Received: September 9, 2021. Revised: January 12, 2022. 24

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Hybrid Gene Selection with Mutable Firefly Algorithm for Feature Selection in Cancer Classification

Mohamed Nisper Fathima Fajila[1]* Yuhanis Yusof[1]

[1]*School of Computing, Universiti Utara Malaysia, Malaysia*
* Corresponding author's Email: fajilanisper@gmail.com

Abstract: Cancer microarray analysis is a challenging and crucial task. Though a massive collection of approaches has been presented, there is still room for improvements particularly in obtaining a higher accuracy while using a smaller set of features. Recently, the deployment of optimization algorithms in cancer research domain has shown to have significant impact. In addition, hybrid gene selection methods have outperformed single methods such as filter and wrapper methods due to their capability in producing better classification accuracy with less number of biomarkers. However, existing hybrid methods with firefly optimization algorithm need to be improved to prevent slow convergence and local optimum. Hence, we propose a new hybrid gene selection approach which integrates the correlation-based feature selection filter and mutable firefly algorithm to determine relevant features for cancer classification. The proposed approach is evaluated on four cancer microarray datasets, where all provided 100% accuracy with minimum biomarkers. The results imply that the proposed hybrid feature selection algorithm is a competitive method in high dimension domain and provide insights in healthcare domain via biomarker identification.

Keywords: Cancer classification, Correlation-based feature selection, Firefly algorithm, Gene selection, Microarray analysis.

1. Introduction

Gene selection is not only important for early cancer diagnosis but also produces genetic information regarding biomarkers which in turn are beneficial in cancer treatment. Recently, cancer microarray analysis has become an active field of research [1-5]. Microarray analysis has enabled researchers to classify cancer samples without biological knowledge. Further, cancer microarray experiments yield thousands of gene expressions concurrently. Typically, microarray datasets consist of thousands of genes compared to a small number of samples [6, 7]. This is a major issue in cancer microarray analysis [7]. The redundant and irrelevant genes in microarray datasets result in low classification performance in addition to computational complexity and overfitting. Removing redundant and irrelevant genes while preserving informative genes would results in successful cancer classification while deploying a simple model. Therefore, gene selection is very important in cancer microarray analysis.

Gene selection approaches are of three types: filters, wrappers, and hybrid approaches. Filter approaches are typically being used for preprocessing rather than in gene selection. Nonetheless, filter-based gene selection approaches [8-10] introduced for cancer classification has yet to report significant classification accuracy. At the same time, wrapper-based gene selection approaches [4, 5] tend to produce large gene subsets which in turn is a drawback. However, hybrid approaches [3, 11] which utilize both filter and wrapper approaches seem to produce competing results over filter and wrapper approaches. Nevertheless, gene subset evaluation from microarray search space becomes strenuous with the increase of features [1]. Therefore, obtaining the exact solution seems impossible under certain circumstances. Hence,

optimal solutions are preferable with the help of meta-heuristic algorithms.

Recently, hybrid approaches with meta-heuristic optimization algorithms [2, 12] have been introduced with significant results in gene selection. Artificial Bee Colony (ABC) algorithm [12], Firefly Algorithm (FA) [2] Genetic Algorithm (GA) [13-15], and Particle Swarm Optimization (PSO) algorithm [16] are some of the meta-heuristic algorithms used for cancer classification. Nevertheless, existing wrapper-based meta-heuristic optimization algorithms have not shown significant performance in terms of accuracy and biomarker selection in microarray analysis. For instance, wrapper based ABC [17], GA [17], PSO [17], Elephant Search Algorithm (ESA) [4], and FA [18, 19] have produced large genes subsets and low accuracy [17, 20] over large feature space in contrast to hybrid meta-heuristic optimization based algorithms [1-3, 12, 13] where both wrapper and filter are utilized to produce better results. Nonetheless, choosing a suitable meta-heuristic optimization algorithm is challenging due to the influence of fitness value, convergence, and exploration and exploitation capabilities of the corresponding swarm based algorithm. FA is a popular demanding optimization algorithm which is actively used in many applications [21]. FA has special characteristics such as automatically subdivision, multimodality, and good balance between exploration and exploitation [22]. However, slow convergence and local optimums [23] are two major drawbacks of FA. Having too much exploration but with limited exploitation lead to a slow convergence [23].

Existing hybrid swarm based algorithms [1, 12, 13, 24] including FA based studies [2, 25, 26] which use fixed size solutions for population generation suffer with slow convergence due to increased exploration and limited exploitation. In other words, defining a large threshold for a fixed size solution will make the algorithm trapped with slow convergence. In addition, fast convergence behaviour has been already demonstrated in the literature with the use of non-fixed size solutions [27]. On the other hand, iteration of the initial population throughout the generations has been identified as an impact for local optimum. Existing studies [1, 2, 12, 13, 24-26] which use the initial population throughout the iterations suffer from local optimum due to lack of population diversity or sufficient exploration. Hence, regeneration formation has been suggested in the literature [28, 29] in order to increase population diversity and thus to prevent local optimum.

Therefore, we propose a hybrid gene selection approach that combines Correlation-based Feature Selection (CFS) filter [30] and a new variant of FA, termed as Mutable Firefly Algorithm (MuFA), and named as CFS-MuFA. The CFS filter is used to preprocess the microarray dataset while the MuFA is introduced to select relevant genes. Notably, CFS-MuFA uses non-fixed size solutions and regeneration formation techniques to overcome slow convergence and local optimum issues respectively. A brief research background is given in Section 2 while the detailed methodology is described in Section 3. The experimental setup and results are provided in Section 4. Moreover, Section 5 discusses the results while the conclusion is presented in Section 6.

2. Research background

This section provides a brief description on meta-heuristic optimization algorithms [31], FA [32], CFS [30] filter, and Support Vector Machine (SVM) [33] classifier. Further, gene selection studies related to aforementioned algorithms are also discussed in this section.

