Pattern Analysis and Applications (2020) 23:1709–1726
https://doi.org/10.1007/s10044-020-00873-w

SHORT PAPER

An instance and variable selection approach in pixel-based classification for automatic white blood cells segmentation

Nesma Settouti[1] · Meryem Saidi[1] · Mohammed El Amine Bechar[1] · Mostafa El Habib Daho[1] · Mohamed Amine Chikh[1]

Received: 7 February 2019 / Accepted: 4 February 2020 / Published online: 14 February 2020

Abstract

Instance and variable selection involve identifying a subset of instances and variables such that the learning process will use only this subset with better performances and lower cost. Due to the huge amount of data available in many fields, data reduction is considered as an NP-hard problem. In this paper, we present a simultaneous instance and variable selection approach based on the Random Forest-RI ensemble methods in the aim to discard noisy and useless information from the original data set. We proposed a selection principle based on two concepts: the ensemble margin and the importance variable measure of Random Forest-RI. Experiments were conducted on cytological images for the automatic segmentation and recognition of white blood cells WBC (nucleus and cytoplasm). Moreover, in order to explore the performance of our proposed approach, experiments were carried out on standardized datasets from UCI and ASU repository, and the obtained results of the instances and variable selection by the Random Forest classifier are very encouraging.

Keywords Instance and variable selection · Random Forest · Data reduction · Small target detection · Automatic segmentation · Pixel-based classification · White blood cells

1 Introduction

Nowadays, the huge amount of data available in many fields makes the search of an optimal subset from a large-size dataset an NP-hard problem. The data reduction process aims to clean the original dataset by removing redundant, missing and useless instances and/or features. The classifier build using this dataset should be as good or nearly good as the one built from the whole dataset.

In the context of medical image segmentation, the aim is to build an algorithm that takes an image as its input and results out the segmentation of the region of interest (ROI). The small target detection problems held the attention of many researchers [19, 27, 31, 32, 61]. Generally, image segmentation was applied by several techniques as thresholding, edge-based segmentation, region-based segmentation or segmentation based on pixel-based classification. However, the segmentation based on pixel-based classification is time-consuming due to the high number of instances and variables (features) which represent each pixel characteristics. It is quite clear that we do not need all the variables to classify all pixels in an image. Specifically, certain relevant features can be conveniently summarized by looking at the relative positioning color or texture of various ROI. However, in image classification, many other potential variables may be used, including color or spatial signatures, textural or contextual information. On the other hand, training samples are usually collected from fieldwork. The different collection strategies used, such as single pixel, seed, and super pixel, would influence classification results, especially for classifications when the regions of interest (ROIs) in a medical image are complex and heterogeneous.

Due to different capabilities in ROIs separability, the use of too many variables and noisy/redundant instances in a classification procedure may decrease classification accuracy and unnecessary increase in the computational cost. It is important to select only the instances and variables that are most useful for separating the different ROIs, especially in medical image processing. Thereby, selecting suitable instances and variables is a critical step for successfully implementing an image segmentation. Through several

✉ Nesma Settouti
nesma.settouti@univ-tlemcen.dz

[1] Biomedical Engineering Laboratory GBM, University of Tlemcen, Tlemcen, Algeria

studies, it has been proven that instance and variable selection can:

- Improve the performance prediction of the model (by removing noisy instances and variables with 'negative' influence for recognition)
- Provide faster and more cost-effective implementations in contexts where datasets have thousands or hundreds of thousands of instances and variables.

In this work, we are concerned by the problem of small target detection for the automatic recognition of white blood cells in cytological images, as well as the recognition of nucleus and cytoplasm which is a big help for hematologists to diagnose leukemia, AIDS, blood cancer and other diseases. The pixel-based classification with the ensemble method Random Forest [9] has proved a great capacity of recognition and segmentation of ROIs in several works [12, 29, 30, 56, 62]. The biggest disadvantage of this image processing scheme is the computational complexity due to the huge amount of data. As a solution, we propose an instance and variable selection approach called IVsel which improves the segmentation performances and reduces the computational cost.

The proposed IVsel algorithm uses the power of ensemble methods to perform instance and variable selection based on two concepts: the ensemble margin and the importance of variables in Random Forest. These concepts allow us to rank the instances and variables in the learning set and evaluate their relevance during the image segmentation process. When performance is important, as it often is, the choice of a fast algorithm that uses the available computing resources efficiently is essential. We shall, in fact, take the efficiency of the compared algorithms by measuring the mean of the amount of time they take. Instead of the evolutionary algorithms, which record the best performances in this field, their major limits are an increasing computation cost in big datasets. By its principle, IVsel saves computing time while maintaining classification performances. Moreover, IVsel can be generalized, and experiments on the UCI [39] and ASU [64] machine learning repository datasets show efficiency in not only segmentation process on clinical images but also in a simple classification task.

The rest of this article is organized in six sections. In the first one, we present a review of some existing methods of instance and variable selection algorithms. In second one, we introduce our instance and variable selection (IVsel) approach and the algorithms used for comparison. In the third section, we explain the different stages (feature extraction selection step and classification method) of white blood cells segmentation using our approach. In the fourth section, we apply our method on several classification problems from UCI and ASU repositories using the random forests classifier. Finally, we present a conclusion that summarized the impact of our work and the tracks defining possible perspectives for future work.

2 Related work for instance and variable selection

The statistic literature contains a whole set of techniques to identify the "relevant" coordinates of a dataset. The instance selection methods are used to extract the most useful set of instances from a database that contains noisy instances. This is the same as variable selection methods that consist to reduce the number of variables, particularly when the variable space is important and computational performance issues are induced. These techniques are rightfully extensively used for image processing as they are relatively easy to implement.

Usually, instance and variable selection can be performed one after the other. In Feature Selection, Instance Selection (FSIS), a variable selection algorithm is applied on the original dataset followed by an instance selection algorithm on the dataset obtained from the first step, Instance Selection, Feature Selection (ISFS) is the opposite approach.

Another possibility is the Feature and Instance Selection (FIS) algorithm where instance and variable selection are applied simultaneously. This approach is more interesting, because the selection algorithm considers all the data.

Tsai et al. [58] conduct a study to examine the performance obtained when both tasks are executed individually or in certain orders with the genetic algorithm (GA), and they conclude that performing variable selection then instance selection (FSIS) provides better classification results than performing instance selection first (ISFS). They also noticed that the use of instance selection or variable selection individually shown better result than the use of ISFS or FSIS in small-scall datasets. In [37], the authors perform an IFIS and a FSIS using the locality-sensitive hashing instance selection F algorithm and a Pearson R test.

Otherwise, the use of instance and variable selection is suitable for large datasets since this step greatly reduces the computational cost of training classifiers.

One of the most widely used techniques for variable and instance selection are evolutionary and coevolutionary algorithm. Different examples can be found in the literature:

Ishibuchi et al. [26] perform an instance and feature selection with a genetic algorithm to improve the classification ability of *k*-NN and neural networks.

In [46], Ramirez-Cruz et al. present an hybrid algorithm called IFS-IBGAES, a combination of a genetic algorithm (GA) and evolution strategies (ES) to solve the problem of instance selection and variable weighting for instance-based methods. GA has the purpose of selecting instances, whereas

ES is used to weight variables, the proposed algorithm increases the predictive accuracy of the *k*-NN classifier. A modification of GA based on the biological evolution (GBA) is proposed by Chen et al. [11] to an instance and feature selection for traffic sign recognition.

In cooperative coevolution, two or more populations called species evolve separately in order to solve a specific problem. Derrac et al. [14] employ this model with the CHC (cross-generational elitist selection strategy, Heterogeneous recombination, cataclysmic mutation) evolutionary algorithm to perform an instance and variable selection in *k*-NN using three different populations. In Derrac et al. [15], they also use the cooperative coevolution model in instance selection, instance weighting, and variable weighting for the nearest neighbor classifiers and obtain good results in both papers. In Perez-Rodriguez et al. [45], a CHC evolutionary algorithm is used to perform a simultaneous instance and feature selection and weighting.

