

# Pap Smear Diagnosis Using a Hybrid Intelligent Scheme Focusing on Genetic Algorithm Based Feature Selection and Nearest Neighbor Classification

Yannis Marinakis<sup>1</sup> and George Dounias<sup>2</sup>

<sup>1</sup> Department of Production Engineering and Management  
Technical University of Crete, University Campus, 73100 Chania  
Phone: +30-28210-37282, E-mail: marinakis@ergasya.tuc.gr

<sup>2</sup> Department of Financial & Management Engineering  
University of the Aegean, 31 Fostini Str., 82100 Chios, Greece  
Phone: +30- 22710-35454, E-mail: g.dounias@aegean.gr

**ABSTRACT:** The term Pap smear refers to samples of human cells stained by the so-called Papanicolaou method. The purpose of the Papanicolaou method is to diagnose pre-cancerous cell changes before they progress to invasive carcinoma. In this paper a metaheuristic algorithm is proposed in order to classify the cells. Two databases are used that consist of 917 images and 500 images of Pap smear cells, respectively. Each cell is described by 20 features, and the cells fall into 7 classes but a minimal requirement is to separate normal from abnormal cells, which is a 2 class problem. For the feature selection problem a genetic algorithm is proposed. Genetic Algorithms are search procedures based on the mechanics of natural selection and offer a particularly attractive approach for problems like feature subset selection since they are generally quite effective for rapid global search of large, non-linear and poorly understood spaces. This algorithm is combined with a number of nearest neighbor based classifiers.

**KEYWORDS:** Genetic Algorithms, Feature Selection Problem, Data Mining, Pap-Smear Classification, Nearest Neighbor based Classifiers

## INTRODUCTION

The proposed method for the solution of the feature selection problem is a Genetic Algorithm (GA). Genetic Algorithms are search procedures based on the mechanics of natural selection and natural genetics. The first GA was developed by John H. Holland in the 1960s to allow computers to evolve solutions to difficult search and combinatorial problems, such as function optimization and machine learning [10]. Genetic algorithms offer a particularly attractive approach for problems like feature subset selection since they are generally quite effective for rapid global search of large, non-linear and poorly understood spaces. Moreover, genetic algorithms are very effective in solving large-scale problems. Genetic algorithms [9] mimic the evolution process in nature. GAs are based on an imitation of the biological process in which new and better populations among different species are developed during evolution. Thus, unlike most standard heuristics, GAs use information of a population of solutions, called individuals, when they search for better solutions. A GA is a stochastic iterative procedure that maintains the population size constant in each iteration, called a generation. Their basic operation is the mating of two solutions in order to form a new solution. To form a new population, a binary operator called crossover, and a unary operator, called mutation, are applied [21],[22]. Crossover takes two individuals, called parents, and produces two new individuals, called offspring, by swapping parts of the parents.

In this paper, a novel approach for the solution of the Feature Subset Selection Problem based on a Genetic Algorithm is presented. In the classification phase of the proposed algorithm, a number of variants of the Nearest Neighbor classification method are used [8]. In order to assess the efficacy of the proposed methodologies, the algorithm is used for the Pap-Smear Cell Classification task. The term "Pap-Smear" refers to samples of human cells stained by the so-called Papanicolaou method. The Papanicolaou method is a medical procedure to detect pre-cancerous cells in the uterine cervix. The performance of the proposed algorithm is tested using two data sets of images of Pap-smear cells distributed unequally on 7 different classes. Each cell is described by 20 features extracted from pictures of single human cells. The rest of the paper is organized as follows: In the next section, a detailed analysis of Algorithms for

Classification is presented. In section 3, an analytical description of the proposed algorithm is given. In section 4 the application of the proposed algorithm to the Pap-Smear Cell Classification task is presented while in the last section conclusions and future research are given.

## ALGORITHMS FOR CLASSIFICATION

In recent years, there has been an increasing need for novel data mining methodologies that can analyze and interpret large volumes of data. Selecting the right set of features for classification is one of the most important problems in designing a good classifier. The basic feature selection problem is an optimization problem, with a performance measure for each subset of features to measure its ability to classify the samples. The problem is to search through the space of feature subsets to identify the optimal or near-optimal one with respect to the performance measure. In the literature many successful feature selection algorithms have been proposed. These algorithms can be classified into two categories based on whether features are done independently of the learning algorithm used to construct the classifier. If feature selection depends on learning algorithm, the approach is referred to as a *wrapper model*. Otherwise, it is said to be a *filter model*. Filters, such as mutual information (MI), are based on the statistical tools. Wrappers assess subsets of features according to their usefulness to a given classifier. Unfortunately, finding the optimum feature subset has been proved to be NP-hard [14]. Many algorithms are, thus, proposed to find the suboptimum solutions in comparably smaller amount of time [11]. The Branch and Bound approaches (BB) [18], the Sequential Forward/Backward Search (SFS/SBE) [1],[6] and the Filters approaches [5] deterministically search for the suboptimum solutions. One of the most important of the filter approaches is the Kira and Rendell's Relief algorithm [13]. Stochastic algorithms, including Simulated Annealing (SA) [23], Scatter Search algorithms [15], ACO [2],[3],[20],[24] and Genetic Algorithms (GA) [6] are of great interest recently because they often yield high accuracy and are much faster.