2.1 Meta-heuristic optimization algorithms

Meta-heuristic optimization algorithms provide optimal solutions for a particular issue [21]. It is computationally expensive to find the absolute solution for a NP-hard optimization problem [34]. Hence, optimal solutions provided by optimization algorithms are more appropriate. Population based meta-heuristic algorithms such as ABC [35], Bat Algorithm (BA) [36], ESA [37], FA [32], and PSO [16] have been widely used in gene selection studies. However, selecting a suitable algorithm out of the pool of swarm based algorithms seems to be dominated by certain factors such as the fitness value, convergence, exploitation, and exploration capabilities.

FA is a simple and popular optimization algorithm with recent demands in many applications [21]. The special properties: automatically subdivision and multimodality provide efficient performance for FA in optimization and classification tasks [22]. In addition, recent studies [38, 39] show that FA has the capabilities for being adapted and hybridized with other swarm algorithms giving more credits for the choice of FA in many applications. Hence, this research focuses on a hybrid gene selection algorithm using a new variant of FA.

2.2 Firefly algorithm

FA imitates the light-emitting nature of fireflies with three rules defined as follows [32].

- Fireflies are unisex and attracted to each other;
- The attraction of a firefly is proportional to its brightness. And the less bright fireflies will be attracted toward the brighter fireflies. The brightness is inversely proportional to the distance.
- The brightness is determined by the fitness or the objective function.

The brightness of a firefly x is given by the fitness function $f(x)$. The attraction is represented by β which decreases with the distance. The pseudo-code of the FA is given in Fig. 1.

Eq. (1) illustrates the position update in a less bright firefly x_i when moves towards a brighter firefly x_j. Parameters β_0, γ, and α represent initial attractiveness, light absorption coefficient, and randomization parameter while ε_i represents a vector of random numbers drawn from a Gaussian distribution or uniform distribution [32].

$$x_i(t+1) = x_i(t) + \beta_0 e^{-\gamma r_{ij}^2}\left(x_j(t) - x_i(t)\right) + \alpha\varepsilon_i \quad (1)$$

where, $x_i(t+1)$ is the new position and r_{ij} is the distance between firefly i and j. The distance r_{ij} is calculated using Cartesian distance [32] as given in Eq. (2).

$$Cartesian\_distance(x_j, x_i) = \sqrt[2]{(x_j - x_i)^2} \quad (2)$$

FA has widely been used in many applications such as face recognition [40], price forecasting [41], speech recognition [42, 43], document clustering [44], and text classification [45]. Further, FA has been also used in many medical applications [46]. For instance, Chaotic Firefly Algorithm (CFA) [47] for brain tissue segmentation, Binary Firefly Algorithm (BFA) [48] for protein detection, Levy-FA for radiotherapy treatment [49], and FA for heart disease diagnosis [50, 51] are few FA based applications used in biomedical systems. Besides, FA has been used for gene selection in cancer classification recently. BFA [25, 47, 52], Recursive Firefly Algorithm (RFA) [19], and composite FA [20] are few alternatives of FA proposed for gene selection in cancer classification.

Gene selection using FA should be adapted in a way so that the optimal solution is produced in concern of the gene interactions. In addition, the standard FA has two major drawbacks: slow convergence and the nature of heavily falling into local optima. Hence, appropriate techniques should be introduced to overcome these issues. Existing wrapper-based firefly algorithms [19, 20] suffer from large genes subsets which cause trouble in identifying informative genes. On the other hand, firefly-based hybrid approaches [2, 26] do not concern gene interactions due to the usage of univariate filters for preprocessing. Meanwhile, firefly-based studies [2, 20, 25] which use fixed-size fireflies for the population are trapped with slow convergence while firefly-based studies [20, 25] which regulate the initially created population over the generations suffer from local optimums. Hence, we propose a hybrid gene selection approach with a multivariate filter: CFS filter and a new variant of FA: MuFA named as CFS-MuFA.

Input: Define parameters –
population size n, β_0, γ, α_0, maximum iteration: tmax
Output: Best firefly
Algorithm:
Generate firefly population randomly:
$x_i,\ i = 1, 2, 3, \ldots, n$;
Evaluate the fitness of each firefly: $f(x)$;
while (t < tmax);
for $i = 1$ to n;
for $j = 1$ to n;
if ($f(x_i) < f(x_j)$);
Move firefly i towards j
Calculate the distance r using Eq. (2)
Calculate the new position x_i using Eq. (1)
Update the fitness of firefly i
end if;
Evaluate new solutions and update light intensity;
end for j;
end for i;
Rank the fireflies and find the best firefly;
end while;

Figure. 1 Pseudo code of firefly algorithm resource. Yang [32]

2.3 Correlation-based feature selection filter

Filter-based preprocessing is an essential step in gene selection. However, univariate filters such as mutual information filter [53] and f-score filter [54] individually evaluate the features in contrast to multivariate filters which evaluate feature subsets [55]. CFS filter [30], and mRMR filter [56] are multivariate filters used in gene selection [1, 12]. CFS filter evaluates the genes subsets in concern to correlations among genes and corresponding class. Genes with large correlations towards the class and small correlations within the genes are highly

prioritized to be present in the selected subset. In concern to preprocessing, irrelevant genes which are less correlated with the class and redundant genes which are highly correlated within genes should be eliminated from the resultant subset [57]. Best First Search (BFS) [58] heuristic strategy is used for searching due to its capability in handling large feature space [12]. Initially, a matrix of gene-class and gene-gene correlations are calculated using the training dataset [59]. Then, a genes subset is evaluated giving a score based on the Eq. (3).

$$Score_s = \frac{D\overline{r_{cg}}}{\sqrt{D+D(D-1)\overline{r_{gg}}}} \quad (3)$$

where $Score_s$ is the score of a genes subset s consists of D number of genes, $\overline{r_{cg}}$ is the average gene-class correlation, and $\overline{r_{gg}}$ is the average gene-gene correlation.