Ros et al. [47] propose the idea of integrating scaling methods with genetic algorithms to feature and instance selection. Garcia-Pedrajas et al. [20] propose the scalable simultaneous instance and feature selection method (SSIFSM) which applies a selection algorithm to subsets of the whole training set and use a voting scheme to combine the results to speed up the selection process on each subset. In [21], they use a scalable memetic algorithm for simultaneous instance and feature selection.

Villuendas-Rey et al. [60] propose a deterministic method of variable and instance selection based on rough set theory and structuralizations of the logical combinatorial approach to pattern recognition to improve nearest neighbor classifiers, and they obtained high data reduction and still maintain the original classifier error. Sakinah et al. [50] perform a variables and instances selection based on the cooperative particle swarm optimization technique on regression problems. In [13], authors select relevant features and retain important instances simultaneously by the construction of the new algorithm based on the combination of FortalFS and DemoIS selection algorithms. Zhang et al. [65] propose a unified criterion for feature and instance selection (UFI), to perform an instance and feature selection in an unsupervised framework. In [57], authors use a simple adaptation of the simulated annealing meta-heuristic to solve the feature and instance selection problems.

After reviewing the principal approaches, we found that evolutionary algorithm obtains the best performances but suffer from an increase in computation cost in big datasets. In this work, we present, IVsel, an approach based on the Random Forest ensemble method which overrides the evolutionary issues. Its principle lies in ranking instances and variables based on the ensemble margin and the importance variable concepts.

3 Methods

In this paper, we present IVsel (instance and variable selection) algorithm, an approach based on the ensemble method: Random Forest-RI (RF-RI). We compare the performances of the three well-known evolutionary algorithms and the IVsel algorithm in terms of accuracy and execution time.

3.1 Instance and variable selection (IVsel) approach

Random forests [9] are among the most popular machine learning methods due to their applicability to a wide range of problems and their relatively good accuracy, robustness, and ease of use. This algorithm is based on the use of two randomization principles namely:

- The Bagging method introduced by Breiman [8]: Its principle is to draw a large number of samples, independently of each other, and to build, applying to each of them the same basic rule, from which results a varied collection predictors. The predictor collection is then aggregated by simply averaging or majority voting.
- The random feature selection introduces randomness in the choice of partitioning rules at each node of the trees, so that each rule is no longer chosen from the complete set of variables M, but from a subset of these characteristics. More specifically, select a number F of features $(1 \leq F \leq M)$ [55] by random sampling without replacement and choose the best possible rules using Gini index on these features.

Thus, the Random Forests is a variant of Bagging, where the difference comes in the construction of individual trees. The draw, at each node, of the F variables is done without replacement and uniformly among all M variables.

The number F is set at the beginning of the forest's construction and is therefore identical for all trees and for all the nodes of the same tree but the F variables involved in the nodes are generally different. The Random Forest process shown in Fig. 1 is summarized as follows:

Let Ntrees be the number of trees to build, for each N iterations:

1. Select a new bootstrap sample from training set
2. Build an un-pruned tree on this bootstrap.
3. At each internal node, randomly select F attributes and determine the best split using Gini index.
4. Save tree constructed using the CART methodology.
5. Output overall prediction as the average response (regression) or majority vote (classification) from all individually trained trees.

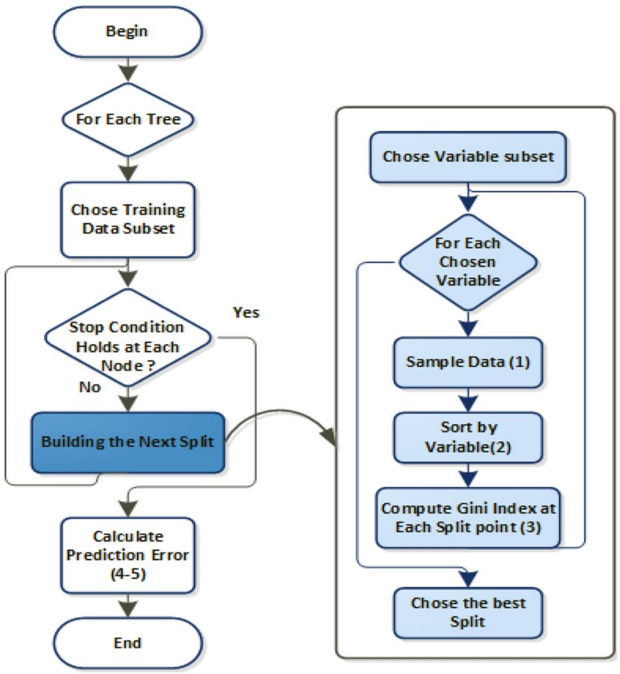


Fig. 1 Flowchart of Random Forest (RF)

The random forest algorithm has become a major data analysis tool used with success in various scientific areas. Indeed, it not only used for prediction, but also to assess how important a variable is by calculating how much out-of-bag performance you lose when you scramble the values of the variable. The profit of this measure has been demonstrated in a large number of studies [23, 28, 34, 40, 51, 63].

On the other hand, random forest also provide an interesting function for evaluating the training instances based on the ensemble margin paradigm [52]. This technique selects the most informative instances based on their margins. Thus, more the margin is close to 1, the confidence in the prediction is great; on the contrary, when the margin is low, confidence in the classification for the instance in question is low.

Therefore, the proposed IVsel algorithm (Algorithm 1) use these two concepts of ensemble methods to perform a variable and instance selection, i.e., the ensemble margin [52] and the importance variable [9]. These concepts allow us to assign a ranking to instances and variables in the training set which assess their relevance to the learning process.

3.1.1 The ensemble margin

The ensemble algorithms use the ensemble margin concept to estimate the performances of the ensemble. The margin expresses the level of disagreement between the learners. An instance correctly classified by the ensemble reaches a high margin; otherwise, the margin value is very small. Generally, central instances have a high margin and carry general information on the cluster. On the other hand, instances near boundaries are more informative and have low margin [52]. In this work, we use the unsupervised ensemble margin as applied by [7, 22, 24, 36, 38, 48]. It is an alternative definition to the classical margin, it ranges from 0 to + 1 and is computed by Eq. 1. An interesting propriety of unsupervised margin is that it does not require the true class labels of instances which is more robust to class noise.

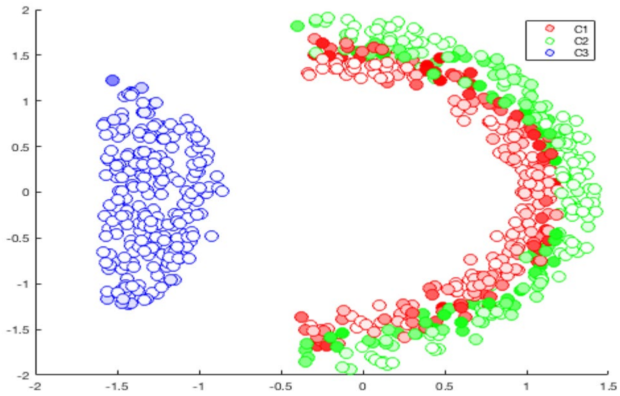


Fig. 2 Ensemble margin representation of an artificial dataset (color figure online)

$$\text{Margin}(x) = \frac{n_{c1} - n_{c2}}{\text{Ntrees}} \tag{1}$$

where n_{c1} represents the number of vote for the most voted class for instance x. n_{c2} represents the number of vote for the second most voted class for instance x. Ntrees represents the number of classifiers in the ensemble.

In our approach, we use the ensemble margin value as metrics to rank the instances of the training set. The instances causing the highest level of disagreement (with low margin value) are considered the most informative for class discrimination, whereas the central instances carry general information about the cluster. Therefore, for the instance reduction process, as applied in our earlier work, [48] is to select a high percentage of border instances α_1 (informative instances, such as samples in class boundaries or those belonging to difficult classes) and a low percentage of central instances α_2 (instances that have been classified by the majority of classifiers in the same class) in order to have a better representation of the dataset.

