Frontiers in Signal Processing, Vol. 4, No. 2, April 2020
https://dx.doi.org/10.22606/fsp.2020.42001

An Internal Clustering Validation Based Fitness Approach for Meta-Heuristic Diagnosis of Cervical Cancer

M. Kerem Un[1], Mustafa Guven[1], Caglar Cengizler[1*], Seyda Erdogan[2] and Aysun Uguz[2]

[1]Faculty of Engineering and Architecture, Department of Biomedical Engineering, Cukurova University, Adana 01330, Balcali, Turkey
[2]Faculty of Medicine, Department of Pathology, Cukurova University, Adana 01330, Balcali, Turkey
Email: ccengizler@cu.edu.tr

Abstract This paper presents an utilization of data clustering with genetic algorithm (GA) approach. Proposed meta-heuristic clustering approach relies on genetic operators and accepts Calinski-Harabasz (CH) measure as fitness criteria where each individual represents a final judgement about existence of malignancy on set of cervical cells. It was aimed to evaluate the performance of fitness criteria on detection of malignancy where classification is performed on salient morphological features. Preferred fitness criteria measures the ability of individuals in a population to form appropriate clusters for normal and abnormal cell samples. Feature space includes data extracted from the previously segmented cervical cell images. Proposed approach is examined with two data sets which contains malignant and healthy cell samples. Preliminary results has shown that preferred fitness criteria for the classification is promising and the presented utilization of GA based clustering approach with CH criteria has a better clustering performance compared to conventional clustering methods.

Keywords: Calinski-Harabasz; Clustering; Cervical Cancer; Meta-Heuristic; Genetic Algorithm

1 Introduction

Evolutionary computing is a nature mimicking mechanism where a search is performed to obtain a sufficiently good solution to a complex optimization problem [1]. In contrast to conventional supervised machine learning methods, genetic algorithms (GA) provide a solution without supervising [2]. They encode the problem at hand to an array of relevant parameters which constitutes the chromosome of an individual (i.e. a candidate solution)[3]. The fitness of an individual is an indicator of how close it is to a solution to the problem at hand. Collection of individuals form populations [4]. Operators like crossing-over and mutation cause populations to evolve and pass the fittest solutions to the next generations [5]. Genetic algorithms have been utilized in a wide variety of applications. One of the most studied computational problems to be solved by genetic algorithms is unsupervised clustering of data [6]. In that branch of studies, genetic operators are applied to constructed genetic structure for forming data clusters on feature space [7]. At that point, one of the crucial factors that directly affects the classification performance of evolutionary algorithms is the utilized fitness criteria [8].

This study is aimed at examining the performance of an internal cluster validation measure, namely Calinski-Harabasz (CH) criterion, as a fitness criteria for determination of abnormality on cervical cells via a GA. Accordingly, a feature space is generated to contain malignant and healthy cervical cells. The algorithm is expected to cluster samples of same groups (i.e. healthy and non-healthy) with respect to similarity of features defining the morphology of each cervical cell. The utilized fitness criteria should be able to evaluate the compactness and distance of the both clusters in each generation [9]. The combination of CH index and GA based meta-heuristic computation presents a novel unsupervised classification mechanism for cervical cell malignancy in this study.

2 Materials and Methods

2.1 Feature Space

The presented problem is a two class (healthy and non-healthy) abnormality classification problem. Consequently, the performance of any possible classifier depends on features chosen for classification. The

search space analyzed in this study consists of four features extracted from sample images (Table 1). Cancer causes significant changes on morphology and appearance of nuclei of cervical cells. The features chosen in this study are frequently utilized for automatic detection of shape abnormality [10][11].

Table 1. Features extracted from cellular and nuclear regions.