CFS filter has been used for feature selection in various fields such as in food grain classification [60], gene expression and proteomic pattern classification [59], quality assessment of retinal images [61], and classification of tear film lipid layer of the eye [62] and many more. In concern to feature selection for microarray analysis, Al-Batah et al. [8] presented a filter-based gene selection approach with CFS filter. However, due to the high dimension of microarray datasets, the approach produced relatively large genes subsets and low accuracy. Besides, Alshamlan [12] hybridized CFS filter with ABC algorithm giving a competing performance in terms of accuracy and genes subset. In addition, Jain et al. [63] proposed CFS filter-based preprocessing with an improved binary PSO algorithm for gene selection. Though relatively acceptable classification accuracy was produced, the approach resulted in slightly large genes subsets. Existing works [12, 63] present the effectiveness of CFS filter in preprocessing. Further, in concern to the gene interactions and their inter-related functionalities [64, 65], the proposed research uses CFS filter for preprocessing.

2.4 Support vector machine

SVM [33] is a popular supervised learning algorithm introduced by Vapnik. SVM has been widely used in data mining applications such as classification [66, 67], prediction [68], forecasting [41] and many more [69]. There are plenty of supervised learning algorithms such as Artificial Neural Network (ANN) [70], Bayesian belief networks [71], instance-based methods [72, 73], and decision tree-based methods [74], etc. However, SVM has the potential to handle large-scale analysis efficiently [69] in contrast to some of the supervised learning algorithms which have the drawback that makes them less efficient over large-scale analysis.

For instance, K-Nearest Neighbor (KNN) is not appropriate for multidimensional feature analysis due to its intuition towards irrelevant features [75]. On the other hand, decision trees don't have the capabilities to handle diagonal separation [75]. Besides, SVM separates the samples in a dataset through a hyperplane drawn concerning the class. Further, SVM is capable of handling both linear and non-linear separations [76]. Existing works have demonstrated competing performance using SVM over large feature space such as microarray [1, 2, 12]. Hence, the proposed research uses SVM classifier for classification purposes.

3. Method

The proposed methodology namely CFS-MuFA consists of a filter approach (i.e. CFS [30]) and a wrapper approach (i.e. MuFA). There are three components in CFS-MuFA: data preprocessing, gene selection, and classification as shown in Fig. 2. Further, each of these components are described in detail.

3.1 Data preprocessing

The standard microarray datasets undergo two processing: normalization and filtering. Initially, the original microarray datasets are normalized using min-max normalization [77]. The values of each feature are normalized in the range of 0 and 1 as denoted in Eq. (4) where the original value y is normalized into y'. Min-max normalization rescales each feature giving equal priority and hence should be performed before filtering. The normalized datasets are passed through the CFS filter [30] in order to eliminate the irrelevant and redundant genes. The employed datasets are reduced in size after deploying the CFS filter, hence producing a CFS filtered dataset. CFS filter is a multivariate filter that evaluates gene subsets concerning gene interactions. Further, Best First Search (BFS) [58] is used to search the best feature subset with features that are highly correlated towards the class while slightly correlated among the features through the feature space. The normalized filtered datasets will then be moved towards the gene selection step.

$$y' = \frac{y-min(y)}{max(y)-min(y)} \quad (4)$$

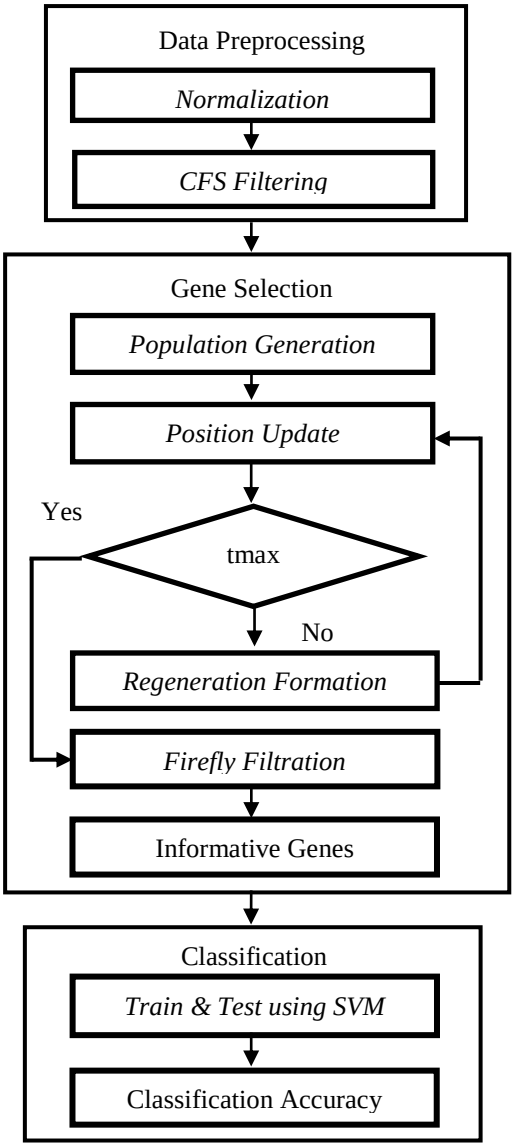

Figure. 2 Flow of CFS-MuFA

3.2 Gene selection

Meta-heuristics optimization algorithms have the ability to find the global optimal solutions [78]. However, high dimensional datasets such as microarray require much attention during the process of feature selection as these datasets consist of a large number of features among which only a few informative features exist. The proposed MuFA consists of four steps: population generation, position update, regeneration formation, and firefly filtration. Each of these steps is described in detail as given below.