## THE PROPOSED GENETIC ALGORITHM

In this paper, as it has already been mentioned, an algorithm for the solution of the feature selection problem based on a Genetic Algorithm is presented. This algorithm is combined with three nearest neighbour based classifiers, the 1-Nearest Neighbor, the k-Nearest Neighbor and the wk-Nearest Neighbor classifier. The steps of the proposed Genetic Algorithm are given in the following.

### ENCODING

Each individual in the population represents a candidate solution to the feature subset selection problem. Let  $m$  be the total number of features (from these features the choice of the features used to represent each individual is done). The individual (chromosome) is represented by a binary vector of dimension  $m$ . If a bit is equal to 1 it means that the corresponding feature is selected (activated); otherwise the feature is not selected. This is the simplest and most straightforward representation scheme.

### INITIAL POPULATION

The initial population is generated randomly. Thus, in order to explore subsets of different numbers of features, the number of 1's for each individual is generated randomly. Only different individuals are allowed. Thus, in the initial population there are not individuals with the same characteristics. With this way we gain the diversity of the initial population.

### FITNESS FUNCTION

The fitness function gives the quality of the produced member of the population. In this problem, the quality is measured with the Root Mean Squared Error (RMSE) and the overall error. Thus, for each individual the classifiers (1-Nearest Neighbor, k-Nearest Neighbor or wk- Nearest Neighbor) are called and the produced RMSE and the overall error give the fitness function.

## Nearest Neighbor Classifiers

Initially, the classic **1-Nearest Neighbor (1-nn)** method is used [8]. The 1-nn works as follows: In each iteration of the feature selection algorithm, a number of features are activated. For each sample of the test set its Euclidean Distance from each sample of the training set is calculated. The Euclidean Distance is calculated as follows:

$$D_{ij} = \sqrt{\sum_{l=1}^d |x_{il} - x_{jl}|^2} \quad (3)$$

where  $D_{ij}$  is the distance between the test sample  $x_{il}$  and the training sample  $x_{jl}$ , and  $l = 1, \dots, d$  is the number of activated features in each iteration. With this procedure the nearest sample from the training set is calculated. Thus, each test sample is classified in the same class that its nearest sample from the training set belongs.

The previous approach may be extended to the **k-Nearest Neighbor (k-nn)** method, where we examine the k-nearest samples from the training set and, then, classify the test sample by using a voting scheme. The most common way is to choose the most representative class in the training set. Thus, the k-nn method makes a decision based on the majority class membership among the k nearest neighbors of an unknown sample. In other words every member among the k nearest has an equal percentage in the vote. However, it is natural to give more weight to those members that are closer to the test samples. This method is called **Weighted k-Nearest Neighbor (wk-nn)**. In this method, the  $i$  neighbor receives weight

$$w_i = \frac{i}{\sum_{i=1}^k i} \quad (4)$$

Thus, the following hold:

$$w_k \geq w_{k-1} \geq \dots \geq w_1 > 0 \quad (5)$$

$$w_k + w_{k-1} + \dots + w_1 = 1. \quad (6)$$

### 3.3.2 Performance Measures

In order to estimate the solution of the classifiers, a number of performance measures are calculated. In a **2-class problem** if the actual class of a sample is the  $N$ , the estimated class is denoted by  $T_N$  (True Negative) if the sample is classified in its actual class and  $F_P$  (False Positive) if the model misclassifies the sample. On the other hand, the estimated class is denoted by  $T_P$  (True Positive) when the sample is classified correctly in class  $P$  and by  $F_N$  (False Negative) if the sample is misclassified in the class  $N$ . In Table I the definitions are presented.

Estimated Class	Actual Class	
	$N$	$P$
$\hat{N}$	$T_N$	$F_N$
$\hat{P}$	$F_P$	$T_P$

Table I: Definitions of the classified and the misclassified samples.

The results of the models are analyzed based on the Root Mean Squared Error, the Error of Group 1, the Error of Group 2 and the Overall Error. The first measure is calculated from the formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^M (y_i - \hat{y}_i)^2}{M}} \quad (7)$$

where  $M$  is the number of samples in the data set, the  $\hat{y}_i$  is the classifier model output and the  $y_i$  is the true class of the test sample  $i$ . Denoting the number of the false negative samples by  $F_N$ , the number of the true negative samples by  $T_N$ , the number of false positive samples by  $F_P$  and the number of true positive samples by  $T_P$ , we define the following relative performance measures in percentage.