Feature	Feature Description
1	Aspect Ratio of Nucleus
2	Eccentricity of Nucleus
3	Area Ratio of Cytoplasm to Nucleus
4	Equivalent Diameter of Nucleus

The first feature involved is the aspect ratio (i.e. major axis/minor axis ratio) of the ellipse fit to cell nucleus (Figure 1):

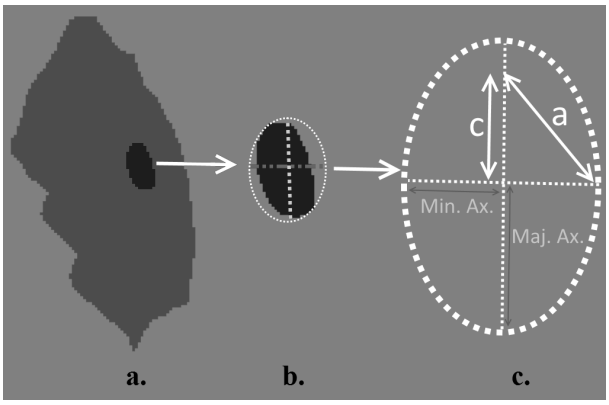

Figure 1. Measurement of some of the morphometric features introduced in text.

$$AspectRatio = \frac{MajorAxis}{MinorAxis} \tag{1}$$

The eccentricity of an elliptic region is defined as:

$$Eccentricity = \frac{c}{a} \tag{2}$$

where c is the distance between the center and the focus of the ellipse and a is the distance between the center and a vertex (Figure 1c).

Note that the above two features are indicators for the deviation of the fitted ellipse from a circle. Nucleus size is typically larger in abnormal cells [12][13]. Accordingly, nuclei to cytoplasm diameter ratio is also accepted as an abnormality indicator (Figure 2).

Shape irregularity of the nucleus is assessed by observing the ratio of the object's equivalent diameter to the actual diameter, which may significantly change in case of serious shape irregularity. In this work, an unsupervised data clustering approach is applied to the four chosen cellular features and each specimen from the combined data set is classified as either atypical or not.

Cluster centroids and standard deviations of abnormal and normal classes for introduced features are given in Table 2.

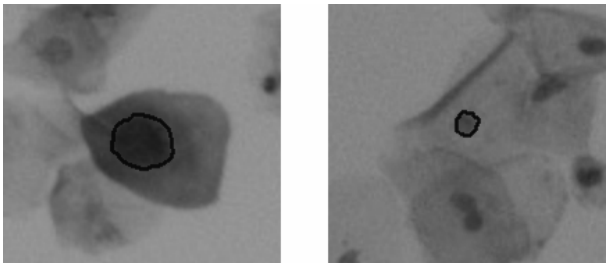

Figure 2. Two nuclei samples from the utilized data set. Nucleus boundaries are indicated with black contours.

Table 2. Cluster centroids and standard deviations are presented for all features.

	Feature 1	**Feature 2**	**Feature 3**	**Feature 4**
Normal Cluster Centroid	1,241	0,553	40,598	14,615
Standard Deviation	0,154	0,136	14,536	2,286
Abnormal Cluster Centroid	1,373	0,638	14,973	18,051
Standard Deviation	0,234	0,143	6,799	3,842

2.2 Structure of the Genetic Algorithm Approach

In this study, we evaluate the convergence of a straightforward GA in case of a internal cluster validation measure used as fitness criteria. The performance of the algorithm is an indication for the effectiveness of the proposed unsupervised clustering approach for abnormality evaluation.

Some important concepts related to genetic algorithms are explained in connection with the current problem below:

Individual: Each generated candidate solution to the the clustering problem, i.e. a specific classification of all Pap-Smear samples as either healthy or unhealthy, is an individual. Individuals consist of genes.

Gene: The classification of each Pap-Smear specimen represents a gene in the candidate solution (i.e. individual). Accordingly, for a given problem, each individual has a fixed number of genes (which equals the total number of specimens in the data set). A gene is a binary variable, which is equal to one if the related sample belongs to abnormal class, and zero if not.

Population: A population is the collection of a predefined number of candidate solutions.