3.2.1. Population generation

The firefly population of the proposed MuFA is generated with mutable size solutions in contrast to the fixed size solutions provided in the standard FA. The slow convergence issue is aimed to be resolved

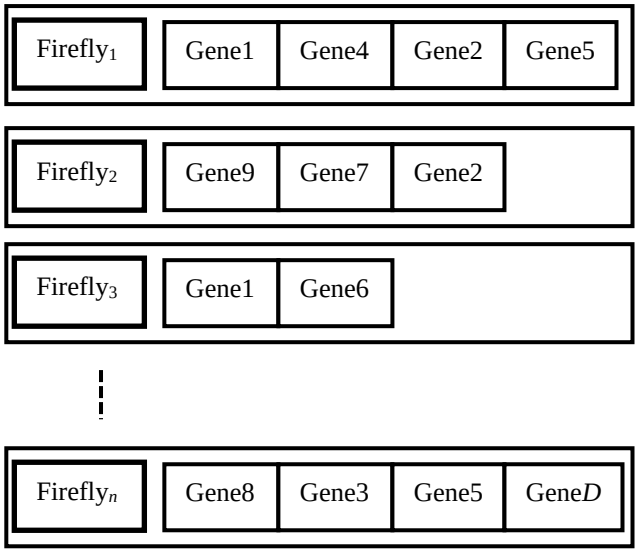

Figure. 3 Population generation in MuFA

with mutable solutions which can hold s number of genes where $s = 1,2,3, \dots, D$ in a D dimensional feature space. A firefly population with n number of mutable size fireflies would be illustrated as in Fig. 3.

3.2.2. Position update

The position of a less bright firefly x_i will be updated according to Eq. (5). It is worthwhile mentioning that the position update takes place if and only if the newly generated firefly results in higher fitness compared to the old firefly.

$$x_i(t+1) = x_i(t) + x_j(t) - \left(x_i(t) \cap x_j(t)\right) \quad (5)$$

3.2.3. Regeneration formation

The repetition of the algorithm on the same population over several iterations seems to results in low exploration which may produce less accuracy. Hence, in the proposed research, the firefly population is regenerated while preserving the best firefly found in the current generation. The regeneration formation process not only increases the exploitation capability as it preserves the best firefly, but also increases the population diversity.

3.2.4. Firefly filtration

The firefly filtration process runs the MuFA on the best fireflies found over the several iterations and hence, increases the exploitation and exploration capabilities. Fig. 4 illustrates the pseudo-code of the proposed CFS-MuFA algorithm.

Input: CFS-filtered cancer microarray dataset
Define parameters –
population size n,
dimension d,
maximum iteration: tmax
Output: Best firefly
Algorithm:
Generate firefly population with mutable size solutions randomly: $x_i,\ i = 1, 2, 3, \ldots, n$;
Evaluate the fitness of each firefly: $f(x)$;
while (t < tmax);
for $i = 1$ to n;
for $j = 1$ to n;
if $(f(x_i) < f(x_j))$;
Calculate the new position x_i using Eq. (5)
Calculate the fitness of new x_i using Eq. (6)
if fitness of new x_i > fitness of old x_i
Move firefly i towards j
Update the position of firefly i
Update the fitness of firefly i
end if;
end if;
Evaluate new solutions and update light intensity;
end for j;
end for i;
Rank the fireflies and find the best firefly;
Regenerate the firefly population;
Find the global best firefly at tmax;
end while;

Figure. 4 Pseudo code of CFS-MuFA

4. Experimental setup and results

The proposed CFS-MuFA was evaluated on four cancer microarray datasets of both binary and multiclass problems. SVM [33] classifier was used for the classification task to observe classification accuracy (as depicted in Eq. (6)). Implementation of the proposed hybrid algorithm was carried out using WEKA and MATLAB platforms in a PC with Intel Core i3 processor, 4.00 GB RAM, and Windows 10 operating system.

$$\text{Classification accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (6)$$

where, TP, TN, FP, and FN denotes true positive, true negative, false positive, and false negative respectively. The higher the value, the better the algorithm is.

4.1 Dataset description

Eight cancer microarray datasets namely; Colon [79], Leukemia2 [80], Leukemia3 [81], and Small Round Blue Cell Tumor (SRBCT) [82] were used for the evaluation of the proposed approach. Two datasets are of a binary classification problem and the remainder represent the multiclass problem. Brief information on the datasets is depicted in Table 1.

Table 1. Cancer microarray datasets information

Dataset	No. of classes	No. of genes	No. of samples
Colon	2	2000	62
Leukemia2	2	7129	72
Leukemia3	3	7129	72
SRBCT	4	2308	83

Table 2. Parameter settings used in this study

Parameter	Value
Population size	80
Dimension	Number of genes
Number of iterations	100
Number of runs	30

The Colon cancer dataset [79] consists of two classes, 2000 genes, and 62 samples. Besides, the Leukemia2 dataset [80] is a two-class dataset whereas the Leukemia3 [81] is a multiclass dataset with three classes. Both Leukemia2 and Leukemia3 datasets consist of the same number of genes and samples which are equal to 7129 and 72 respectively. Further, the SRBCT [82] dataset is a four-class dataset consists of 2308 genes and 83 samples. All of these datasets consist of a large number of gene expressions that could be analyzed with the purpose of finding biomarkers that are crucial in healthcare applications such as cancer classification and therapeutics [83]. Furthermore, these cancer microarray datasets have been widely used in the evaluation of existing algorithms [1, 2, 12, 63, 84]. Hence, the proposed hybrid algorithm is evaluated on these cancer microarray datasets.

4.2 Parameter settings

In order to have a fair evaluation, parameters of the proposed algorithm are assigned based on preliminary studies and existing work [1, 12]. Table 2 illustrates the parameters along with the assigned values.

Population size represents the size of the firefly population. More specifically, there are 80 fireflies in the population according to the proposed setting. The dimension of the feature space equals to the number of genes in the dataset. Further, the proposed MuFA is iterated 100 times over 30 independent runs to obtain better results.

4.3 Results

The experimental results (i.e. accuracy) obtained

over the four cancer microarray datasets are tabulated in Table 3-7. The values given between parentheses in Table 3-7 provide the gene count in the dataset. Further, results were presented separately; classification accuracy obtained on CFS filtered datasets is given in Table 3 whereas classification accuracy obtained on standard microarray datasets using MuFA is given in Table 4. Further, the classification accuracy obtained on filtered microarray datasets using CFS-MuFA is given in Table 5. Moreover, the informative genes selected using the proposed CFS-MuFA are tabulated in Table 6. Notably, Table 5 presents the best results obtained over 30 runs.