The Error for Group 1: 
$$F_N \% = \frac{F_N}{T_P + F_N} 100 \quad (8)$$

The Error for Group 2: 
$$F_P \% = \frac{F_P}{T_N + F_P} 100 \quad (9)$$

The Overall Error is: 
$$OE \% = \frac{F_N + F_P}{T_P + F_N + T_N + F_P} 100. \quad (10)$$

In a classification problem having more than two classes, the classification errors  $F_N\%$ ,  $F_P\%$  and  $OE\%$  using the definitions above do not make sense. The error of a C class problem can be described using a CxC confusion matrix. The element  $x_{ij}$  in row  $i$  and column  $j$  describes the number of samples, of true class  $j$  classified as class  $i$ , i.e. all truly classified samples are placed in the diagonal and the remaining misclassified samples in the upper and lower triangular parts. The confusion matrix describes the number of errors, but an error rate is obtained by scaling each column to a norm of 100%. Table II shows the confusion matrix of a **7-class** problem where the classes 1,2, and 3 belong to class  $N$  and the classes 4, 5, 6, and 7 belong to class  $P$  and, thus, the classification errors  $F_N\%$ ,  $F_P\%$  and  $OE\%$  of the 2-class problem can easily identified for the 7-class problem.

Estimated Class	Actual Class	
	1-3	4-7
1-3	$T_N$	$F_N$
4-7	$F_P$	$T_P$

Table II: Confusion matrix for a 7-class problem

#### SELECTION OF THE PARENTS

The selection mechanism is responsible for selecting the parent chromosome from the population and forming the mating pool. The selection mechanism emulates the survival of- the-fittest mechanism in nature. It is expected that a fitter chromosome has a higher chance of surviving on the subsequent evolution. In this work, we are using the roulette wheel selection [16] which is one of the most common and easy to implement selection mechanism. Basically, it works as follows: each individual in the population is associated with a sector in a virtual wheel. According to the fitness value of the individual, the sector will have a larger area when the corresponding individual has a better fitness value while a lower fitness value will lead to a smaller sector.

#### CROSSOVER OPERATOR

The most classic crossover operator, the 1-point crossover, is used in the crossover phase of the algorithm. In the 1-point crossover, the two parents are separated in two parts in a specific point. Then the two offsprings takes the 1-part from the one parent and the second part of the other parent. Afterwards for each offspring its fitness function is calculated. For example:

Parent 1: 1 0 0 1 | 1 0 1 0 1

Parent 2: 1 1 1 1 | 0 0 0 0 0

The two parents are separated between the fourth and the fifth feature and the two produced offsprings are the following:

Offspring 1: 1 0 0 1 0 0 0 0 0

Offspring 2: 1 1 1 1 1 0 1 0 1

#### MUTATION OPERATOR

In a specific percentage of the offsprings (25% in our case) a mutation phase is applied. Mutation operates on a single string and generally changes a bit at random. Afterwards for each offspring its fitness function is calculated. For example:

Before mutation: 1 1 1 1 1 0 1 0 1  
 After mutation: 1 1 1 0 1 0 1 0 1

## NEXT GENERATION – SURVIVAL OF THE FITTEST

In the next generation, the fittest from the whole population survives. With the term whole population we mean the initial population and the offsprings from both mutation and crossover phases. Thus, the population is sorted based on the fitness function of the individuals and in the next generation the fittest individuals survived. It must be mentioned that the size of the population of each generation is equal to the initial size of the population.

## STOPPING CRITERIA

There are two stopping criteria for the genetic algorithm. The one is the maximum number of generations, which is a variable of the problem, and the other is the genetic convergence, which means that whenever the solutions of the genetic algorithm converge to one solution the genetic algorithm stops.

## APPLICATION EXAMPLE

### DATA AND PARAMETER DESCRIPTION

The metaheuristic algorithm presented in the previous sections is used for the classification of cells in cervical smears. Cervical smears are cytology specimens taken from the uterine cervix. A specimen is taken from the uterine cervix using a small brush, a cotton stick or wooden stick and transferred onto a thin, rectangular glass plate (slide). The specimen is stained using the Papanicolaou method which is a medical procedure to find pre-cancerous cells in the uterine cervix. This makes it possible to see characteristics of cells more clearly in a microscope. The purpose of the smear screening is to diagnose premalignant cell changes before they progress to cancer [4],[12],[17],[19]. The specimens most often contain cells from the *columnar epithelium* and the *squamous epithelium*. The columnar epithelium is located in the upper part of the cervix and the squamous epithelium in the lower part. Between these two is the *metaplastic epithelium*, also called the transformation zone or the **squamo-columnar junction**. It takes a skilled cyto-technician to differentiate between the different kinds of cells and it is a time consuming job as every glass slide can contain up to 300,000 cells. In the *squamous epithelium* there are 4 layers of cells: the basal, the parabasal, the intermediate, and finally the superficial layer. The *columnar epithelium* only contains a single layer of cells containing columnar cells and reserve cells. The *metaplastic epithelium* consists of reserve cells from the columnar epithelium. In *dysplastic* cells, the genetic information is somehow changed, and the cell will not divide as it should. This is a precancerous cell. Depending on which kind of cell that divides incorrectly, it is given diagnoses like *dysplasia* (i.e. disordered growth) and *carcinoma in situ*. The dysplastic cells are divided into *mild*, *moderate* and *severe* dysplastic.