Generation: Population modified by evolutionary operators forms the next generation in the GA.

In our approach, the classification process is based on the evolution of an initial population. Five example specimens from a given solution set are shown in Figure 3.

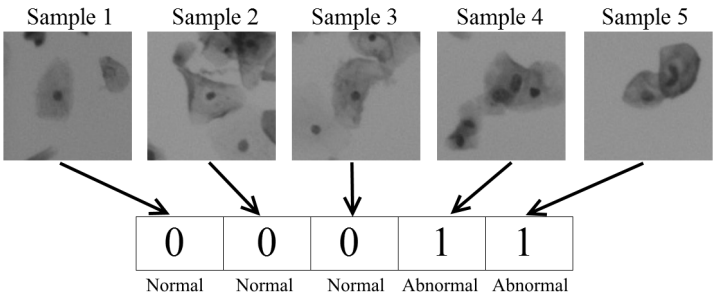

Figure 3. A sample individual representing a five specimen solution.

The evolution of a generation is realized by applying two genetic operators to the population at every iteration (representing a generation). The crossing-over operator allows the best (i.e. fittest) solutions to

be passed to the subsequent iterations [14]. The mutation operator increases the chance of finding fitter solutions by providing a larger search space with diversified individuals [15] (Figure 4).

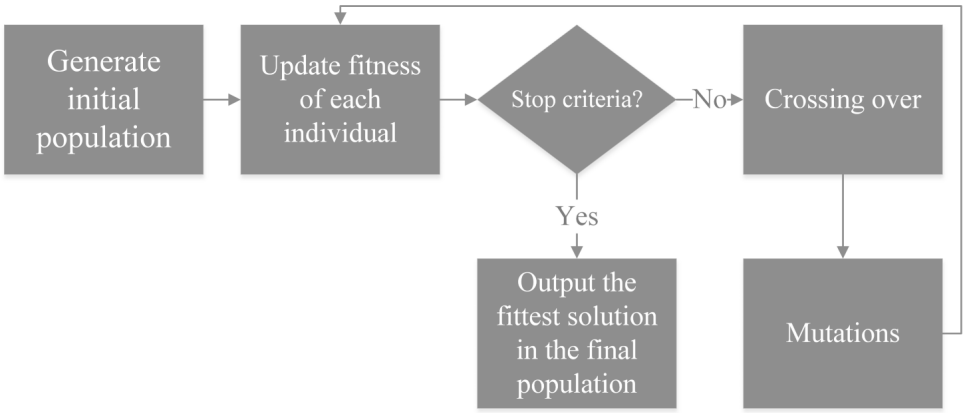

Figure 4. General flow diagram of applied genetic operators during iterations.

2.3 Population of Solutions

With N being the number of samples to be classified, in the beginning, n random classifications are imposed on the sample pool. In other words, n individuals each with N genes is created as the initial population which then evolves according to the GA. The population at each generation can be represented by an nxN matrix P where each row contains the N binary values associated with the genes. The matrix entry $x_{i,j}$ represents j-th gene of the i-th individual in population. The genome of an individual with five genes (i.e. the classification of a five sample set) is shown in Figure 5.

Ind1 (Fittest)	0	1	1	0	0
Ind2	1	1	0	1	1
Ind3	1	0	0	1	1

Actual Solution	1	1	1	0	0

p

Figure 5. An example classification of a five sample.

Each gene of a solution holds a binary value for representing a class described by:

$$w_{k,i} = \begin{cases} 1, \text{ if pattern } x_i \in \text{cluster } C_k \\ 0, \text{ Otherwise} \end{cases} \tag{3}$$