An illustration of the ensemble margin value of an artificial dataset is presented in Fig. 2. This dataset contains three clusters. The instances were colored based on their margin value in shades of red (class 1), green (class 2) and blue (class 3). White points (empty circles) represent the high margin instances and the dark points (filled circles) are the low margin instances. We notice that the darkest points (low margin) are those belonging to boundaries and the white points are the central instances.

Figure 3 shows that if we decide to select only low margin instances, the cluster representing “class 3” will be entirely removed. On the other hand, retaining a high number of low margin instances allow good discrimination between classes. So, we conclude that the best combination is to maintain a high percentage of low margin instances and a small percentage set of high margins instances.

3.1.2 The importance variable measure

The importance variable is evaluated for each variable VI by removing the association between that variable and the target Y. To break the link between X_i and Y, Breiman [9] proposed to randomly permute the observations of the VI and measure how much the permutation decreases the accuracy of the model. This process is achieved by randomly permuting the values of the variable.

The formalism of this approach as presented in [9] is defined as follows:

Let consider an out-of-bag (OOB) set $\overline{D}_n^t = D_n/D_n^t$ where $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is a learning set of n, and $D_n^t, t = 1, \ldots,$ Ntrees which contains the observations selected in the bootstrap subsets. $\overline{D}_n^{tj}, t = 1, \ldots,$ Ntrees is the permuted out-of-bag samples obtained by the permutation of the values of the jth variable in each out-of-bag subset.

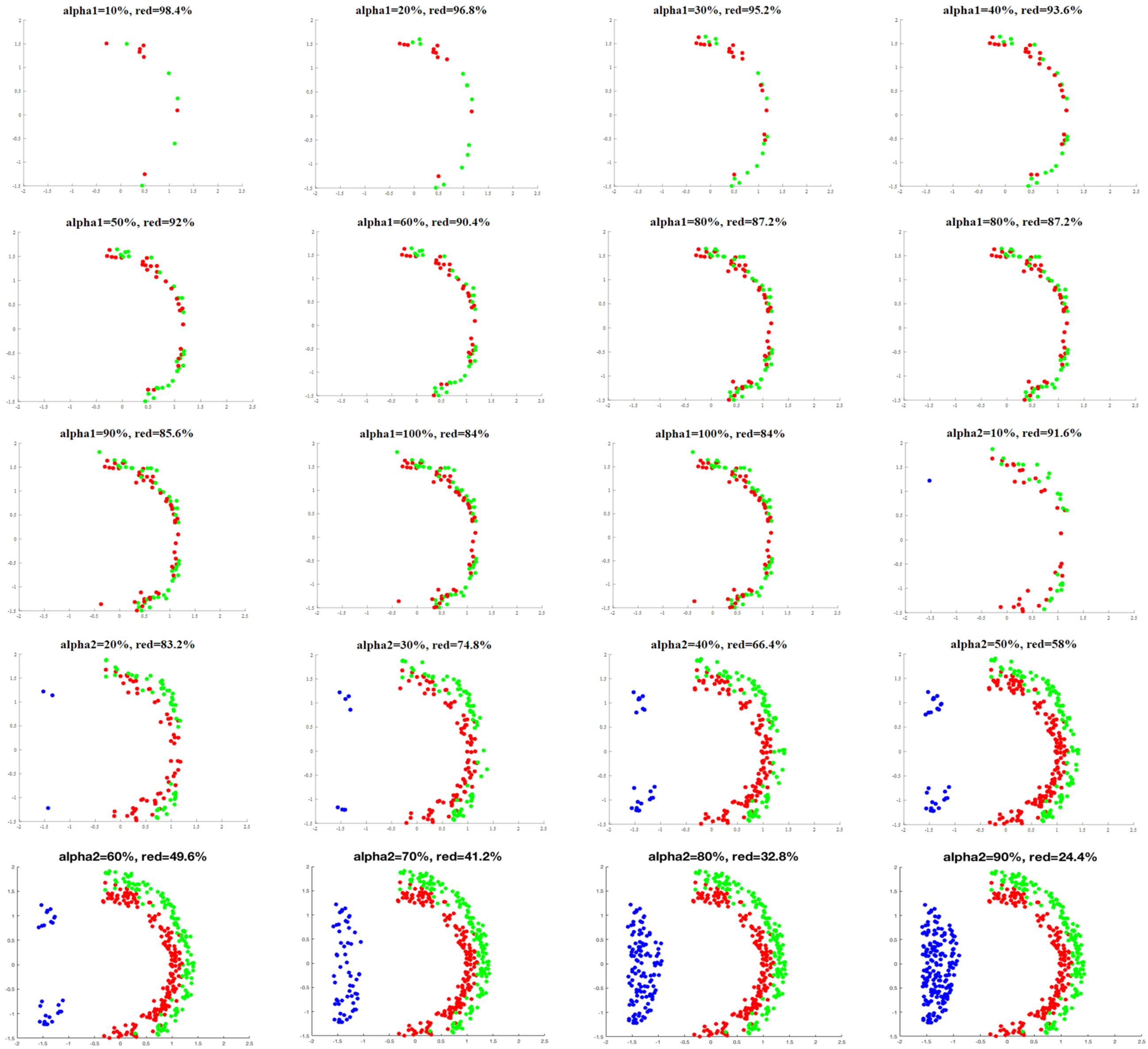


Fig. 3 Selected instances for different values of α_1 and α_2

The importance measure of the variable X_j is calculated by Eq. 2 as formalized in Breiman [9]:

$$\mathrm{VI}(X_j) = \frac{1}{\mathrm{Ntrees}} \sum_{t=1}^{\mathrm{Ntrees}} [R(h_t, \overline{D}_n^{tj}) - R(h_t, \overline{D}_n^{t})] \quad (2)$$

with $\overline{D}_n^t$ out-of-bag of D_n^t, where D_n^t is the bootstrap samples of the training data D_n used for building the trees over Ntrees. $\overline{D}_n^{tj}$ the permuted out-of-bag samples on the *j*th variable. $h_t$the hypothesis of prediction the *t*th tree. $R(h_t)$the prediction error of h_t.

To estimate the importance of a specific variable X_j of a *t*th tree as shown in Fig. 4:

- First, the prediction error on the out-of-bag samples $\overline{D}_n^t$ is measured.
- Then, the values of the variable in the out-of-bag samples $\overline{D}_n^t$ are randomly permuted, keeping all other variables the same $(\overline{D}_n^{tj})$.
- Finally, the prediction error difference between the permuted data $\overline{D}_n^{tj}$ and the original out-of-bag samples $\overline{D}_n^t$ is measured. The mean increase in error prediction across the *t*th trees is reported.

The proposed IVsel algorithm (Algorithm 1) proceeds in the following steps:

- At first, a random forest-RI equals to Ntrees is built, and during the evaluation phase, the variable importance measure is assessed.
- Second, the margin ensemble is calculated according to Eq. 1.
- Subsequently, variables are ranked in descending order of importance and instances are ranked by margin value, with the references: α_1 low margin instances, and α_2 as high margin.
- Finally, we obtain a database with variables and relevant instances.