The classification task is performed using 2 databases, an old one with 500 cells and a new one with 917 cells, built by the Herlev University Hospital. The images were prepared and analyzed by the staff at the hospital using a commercial software package CHAMP (Dimac Imaging – www.dimac-imaging.com) for segmenting the images. The cells were selected, not to collect a natural distribution, but to make a good collection of the important classes (7 classes). For both databases, 20 features are extracted (Table IV). The Pap-smear data fall into 7 classes (Table III), but a minimal requirement is to separate normal from abnormal, which is a 2-class problem.

<b>Old Data</b>	<b>New Data</b>
The old data contains 500 cells with the following distribution:	The new data contains 917 cells with the following distribution:
1. Normal - Columnar epithelial, 50 cells.	1. Normal - Super cial squamous epithelial, 74 cells.
2. Normal - Parabasal squamous epithelial, 50 cells.	2. Normal - Intermediate squamous epithelial, 70 cells
3. Normal - Intermediate squamous epithelial, 50 cells.	3. Normal - Columnar epithelial, 98 cells.
4. Normal - Super cial squamous epithelial, 50 cells.	4. Abnormal - Mild squamous non-keratinizing dysplasia, 182 cells.
5. Abnormal - Mild squamous non-keratinizing dysplasia, 100 cells.	5. Abnormal - Moderate squamous non-keratinizing dysplasia, 146 cells.

6. Abnormal - Moderate squamous non-keratinizing dysplasia, 100 cells.	6. Abnormal - Severe squamous non-keratinizing dysplasia, 197 cells.
7. Abnormal - Severe squamous non-keratinizing dysplasia, 100 cells.	7. Abnormal - Squamous cell carcinoma in situ intermediate, 150 cells.

Table III: Description of classes.

Feature number	Feature	Feature number	Feature
1	Nucleus area	11	Cytoplasm longest diameter
2	Cytoplasm area	12	Cytoplasm elongation
3	N/C ratio (Size of nucleus relative to cell size)	13	Cytoplasm roundness
4	Nucleus brightness	14	Nucleus perimeter
5	Cytoplasm brightness	15	Cytoplasm perimeter
6	Nucleus shortest diameter	16	Nucleus position
7	Nucleus longest diameter	17	Maxima in nucleus
8	Nucleus elongation	18	Minima in nucleus
9	Nucleus roundness	19	Maxima in cytoplasm
10	Cytoplasm shortest diameter	20	Minima in cytoplasm

Table IV: Features describing each cell.

It should be noted that from a medical point of view, it is worse to misclassify an abnormal cell as normal, than oppositely. Since we are looking for abnormal cells this is called a positive finding, while a normal cell is called a negative finding. An abnormal cell misclassified as normal is called a false negative finding, and it is important that a classifier minimizes the number of false negative. By analogy, a normal cell misclassified as abnormal is called false positive.

#### PARAMETERS OF THE PROPOSED ALGORITHM

The data set is divided in two categories the training set and the test set. To test the efficiency of the proposed methods a s-fold validation procedure is utilized. Initially, the data set is divided in s disjoint groups containing approximately M/s samples each, where M is the number of the samples in the data set (for example if a 10-fold validation is used the data set is divided in 10 equal size disjoint groups). Next, each of these groups is systematically removed from the data set, a model is building from the remaining groups (the training set) and, then, the accuracy of the model is calculated using the hidden group (the test set). If the 10-fold validation is used, the procedure is repeated for ten times having different group hidden, and, then, the average accuracies are measured. In this paper, the accuracy of the proposed algorithm is measured using the 10-fold cross validation procedure. In the proposed algorithms we tested different k's. The first choice is the 2-fold validation where the set is divided in two equal parts. The other choices are the 3-fold (66% training set - 33% test set), 4-fold (75% - 25%), 5-fold (80% - 20%), 10-fold (90% - 10%) and the last is 20-fold (95% - 5%). The algorithm was implemented in Fortran 90 and was compiled using the Lahey f95 compiler on a Centrino Mobile Intel Pentium M 750 at 1.86 GHz, running Suse Linux 9.1. As it has already been mentioned two approaches that use different classifiers, the 1-nn, the k-nn and the wk-nn, are used. The first approach is called the GEN-1nn, the second is called GEN-knn and the third is called GEN-wknn. In GEN-wknn and in GEN-knn the value of k is changed dynamically depending on the number of iteration. Each generation uses different k. The reason that k does not have a constant value is that we would like to ensure the diversity of the individuals in each generation. Thus, the genetic convergence will not be achieved too soon and this will probably lead to a better solution. The parameter settings for the genetic based metaheuristic are:

- Population size equal to 1000.
- Number of generations equal to 50.
- Probability of crossover equal to 0.8.
- Probability of mutation equal to 0.25.