The classification accuracy obtained using the proposed CFS-MuFA approach (refer to Table 5) signifies the contribution of CFS filter for gene selection in contrast to the results obtained on the standard microarray datasets (refer to Table 4) without data preprocessing. At the same time, the proposed CFS-MuFA has improved the classification accuracy together with few informative genes indicating the significant contribution of the proposed algorithm in gene subset selection.

Table 3. Classification accuracy obtained on CFS filtered datasets

Dataset	Classification Accuracy (%) (Number of Genes)
Colon(26)	100(26)
Leukemia2(81)	100(81)
Leukemia3(104)	100(104)
SRBCT(111)	100(111)

Table 4. Classification accuracy obtained on standard microarray datasets using MuFA

Dataset	Classification Accuracy (%) (Number of Genes)
Colon(2000)	94.74(338)
Leukemia2(7129)	100(55)
Leukemia3(7129)	100(422)
SRBCT(2308)	100(84)

Table 5. Classification accuracy obtained on filtered datasets using CFS-MuFA

Dataset	Classification Accuracy (%) (Number of Genes)
Colon(26)	100(5)
Leukemia2(81)	100(1)
Leukemia3(104)	100(2)
SRBCT(111)	100(7)

Table 6. Informative genes subsets selected using CFS-MuFA

Dataset	Genes
Colon(5)	A3, A5, A13, A15, A17
Leukemia2(1)	attribute3252
Leukemia3(2)	U05259_rna1_at, M83652_s_at
SRBCT(7)	gene229, gene153, gene1708, gene1962, gene1377, gene1613, gene867

5. Discussion

This study compares the results of CFS-MuFA with those produced by an existing approach that uses FA [2]. Besides, other benchmark techniques [1, 12, 13, 24, 63, 84, 85] are also compared for the evaluation of the proposed approach. The classification results obtained for CFS filtered datasets show the efficiency of the filter giving 100% accuracy for all the datasets. On the other hand, the classification results obtained for standard microarray datasets denote the need for gene selection as even though some datasets have produced 100% accuracy the genes subset size is very large. Besides, the classification results obtained for all the four cancer microarray datasets (refer Table 5 and Table 6) indicate the efficiency of the proposed algorithm in gene selection. All the datasets are classified with 100% classification accuracy with only a few informative genes.

In regards to Colon cancer classification, CFS-MuFA has produced 100% accuracy with only 5 informative genes. The classification results depict the efficiency of the proposed algorithm compared to the existing results [1, 2, 12, 13, 24, 63, 85] as the classification accuracy is 100% with only 5 genes. As for the Leukemia2 classification, Alshamlan et al. [24], Alshamlan et al. [13], Alshamlan [12], Jain et al. [63], Almugren and Alshamlan [2], and Al-Betar et al. [1] produced 100% accuracy with 14, 4, 3, 4.3, 5, and 4.07 genes respectively. However, the proposed CFS-MuFA has obtained the same accuracy with a single gene.

Further, the classification accuracy obtained in this study for Leukemia3 is greater than the ones reported by Almugren and Alshamlan [2] and Mazumder and Veilumuthu [84]. At the same time, when compared to the outcomes of Alshamlan et al. [24], Alshamlan et al. [13], Alshamlan [12], Jain et al. [63], Al-Betar et al. [1], and Fajila and Yusof [85], even though the accuracy is the same, Alshamlan [12], Al-Betar et al. [1], and Fajila and Yusof [85] have produced 20, 8, 6, 6, 5.33, and 4 genes respectively whereas CFA-MuFA produced only 2 genes. Moreover, compared to earlier studies [1, 2, 24, 63, 85], CFS-MuFA has produced a small

Table 7. The classification performance comparison

Reference	Colon	Leukemia2	Leukemia3	SRBCT
CFS-MuFA	**100(5)**	**100(1)**	**100(2)**	100(7)
ISIG [85]	95.23(4)	-	100(4)	100(8)
rMRMR-MBA [1]	97.85(12.27)	100(4.07)	100(5.33)	100(9.13)
FFF-SVM [2]	94.3(15)	100(5)	97.8(10)	100(8)
Co-ABC [12]	96.77(9)	100(3)	100(6)	**100(4)**
Mazumder and Veilumuthu [84]	-	98.61(3)	98.61(3)	100(6)
Jain et al. [63]	94.89(4.2)	100(4.3)	100(6)	100(34.1)
Alshamlan et al. [24]	96.77(15)	100(14)	100(20)	100(10)
Alshamlan et al. [13]	98.38(10)	100(4)	100(8)	100(6)

number of genes in the classification of SRBCT. However, the number of genes produced by Alshamlan et al. [13], Alshamlan [12], and Mazumder and Veilumuthu [84] in SRBCT is smaller than the results obtained using the proposed algorithm. Table 7 compares the classification performance of CFS-MuFA related to existing methods. The performance is compared with respect to classification accuracy and the number of biomarkers obtained over each approach. It is worthwhile mentioning that the proposed CFS-MuFA has contributed significantly to gene selection for cancer classification as three out of four datasets have produced the best results as highlighted in Table 7.

It is believed that the CFS-based filtering has prepared the high dimensional microarray datasets with more relevant genes which would have assisted informative genes subset selection using MuFA successfully. Further, the slow convergence issue in the standard FA is resolved with mutable size fireflies giving a balance between exploration and exploitation in contrast to fixed-size solutions [1, 2, 12, 13]. Moreover, the regeneration formation and firefly filtration also have enhanced the exploitation and exploration capabilities of the proposed MuFA. Thus, the search space has been utilized with the use of mutable solutions and diversification producing highest accuracy with less number of biomarkers compared to existing algorithms [1, 2, 12, 13]. In addition, the position update strategy has given more insights into gene selection in contrast to the position update method in standard FA [2]. Thus, it is anticipated that the proposed new CFS-MuFA will be beneficial in feature selection and classification over a large feature space in addition to gene selection in cancer classification.