Algorithm 1 IVsel Algorithm

```
Input: $S = ((x_1, y_1), \ldots, (x_m, y_m)))$.
Execute the RF-RI algorithm.
$Mg(S)$ = Calculate margin value.
The Variable Importance : Ranking variables in decreasing order of importance.
The Ensemble Margin: Ranking instances in increasing order of margin value.
    $S1 \leftarrow \alpha_1\%$ of low margin instances
    $S2 \leftarrow \alpha_2\%$ of high margin instances
$S' \leftarrow S1 \cup S2$
$S' \leftarrow S'$ with the most important variables.
```

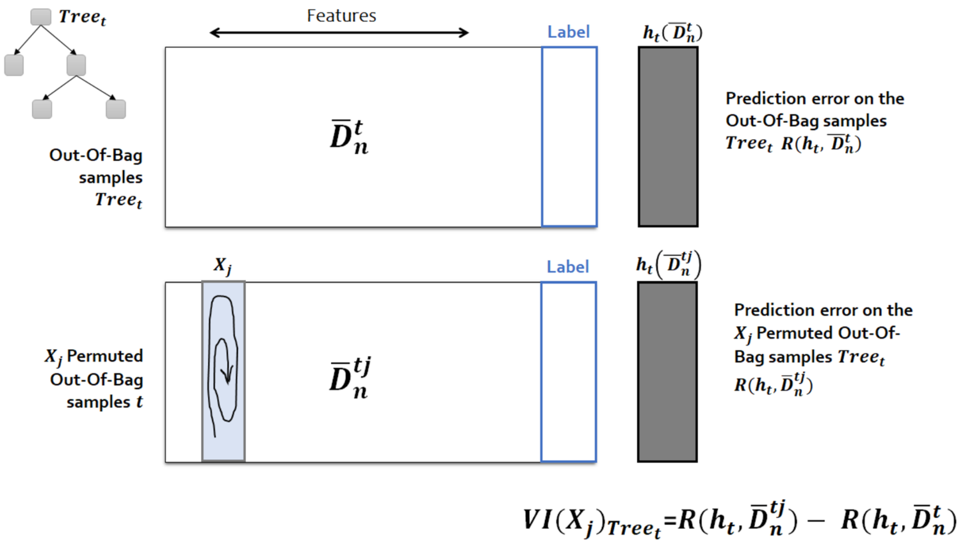

$Tree_t$
Features
Label
$h_t(\overline{D}_n^t)$
Out-Of-Bag samples $Tree_t$
$\overline{D}_n^t$
Prediction error on the Out-Of-Bag samples $Tree_t$ $R(h_t, \overline{D}_n^t)$
X_j
Label
$h_t(\overline{D}_n^{tj})$
X_j Permuted Out-Of-Bag samples t
$\overline{D}_n^{tj}$
Prediction error on the X_j Permuted Out-Of-Bag samples $Tree_t$ $R(h_t, \overline{D}_n^{tj})$
$VI(X_j)_{Tree_t}=R(h_t, \overline{D}_n^{tj}) - R(h_t, \overline{D}_n^t)$

Fig. 4 Measure importance principle of X_j for a *t*th tree

A deep look in the literature shows that the majority of approaches of instances and variables selection are time-consuming. Previous researches [26, 44, 58] show that evolutionary models generally outperform classical instance and/or feature selection algorithms in very complex problems reducing data in machine learning field without increasing classification error. In this paper, we select the three most accurate evolutionary algorithms, PBIL [2, 4], CHC [10] and IFS_CoCo [14, 15], to perform a comparison with the proposed reduction approach.

3.1.3 The population-based incremental learning algorithm

The population-based incremental learning algorithm (PBIL) is an evolutionary algorithm which combines genetic algorithm and competitive learning [2, 4]. Contrary to standard genetic algorithms (GAs), PBIL (Algorithm 2) use a probability vector ($V_p = \{p_1, p_2, \ldots, p_i\}$, where p_n represents the probability of obtaining a value of 1 in the ith component from which samples can be drawn to produce the next generation's population.

```
Input A training set $T$, Number of variables $n$, Population size $L$, Number of generations $G$,
   Learning rate $\lambda$
m ← |T|
$P$ ←initialize probability vector. % (Each position = 0.5)
for $j = 1 \ldots G$ do
    for $i = 1 \ldots L$ do
        $x_i$ ← generate sample vector according to probabilities in $P$.
        $evaluation_i \leftarrow evaluate(x_i)$
    end for
    $max$ ← find vector corresponding to maximum $evaluation$ % Find Best Sample
    for $i = 1 \ldots m + n$ do
        % Update Probability Vector
        $P_i \leftarrow P_i * (1.0 - \lambda) + max_i * (\lambda)$
    end for
end for
```

Initially, the values of V_p are set at 0.5. Then, at each generation, we generate M solutions based on the probabilities in the probability vector $pl(x)$. The N best solutions ($N \leq M$) are selected as the best solutions set and used to update the probability vector with: $P_i \leftarrow P_i * (1.0 - \lambda) + \max_i * (\lambda)$, where λ is the learning rate. After, a new population is generated from the updated probability vector. The process is repeated for a G number of generations.

Quickly, the values of V_p will be changed to favor either 0.0 or 1.0 through the search's progression. For example, a final probability vector of a good solution of the proposed problem would be 0.01, 0.98, 0.02, 0.99, etc [3, 25]. Notice that the population represented by a probability vector is not unique, which aids in maintaining diversity in search.

We assume that the training set is formed by m labeled instances of n variables. To perform instance and variable selection, each chromosome is coded as indicated by Eq. 3:

$$C = a_1 a_2 \ldots a_m a_{m+1} \ldots a_{m+n} \tag{3}$$

3.1.4 CHC adaptive search algorithm

The CHC algorithm is an evolutionary algorithm proposed by Eshelman [18] (cross-generational elitist selection strategy, heterogeneous recombination, cataclysmic mutation). CHC generates the offspring by exchanging half of the bits that differ between parents separated by a threshold Hamming distance (incest prevention). Then, the parent and the offspring are merged, and only the N best individuals are selected for the new population. In case that a parent and an offspring have the same fitness value, the offspring is selected.

No mutation is applied during the recombination phase. Instead, when the population converges or the search stops making progress, the population is reinitialized by randomly changing 35% of the bits of the best solution.

```
Input A training set $T$, number of generations $G$, a base learning algorithm
    $L$, population size $N$
$S \leftarrow$ initialize population
for $j = 1...G$ do
    Obtain new individuals using HUX crossover
    Apply random mutation with probability
    % Evaluation of individuals
    for $i = 1...N$ do
        Train a classifier $f = L(S_i)$
        Evaluate error $e$ of $f$
    end for
    Select best individuals for the next generation.
end for
Return best individual
```

CHC also employs a method of incest prevention. Only different individuals separated by a threshold Hamming distance are allowed to mate. No mutation is applied during the recombination phase. Instead, when the population converges or the search stops making progress, the population is reinitialized by randomly changing 35% of the bits of the best solution.

3.1.5 Cooperative coevolution model

The IFS_CoCo algorithm can handle two or more populations simultaneously. Indeed, each population is responsible for solving a part of the original problem using a divide-and-conquer strategy. Each spice evolve with its own evolutionary algorithm without interaction between them which makes parallel implementation possible. Finally, the global solution is the combination of the solutions presented by representative individuals taken from each population.

The IFS CoCo model manage three populations, each one performs a specific selection task:

- *IS population* Performs an instance selection.
- *FS population* Performs a features selection.
- *FIS population* Performs a features and instances selection.

In IFS_CoCo, the CHC algorithm is applied to evolve the three populations. The individuals of the three populations use the same binary structure. In the IS and FS population, each chromosome of the phenotypes represents an instance and a feature, respectively. In IFS population, the first N chromosomes represent the instances and the M last ones represent the features.