2.4 Fitness of an Individual

In our experimental set-up, every individual consists of two clusters of healthy and unhealthy samples. CH index is used as an internal cluster validity measure as well as the fitness measure of the individual in the GA. For n individuals and two clusters, the CH index is given by:

$$CH = \frac{B_c}{k-1} . \frac{n-k}{W_c} \tag{4}$$

where n is the total number of observations and k the number of clusters, which equals two in our problem. Between-cluster and within-cluster sums of squares are denoted by B_c and W_c respectively. B_c is calculated by:

$$B_c = \sum_{i=1}^{2} |C_i| \left\| \overline{C_i} - \overline{x} \right\|^2 \tag{5}$$

where $\overline{C_i}$ is the center of ith cluster, $\overline{x}$ is the center of all observations and $|C_i|$ is the number of points in ith cluster. W_c is also calculated by:

$$W_c = \sum_{i=1}^{2} \sum_{x \in c_i} \left\| x - \overline{C_i} \right\|^2 \tag{6}$$

where x is an observation belonging to ith cluster.

Criteria given in Eq. (5) is the evaluation of the quality of each possible cluster solution. Which depends on cluster compactness and proximity of intra-cluster distances [16].

2.5 Crossover

Crossover operation is applied to the population of solutions for generating new solutions from the fittest of existing solutions. The operation starts with the selection of two random parent solutions from the population. To create a bias toward the selection of fittest, the probability for a solution to be selected is set proportional to its fitness value. In this fitness proportionate selection method, also known as roulette wheel selection method, a solution with a fitness value of, say, 2c is twice as likely to be selected for the crossover operation compared to a solution with a fitness of c (Figure 6).

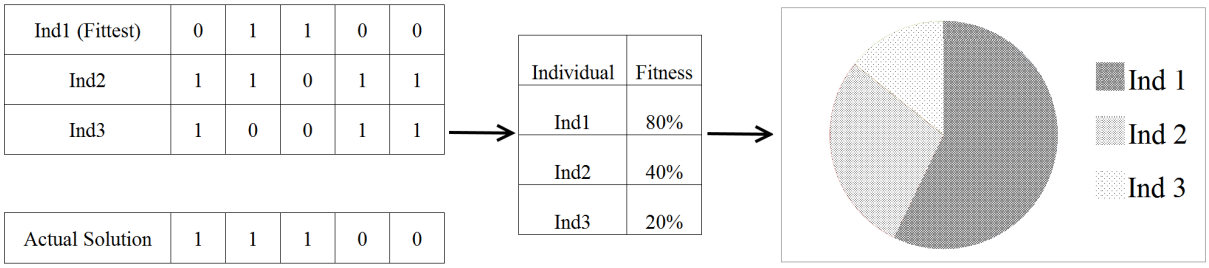

Figure 6. Roulette wheel representation of a sample population for crossover process.

After the parent solutions are picked, a new solution is formed by combining blocks of chromosomes. To achieve that, the genome of each parent is divided into two blocks at a random spot. One block is taken from each parent to form a complete new individual (Figure 7).

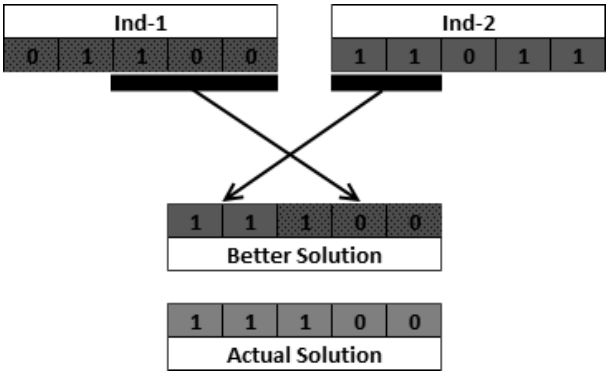

Figure 7. Crossover operation applied on two sample solutions to generate a new solution.

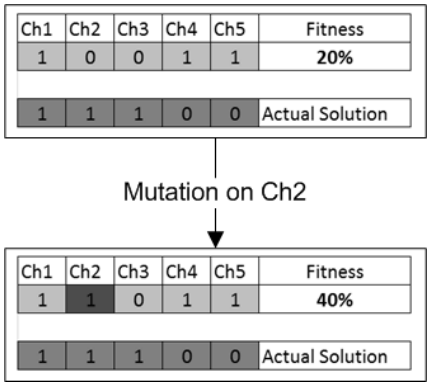

Figure 8. An example to mutation process and its effect is shown on two individuals with five chromosomes.