#### RESULTS OF 2-CLASS PROBLEM

Table V shows the errors (the Root Mean Square Error, the Error for group 1 and for group 2,  $F_N\%$  and  $F_P\%$ , and the Overall Error) for the proposed algorithms in the 2-class problem for both data sets. We observe that the proposed algorithms give, for example, in the 2-fold Cross Validation, where the data set is divided in two equal parts and the half samples belongs to the training set while the other samples belongs to the test set, very good results as the RMSE is between 0,218828 - 0,242629 for the new data set and is between 0,089443 - 0,107967 for the old data set. In this case, in the new data set although GEN – 1nn has the best performance based on the RMSE, the  $F_P\%$  and the  $OE\%$ , the value of  $F_N\%$  is not as good as in the other method, meaning is equal to 2,669306 in GEN-1nn while for GEN – knn is equal to 1,333556. As the value of s in the s-fold Cross Validation is increasing the results are improved taking into account the values of the four performance measures. For example, the best solution in the 10-fold Cross Validation has in the new data set the RMSE equal to 0,126588 and in the old data set has the RMSE equal to 0 while in the 20-fold Cross Validation in the new data set the RMSE has been improved and is equal to 0,022116 and in the old data set in all methods all the performance measures are equal to 0. From this Table it can be observed that all methods give very good results and these results are almost identical for all methods. However, a higher performance is observed for the GEN-1nn. It should, also, be noted that the value of the most significant performance measure, the  $F_N\%$ , is always smaller than 2,669306 and especially in 10-fold and 20-fold Cross Validation is smaller than 1 and for a number of methods is equal to 0 in the new data set while in the old data is smaller than 1 in all methods independent of the selected cross validation.

	NEW DATA				OLD DATA			
	RMSE	$F_N\%$	$F_P\%$	$OE\%$	RMSE	$F_N\%$	$F_P\%$	$OE\%$
<i>2-fold Cross Validation (50% Training Set – 50% Test Set)</i>								
GEN – 1nn	0,218828	2,669306	10,7438	4,798737	0,107967	1	1,5	1,2
GEN – knn	0,235818	1,333556	17,35537	5,561502	0,089443	0	2	0,8
GEN – wknn	0,242629	1,927028	16,94215	5,889013	0,094868	0,666667	1,5	1
<i>3-fold Cross Validation (66% Training Set – 33% Test Set)</i>								
GEN – 1nn	0,189321	1,62963	9,084362	3,598343	0,07746	0,333333	1,002563	0,600005
GEN – knn	0,192821	1,62963	9,907407	3,816208	0,081155	1	0,995025	0,998004
GEN – wknn	0,197085	1,333333	11,1677	3,925855	0,072957	0,666667	0,995025	0,798403
<i>4-fold Cross Validation (75% Training Set – 25% Test Set)</i>								
GEN – 1nn	0,194807	2,073648	8,674863	3,816214	0,031623	0	1	0,4
GEN – knn	0,19909	1,038145	12,39071	4,033606	0,031623	0	1	0,4
GEN – wknn	0,194807	2,073648	8,674863	3,816214	0,022361	0	0,5	0,2
<i>5-fold Cross Validation (80% Training Set – 20% Test Set)</i>								
GEN – 1nn	0,176904	1,481482	7,87415	3,164053	0,04	0	1	0,4
GEN – knn	0,183068	0,740741	10,76531	3,380851	0	0	0	0
GEN – wknn	0,177229	0,888889	9,532313	3,163459	0,02	0,333333	0	0,2
<i>10-fold Cross Validation (90% Training Set – 10% Test Set)</i>								
GEN – 1nn	0,126588	0,741879	4,95	1,853798	0	0	0	0
GEN – knn	0,152297	0	9,1	2,398471	0	0	0	0
GEN – wknn	0,159971	0	9,95	2,618251	0	0	0	0
<i>20-fold Cross Validation (95% Training Set – 5% Test Set)</i>								
GEN – 1nn	0,022116	0	1,25	0,326087	0	0	0	0
GEN – knn	0,073721	0	4,102564	1,086957	0	0	0	0
GEN – wknn	0,076775	0	4,519231	1,195652	0	0	0	0

Table V: Results of the algorithms in the 2-class problem

The selection of a set of appropriate input feature variables is an important issue in building a good classifier. The purpose of feature variable selection is to find the smallest set of features that can result in satisfactory predictive performance. Because of the curse of dimensionality, it is often necessary and beneficial to limit the number of input features in a classifier in order to have a good predictive and less computationally intensive model. In the Pap-smear cell classification problem analysed in this paper, there are  $2^{20}$  possible feature combinations. The objective of the computational experiments is to show the performance of the proposed algorithm in searching for a reduced set of features with low RMSE. In Table VI, the average number of features that each algorithm selects for both data bases is presented. It can be seen that in all cases the number of features selected are fewer than the total number of features used in order to describe a cell. For the new data set, the minimum average number of features used is equal to 8,5 observed in GEN-knn in 2-fold cross validation while the maximum average number of features used is equal to 11,33

observed in GEN-wknn in 3-fold cross validation. For the old data set, the minimum average number of features used is equal to 8,33 observed in GEN-1nn in 3-fold cross validation while the maximum average number of features used is equal to 13,5 observed in GEN-knn in 2-fold cross validation.