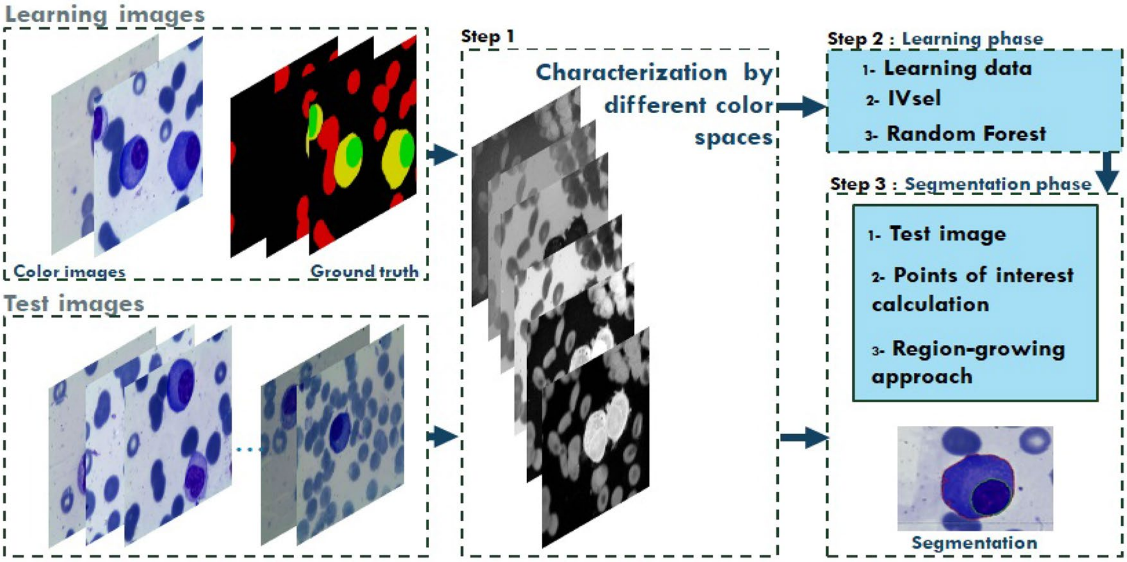


Fig. 5 Proposed approach for automatic segmentation by pixel-based classification

The final reduced dataset is obtained by majority vote between the best chromosomes of each population.

4 Application

To evaluate our proposal, we first performed in the automatic detection of white blood cells (nucleus and cytoplasm) in cytological images. By next we launch experiments on standardized datasets from the UCI [39] and ASU [64] repository to prove the universality of the proposed approach. The main objective is to execute this process in minimum computational time while saving the predictive power of segmentation or Classification.

For this purpose of WBC segmentation, we use an intelligent region-growing approach beginning with a point of interest and sorting the neighboring pixels to construct the region of interest. The process applied for pixel-based classification involves first the intervention of an expert hematologist to identifying nucleus and cytoplasm in cytological images. We already proposed in [5, 48, 49, 54] a principle inspired by Reza et al. [1], where the expert intervenes for windowing the regions of interest of a minimum number of images from the database. The selection is made in four windows (nucleus, cytoplasm, plasma and red blood cell). Thereafter, a features extraction step is used, to represent each pixel in the image by a vector parameter.

Our contribution is the application of an instance (pixel) and variable selection (feature extraction) step to build a reliable pixel-based classification segmentation model that offers the best trade-off between computing time and segmentation performance. The proposed approach follows three steps: characterization, learning and segmentation step.

- *Step 1* The pixels of the images are characterized by a parameter vector of different color spaces (Table 1) obtained during the feature extraction phase.
- *Step 2* In this phase, the IVsel algorithm reduces the original dataset by selecting only the most interesting pixels and variables. Thereafter, we construct a classifier using the reduced dataset and Random Forest algorithm that predicts the label of each pixel.
- *Step 3* The ultimate erosion calculates the points of interest of each test image. Following, we applied the previously learned hypothesis to classify pixels near the point of interest and so we perform the region-growing approach. The segmentation process is shown in Fig. 5.

4.1 Features extraction

The color reference in the pixel-based classification is the RGB color space, because this is the format most used in the acquisition system. The RGB can provide good color discrimination under controlled illumination conditions. Various color spaces are proposed in the literature which can be useful to represent the color of each pixel, as studied in [16, 42]. This diversity opens a question related to the choice of the relevant color space in the pixel-based classification. According to the study of [17], a good choice may bring considerable improvements in certain applications of image processing. In the same application framework of white blood cells segmentation, Benazzouz et al. [6] have studied the importance of the feature selection technique in the choice of the color spaces to obtain a good pixel-based classification. However, since there are a wide variety of color spaces, it can be grouped into four main families [59]:

The primary spaces are based on the trichromatic theory. This family assumes that mixing appropriate amounts of R, G and B can produce any existing color. The normalized primary color can be obtained by dividing each primary color component value by the sum of the other three.

The luminance–chrominance spaces are characterized by the luminance component which represents an achromatic information two chrominance components which represent the chromatic information. The color spaces of this family are calculated from the primary components by linear or nonlinear transformation.

The perceptual spaces can present the subjective perceptual quantification of human color using intensity, hue, and saturation components.

The independent axis spaces provide the least correlated components as possible between color resulting, using different statistical methods like principal component analysis (PCA).

4.2 Points of interest detection

In this work, we target the identification of white blood cells in a cytological image which also contains red blood cells and plasma. In a previous work [48], we proposed a fast pixel classification treatment using ultimate erosion. This treatment proves that it can minimize computation time by starting the classification with pixels of the region of interest [53]. Thereby, we reproduce in the proposed outline, the same mathematical morphology for points of interest detection as a preprocessing step for the pixel-based classification.

5 Experiments and results

This section aims to carry out experiments using a methodology in order to investigate whether the proposed IVsel algorithm contributes to reduce the computational cost while improving the segmentation performances. The

Table 1 Different color spaces features

Family	Color spaces	Computational formula
Primary spaces	*RGB*	$R(i,j)$ $G(i,j)$ $B(i,j)$
Luminance–chrominance spaces	*LUV*	$L = 116(\frac{Y}{Y_n})^{1/3} - 16 \text{ If } \frac{Y}{Y_n} > 0.008856$ $= 903.3(\frac{Y}{Y_n}) \text{ If } \frac{Y}{Y_n} \le 0.008856$ $U = 13L(U' - U'_n)$ $V = 13L(V' - V'_n)$
	Lab	$L = 116(\frac{Y}{Y_n})^{1/3} - 16 \text{ If } \frac{Y}{Y_n} > 0.008856$ $= 903.3(\frac{Y}{Y_n}) \text{ If } \frac{Y}{Y_n} \le 0.008856$ $a = 500 * (f(X/X_n) - f(Y/Y_n))$ $b = 200 * (f(Y/Y_n) - f(Z/Z_n))$
	YUV	$K = 0.299 * R + 0.587 * G + 0.114 * B$ $Y = (0.859 * K) + 16$ $U = (0.496 * (B - K)) + 128$ $V = (0.627 * (R - K)) + 128$
	YIQ	$Y = 0.299R + 0.587G + 0.114B$ $I = 0.596 * R - 0.274 * G + 0.322 * B$ $Q = 0.212 * R - 0.523 * G - 0.311 * B$
	YCbCr	$Y = 0.299R + 0.587G + 0.114B$ $\text{Cb} = -0.169 * R - 0.331 * G + 0.500 * B$ $\text{Cr} = 0.500 * R - 0.419 * G - 0.081 * B$
Perceptual spaces	*HSL*	$H = \frac{G-B}{(\text{Max}-\text{Min})} \text{ If } R = \text{Max}$ $= \frac{B-R}{(\text{Max}-\text{Min})} + 2 \text{ If } G = \text{Max}$ $= \frac{R-G}{(\text{Max}-\text{Min})} + 4 \text{ If } B = \text{Max}$ $S = \frac{\text{Max}(R,G,B) - \text{Min}(R,G,B)}{\text{Max}(R,G,B)}$ $L = \frac{\text{Max}(R,G,B) + \text{Min}(R,G,B)}{2}$
	HSV	$\text{H} = \frac{G-B}{(\text{Max}-\text{Min})} \text{ If } R = \text{Max}$ $= \frac{B-R}{(\text{Max}-\text{Min})} + 2 \text{ If } G = \text{Max}$ $= \frac{R-G}{(\text{Max}-\text{Min})} + 4 \text{ Si } B = \text{Max}$ $S = \frac{\text{Max}(R,G,B) - \text{Min}(R,G,B)}{\text{Max}(R,G,B)}$ $V = \text{Max}(R, G, B)$
Independent axis spaces	*I1I2I3*	$I1 = (R + G + B)/3$ $I2 = (G - B)/2$ $I3 = (2G - R - B)/4$

three algorithms were evaluated under an i7-4820 CPU @ 3.7 GHz, 56Go RAM, MATLAB R2013a environment.

