An Effective Hybrid Feature Selection Method Based on an Improved Artificial GTO Algorithm for Medical Datasets

Abd Al-Baset Rashed Saabia [1], Mondher Frikha [2]

[1] *ENETCOM, ATISP Research Lab, University of Sfax, Sfax, Tunisia*
[2] *Department of Electronics, National School of Electronics and Telecommunications of Sfax, University of Sfax, Tunisia*

Abstract **– Feature subset selection is considered as the most essential pre-processing step. Metaheuristic approaches may be employed to discover a solution to difficulties in feature selection, which can be viewed as an optimisation problem. The aim of the system is to provide a hybrid binary metaheuristic algorithm that combines gorilla troop optimisation and genetic algorithm to handle the feature selection issue effectively. This new method is called GTO-GA. To ensure that the optimisation technique converges fast and properly and to enhance the exploration process, the GA were used. The suggested technique is tested for stability and robustness using 16 medical datasets taken from the Kaggle and UCI repositories. To evaluate the chosen features' performance in classification issues further. The results show that the algorithm outperforms 10 top-tier optimisation methods, including PSO, ALO, the original GTO and the SCA algorithm. The results highlighted the statistical difference, superiority and importance of the suggested feature selection strategies.**

Keywords **– Data mining, hybrid feature selection, machine learning, genetic algorithm, artificial gorilla troops optimizer.**

DOI: 10.18421/TEM134-09
https://doi.org/10.18421/TEM134-09

Corresponding author: Abd Al-Baset Rashed Saabia,
ENETCOM, ATISP Research Lab, University of Sfax, Sfax, Tunisia
Email: abdalbast.rsheed@gmail.com

Received: 28 April 2024.
Revised: 29 August 2024.
Accepted: 14 October 2024.
Published: 27 November 2024.

1. Introduction

Data mining is an umbrella phrase for the practice of semi-automatically exploring massive datasets for valuable patterns [1].

It is an endeavour that involves finding patterns and rules in data, which is similar to the knowledge discovery process in statistical analysis or artificial intelligence (also known as machine learning) [2].

The accuracy and performance of the system are often negatively affected by high-dimensional datasets because of the challenge of dimensionality. Notably, classifying with high-dimensional features requires a large amount of time and difficult computations [3]. Feature selection (FS) can help with data that have a lot of dimensions. When dealing with high-dimensional datasets, feature selection becomes an essential part of data mining. As a standard procedure in machine learning, feature selection involves extracting subsets of data features for use by a learning algorithm. To increase machine learning performance, feature selection seeks to minimise the number of dimensions that contribute most to accuracy [4]. The optimal subset comprises these dimensions [4]. Wrapper, filter, and embedded models for feature selection approaches are the three main types [5]. Filter-based approaches are rankers. Features are ranked and assessed based on metrics that are directly extracted from the data without the need for predictors. Embedded techniques are an extension of the approach that works with linear classifiers, such as support vector machines (SVMs). The FS approach is a search issue. Several types of search algorithms can be used, including heuristic, probabilistic, exhaustive, and automated hybrid. A heuristic search for the best neighbour takes far less time and effort, but it only looks in one specific direction [5]. Heuristics are often used by a diverse group of computer scientists to solve practical challenges [9].

In addition, heuristics can effectively handle other features of big data, such as diversity and velocity. The two main requirements of a heuristic-based search strategy are the exploration and the exploitation strategy [6].

The selection of qualities has been addressed using a variety of heuristics, and an overview of these methods is available in [7]. One of the most popular metaheuristics is the genetic algorithm (GA) [8]. The heuristic-based approaches that have been proposed are either population-based or single-based. The former involves simulated annealing [9], hill climbing [10], tabu search [11], and harmony search [12]. One of the main problems with hill climbing is that it frequently falls into local optima and is also highly sensitive to the initial solution. From the original mimetic algorithm [10] and genetic programming [13] to the bat algorithm [14] and ACO [15], every method has been employed in population-based heuristics. For data-hard combinatorial optimisation issues, hybrid heuristic approaches have shown effectiveness. In addition, several methods use local based search algorithms as internal operators to strike a balance between intensification and diversity. Aggregation approaches, such as ACO with GA [17] and hybrids between GA and PSO [16] have also been suggested [18]. Moreover, in [19], the most current hybrid approaches have been proposed, such as the suggested combination of GA and PSO with the SVM as the classifier [20].

