorithm’s effectiveness in detecting specific types of abnormalities. In addition, testing the algorithm in a clinical setting with additional variables could help determine its applicability in real-world settings. Lastly, comparing DTGA with other ECG data-mining algorithms could assist in determining the most effective strategy. Even though DTGA outperformed CGA in this study, additional research and algorithm comparisons are required to determine the most effective method for ECG DM.

REFERENCES

[1] Al Ameen, M., Liu, J. and Kwak, K., 2012. Security and privacy issues in wireless sensor networks for healthcare applications. *Journal of medical systems*, *36*(1), pp.93-101.

[2] Act, A., 1996. Health insurance portability and accountability act of 1996. *Public law*, *104*, p.191.

[3] Edward Jero, S., Ramu, P. and Ramakrishnan, S., 2014. Discrete wavelet transform and singular value decomposition based ECG steganography for secured patient information transmission. *Journal of medical systems*, *38*(10), pp.1-11.

[4] Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M. and Webster, S., 2000. An investigation of machine learning based prediction systems. *Journal of systems and software*, *53*(1), pp.23-29.

[5] O'Neill, M., Vanneschi, L., Gustafson, S. and Banzhaf, W., 2010. Open issues in genetic programming. *Genetic Programming and Evolvable Machines*, *11*(3-4), pp.339-363.

[6] Kobashigawa, J., Youn, H.S., Iskander, M. and Yun, Z., 2009, June. Comparative study of genetic programming vs. neural networks for the classification of buried objects. In *2009 IEEE Antennas and Propagation Society International Symposium* (pp. 1-4). IEEE.

[7] Brezocnik, M., Kovacic, M. and Gusel, L., 2005. Comparison between genetic algorithm and genetic programming approach for modeling the stress distribution. *Materials and Manufacturing Processes*, *20*(3), pp.497-508.

[8] Ayar, M. and Sabamoniri, S., 2018. An ECG-based feature selection and heartbeat classification model using a hybrid heuristic algorithm. *Informatics in Medicine Unlocked*, *13*, pp.167-175.

[9] Guo, H., Jack, L.B. and Nandi, A.K., 2005. Feature generation using genetic programming with application to fault classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *35*(1), pp.89-99.

[10] Wang, L., Tang, F. and Wu, H., 2005. Hybrid genetic algorithm based on quantum computing for numerical optimization and parameter estimation. *Applied Mathematics and Computation*, *171*(2), pp.1141-1156.

[11] Doan, K., Quang, M.N. and Le, B., 2017, December. Applied cuckoo algorithm for association rule hiding problem. In *Proceedings of the Eighth International Symposium on Information and Communication Technology* (pp. 26-33).

[12] Kenthapadi, K., Mironov, I. and Thakurta, A.G., 2019, January. Privacy-preserving data mining in industry. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 840-841).

[13] Dorigo, M. and Gambardella, L.M., 1997. Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Transactions on evolutionary computation*, *1*(1), pp.53-66.

[14] Jothi, N. and Husain, W., 2015. Data mining in healthcare–a review. *Procedia computer science*, *72*, pp.306-313.

[15] Abdollahi, J. and Nouri-Moghaddam, B., 2022. A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation. *Iran Journal of Computer Science*, *5*(3), pp.229-246.

[16] Shana, J. and Venkatachalam, T., 2021. Efficient Data-Mining Classification Approach For Ecg Data In Health Care Application.

[17] Alarsan, F.I. and Younes, M., 2019. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. *Journal of big data*, *6*(1), pp.1-15.

[18] Rostami, M., Berahmand, K. and Forouzandeh, S., 2021. A novel community detection based genetic algorithm for feature selection. *Journal of Big Data*, *8*(1), pp.1-27.

[19] Tuba, E., Strumberger, I., Bezdan, T., Bacanin, N. and Tuba, M., 2019. Classification and feature selection method for medical datasets by brain storm optimization algorithm and support vector machine. *Procedia Computer Science*, *162*, pp.307-315.