2.6 Mutation

Mutation operator is applied at each generation to increase the genetic diversity within the solution population. It is achieved by altering a randomly selected gene (i.e. switching its value from 1 to 0 or vice versa) of randomly selected specimen (Figure 8).

Mutation may change a solution entirely and increase its fitness. In this study, the rate at which a mutation is applied to the solution population is a predefined parameter, which has been tuned to maximize diversity in the population. Mutation also prevents population generations from getting stuck in a state of local optimum where fitness of individuals cannot be further improved. On the other hand, excessive mutation may have a negative effect where new individuals are generated without an apparent improvement in fitness.

3 Results

3.1 Dataset

Two separate Pap-Smear data sets are utilized to examine in the experimental setting. First data set includes 150 normal and 150 abnormal specimens provided by the Department of Pathology at Cukurova University, Adana-Turkey and the study is performed in accordance with the Declaration of Helsinki and approved by institutional ethics committee. Obtained samples are delivered from slides processed with Papanicolaou staining. Digital images of slides are taken via a Nikon microscope (Burgerweeshuispad, Amsterdam) equipped with 100x magnification. Samples are down-sized from 2560 x 1920 pixel to 1280x960 pixel resolution and stored in RGB color space in JPEG format. Nucleus and cytoplasm of each specimen in the set are manually segmented by a pathologist in the Department of Pathology of

Cukurova University to serve as ground truth. Each ground truth is prepared in digital binary mask form. To evaluate the robustness of the developed algorithm, the data set is expanded with additional 200 abnormal and 200 normal samples from Herlev Data Set, which consists of single cell images that are ranked in seven groups according to the level of abnormality they display [17]. Sample images from Cukurova and the Herlev sets are displayed in Figure 9.

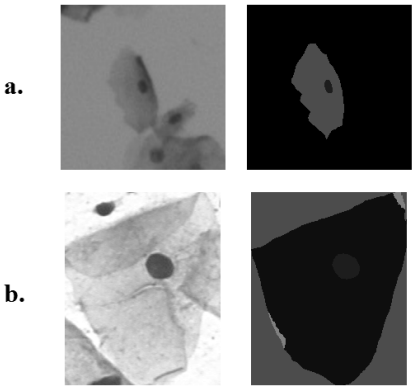

Figure 9. a)A sample image and its empiric area from the utilized data set. b) A sample image and its automated segmented area from Herlev Data Set

Furthermore, Fisher's Iris data base is utilized for examination of performance. It is one of the widely known datasets for classification tests. It consists of 3 classes and each class contains 50 specimens. Each specimen is represented by four features, namely sepal length, sepal width, petal length and petal width [18].

3.2 Performance

Several experimental settings are utilized for objective evaluation of the proposed methodology. In one setting CH and Davies-Bouldin index are compared on clustering process of 60 abnormal and 40 normal randomly selected specimens from our data set. Davies-Bouldin index is another widely used metric for evaluating clusters [19]. In that setting introduced conventional evolutionary algorithm is performed with both of the indexes. Moreover, the algorithm is tested with a case where the best possible solution is known and the genetic operations are performed in order to approach this known solution. Clearly, in this case, the algorithm produces the correct solution faster than the case where the solutions are unknown. The performance of the algorithm in this hypothetical best case is referred to as "gold standard" in this manuscript and our algorithm, together with other available algorithms, is compared to this gold standard case. Accordingly F-score is utilized as golden standard fitness criteria as well as a success criteria [20] which is described below with other measures utilized to evaluate the performance:

$$Accuracy = \frac{Tp + Tn}{(N)} \tag{7a}$$

$$Sensitivity = \text{True Positive Rate} \tag{7b}$$

$$Specificity = \text{True Negative Rate} \tag{7c}$$

$$Precision = \frac{Tp}{(Tp + Fp)} \tag{7d}$$

$$Recall = \frac{Tp}{(Tp + Fn)} \tag{7e}$$

$$Fscore = 2\frac{Precision * Recall}{(Precision + Recall)} \tag{7f}$$

$$Gmean = \sqrt{Sensitivity * Specificity} \tag{7g}$$

where, Tp indicates number of true positives, Fp indicates number of false positives, Fn indicates number of false negatives and sensitivity and specificity indicates true positive and true negative rates respectively. At the end of iteration best fitting individual in most recent generation is considered as solution and classification performance of that individual is evaluated. 100 randomly chosen individuals with 60% abnormality rate is classified with both indexes and F-score as fitness criteria. Results are plotted in Figure 10.

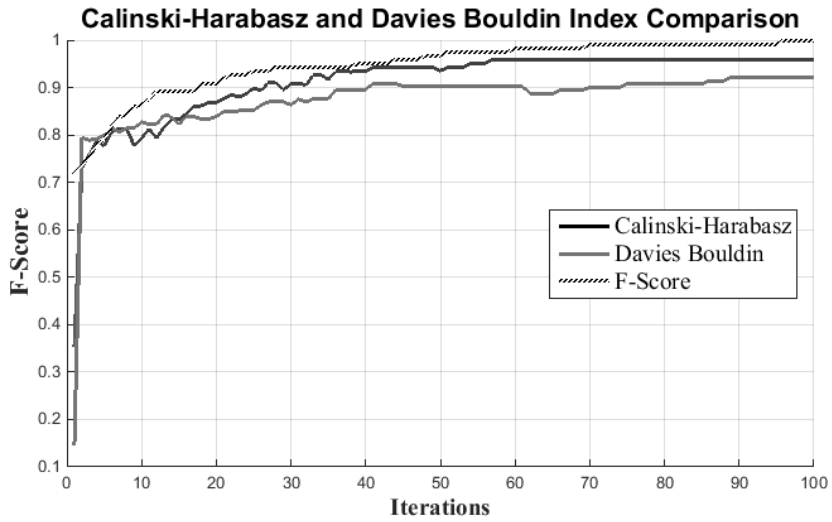

Figure 10. Performance comparison of CH and Davies-Bouldin indexes with F-score.

In another setting, classification performance of the proposed combination of evolutionary computing and CH index is compared with two well-known clustering algorithms. One of them is K-means where each of the n observations (i.e. specimens) is divided into two clusters (normal and abnormal) by assigning them to the nearest mean obtained in the previous iteration [21]

To formulate K-means, let the arrays of observations and centers be defined, respectively, as:

$$X = \{x_1, x_2, x_3..., x_n\} \tag{8}$$

$$V = \{v_1, v_2, v_3..., v_c\} \tag{9}$$

where c is the number of clusters. Center of the jth cluster (v_j) is calculated with observations as:

$$v_j = (\frac{1}{c_j})\sum_{i=1}^{c_j} x_i \tag{10}$$

where c_j stands for the number of observations in jth cluster. Clustering iteration starts with forming random clusters and finding their centers. Then, each observation is assigned to the closer one of the two centers and cluster centers are recalculated. The procedure is repeated until square sums of distance to the center of each cluster attains a minimum:

$$J(V) = \sum_{j=1}^{k}\sum_{i=1}^{n} \|x_i - v_j\|^2 \tag{11}$$

where, k, n, x_i and v_j indicates number of clusters, number of observations, ith observation and centroid of the jth cluster respectively.