For FS, no heuristic-based approach can guarantee a perfect solution. However, the current search space technique may be enhanced to acquire search areas with excellent performance. This work presented a novel hybrid approach called GTO-GA to improve the exploitation ability of the original GTO. It is based on a hybridisation of global search and local search algorithms and is the basis for most of the efforts to develop a predictive model.

The remaining parts are structured as follows: Sections 2 and 3 provide a concise overview of the necessary algorithms for GA, GTO, and hybrid feature selection. Datasets, parameters, and experimental findings are reported in Section 5, whereas Section 4 thoroughly describes the basic motivation and recommended strategy of this study. The work, the primary success and the recommendations for future research projects are finally addressed in Section 6.

2. Artificial Gorilla Troops Optimiser

The following explanation is the philosophy behind the adaptable GTO, which is a kind of metaheuristic algorithm. The GTO metaheuristic algorithm, which draws inspiration from the group behaviours of gorillas, is shown here.

It provides comprehensive mathematical algorithms for the exploration and exploitation phases [21]. Two crucial steps in this method, the exploration and exploitation stages are modelled after the way gorillas hunt for the optimal solution. In the exploratory phase, researchers discovered that gorillas follow the lead of a dominant male, called a silverback and that a gorilla sometimes opts to spend time alone instead of with the group.

Separated gorillas may venture to different areas of the forest, where they can encounter other primates. Every gorilla is considered a potential solution by the algorithm. The one that performs optimally at each optimisation stage is referred to as a silverback gorilla. Each of the three primary processes used during the exploration phase—relocation, exploration, and movement to a new site or closer to another gorilla is represented by equation (1), which elucidates these mechanisms [21].

$$GX(t+1) = \begin{cases} (UB - LB) \times r_i + LB, & rand < p, \\ (r_2 - C) \times X_r(t) + L \times H, & rand \geq 0.5, \\ X(i) - L \times (L \times \Big(X(t) - GX_r(t) + r_3 \times \big(X(t) - GX_r(t)\big)\Big), & rand < 0.5, \end{cases} \quad (1)$$

Where, $GX(t+1)$ is the new gorilla position, and $X(t)$ is the current position. Moreover (r_1, r_2, r_3) are random values ranging between 0 and 1, that update in each iteration. Finally, C, L and H are calculated using Equations (2), (3) and (4), respectively.

$$C = F \times (1 - \frac{It}{MaxIt}), \quad (2)$$

$$L = C \times l, \quad (3)$$

and

$$H = Z \times X(t), \quad (4)$$

Where (It) is the current iteration, $MaxIt$ is the max iteration , and F is compute by cos(2 ×r_4) + 1.

In equation (3), l is a random value and L is used to simulate the silverback leadership. Z in equation (4) is a random value in the problem dimensions.

Following the silverback and competing for adult females are two behaviours that emerge during the exploitation stage. Equation (5) is used to simulate this first behaviour and equation (6) is used to simulate the second.

Being the group's alpha male, the silverback gorilla makes all the critical decisions, plots the best routes, and directs the others to the most nutritious food.To ensure the safety and well-being of the group, all the gorillas submit to the silverback and follow his or her orders.