6. Conclusion

A hybrid gene selection approach named CFS-MuFA is proposed in this study. The proposed approach introduces a mutable property for firefly population giving significant results in gene selection. Four cancer microarray datasets of both binary and multiclass are evaluated on the proposed approach. It is noticeable that CFS-MuFA has contributed to cancer classification with higher accuracy and fewer informative genes. More specifically, 100% accuracy was produced on all the four datasets with few biomarkers where the smallest subset size was one while the largest subset was only seven. Three out of four datasets produced the best results compared to existing techniques reflecting the efficiency of the proposed algorithm. Hence, the future task will focus on more techniques to increase the exploitation and exploration capabilities of the proposed CFS-MuFA. In addition, the approach will be evaluated on various microarray datasets to validate the robustness of the approach in gene selection.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, methodology, software, formal analysis, resources, data curation, and writing-original draft preparation: Mohamed Nisper Fathima Fajila. Writing-review, editing, and supervision: Yuhanis Yusof.

References

[1] M. A. A. Betar, O. A. Alomari, and S. M. A. Romman, “A TRIZ-inspired bat algorithm for gene selection in cancer classification”, *Genomics*, Vol. 112, No. 1, pp. 114-126, 2020.

[2] N. Almugren and H. M. Alshamlan, “New bio-marker gene discovery algorithms for cancer gene expression profile”, *IEEE Access*, Vol. 7, pp. 136907-136913, 2019.

[3] R. Dash, “An Adaptive Harmony Search Approach for Gene Selection and Classification of High Dimensional Medical Data”, *Journal of King Saud University – Computer and*

Information Sciences, Vol. 33, No. 2, pp. 195-207, 2021.

[4] M. Panda, "Elephant search optimization combined with deep neural network for microarray data analysis", *Journal of King Saud University-Computer and Information Sciences*, Vol. 32, No. 8, pp. 940-948, 2020.

[5] B. H. Shekar and G. Dagnew, "L1-regulated feature selection and classification of microarray cancer data using deep learning", In: *Proc. of 3rd International Conference on Computer Vision and Image Processing*, pp. 227-242, 2020.

[6] A. E. Akadi, A. Amine, A. E. Ouardighi, and D. Aboutajdine, "A two-stage gene selection scheme utilizing MRMR filter and GA wrapper", *Knowledge and Information Systems*, Vol. 26, No. 3, pp. 487-500, 2011.

[7] C. M. Lai, W. C. Yeh, and C. Y. Chang, "Gene selection using information gain and improved simplified swarm optimization", *Neurocomputing*, Vol. 218, pp. 331-338, 2016.

[8] M. A. Batah, B. Zaqaibeh, S. A. Alomari, and M. S. Alzboon, "Gene Microarray Cancer Classification using Correlation Based Feature Selection Algorithm and Rules Classifiers", *International Journal of Online & Biomedical Engineering*, Vol. 15, No. 8, pp. 62-73, 2019.

[9] D. H. Mazumder and R. Veilumuthu, "An enhanced feature selection filter for classification of microarray cancer data", *ETRI Journal*, Vol. 41, No. 3, pp. 358-370, 2019.

[10] Y. Wang, X. G. Yang, and Y. Lu, "Informative gene selection for microarray classification via adaptive elastic net with conditional mutual information", *Applied Mathematical Modelling*, Vol. 71, pp. 286-297, 2019.

[11] V. Elyasigomari, D. A. Lee, H. R. C. Screen, and M. H. Shaheed, "Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification", *Journal of Biomedical Informatics*, Vol. 67, pp. 11-20, 2017.

[12] H. M. Alshamlan, "Co-ABC: Correlation artificial bee colony algorithm for biomarker gene discovery using gene expression profile", *Saudi Journal of Biological Sciences*, Vol. 25, No. 5, pp. 895-903, 2018.

[13] H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, "Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification", *Computational Biology and Chemistry*, Vol. 56, pp. 49-60, 2015.

[14] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, "A hybrid feature selection algorithm for gene expression data classification", *Neurocomputing*, Vol. 256, pp. 56-62, 2017.

[15] H. Motieghader, A. Najafi, B. Sadeghi, and A. M. Nejad, "A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata", *Informatics in Medicine Unlocked*, Vol. 9, pp. 246-254, 2017.

[16] C. S. Yang, L. Y. Chuang, C. H. Ke, and C. H. Yang, "A Hybrid Feature Selection Method for Microarray Classification", *IAENG International Journal of Computer Science*, Vol. 35, No. 3, 2008.

[17] H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, "Abc-svm: artificial bee colony and svm method for microarray gene selection and multi class cancer classification", *Int. J. Mach. Learn. Comput*, Vol. 6, No. 3, p. 184, 2016.

[18] N. Almugren and H. Alshamlan, "FF-SVM: New FireFly-based Gene Selection Algorithm for Microarray Cancer Classification", In: *Proc. of 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 1-6, 2019.

[19] N. Dif and Z. Elberrichi, "An Enhanced Recursive Firefly Algorithm for Informative Gene Selection", *International Journal of Swarm Intelligence Research*, Vol. 10, No. 2, pp. 21-33, 2019.

[20] H. Peng, W. Zhu, C. Deng, K. Yu, and Z. Wu, "Composite firefly algorithm for breast cancer recognition", *Concurrency and Computation: Practice and Experience*, Vol. 33, No. 5, 2020.

[21] T. Dokeroglu, E. Sevinc, T. Kucukyilmaz, and A. Cosar, "A survey on new generation metaheuristic algorithms", *Computers & Industrial Engineering*, Vol. 137, p. 106040, 2019.

[22] X. S. Yang and X. He, "Firefly algorithm: recent advances and applications", *International Journal of Swarm Intelligence*, Vol. 1, No. 1, pp. 36-50, 2013.