	2-fold	3-fold	4-fold	5-fold	10-fold	20-fold
<b>NEW DATA</b>						
GEN – 1nn	10	10	11	11	10,2	9,4
GEN – knn	<b>8,5</b>	12	10,75	9,2	10,5	9,4
GEN – wknn	9,5	11,33	11,25	10	9,9	10,55
<b>OLD DATA</b>						
GEN – 1nn	11,5	<b>8,33</b>	10	11,2	10,3	9,85
GEN – knn	13,5	11,33	9,25	10	9,8	10,7
GEN – wknn	11,5	12	10,75	9,6	10,1	9,75

Table VI: Results of the algorithms (average number of features used) for the 2-class problem.

## RESULTS OF 7-CLASS PROBLEM

Table VII shows the errors (the Root Mean Square Error and the Overall Error) for the proposed algorithms in the 7-class problem for both data sets (as it has already been mentioned the  $F_N\%$  and the  $F_P\%$  do not make sense in the 7-class problem). We observe that the proposed algorithms give, for example, in the 2-fold Cross Validation, where the data set is divided in two equal parts and the half samples belongs to the training set while the other samples belongs to the test set, very good results as the RMSE is between 0,9989 – 1,019975 for the new data set and between 0,860093 – 0,910777 for the old data set. As the value of s in the s-fold Cross Validation is increasing the results are improved taking into account the values of the four performance measures. For example, the best solution in the 10-fold Cross Validation has in the new data set the RMSE equal to 0,761857 and in the old data set has the RMSE equal to 0,499766 while in the 20-fold Cross Validation in the new data set the RMSE has been improved and is equal to 0,624011 and in the old data set is equal to 0,34175. From this Table it can be observed that all methods give very good results and these results are almost identical for all methods. However, a higher performance is observed for the GEN-1nn.

	NEW DATA		OLD DATA	
	RMSE	OE%	RMSE	OE%
<i>2-fold Cross Validation (50% Training Set – 50% Test Set)</i>				
GEN – 1nn	1,019975	5,125534	0,883509	1,4
GEN – knn	0,9989	6,543083	0,910777	2
GEN – wknn	1,011858	6,107591	0,860093	1,4
<i>3-fold Cross Validation (66% Training Set – 33% Test Set)</i>				
GEN – 1nn	0,910197	4,362299	0,803111	1,798812
GEN – knn	1,000221	5,779492	0,758028	1,598009
GEN – wknn	1,024441	5,671631	0,770776	1,20001
<i>4-fold Cross Validation (75% Training Set – 25% Test Set)</i>				
GEN – 1nn	0,893888	4,687678	0,697431	0,4
GEN – knn	0,932513	5,124359	0,658152	0,8
GEN – wknn	0,951439	4,905544	0,684904	0,8
<i>5-fold Cross Validation (80% Training Set – 20% Test Set)</i>				
GEN – 1nn	0,894883	4,253386	0,682463	0,8
GEN – knn	0,871883	4,034212	0,635022	0,4
GEN – wknn	0,875094	3,926111	0,615672	0,4
<i>10-fold Cross Validation (90% Training Set – 10% Test Set)</i>				
GEN – 1nn	0,782835	3,054228	0,499766	0
GEN – knn	0,761857	3,27162	0,529694	0,2
GEN – wknn	0,790183	3,059006	0,500224	0
<i>20-fold Cross Validation (95% Training Set – 5% Test Set)</i>				
GEN – 1nn	0,624011	1,852657	0,34175	0
GEN – knn	0,66072	1,637681	0,432793	0
GEN – wknn	0,67809	1,958937	0,372304	0

Table VII: Results of the algorithms in the 7-class problem

In the 7-class problem, the same objective with the one of the 2-class problem concerning the number of features selected (i.e., to have a high performance of the proposed algorithm in searching for a reduced set of features with low RMSE) is tested. In Table VIII, the average number of features that each algorithm selects for both data bases are presented. It can be seen that in all cases the number of features selected are fewer than the total number of features used in order to describe a cell. For the new data set, the minimum average number of features used is equal to 8 observed in GEN-knn in 2-fold cross validation while the maximum average number of features used is equal to 12,6 observed in GEN-wknn in 5-fold cross validation. For the old data set, the minimum average number of features used is equal to 8,75 observed in GEN-1nn in 4-fold cross validation while the maximum average number of features used is equal to 13,6 observed in GEN-knn in 5-fold cross validation.