The silverback gorilla may become weak and elderly, thus allowing the black gorilla to take over; alternatively, other gorillas may successfully challenge the silverback and assume leadership [21].

$$GX(t + 1) = L \times M \times (X(t) - X_{silverback}) + X(t), \quad (5)$$

where $X(t)$ is the new gorilla location and $X_{silverback}$ is the best solution (silverback location). and M is determined by

$$(\left|\frac{1}{N}\sum_{I=1}^{N} GX_i(t)\right|^{g})^{\frac{1}{g}}$$

g is given by 2^l.

$$GX(i) = X_{silverback} - (X_{silverback} \times Q - X(t) \times Q) \times A, \quad (6)$$

where $X_{silverback}$ is the best solution (silverback) and $X(t)$ is the current location of gorilla. Q is determined by $Q = 2 \times r_5 - 1.A$ coefficient to detect the degree of violence in conflicts is calculate using $\beta \times E$, where E is given by { N_1, rand ≥ 5, N_2, rand ≤ 5} [21].

3. Genetic Algorithm

According to Holland and Goldberg, GAs were first described in the 1960s [22]. A GA, or a randomised global search, attempts to find solutions to problems by modelling them after evolutionary processes. Natural selection drives the GA to always seek improved solutions, disregarding assumptions such as continuity and unimodality, to reproduce and thrive. The GA generates a population of candidate solutions and has been effectively used to solve several complicated optimisation problems. Thus, the approach demonstrates its superiority over traditional optimisation approaches, particularly in cases when the system being studied contains multiple locally optimal solutions. In most cases, a chromosome is used to encode each answer as a binary string. Once a chromosome has been decoded, its fitness is assessed using a performance function. After the test, a biased roulette wheel is used to pick a couple of the best chromosomes at random to undertake natural-looking genetic operations, such as mutation and crossover.

This evolutionary process will continue until the stopping requirements are met, at which point the stronger chromosomes from the next generation will supersede the weaker ones [22].

4. Methodology

This section presents the design of the proposed hybrid algorithm, which includes the steps for initialization, population generation, and fitness evaluation. The implementation details, such as pseudocode and parameter configurations, are also included.

4.1. Proposed Mode

Feature subset selection uses the GTO to categorise issues using the wrapper-based mode. The feature subset selection strategy in the wrapper-based methodology is based on a few optimising algorithms. Moreover, it uses the classification approach as evidence. The GTO population is thought of as a binary bound of dimension in a binary search issue. A binary version of GTO should be sophisticated when utilised as a feature subset selection strategy. It is assumed that the dataset's dimensions are directly proportional to the length of the one-dimensional vector that represents each population solution. One or zero represents each vector cell: with a value of 1, the matching attribute is selected; with a value of 0, it is ignored.

For the classification task in this article, employed KNN algorithm with (k = 3). KNN is non parametric approach can find the best answers based on Euclidean a distance equation, which is one of the simplest supervised learning approaches [23]. Each, native GTO, ALO and PSO position's search agents are evaluated using the fitness function in Eq. (7), which aims to obtain high classification accuracy with minimum number of selected features in each iteration:

$$F = \alpha * ERR(D) + \beta \frac{|L|}{|T|}, \qquad (7)$$

where ERR(D) is error rate , |L| is the length , |T| is the total number of features ,α and β are constant parameters respectively, α be between 0 and 1, and β = 1–α adopted by [23].

```
INPUT: max iteration ,Search agent number,
upper bound (ub), lower bound (lb), dataset
dimension, , and parameters β and p
OUTPUT: best vector with optimal 20 solutions,
The position of Gorilla and its cost value
  For q=1 to 20
    Initialise the first population Xi (i = 1,
  2, …, N)
    compute the cost value for all current
position .
    while (condition) do
       Update the (C) using Equ. (2)
      Update the (L) using Equ.(3)
  for (each Gorilla (Xi)) do
    Update the position Gorilla using Equ (1)
    Update the search space of current position
  using proposed mutation Agent.
  end for
    compute the cost value of Gorilla
    if (GX is better than X) , replace them
      Set (X silverback) as the best location
  for (each_Gorilla (Xi)) do
    if (|C| >= one) then
      Update the current position using Equ.
  (5)
    Else
      Update the current position using Equ.
  (6)
    End if
    Update best current position using proposed
  crossover agent.
  end for
   compute the objective value of new position
   if New Solution are better than previous
  solutions, replace them
   Set (Xsilverback) as the position of
  silverback (best location)
end while
  Return (XBestGorilla), (bestobjective value)
  End For
```

The two main steps of the GTO algorithm, which is based on how gorillas look for the optimal answer, are the exploration and exploitation stages. Each gorilla in the algorithm symbolizes a potential solution. The gorilla who demonstrates the highest level of performance at each step of optimization is referred to as the silverback gorilla. Throughout the exploration phase, the silverback gorilla undergoes two pivotal stages in which the genetic algorithm is employed to guarantee the improvement and optimization of the answer. The mutation function serves as an internal function used by the GTO algorithm to enhance exploration capabilities.