[23] X. S. Yang, "Swarm intelligence based algorithms: a critical analysis", *Evolutionary Intelligence*, Vol. 7, No. 1, pp. 17-28, 2014.

[24] H. Alshamlan, G. Badr, and Y. Alohali, "mRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling", *BioMed Research International*, Vol. 2015, 2015.

[25] N. Emami and A. Pakzad, "A New Knowledge-Based System for Diagnosis of Breast Cancer by a combination of the Affinity Propagation

and Firefly Algorithms", *Journal of AI and Data Mining*, Vol. 7, No. 1, pp. 59-68, 2019.

[26] S. F. Jabbar, "A classification model on tumor cancer disease based mutual information and firefly algorithm", *Periodicals of Engineering and Natural Sciences*, Vol. 7, No. 3, pp. 1152-1162, 2019.

[27] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts", *Genomics*, Vol. 109, No. 2, pp. 91-107, 2017.

[28] S. Cheng, Y. Shi, Q. Qin, Q. Zhang, and R. Bai, "Population diversity maintenance in brain storm optimization algorithm", *Journal of Artificial Intelligence and Soft Computing Research*, Vol. 4, No. 2, pp. 83-97, 2014.

[29] R. Salgotra, U. Singh, and S. Saha, "On some improved versions of whale optimization algorithm", *Arabian Journal for Science and Engineering*, Vol. 44, No. 11, pp. 9653-9691, 2019.

[30] M. A. Hall, "Correlation-based feature selection for machine learning [Doctoral dissertation, University of Waikato]", *University of Waikato*, 1999.

[31] S. Das, A. Abraham, and A. Konar, "Metaheuristic Clustering. Studies in Computational Intelligence", Vol. 178, 2009.

[32] X. S. Yang, "Nature-inspired Metaheuristic Algorithms", 2010.

[33] V. Vapnik, S. E. Golowich, and A. J. Smola, "Support vector method for function approximation, regression estimation and signal processing", In: *Proc. of Advances in neural information processing systems 9*, pp. 281-287, 1997.

[34] D. T. Hoang, "Metaheuristics for NP-hard combinatorial optimization problems [Doctoral dissertation, National University of Singapore]", 2008.

[35] D. Karaboga, "An Idea Based on Honey Bee Swarm for Numerical Optimization", *Technical report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department*, Vol. 200, pp. 1-10, 2005.

[36] X. S. Yang, "A new metaheuristic bat-inspired algorithm", In: *Proc. of Nature Inspiredcooperative Strategies for Optimization (NICSO 2010)*, pp. 65-74, 2010.

[37] S. Deb, S. Fong, and Z. Tian, "Elephant search algorithm for optimization problems", In: *Proc. of 2015 Tenth International Conference on Digital Information Management*, pp. 249-255, 2015.

[38] N. A. A. Thanoon, O. S. Qasim, and Z. Y. Algamal, "A new hybrid firefly algorithm and particle swarm optimization for tuning parameter estimation in penalized support vector machine with application in chemometrics", *Chemometrics and Intelligent Laboratory Systems*, Vol. 184, pp. 142-152, 2019.

[39] M. Pyingkodi, S. Shanthi, M. Muthukumaran, K. Nanthini, and K. Thenmozhi, "Hybrid bee colony and weighted ranking firefly optimization for cancer detection from gene regulatory sequences", *International Journal of Scientific & Technology Research*, Vol. 9, No. 1, pp. 2459-2465, 2020.

[40] V. Agarwal and S. Bhanot, "Firefly inspired feature selection for face recognition", *In: Proc. of 2015 Eighth International Conference on Contemporary Computing*, pp. 257-262, 2015.

[41] A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi, and O. K. Hussain, "Support vector regression with chaos-based firefly algorithm for stock market price forecasting", *Applied Soft Computing*, Vol. 13, No. 2, pp. 947-958, 2013.

[42] T. Hassanzadeh, K. Faez, and G. Seyfi, "A speech recognition system based on structure equivalent fuzzy neural network trained by firefly algorithm", In: *Proc. of 2012 International Conference on Biomedical Engineering*, pp. 63-67, 2012.

[43] H. H. Nuha and M. Abido, "Firefly algorithm for log-likelihood optimization problem on speech recognition", In: *Proc. of 2016 4th International Conference on Information and Communication Technology*, pp. 1-6, 2016.

[44] A. J. Mohammed, Y. Yusof, and H. Husni, "Weight-based Firefly algorithm for document clustering", In: *Proc. of the First International Conference on Advanced Data and Information Engineering*, pp. 259-266, 2014.

[45] S. L. M. Sainte and N. Alalyani, "Firefly algorithm based feature selection for Arabic text classification", *Journal of King Saud University-Computer and Information Sciences*, Vol. 32, No. 3, pp. 320-328, 2020.

[46] J. Nayak, B. Naik, P. Dinesh, K. Vakula, and P. B. Dash, "Firefly Algorithm in Biomedical and Health Care: Advances, Issues and Challenges", *SN Computer Science*, Vol. 1, No. 6, pp. 1-36, 2020.

[47] P. Ghosh, K. Mali, and S. K. Das, "Chaotic firefly algorithm-based fuzzy C-means algorithm for segmentation of brain tissues in magnetic resonance images", *Journal of Visual*

Communication and Image Representation, Vol. 54, pp. 63-79, 2018.

[48] J. Zhang, B. Gao, H. Chai, Z. Ma, and G. Yang, "Identification of DNA-binding proteins using multi-features fusion and binary firefly optimization algorithm", *BMC Bioinformatics*, Vol. 17, No. 323, pp. 1-12, 2016.

[49] G. Kalantzis, C. Shang, Y. Lei, and T. Leventouri, "Investigations of a GPU-based levy-firefly algorithm for constrained optimization of radiation therapy treatment planning", *Swarm and Evolutionary Computation*, Vol. 26, pp. 191-201, 2016.