We first begin by describing the cytological images dataset used, then, a summary of the fixed parameters of each algorithm in the experiments. Thereafter, the experiments conducted and the obtained results to determine the best compromise of the parameters α_1, α_2 border and central instances, respectively, and the best numbers of variables, to have the closest representation with a minimum dimension. Finally, we show the comparison between IVsel (with the best parameters) and the three evolutionary algorithms from the literature.

In order to better assess the obtained results for each algorithm, we complete these experiments with other tests

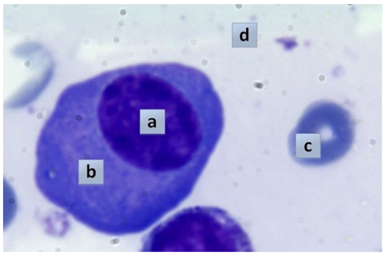


Fig. 6 (a) Nucleus, (b) cytoplasm, (c) red blood cells (d) plasma

Table 2 Pixel selection algorithm parameters

Methods	Parameters
IVsel	Number of trees (Ntrees) = 50
CHC	Population = 50, Generation = 50
PBIL	Population = 50, Generation = 50, Learning rate = 0.1
IFS_CoCo	Population = 50, Generation=50

carried out on standardized datasets from the UCI [39] and ASU [64] repository, to confirm that IVsel proposed the best compromise toward reduction rate/performances and computational cost.

5.1 Database

In our experiments, we use a database acquired in the Haemobiology Service (CHU Tlemcen) with MGG staining (May Grunwald Giemsa) [6] by the LEICA environment (camera and microscope) which provides RGB color images (768 × 1024 pixels).

In this work, we have chosen to partitioning the database into ten images for the learning dataset and 60 images for the test dataset, so that, we could evaluate our model in the presence of two major factors that influence the classification performance, namely class imbalance which introduces a bias toward the majority class; and the sample representativeness issues of the training set that affects model's performance, by not presenting the relevant examples. For the training set, the regions of interest are labeled by an expert in the field, which contains four regions, namely nucleus, cytoplasm, red blood cells and plasma, as represented in Fig. 6.

5.2 Parameters

Table 2 lists related parameters used in the instance and variable selection algorithms. A Random Forest with a number of trees fixed to 100 is built in the learning step. In the experimental tests, a five-cross-validation is applied with

Table 3 Classes distribution for the different subsets

Parameters	Before selection	Selection data 1	Selection data 2	Selection data 3	Selection data 4
α_1 (border), α_2 (center)	(100%, 100%)	(60%, 40%)	(70%, 30%)	(80%, 20%)	(90%, 10%)
Reduction rate (%)	0	59.41	68.83	78.25	87.67
Size data (pixels)	4,363,984	1,770,916	1,359,839	948,763	537,687
$\text{Distribution}_{\text{Nucleus}}(\%)$	11.61	11.38	11.42	11.47	11.54
$\text{Distribution}_{\text{Cytoplasm}}(\%)$	16.15	13.36	13.88	14.90	17.48
$\text{Distribution}_{\text{Red-cells}}(\%)$	49.31	50.92	50.51	49.81	48.02
$\text{Distribution}_{\text{Background}}(\%)$	22.91	24.32	24.16	23.80	22.93

Fig. 7 Importance variable plot by IVsel

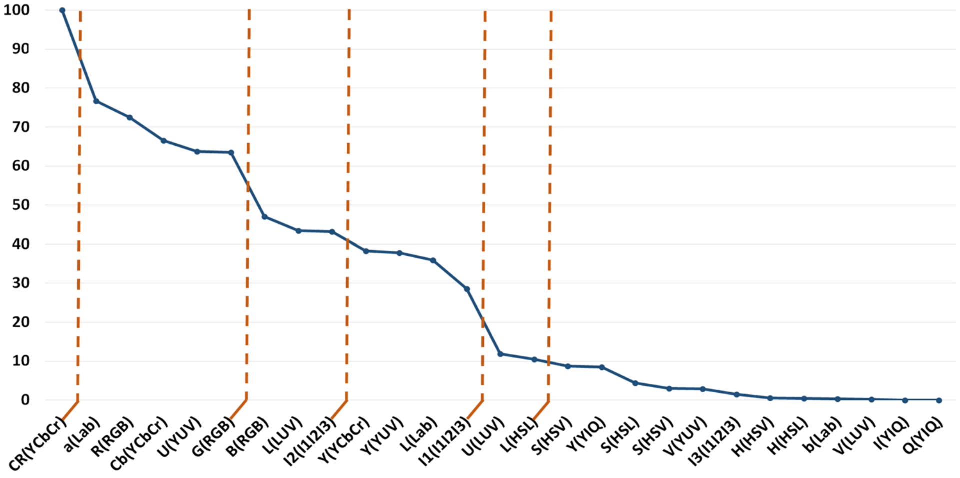


Table 4 Segmentation accuracy of WBC by the different subset variables

Subset of variables	Color spaces	Importance degree (%)	Nucleus Acc (%)	Cytoplasm Acc (%)
1	CR	= 100	96.53	94.23
6	CR, a, R, Cb, U, G	> 60	99.10	94.99
9	CR, a, R, Cb, $U, G, B, L, I2$	> 40	99.08	93.30
13	CR, a, R, Cb, $U, G, B, L, I2, Y, Y, L, I1$	> 20	99.01	94.75
15	CR, a, R, Cb, $U, G, B, L, I2, Y, Y, L, I1, U, L$	> 10	99.05	94.68

100 iterations for the region-growing approach classification, and this number depends on the size of cytoplasm in the image.

The classification performances are evaluated based on the accuracy and F-score measurements:

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$F\text{-score} = 2\,\frac{\text{Accuracy} \cdot \text{Recall}}{\text{Accuracy} + \text{Recall}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}$$

where TP is the number of true positive, FP is the number of false positive and FN is the number of false negative pixels. F-score define a compromise of accuracy and recall giving the performance of the system. This compromise is given in a simple way by the harmonic mean of precision and recall.

5.3 IVsel results

As shown in Sect. 3, the IVsel algorithm constructs an instance ranking based on the margin value, this technique selects the most informative instances based on their margins. Thereby, in order to have a good representation of the data, our algorithm selects some instances from two ranges in the margin value: the lowest margin (the most informative instances) and the medium and highest (instances which carry general information).

In order to choose the best values of (α_1, α_2) for the dataset, we perform a series of experiments varying from $\{(90\%, 10\%), (80\%, 20\%), (70\%, 30\%), (60\%, 40\%)\}$. Table 3 presents the reduction rate and the distribution of classes in the data for the different percentage of margin instances. The results show that the classes distribution is maintaining, when α_2 decreases, while the reduction rate increases. In the light of these results, we choose (90%, 10%) as the best values for the parameters (α_1, α_2).

Moreover, the IVsel method used the importance variable (VI) measure provided by random forests to identify the most important predictor variables. The VI degrees of a variable can be measured when the values of the concerned variable are randomly permuted, and it reflects the average decrease in model accuracy on the OOB samples. Figure 7 illustrates the importance variable plot. The plot shows each variable on the x-axis, and their importance on the y-axis. They are ordered as most to least important. Typically, RF-RI measures the importance of all variables and the ability of each variable to classify the data appropriately. To decide how many important variables to choose, we look for a large break between variables. In Fig. 7, the ranking of variables clearly show five subsets of 1, 6, 9, 13, 15 variables with an VI > 10% that contributes to the good recognition of ROIs in WBD segmentation.