To avoid local minima and keep diversity high, the GTO incorporates a mutation operator that prevents population members from being too similar to one another. This operator allows the GTO to explore various locations. An amplification weight factor, also known as a mutation rate, is user defined and controls the mutation operation. This factor lies between 1 and 0. A typical issue in this area is determining the ideal mutation rate, which should remain low. Meanwhile, a high number for this rate will cause the search to diverge into a random one. Hence, the algorithm cannot converge to an optimal answer.

The proposed crossover is utilised as an internal agent within the GTO algorithm to improve the exploration ability. To improve the worst solution selected by the GTO algorithm, the crossover agent recombines the worst solution with the best one acquired from earlier iterations by employing three crucial formulas: single, double and uniform. The formula that works best is selected based on the roulette. When using wrapper-based mode for issue classification, the GTO is employed in a feature subset selection. The feature subset selection strategy in the wrapper-based methodology is based on a few optimising algorithms, and it uses the classification approach as evidence. The GTO population is thought of as a dimensionality that is bound in a binary search issue. An advanced binary version of the GTO is required when using the algorithm as a feature subset selection strategy.

5. Experimental Results and Discussion

This section discusses the results achieved through the utilization of the wrapper based pattern. Subsequently, the performance of the proposed method is evaluated on different 16 medical datasets, offering valuable insights into its performance. The detailed evaluations and comparisons with other FS approaches will be presented in the following section.

5.1. Datasets and Parameters

Sixteen benchmark dataset from the Kaggle and UCI Machine Learning repositories were used to assess the efficacy of the suggested methods [24]. Given that GTO-GA is a population-based method, it is assumed that every member of the population stands in for a feature index vector. Only the most exceptional person and its fitness were retained in the basic GTO-GA after each iteration were evaluated the remaining solutions based on the quality of the feature set.

This research contrasted the findings with native (GTO) and other algorithms, specifically PSO and ALO, to extract the best features from the whole datasets and demonstrate the performance of all the approaches. Table 2 presents a summary of all the parameters.

Table 1. Description of all utilize datasets

Dataset	**Feature number**	**Instances number**
Breast_cancer	9	699
Breast_EW	30	699
Colon	2001	62
fetal_health	21	2126
Heart_EW	13	270
Heart Failure	13	271
Ionosphere_EW	34	351
Leukaemia	7130	28
Lymphography	19	148
Brain Tumour	7466	36
Prostate_Cancer	9	100
SonarEW	60	208
SpectEW	22	267
Stroke	11	110
Lung cancer	57	32
Hepatitis C	13	615

Table 2. Parameters setting

Parameters	**Value**
Repetitions of runs	20
Iteration number	100
No.of search agent	5
Dimension	attributes number
Domain reang	0 ,1
α	0.01
β	1-α
Mutation rate	0.05

A training dataset was used to fit the subset selection approach. In addition, a validation set was used to infer prediction error, and a testing dataset was used to evaluate the final model fairly.