[50] N. Mülayim and A. Alaybeyoğlu, "Designing of an expert system based on firefly algorithm for diagnosis of Heart Disease", In: *Proc. of 2016 20th National Biomedical Engineering Meeting*, pp. 1-4, 2016.

[51] B. R. Rajakumar and A. George, "On hybridizing fuzzy min max neural network and firefly algorithm for automated heart disease diagnosis", In: *Proc. of 2013 Fourth International Conference on Computing, Communications and Networking Technologies*, pp. 1-5, 2013.

[52] R. Sawhney, P. Mathur, and R. Shankar, "A firefly algorithm based wrapper penalty feature selection method for cancer diagnosis", In: *Proc. of International Conference on Computational Science and Its Applications*, pp. 438-449, 2018.

[53] J. R. Vergara, and P. A. Estévez, "A review of feature selection methods based on mutual information", *Neural Comput & Applic*, Vol. 24, No. 1, pp. 175-186, 2014.

[54] S. Wright, "The interpretation of population structure by F-statistics with special regard to systems of mating", *Evolution*, Vol. 19, No. 3, pp. 395-420, 1965.

[55] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics", *Bioinformatics*, Vol. 23, No. 19, pp. 2507-2517, 2007.

[56] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp. 1226-1238, 2005.

[57] V. B. Canedo, N. S. Marono, A. A. Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods", *Information Sciences*, Vol. 282, pp. 111-135, 2014.

[58] J. Pearl, "Heuristics: Intelligent Search Strategies for Computer Problem Solving", *United States*, 1984.

[59] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns", *Genome Informatics*, Vol. 13, pp. 51-60, 2002.

[60] K. R. Pushpalatha and A. G. Karegowda, "CFS based feature subset selection for enhancing classification of similar looking food grains-a filter approach", In: *Proc. of 2017 2nd International Conference on Emerging Computation and Information Technologies*, pp. 1-6, 2017.

[61] B. Remeseiro, A. M. Mendonça, and A. Campilho, "Objective quality assessment of retinal images based on texture features", In: *Proc. of 2017 International Joint Conference on Neural Networks*, pp. 4520-4527, 2017.

[62] B. Remeseiro, V. B. Canedo, A. A. Betanzos, and M. G. Penedo, "Learning features on tear film lipid layer classification", In: *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 195-200, 2015.

[63] I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification", *Applied Soft Computing*, Vol. 62, pp. 203-215, 2018.

[64] R. Kohavi and G. H. John, "Wrappers for feature subset selection", *Artificial Intelligence*, Vol. 97, No. 1-2, pp. 273-324, 1997.

[65] J. G. Zhang and H. W. Deng, "Gene selection for classification of microarray data based on the Bayes error", *BMC Bioinformatics*, Vol. 8, No. 1, pp. 1-9, 2007.

[66] S. Li, J. T. Kwok, H. Zhu, and Y. Wang, "Texture classification using the support vector machines", *Pattern Recognition*, Vol. 36, No. 12, pp. 2883-2893, 2003.

[67] Y. H. Shao and N. Y. Deng, "A coordinate descent margin based-twin support vector machine for classification", *Neural Networks*, Vol. 25, pp. 114-121, 2012.

[68] Z. Mustaffa, Y. Yusof, and S. S. Kamaruddin, "Enhanced Abc-Lssvm for Energy fuel price prediction", *Journal of Information and Communication Technology*, Vol. 12, pp. 73-101, 2013.

[69] J. Nayak, B. Naik, and H. Behera, "A comprehensive survey on support vector machine in data mining tasks: applications & challenges", *International Journal of Database*

Theory and Application, Vol. 8, No. 1, pp. 169-186, 2015.

[70] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity", *The Bulletin of Mathematical Biophysics*, Vol. 5, No. 4, pp. 115-133, 1943.

[71] J. Pearl, "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference", *Elsevier*, 2014.

[72] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms", *Mach Learn*, Vol. 6, pp. 37-66, 1991.

[73] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, Vol. 13, No. 1, pp. 21-27, 1967.

[74] J. R. Quinlan, "Induction of decision trees", *Mach Learn*, Vol. 1, No. 1, pp. 81-106, 1986.

[75] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques", *Emerging Artificial Intelligence Applications in Computer Engineering*, Vol. 160, No. 1, pp. 3-24, 2007.

[76] D. M. Abdulqader, A. M. Abdulazeez, and D. Q. Zeebaree, "Machine Learning Supervised Algorithms of Gene Selection: A Review", *Machine Learning*, Vol. 62, No. 03, pp. 233-244, 2020.

[77] L. A. Shalabi, Z. Shaaban, and B. Kasasbeh, "Data mining: A preprocessing engine", *Journal of Computer Science*, Vol. 2, No. 9, pp. 735-739, 2006.

[78] S. Fong, R. P. B. Aghai, and R. C. Millham, "Swarm search methods in Weka for data mining", In: *Proc. of the 2018 10th International Conference on Machine Learning and Computing*, pp. 122-127, 2018.

[79] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *PNAS*, Vol. 96, No. 12, pp. 6745-6750, 1999.

[80] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, L. Coller, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, Vol. 286, No. 5439, pp. 531-537, 1999.

[81] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. D. Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia", *Nat Genet.* Vol. 30, No. 1, pp. 41-47, 2001.

[82] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", *Nat. Med*. Vol. 7, No. 6, pp. 673-679, 2001.

[83] S. Ramasubramanian, "Role of Microarray in Cancer Biology", *Journal of Complementary Medicine Research*, Vol. 11, No. 3, pp. 262-268, 2020.

[84] D. H. Mazumder and R. Veilumuthu, "Cancer Classification with a Novel Hybrid Feature Selection Technique", *International Journal of Simulation: Systems, Science & Technology*, Vol. 19, No. 2, 2018.

[85] F. Fajila and Y. Yusof, "Incremental Search for Informative Gene Selection in Cancer Classification", *Annals of Emerging Technologies in Computing*, Vol. 5, No. 2, pp. 15-21, 2021.