[20] Yan, K., Ma, L., Dai, Y., Shen, W., Ji, Z. and Xie, D., 2018. Cost-sensitive and sequential feature selection for chiller fault detection and diagnosis. *International Journal of Refrigeration*, *86*, pp.401-409.

[21] Li, S., Tang, C., Liu, X., Liu, Y. and Chen, J., 2019. Dual graph regularized compact feature representation for unsupervised feature selection. *Neurocomputing*, *331*, pp.77-96.

[22] Jayaraman, V. and Sultana, H.P., 2019. Artificial gravitational cuckoo search algorithm along with particle bee optimized associative memory neural network for feature selection in heart disease classification. *Journal of Ambient Intelligence and Humanized Computing*, pp.1-10.

[23] Zhang, Y., Gong, D.W., Gao, X.Z., Tian, T. and Sun, X.Y., 2020. Binary differential evolution with self-learning for multi-objective feature selection. *Information Sciences*, *507*, pp.67-85.

[24] Emary, E., Zawbaa, H.M. and Hassanien, A.E., 2016. Binary grey wolf optimization approaches for feature selection. *Neurocomputing*, *172*, pp.371-381.

[25] Neggaz, N., Ewees, A.A., Abd Elaziz, M. and Mafarja, M., 2020. Boosting salp swarm algorithm by sine cosine algorithm and disrupt operator for feature selection. *Expert Systems with Applications*, *145*, p.113103.

[26] Rostami, M., Berahmand, K. and Forouzandeh, S., 2020. A novel method of constrained feature selection by the measurement of pairwise constraints uncertainty. *Journal of Big Data*, *7*(1), pp.1-21.

[27] Arowolo, M.O., Abdulsalam, S.O., Isiaka, R.M. and Gbolagade, K.A., 2017. A hybrid dimensionality reduction model for classification of microarray dataset. *Int J Inf Technol Comput Sci*, *9*(11), pp.57-63.

[28] Neri, L., Oberdier, M.T., Augello, A., Suzuki, M., Tumarkin, E., Jaipalli, S., Geminiani, G.A., Halperin, H.R. and Borghi, C., 2023. Algorithm for Mobile Platform-Based Real-Time QRS Detection. *Sensors*, *23*(3), p.1625.

[29] Liao, T., Socha, K., de Oca, M.A.M., Stützle, T. and Dorigo, M., 2013. Ant colony optimization for mixed-variable optimization problems. *IEEE Transactions on Evolutionary Computation*, *18*(4), pp.503-518.

[30] Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K. and Stanley, H.E., 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, *101*(23), pp.e215-e220.

[31] Moody, G.B. and Mark, R.G., 1990, September. The MIT-BIH arrhythmia database on CD-ROM and software for use with it. In *[1990] Proceedings Computers in Cardiology* (pp. 185-188). IEEE.

[32] Liao, T., Socha, K., de Oca, M.A.M., Stützle, T. and Dorigo, M., 2013. Ant colony optimization for mixed-variable optimization problems. *IEEE Transactions on Evolutionary Computation*, *18*(4), pp.503-518.

[33] Loukhaoukha, K., Chouinard, J.Y. and Taieb, M.H., 2011. Optimal Image Watermarking Algorithm Based on LWT-SVD via Multi-objective Ant Colony Optimization. *J. Inf. Hiding Multim. Signal Process.*, *2*(4), pp.303-319.

[24] Zhang, Q., Xu, X. and Liang, Y.C., 2006. An improved artificial immune algorithm with a dynamic threshold. *Journal of Bionic Engineering*, *3*(2), pp.93-97.

[25] Dasgupta, D. ed., 2012. *Artificial immune systems and their applications*. Springer Science & Business Media.

[26] Darwish, S.M., Shendi, T.A. and Younes, A., 2019. Chemometrics approach for the prediction of chemical compounds' toxicity degree based on quantum inspired optimization with applications in drug discovery. *Chemometrics and Intelligent Laboratory Systems*, *193*, p.103826.