All datasets were divided into these three equal portions. The various optimisation approaches were compared with the suggested GTO-based strategy and utilised the following indicators to find the best one. Average selection size, Best, worst, mean, and standard deviation equation, these are represented as the following equations:

$$\text{Bestfit} = min_i^M \ = 1g_*^i, \tag{8}$$

$$\text{Meanfit} = \frac{1}{\text{M}}\sum_{\text{i}=1}^{\text{M}} \text{g}_*^{\text{i}} \tag{9}$$

$$\text{Worstfit} = max_i^M = 1g_*^i \tag{10}$$

$$\text{Std} = \sqrt{\frac{1}{\text{M}-1}}\sum(g_*^i - \text{mean})2 \tag{11}$$

$$\text{AVGselectionSIZE} = \frac{1}{\text{M}}\sum_{\text{i}=1}^{\text{M}} \frac{\text{size}(\text{g}_*^{\text{i}})}{\text{D}} \tag{12}$$

According to the suggested method, GTO-GA functions as an internal operator by incorporating a mutation operator into GTO., statistically (best, worst, mean) fitness, classification accuracy, average selection size and standard deviation (std) were the assessment criteria used to compare GTO-GA to the native GTO and other feature selection approaches, such as ALO and PSO. The MATLAB framework were used to calculate all of the findings from the assessment criteria with an average of 20 runs. Two goals, classification accuracy and average selected size were used to evaluate the performance of GTO-GA to that of native GTO. Table 3 shows that the hybrid mutation agent with in GTO performs much better than the native GTO when comparing the classification accuracy and the number of the selected features. In terms of classification accuracy, GTO-GA outperforms the original GTO across all datasets. Over a wide range of datasets, GTO-GA also achieves better results than the original GTO with regard to the average selected features. The performance of GTO-GA was also compared with that of PSO and ALO, which are two other relevant techniques from the literature. The accuracy performance when utilising entire features is lower than when using the suggested methods to select optimal features, as shown in Table 3. On every dataset, GTO-GA beats all other optimisers.

Table 3. Comparison and classification accuracy of proposed approach and other approaches in literature.

Medical Datasets	GTO	GTO_GA	SCA	ALO	PSO
breast-cancer	0.923509	0.939298	0.926316	0.922105	0.917895
BreastEW	0.942807	0.958070	0.942281	0.939649	0.930877
Colon	0.740323	0.772581	0.735484	0.730645	0.695161
fetal_health	0.892897	0.906773	0.894309	0.891110	0.887770
HeartEW	0.792963	0.828889	0.801852	0.792963	0.754074
Heart Failure	0.776000	0.832333	0.773000	0.773667	0.687333
IonosphereEW	0.880398	0.905682	0.878693	0.877557	0.853693
Leukaemia	0.894444	0.912500	0.890278	0.886111	0.865278
Lymphography	0.802134	0.856081	0.784996	0.782275	0.745946
Brain Tumour	0.952778	0.961333	0.950000	0.952778	0.925000
Prostate_Cancer	0.773000	0.811000	0.785000	0.780000	0.738000
SonarEW	0.858173	0.906731	0.851442	0.844712	0.836538
SpectEW	0.792537	0.838060	0.803731	0.783955	0.773134
Stroke	0.933386	0.943033	0.925714	0.930372	0.917417
Lung cancer	0.859375	0.937500	0.862500	0.862500	0.740625
Hepatitis C	0.920130	0.930195	0.921591	0.918344	0.918344

Summary findings from the best, worst, and mean measurements for all datasets and all 20 approach runs are shown in Tables 4. Various FS approaches, including native GTO, were contrasted with the GTO-GA approach. On every dataset, GTO-GA achieved better results than native GTO, ALO and PSO according to statistical best, mean, and worst criteria.

Across all datasets, the suggested method achieved the lowest values. Table 5, which displays the average selected size, shows that GTO-GA outperforms the other methods across 14 datasets. Table 6 shows that across 11 datasets, GTO-GA outperforms other methods in terms of the average selected size.