	<i>2-fold</i>	<i>3-fold</i>	<i>4-fold</i>	<i>5-fold</i>	<i>10-fold</i>	<i>20-fold</i>
<b>NEW DATA</b>						
GEN – 1nn	10	10,33	10	9,2	11,5	9,5
GEN – knn	<b>8</b>	9,66	9	12	10,2	9,45
GEN – wknn	9	11	11	12,6	10,4	10,5
<b>OLD DATA</b>						
GEN – 1nn	11,5	10,7	<b>8,75</b>	9,6	10,4	9,4
GEN – knn	12,5	11,7	13	13,6	11,1	10,3
GEN – wknn	13,5	12	13,3	12,6	11,1	9,5

Table VIII: Results of the algorithms (average number of features used) for the 7-class problem.

In Table IX, the times that each feature was selected in the optimal solutions of all algorithms are presented. The five most important features that are used in the algorithms for the two Datasets and for both 2-class and 7-class problems are typed with bold letters. As it can be seen, the three most important features are the fifth feature (Cytoplasm brightness) as it was selected totally 408 times, i.e. in the 77,27% in all solutions, the third feature (N/C ratio (Size of nucleus relative to cell size)) as it was selected totally 366 times, i.e. in the 69,2% in all solutions and the fourth feature (Nucleus brightness) as it was selected totally 328 times, i.e. in the 62,12% in all solutions. The five less important features that are used in the algorithms for the two Datasets and for both 2-class and 7-class problems are typed with italics letters. As it can be seen, the three less important features are the ninth feature (Nucleus roundness) as it was selected totally only 154 times, i.e. in the 29,16% in all solutions, the eighth feature (Nucleus elongation) as it was selected totally 170 times, i.e. in the 32,19% in all solutions and the twelfth feature (Cytoplasm elongation) as it was selected 216 times, i.e. in the 40,9% in all solutions.

Features	<i>2-class problem</i>				<i>7-class problem</i>			
	<b>NEW DATA</b>		<b>OLD DATA</b>		<b>NEW DATA</b>		<b>OLD DATA</b>	
	Times Selected	Average (%)	Times Selected	Average (%)	Times Selected	Average (%)	Times Selected	Average (%)
1	<b>80</b>	<b>60,60606</b>	68	51,51515	79	59,84848	63	47,72727
2	66	50	65	49,24242	68	51,51515	70	53,0303
3	<b>92</b>	<b>69,69697</b>	71	53,78788	<b>117</b>	<b>88,63636</b>	<b>86</b>	<b>65,15152</b>
4	<b>93</b>	<b>70,45455</b>	57	43,18182	<b>84</b>	<b>63,63636</b>	<b>94</b>	<b>71,21212</b>
5	<b>92</b>	<b>69,69697</b>	<b>109</b>	<b>82,57576</b>	<b>94</b>	<b>71,21212</b>	<b>113</b>	<b>85,60606</b>
6	69	52,27273	69	52,27273	<i>51</i>	<i>38,63636</i>	74	56,06061
7	<b>90</b>	<b>68,18182</b>	<b>77</b>	<b>58,33333</b>	<b>83</b>	<b>62,87879</b>	74	56,06061
8	<i>37</i>	<i>28,0303</i>	69	52,27273	<i>33</i>	<i>25</i>	<i>31</i>	<i>23,48485</i>
9	28	<i>21,21212</i>	<i>50</i>	<i>37,87879</i>	<i>40</i>	<i>30,30303</i>	<i>36</i>	<i>27,27273</i>
10	62	46,9697	69	52,27273	73	55,30303	71	53,78788
11	71	53,78788	<i>54</i>	<i>40,90909</i>	67	50,75758	<b>86</b>	<b>65,15152</b>
12	<i>56</i>	<i>42,42424</i>	60	45,45455	<i>43</i>	<i>32,57576</i>	<i>57</i>	<i>43,18182</i>
13	<i>48</i>	<i>36,36364</i>	<i>54</i>	<i>40,90909</i>	63	47,72727	69	52,27273
14	76	57,57576	<i>49</i>	<i>37,12121</i>	<b>88</b>	<b>66,66667</b>	61	46,21212
15	61	46,21212	<b>73</b>	<b>55,30303</b>	68	51,51515	<i>57</i>	<i>43,18182</i>
16	64	48,48485	<b>105</b>	<b>79,54545</b>	38	28,78788	<b>109</b>	<b>82,57576</b>
17	61	46,21212	<i>53</i>	<i>40,15152</i>	60	45,45455	<i>56</i>	<i>42,42424</i>
18	64	48,48485	67	50,75758	67	50,75758	63	47,72727
19	<i>57</i>	<i>43,18182</i>	<b>72</b>	<b>54,54545</b>	<i>57</i>	<i>43,18182</i>	74	56,06061
20	65	49,24242	59	44,69697	73	55,30303	62	46,9697

Table IX. Results of the algorithms (times each feature was selected).