To select the best subset of variables, we study the impact of these five subsets on the accuracy of WBC segmentation process. Table 4 resumes each color spaces selected for each subset, their importance degree and accuracy segmentation of nucleus and cytoplasm. We can clearly see that with a

Table 5 Performances, running times (min) and reduction rates for each algorithm

Regions	Methods	Benchmark	IVsel	CHC	PBIL	IFS_CoCo
Cytoplasm	Accuracy	94.70	94.99	94.79	95.41	94.33
	F-score	0.4297	0.4518	0.4449	0.4640	0.4567
Nucleus	Accuracy	99.08	99.10	99.10	99.10	97.49
	F-score	0.8388	0.84	0.8376	0.8413	0.7292
Running time	Selection	0	7245	313,245	183,657	14,235
	Learning	97,204	7235	28,831	18,039	10,245
Reduction Rate	Instance	0	87.67	55.04	54.98	22.52
	Variable	0	78.57	53.57	59.25	70.22

100% VI degree, Cr which belongs to the YCbCr color space achieved on its own a good segmentation of nucleus and cytoplasm. However, the best performance is reached by the second subset formed by variables with an importance degree greater than 60%. This subset contains only six variables: CR, a, R, Cb, U and G.

The IVsel method, by its different results for instance and variable selection, prove its efficiency to deal with high-dimensional image dataset by reducing 87.67% of redundant instances with respect of the original distribution dataset, and by indicating the most accurate variable set with six color spaces for the WBC segmentation, which represents a reduction of 78.57%.

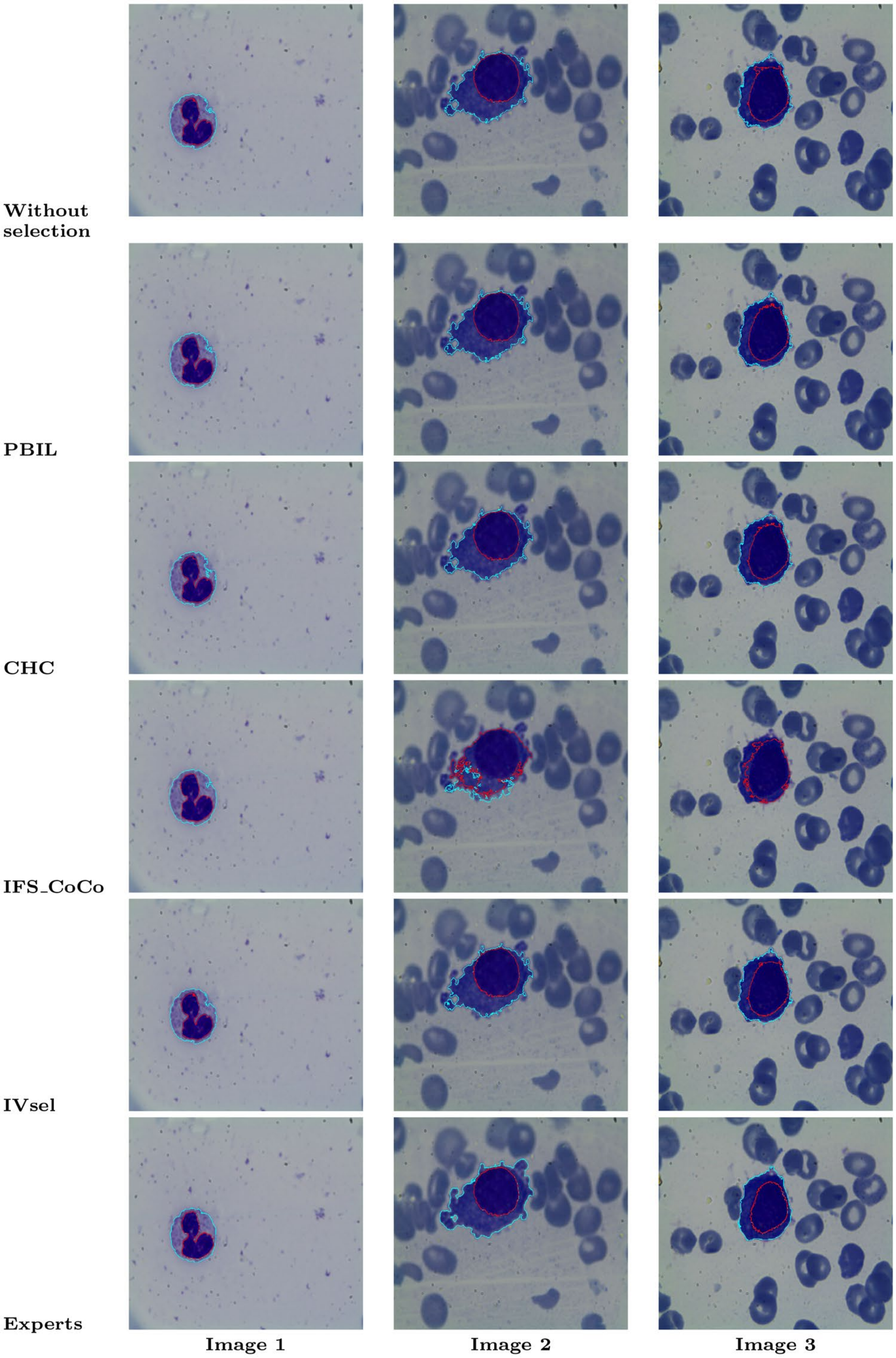
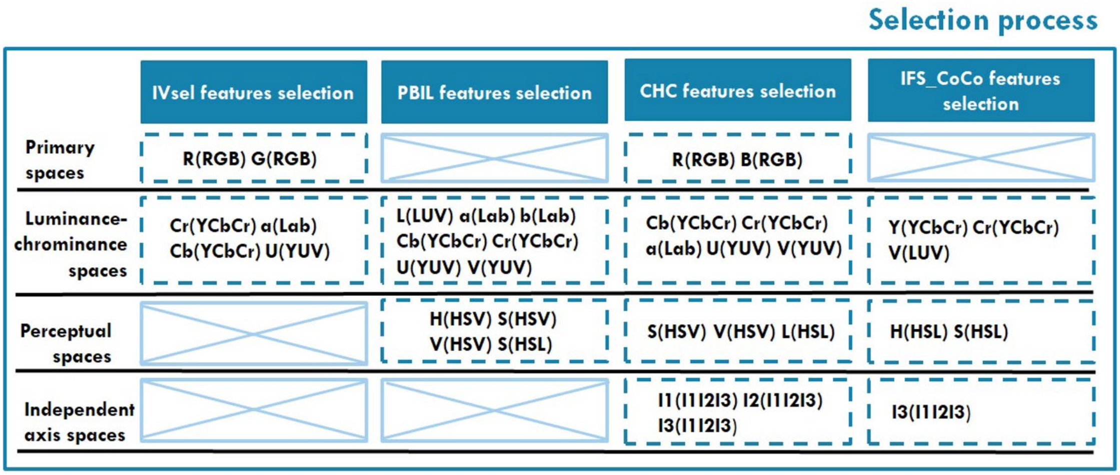


Fig. 8 Results of automatic segmentation by the Random Forest classifiers with and without instance and variable selection



Selection process

	IVsel features selection	PBIL features selection	CHC features selection	IFS_CoCo features selection
Primary spaces	R(RGB) G(RGB)		R(RGB) B(RGB)	
Luminance-chrominance spaces	Cr(YCbCr) a(Lab) Cb(YCbCr) U(YUV)	L(LUV) a(Lab) b(Lab) Cb(YCbCr) Cr(YCbCr) U(YUV) V(YUV)	Cb(YCbCr) Cr(YCbCr) a(Lab) U(YUV) V(YUV)	Y(YCbCr) Cr(YCbCr) V(LUV)
Perceptual spaces		H(HSV) S(HSV) V(HSV) S(HSL)	S(HSV) V(HSV) L(HSL)	H(HSL) S(HSL)
Independent axis spaces			I1(I1I2I3) I2(I1I2I3) I3(I1I2I3)	I3(I1I2I3)

Fig. 9 Selected variables by each algorithm

5.4 Comparison analysis

In this section, we perform a comparison between the proposed approach IVsel and the evolutionary algorithms: PBIL, CHC and IFS_CoCo. Table 5 presents the performances, running times and reduction rates for each algorithm and the benchmark, which are the results achieved without any selection.

When comparing the three approaches, we notice that neither algorithm consistently outperformed the others in classification accuracy and *F*-score. However, all of them improve the results of the original datasets. This can be explained by the fact that the reduction process eliminates noisy instances.