Table 4. Statistical best, worst, and mean fitness results

Medical Datasets	Statistical Fitness	GTO	GTO_GA	SCA	ALO	PSO
breast-cancer	best	0.052825	0.047158	0.051825	0.062579	0.053632
	worst	0.102789	0.086035	0.099123	0.102596	0.099316
	mean	0.081743	0.063361	0.078981	0.083116	0.080394
BreastEW	best	0.046211	0.029982	0.049158	0.055158	0.045877
	worst	0.084895	0.056965	0.076947	0.081088	0.081421
	mean	0.063871	0.047527	0.063859	0.066714	0.063083
Colon	best	0.132852	0.100391	0.132737	0.164747	0.164547
	worst	0.422541	0.387786	0.420146	0.420266	0.451947
	mean	0.262547	0.229726	0.267152	0.271872	0.263627
fetal_health	best	0.084525	0.072987	0.094182	0.096113	0.093251
	worst	0.138886	0.114759	0.133229	0.129641	0.129072
	mean	0.112182	0.096369	0.110935	0.113601	0.110569
HeartEW	best	0.177128	0.137385	0.167487	0.157077	0.162462
	worst	0.259333	0.219590	0.252000	0.259333	0.275949
	mean	0.213236	0.176554	0.203590	0.212121	0.220074
Heart Failure	best	0.162567	0.114700	0.161733	0.162567	0.162567
	worst	0.334167	0.219467	0.349033	0.270667	0.335000
	mean	0.226843	0.169032	0.230147	0.229278	0.210020
IonosphereEW	best	0.092316	0.062721	0.089375	0.077831	0.079890
	worst	0.173750	0.158934	0.182941	0.173162	0.173750
	mean	0.124406	0.098846	0.125550	0.127072	0.128445
Leukaemia	best	0.032457	0.032191	0.059863	0.032572	0.032442
	worst	0.172488	0.170431	0.197440	0.197501	0.170047
	mean	0.110091	0.091590	0.113704	0.117718	0.112361
Lymphography	best	0.151051	0.089159	0.152718	0.103093	0.153273
	worst	0.272568	0.206787	0.285390	0.277012	0.260300
	mean	0.201804	0.148813	0.218743	0.222687	0.211243

Brain Tumour	best	0.004836	0.004622	0.004910	0.004931	0.004892
	worst	0.169970	0.115770	0.169960	0.169959	0.170013
	mean	0.051698	0.045993	0.054441	0.051965	0.054574
Prostate_Cancer	best	0.141100	0.139850	0.141100	0.162150	0.142350
	worst	0.320550	0.259900	0.284700	0.341600	0.323050
	mean	0.229418	0.189735	0.217913	0.223112	0.215808
SonarEW	best	0.104192	0.073801	0.110045	0.110212	0.112545
	worst	0.179179	0.130083	0.223609	0.225276	0.203571
	mean	0.147042	0.099162	0.153005	0.160452	0.145757
SpectEW	best	0.157877	0.122300	0.145373	0.179925	0.154579
	worst	0.256194	0.227551	0.251079	0.275516	0.265855
	mean	0.210684	0.165935	0.199874	0.219793	0.203034
Stroke	best	0.049659	0.044397	0.051759	0.048659	0.051759
	worst	0.085045	0.080395	0.085432	0.087082	0.086757
	mean	0.069548	0.058747	0.078193	0.073182	0.069506
Lung cancer	best	0.066696	0.002500	0.065446	0.065804	0.004107
	worst	0.254643	0.128929	0.252321	0.252321	0.252500
	mean	0.145237	0.065982	0.141313	0.141107	0.144692
Hepatitis C	best	0.061310	0.057262	0.069405	0.072500	0.066905
	worst	0.108810	0.095000	0.112738	0.119286	0.118452
	mean	0.086571	0.074440	0.085625	0.088548	0.085202