The second classifier implemented in the study is the fuzzy C-means (FCM) classifier. FCM algorithm runs on similar principles as the K-means algorithm. However, unlike to K-means, FCM operates on partial membership, where an observation has a degree of belongingness to each cluster [22]. Objective function of FCM with membership degree is given as:

$$j_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\|x_i - c_j\right\|^2 ; 1 < m < \infty \tag{12}$$

where u_{ij} is the degree of membership, x_i is ith observation of data, and c_j is center of the jth cluster. m is a finite real value larger than 1 that sets the level of fuzziness. Accordingly u_{ij} is calculated by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}} \tag{13}$$

where center of the clusters are calculated by:

$$c_j = \frac{\sum_{j=1}^{N} u_{ij}^{m}.x_i}{\sum_{j=1}^{N} u_{ij}^{m}} \tag{14}$$

Comparison tests are performed with both our and Herlev data sets. Results according to performance measures with test parameters are given in Table 3.

Table 3. Performance comparison on pap-smear data

Herlev Data							
Parameters:	Individuals: 100; Abnormal Rate:60; Specimen: 100; Iterations:100; Death Rate: 90; Mutate Rate: 90; Mutate Perc:90;						
Performance							
	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F-measure**	**G-mean**
Genetic Clustering	0,890	0,983	0,750	0,855	0,983	0,914	0,858
FCM	0,850	1	0,625	0,800	1	0,888	0,790
K-Means	0,840	1	0,600	0,789	1	0,882	0,774
Our Data							
Parameters:	Individuals: 100; Abnormal Rate:60; Specimen: 100; Iterations:100; Death Rate: 90; Mutate Rate: 90; Mutate Perc:90;						
Performance							
	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F-measure**	**G-mean**
Genetic Clustering	0,940	0,950	0,925	0,95	0,950	0,950	0,937
FCM	0,920	0,983	0,825	0,893	0,983	0,936	0,900
K-Means	0,910	0,983	0,800	0,880	0,983	0,929	0,886

In addition to pap-smear data, method is also examined on IRIS set. 4 different test groups with 50 randomly selected specimens are derived from first and third classes of IRIS data. Test groups have 13, 15, 17 and 20 random specimens from class 1 of iris database respectively. Each test is repeated 10 ten times and each group is also classified with FCM and K-means clustering algorithms for comparison. Mean and standard deviation values for each specimen group is given in Table 4 with all test parameters.

Table 4. IRIS Data set performance comparison

Parameters:	Individuals: 100;,Iteration:100; Mutate Rate:30; Specimen:50; Death Rate:50; Mutate Perc:10;			
	Mean F-Score	**Std F-Score**	**FCM**	**K-Means**
Test Group 1	0,740	1,17E-16	0,740	0,740
Test Group 2	0,888	1,17E-16	0,857	0,888
Test Group 3	0,967	2,34E-16	0,882	0,857
Test Group 4	0,952	0	0,930	0,888

4 Discussion and Conclusions

In this study, a meta-heuristic classification scheme for malignant cell discrimination is implemented. In the experimental setting, each individual in a generation represents a clustering result where abnormal and normal specimens are tagged with a binary value. At that point internal cluster validation index is utilized as fitness function for determining without any prior information how an individual fits to recent clustering problem. It should be noted that, CH index is utilized for judging the form and compactness of clusters [23] and tested innovatively for malignancy detection.

We have compared the efficiency of the CH index with Davies-Bouldin index, another widely used cluster validation measure which is succesfully utilized as a fitness criteria on occasion [9]. As a minor novelty, the fitness of an individual calculated with F-score is defined as gold standard which allows absolute evaluation of both indexes besides a relative comparison.

Our results (Figure 10) show that CH index has the capability to discriminate better cluster representations. Also, it is possible to conclude that CH performs better than Davies-Bouldin index as a fitness function with respect to the proposed genetic approach for identifying abnormal cervical cells.

Proposed identification methodology is based on a binary unsupervised clustering approach. Accordingly, classification ability of introduced novel combination is compared with two well known unsupervised clustering methods. Analysis (Table 3) on Pap-smear data from two different databases show that the proposed approach outperforms both of the clustering methods, namely FCM and K-means, which leads us to conclude that the method is promising.