## CONCLUSIONS

In this paper, a classification algorithm is proposed for solving the Pap-smear cell classification problem. Different classifiers are used for the classification problem, based on the Nearest Neighbor classification rule (the 1-Nearest Neighbor, the k-Nearest Neighbor and the wk-Nearest Neighbor) and a Genetic Algorithm method is used for the Feature Selection Problem. The performance of the proposed algorithm is tested using two data sets of Pap-smear cells. The obtained results indicate the high performance of the proposed algorithm in searching for a reduced set of features (in almost all cases less than 50% of all features are used) with high accuracy and in achieving excellent classification of Pap-smear cells both in 2 classes and in 7 classes. Future research is intended to be focused in using different than the Nearest Neighbor classifiers and different algorithms for the Feature Selection Problem.

## REFERENCES

- [1] Aha D.W. and Bankert R.L., 1996, "A Comparative evaluation of sequential feature selection algorithms". In *Artificial Intelligence and Statistics*, Fisher D. and Lenx J.-H. (Eds.), Springer-Verlag, New York.
- [2] Al-Ani A., 2005a, "Feature subset selection using ant colony optimization". *International Journal of Computational Intelligence*, 2(1), pp. 53-58.
- [3] Al-Ani A., 2005b, "Ant colony optimization for feature subset selection". *Transactions on Engineering, Computing and Technology*, 4, pp. 35-38.
- [4] Byriel, J., 1999, "Neuro-fuzzy classification of cells in cervical smears". Master's Thesis, Technical University of Denmark: Oersted-DTU, Automation.
- [5] Cantu-Paz E., 2004, "Feature subset selection, class separability, and genetic algorithms". *Genetic and Evolutionary Computation Conference*, pp. 959-970.
- [6] Cantu-Paz E., Newsam S. and Kamath C., 2004, "Feature selection in scientific application". *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 788-793.
- [7] Dorigo M. and Stützle T., 2004, *Ant Colony Optimization*. A Bradford Book, The MIT Press Cambridge, Massachusetts, London, England.
- [8] Duda R.O. and Hart P. E., 1973, *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York,.
- [9] Goldberg, D. E., 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Publishing Company, INC, Massachussets.
- [10] Holland, J. H., 1975, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI.
- [11] Jain A. and Zongker D., 1997, "Feature selection: Evaluation, application, and small sample performance". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, pp. 153-158.
- [12] Jantzen, J., Norup, J., Dounias, G. and Bjerregaard B., 2006, "Pap-smear benchmark data for pattern classification". (submitted).
- [13] Kira K. and Rendell L., 1992, "A practical approach to feature selection". *Proceedings of the Ninth International Conference on Machine Learning*, Aberdeen, Scotland, pp. 249-256.
- [14] Kohavi R. and John G., 1997, "Wrappers for feature subset selection". *Artificial Intelligence*, 97, pp. 273-324.
- [15] Lopez F.G., Torres M.G., Batista B.M., Perez J.A.M. and Moreno-Vega. J.M., 2006, "Solving feature subset selection problem by a parallel scatter search". *European Journal of Operational Research*, 169, pp. 477-489.
- [16] Marinakis Y., Migdalas, A., and P. M. Pardalos, 2005, "A Hybrid Genetic-GRASP algorithm Using Langrangean Relaxation for the Traveling Salesman Problem", *Journal of Combinatorial Optimization*, 10, pp. 311-326.
- [17] Martin, E., 2003, "Pap-smear classification", Master's Thesis, Technical University of Denmark: Oersted-DTU, Automation,.
- [18] Narendra P.M. and Fukunaga K., 1977, "A branch and bound algorithm for feature subset selection". *IEEE Transactions on Computers*, 26(9), pp. 917-922.
- [19] Norup, J., 2005, "Classification of pap-smear data by transductive neuro-fuzzy methods", Master's Thesis, Technical University of Denmark: Oersted-DTU, Automation,.
- [20] Parpinelli R.S., Lopes, H.S., and Freitas, A.A., 2002, "An ant colony algorithm for classification rule discovery". In *Data mining: A heuristic approach*, Abbas H., Sarker R. and Newton C. (Eds.), London, UK: Idea group publishing, pp. 191-208.
- [21] Reeves, C. R., 1995, "Genetic Algorithms", *Modern Heuristic Techniques for Combinatorial Problems*, Reeves, C. R. (Ed.), McGraw - Hill, London, 151-196.

- [22] Reeves, C. R., 2003, "Genetic Algorithms", Handbooks of Metaheuristics, Glover, F. and G. A. Kochenberger (Eds.), Kluwer Academic Publishers, Dordrecht, 55-82.
- [23] Siedlecki W. and Sklansky J., 1988, "On automatic feature selection". International Journal of Pattern Recognition and Artificial Intelligence, 2(2), pp. 197-220.
- [24] Shelokar P.S., Jayaraman V.K. and Kulkarni B.D., 2004, "An ant colony classifier system: application to some process engineering problems". Computers and Chemical Engineering, 28, pp. 1577-1584.
- [25] Zhang C. and Hu H., 2005, "Ant colony optimization combining with mutual information for feature selection in support vector machines". In AI 2005, LNAI 3809, Zhang S. and R. Jarvis (Eds.), pp. 918-921.