Table 5. Average selected size for the different approaches

Medical Datasets	GTO	GTO_GA	SCA	ALO	PSO
breast-cancer	0.601667	0.326667	0.603333	0.600000	0.765000
BreastEW	0.725000	0.601667	0.671667	0.696667	0.855000
Colon	0.546600	0.458100	0.528125	0.521025	0.712425
fetal_health	0.615000	0.407500	0.630000	0.580000	0.767500
HeartEW	0.826923	0.715385	0.742308	0.715385	0.811538
Heart Failure	0.508333	0.304167	0.541667	0.520833	0.762500
IonosphereEW	0.600000	0.547059	0.545588	0.585294	0.836765
Leukaemia	0.559139	0.496451	0.507904	0.496774	0.731140
Lymphography	0.591667	0.633333	0.588889	0.713889	0.805556
Brain Tumour	0.494762	0.474253	0.494052	0.521507	0.703664
Prostate_Cancer	0.468750	0.262500	0.506250	0.531250	0.712500
SonarEW	0.663333	0.682500	0.593333	0.671667	0.748333
SpectEW	0.529545	0.561364	0.556818	0.590909	0.759091
Stroke	0.360000	0.235000	0.465000	0.425000	0.675000
Lung cancer	0.601786	0.410714	0.518750	0.498214	0.771429
Hepatitis C	0.750000	0.533333	0.800000	0.770833	0.837500

Table 6. Standard deviation for the different approaches

Medical Datasets	GTO	GTO_GA	SCA	ALO	PSO
breast-cancer	0.012901	0.011252	0.012314	0.011624	0.012510
BreastEW	0.011184	0.006493	0.007024	0.007357	0.009350
Colon	0.085653	0.085078	0.081765	0.081505	0.085975
fetal_health	0.013738	0.009041	0.010288	0.010169	0.008977
HeartEW	0.021339	0.021443	0.024696	0.025271	0.028031
Heart Failure	0.043403	0.023437	0.048374	0.031349	0.040711
IonosphereEW	0.024534	0.024608	0.026135	0.028166	0.024152
Leukaemia	0.038785	0.039290	0.042310	0.043594	0.045327
Lymphography	0.035564	0.030694	0.030784	0.042511	0.031236
Brain Tumour	0.040993	0.030408	0.043346	0.041052	0.043389
Prostate_Cancer	0.049740	0.036161	0.046037	0.045355	0.051807
SonarEW	0.019584	0.018443	0.027244	0.029740	0.026639
SpectEW	0.029091	0.028721	0.029936	0.029936	0.030297
Stroke	0.013685	0.013050	0.008699	0.013694	0.011110
Lung cancer	0.059504	0.039917	0.065153	0.062159	0.076804
Hepatitis C	0.010775	0.009911	0.011623	0.012629	0.011820

For this study, three massive high-dimensional datasets were drawn: Colon (2001 features), Leukaemia (7130 features) and Brain Tumour (7466 features). The GTO-GA method outperforms competing FS algorithms according to results from several metrics. Regarding the solution found in the standard deviation, this strategy is likewise more superior than native GTO, PSO, and ALO. In general, the final findings show that the GTO-GA model significantly improves the performance of the original GTO. In addition to enhancing exploration capabilities in SSA, GTO-GA investment prevents population similarity, which means that this approach can avoid local minima. The results demonstrate that GTO-GA can successfully identify search spaces with high-performance areas.

6. Conclusion

The goal of this study was to improve classification accuracy by using all dataset characteristics and to utilise the GTO to minimise dimensionality by picking an ideal feature subset based on specified parameters. To address the FS issue in data mining activities, GTO was used in a wrapper-based manner. The native GTO and additional feature selection modes, such as ALO and PSO, were pitted against the suggested GTO-GA method using standard evaluation metrics. All of the characteristics that were considered demonstrated that SSA-GTO performed exceptionally well. The suggested method enhanced the exploration potential inside GTO while preserving variety. GTO can be used in the future as a filter approach or combined with single based algorithm such as SA.

References:

[1]. Raja, R., Nagwanshi, K. K., Kumar, S., & Laxmi, K. R. (Eds.). (2022). *Data mining and machine learning applications*. John Wiley & Sons.

[2]. Bobra, M.G., & Mason, J. (2019). Machine Learning, Statistics, and Data Mining for Heliophysics. In AGU Fall Meeting Abstracts, *2019*.

[3]. Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. Cybernetics and information technologies, *19*(1), 3-26.

[4]. Khamees, M., Albakry, A., & Shaker, K. (2018). Multi-objective feature selection: Hybrid of salp swarm and simulated annealing approach. In *International conference on new trends in information and communications technology applications*, 129-142. Cham: Springer International Publishing.