Results of another setting (Table 4) is showed that the proposed combination of internal cluster validation and evolutionary algorithm is robust and performing with high sensitivity. It is also possible to conclude that, the performance is independent from extracted features and classified data.

In the future, we intend to further refine the presented classification approaches by optimising CH index for introduced problem. Additionally more specialised and distinctive features can improve the overall success of the methodology.

Abbreviations

CH: Calinski-Harabasz
FCM: Fuzzy C-Means
GA: Genetic Algorithm

Competing interests

The authors declare that they have no competing interests.

Author's contributions

All authors are contributed equally.

Consent for publication

All authors read and approved the manuscript.

References

1. U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," *Pattern recognition*, vol. 33, no. 9, pp. 1455–1465, 2000.
2. T. Jiang and S. De Ma, "Cluster analysis using genetic algorithms," in *Signal Processing, 1996., 3rd International Conference on*, vol. 2. IEEE, 1996, pp. 1277–1279.
3. C. A. Murthy and N. Chowdhury, "In search of optimal clusters using genetic algorithms," 1996.
4. J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.
5. C. Raposo, C. H. Antunes, and J. P. Barreto, "Automatic clustering using a genetic algorithm with new solution encoding and operators," in *International Conference on Computational Science and Its Applications*. Springer, 2014, pp. 92–103.
6. E. R. Hruschka, R. J. Campello, A. A. Freitas *et al.*, "A survey of evolutionary algorithms for clustering," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 39, no. 2, pp. 133–155, 2009.
7. P. Scheunders, "A genetic c-means clustering algorithm applied to color image quantization," *Pattern recognition*, vol. 30, no. 6, pp. 859–866, 1997.
8. A. Li, "The operator of genetic algorithms to improve its properties," *Modern Applied Science*, vol. 4, no. 3, p. 60, 2010.
9. S. Bandyopadhyay and U. Maulik, "Genetic clustering for automatic evolution of clusters and application to image classification," *Pattern recognition*, vol. 35, no. 6, pp. 1197–1208, 2002.
10. V. Roth and T. Lange, "Feature selection in clustering problems," in *Advances in neural information processing systems*, 2004, pp. 473–480.
11. M. E. Plissiti, C. Nikou, and A. Charchanti, "Combining shape, texture and intensity features for cell nuclei extraction in pap smear images," *Pattern Recognition Letters*, vol. 32, no. 6, pp. 838–853, 2011.
12. M. Guven and C. Cengizler, "Data cluster analysis-based classification of overlapping nuclei in pap smear samples," *Biomedical engineering online*, vol. 13, no. 1, p. 159, 2014.
13. E. Bengtsson and P. Malm, "Screening for cervical cancer using automated analysis of pap-smears," *Computational and mathematical methods in medicine*, vol. 2014, 2014.
14. P. W. Poon and J. N. Carter, "Genetic algorithm crossover operators for ordering applications," *Computers & Operations Research*, vol. 22, no. 1, pp. 135–147, 1995.
15. D. M. Deaven and K.-M. Ho, "Molecular geometry optimization with a genetic algorithm," *Physical review letters*, vol. 75, no. 2, p. 288, 1995.
16. T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.
17. J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear benchmark data for pattern classification," *Nature inspired Smart Information Systems (NiSIS 2005)*, pp. 1–9, 2005.
18. R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
19. S. Petrovic, "A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters," in *Proceedings of the 11th Nordic Workshop of Secure IT Systems*. sn, 2006, pp. 53–64.
20. S. Ding, "Feature selection based f-score and aco algorithm in support vector machine," in *2009 Second International Symposium on Knowledge Acquisition and Modeling*, vol. 1. IEEE, 2009, pp. 19–23.
21. T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 881–892, 2002.
22. R. L. Cannon, J. V. Dave, and J. C. Bezdek, "Efficient implementation of the fuzzy c-means clustering algorithms," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 248–255, 1986.
23. U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