[5]. Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in biology and medicine*, *112*, 103375.

[6]. Braik, M., Sheta, A., & Al-Hiary, H. (2021). A novel meta-heuristic search algorithm for solving optimization problems: capuchin search algorithm. *Neural computing and applications*, *33*(7), 2515-2547.

[7]. Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and information technologies*, *19*(1), 3-26.

[8]. Sayed, S., Nassef, M., Badr, A., & Farag, I. (2019). A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. *Expert Systems with Applications*, *121*, 233-243.

[9]. Abdel-Basset, M., Ding, W., & El-Shahat, D. (2021). A hybrid Harris Hawks optimization algorithm with simulated annealing for feature selection. *Artificial Intelligence Review*, *54*(1), 593-637.

[10]. Goswami, S., Chakraborty, S., Guha, P., Tarafdar, A., & Kedia, A. (2019). Filter-based feature selection methods using hill climbing approach. *Natural computing for unsupervised learning*, 213-234.

[11]. Benito-Epigmenio, L., Ibarra-Martínez, S., Ponce-Flores, M., & Castán-Rocha, J. A. (2023). Feature Selection: Traditional and Wrapping Techniques with Tabu Search. In *Innovations in Machine and Deep Learning: Case Studies and Applications*, 21-38. Cham: Springer Nature Switzerland.

[12]. Gholami, J., Pourpanah, F., & Wang, X. (2020). Feature selection based on improved binary global harmony search for data classification. *Applied Soft Computing*, *93*, 106402.

[13]. Kılıç, F., Kaya, Y., & Yildirim, S. (2021). A novel multi population based particle swarm optimization for feature selection. *Knowledge-Based Systems*, *219*, 106894.

[14]. Ibrahim, A. M., & Tawhid, M. A. (2021). A new hybrid binary algorithm of bat algorithm and differential evolution for feature selection and classification. *Applications of bat algorithm and its variants*, 1-18.

[15]. Nayar, N., Gautam, S., Singh, P., & Mehta, G. (2021). Ant colony optimization: A review of literature and application in feature selection. *Inventive Computation and Information Technologies: Proceedings of ICICIT 2020*, 285-297.

[16]. Premalatha, K., & Natarajan, A. M. (2009). Hybrid PSO and GA for global maximization. *Int. J. Open Problems Compt. Math*, *2*(4), 597-608.

[17]. Nemati, S., Basiri, M. E., Ghasem-Aghaee, N., & Aghdam, M. H. (2009). A novel ACO–GA hybrid algorithm for feature selection in protein function prediction. *Expert systems with applications*, *36*(10), 12086-12094.

[18]. Kabir, M. M., Shahjahan, M., & Murase, K. (2011). A new local search based hybrid genetic algorithm for feature selection. *Neurocomputing*, *74*(17), 2914-2928.

[19]. Alhenawi, E. A., Alazzam, H., Al-Sayyed, R., AbuAlghanam, O., & Adwan, O. (2022). Hybrid feature selection method for intrusion detection systems based on an improved intelligent water drop algorithm. *Cybernetics and Information Technologies*, *22*(4), 73-90.

[20]. Dokeroglu, T., Deniz, A., & Kiziloz, H. E. (2022). A comprehensive survey on recent metaheuristics for feature selection. *Neurocomputing*, *494*, 269-296.

[21]. Abdollahzadeh, B., Soleimanian Gharehchopogh, F., & Mirjalili, S. (2021). Artificial gorilla troops optimizer: a new nature-inspired metaheuristic algorithm for global optimization problems. *International Journal of Intelligent Systems*, *36*(10), 5887-5958.

[22]. Lambora, A., Gupta, K., & Chopra, K. (2019). Genetic algorithm-A literature review. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, 380-384. IEEE.

[23]. Begum, S., Chakraborty, D., & Sarkar, R. (2015). Data classification using feature selection and kNN machine learning approach. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN),* 811-814. IEEE.

[24]. Asuncion, A., & Newman, D. (2007). UCI machine learning repository. Irvine, CA, USA, 2007.