On the other hand, we notice that IVsel also obtains a higher reduction rate. CHC and PBIL reached a reduction rate lower than 60% in either instance or variable dimension. Likewise, IFS_CoCo reach a small reduction rate (22.52%) on instance but an interesting one with 70.22% in feature, while IVsel performs an instance reduction of 87.67% and a variable reduction of 78.57% which result in a higher gain of time in the learning step, performed by RF-RI classifier, with respect to PBIL and CHC (7235 min for IVsel vs 18,039, 28,831 and 10,245 for PBIL, CHC and IFS_CoCo, respectively).

Moreover, Table 5 shows that like expected evolutionary algorithm is the slowest ones with IFS_CoCo faster than PBIL and CHC. By contrast, the execution time of the selection process of IVsel is considerably lower than the three evolutionary algorithms. Indeed, IVsel is more than 10 times faster than IFS_CoCo, 20 times faster than PBIL and more than 40 times faster than CHC.

In addition, Table 5 shows that the computational time of CHC, PBIL and IFS CoCo in learning phase exceeds the execution time of using the entire dataset. Thus, the increase in performance results in an increase in the cost of calculation. We can therefore conclude that when it comes to large databases, our approach is the most appropriate, since IVsel obtains comparable results to the ones obtained by the best algorithms of the state of the art with lower cost.

Table 6 Datasets description

Datasets	#Instances	#Variables	#Classes
BaseHock	1993	4862	2
Brieman	5000	40	3
CNAE-9	1080	856	9
Ionosphere	351	34	2
Madelon	2598	500	2
Musk	476	166	2
PCMAC	1943	3289	2
pendigits	7494	16	9
Pima	768	9	2
Relathe	1427	4322	2
Segmentation	2310	19	7
wdbc	569	30	2

The computational complexity analysis can explain these results, where the computational complexity of CHC and PBIL is $O(M^2)$ with M the number of instances, while the complexity of IFS_CoCo algorithm is $O(M^2)$ for each population. This means that the algorithms run in a quadratic polynomial time. So, since the population size grows, the problem size also grows and therefore needs a much larger computing resource. However, the complexity of the random forest used in the construction of IVsel is $O(LM\log(M))$, where M is the number of instances and L is the number of trees in the forest. On the other hand, a random forest is a bagging of trees which allows a parallel implementation. Regarding these pieces of information, we can say that IVsel

Table 7 Performances, running times (s) and reduction rates for each algorithm on standardized datasets

Dataset	Methods	Accuracy%	Var Reduction Rate%	Ins Reduction Rate%	Running time (s)
BaseHock	PBIL	94.30	51.08	46.61	34,944.14
	CHC	93.89	50.15	43.50	638.66
	IFS_CoCo	88.15	25.05	22.88	179.33
	IVsel	96.85	50.03	57.30	1118.12
Brieman	PBIL	85.75	58.54	43.50	1608.46
	CHC	80.33	60.98	45.06	20.08
	IFS_CoCo	78.49	26.83	22.70	21.63
	IVsel	86.98	53.66	50.84	27.05
CNAE-9	PBIL	81.22	49.24	44.35	3981.64
	CHC	73.75	50.18	44.44	132.95
	IFS_CoCo	59.96	26.49	20.56	31.38
	IVsel	88.68	50.18	54.81	158.67
Ionosphere	PBIL	92.02	51.43	46.72	31.02
	CHC	92.53	65.71	37.89	4.41
	IFS_CoCo	90.89	37.14	21.65	3.02
	IVsel	97.00	54.29	56.98	3.03
Madelon	PBIL	70.71	50.90	45.73	1724.60
	CHC	63.57	52.10	44.69	83.59
	IFS_CoCo	55.99	23.75	22.83	29.03
	IVsel	68.26	50.30	42.57	133.31
Musk	PBIL	86.77	50.30	46.01	135.97
	CHC	79.39	52.10	45.80	21.27
	IFS_CoCo	73.27	26.35	21.64	1.89
	IVsel	86.66	50.90	52.73	13.10
PCMAC	PBIL	91.01	50.73	46.99	26,854.33
	CHC	88.98	51.16	45.75	533.51
	IFS_CoCo	82.03	24.32	24.65	560.02
	IVsel	91.07	50.03	53.58	461.25
pendigits	PBIL	87.44	82.35	47.16	337.93
	CHC	86.85	70.59	45.37	11.63
	IFS_CoCo	78.30	35.29	22.75	3.29
	IVsel	93.36	58.82	53.86	18.40
Pima	PBIL	76.05	44.44	45.18	44.10
	CHC	75.06	66.67	45.44	5.05
	IFS_CoCo	70.16	55.56	23.18	8.42
	IVsel	77.83	66.67	51.69	4.18
Relathe	PBIL	87.71	49.34	46.18	15,900.39
	CHC	86.31	50.20	45.55	495.30
	IFS_CoCo	80.46	25.10	22.56	662.83
	IVsel	88.32	50.03	53.33	527.27
Segmentation	PBIL	96.69	40.00	45.76	74.42
	CHC	95.70	55.00	46.32	4.00
	IFS_CoCo	88.86	30.00	22.60	0.93
	IVsel	96.81	55.00	58.35	4.13
wdbc	PBIL	96.34	48.39	47.98	27.07
	CHC	93.90	58.06	43.41	2.18
	IFS_CoCo	92.31	29.03	22.67	4.30
	IVsel	97.58	54.84	58.00	1.54

performs as well as the other state-of-the-art approaches with lower computational cost.

To discuss the performance and quality of the segmentation approaches, we randomly select three images from the test database (Fig. 8). A qualitative comparison shows a successful and same recognition of the nucleus from the cytoplasm before and after selection which is basically the same of the expert annotation. These results allow us to confirm the superiority of our approach compared to other approaches of growing-region approach by pixel-based classification. The use of the IVsel approach reduces the learning set, which saves computing time while maintaining segmentation performance.

On the other hand, the variables selected by each algorithm are substantially different. Figure 9, resumes the space color selected by each algorithm. To find out which spaces are most relevant for better segmentation, several studies [33, 35, 41, 43] have demonstrated the utility of primary spaces in color image segmentation applications, as has been well-confirmed by IVsel and CHC. Notably, the space G which brings a better discrimination of the regions, unlike PBIL and IFS_CoCo which eliminated the primary spaces. IVsel with six variables demonstrates that even without the perceptual and independent spaces, the results are competitive and almost identical to those of PBIL, CHC and IFS_CoCo which use these characteristics.

5.5 Experiments on UCI datasets

In order to prove the universality of the presented method, we launch experiments on twelve databases from the UCI [39] and ASU [64] repository. The used datasets are characterized by a variant number of variables and instances ranging from small to large. The details of these datasets are shown in Table 6.

For each dataset, a ten-cross-validation is carried out for evaluation. The same related parameters used in instance and variable selection algorithms for the pixel-based classification are conducted for the classification of UCI and ASU datasets. For the values of (α_1, α_2) i.e., low and high margin instances, we fixed them at (60%, 40%).

The obtained results are very interesting, and their differences depend on the databases. For example, in Table 7, the performance of IVsel on Breast, CNAE9, Ionosphere and pendigits are quite remarkable. Furthermore, IVsel gives a slight improvement or equivalent accuracy on the remaining datasets. This can be explained by a lower reduction rate of variables and instances compared to those of the evolutionary algorithm PBIL, CHC and IFS_CoCo. Moreover, we notice in the running time column (Table 7), IVsel records slightly less running time than the three other approaches, especially in comparison with PBIL which present a comparable reduction rates to IVsel but a much longer time.

These experiments can confirm the effectiveness of IVsel and prove that its principal based on the two ensemble concepts: the ensemble margin and the importance variable which provide a good compromise between performance and running time.

6 Conclusion and perspectives

The main problem when dealing with huge datasets, which is currently the case in image analysis, is the high computational cost. We consider the use of reduction techniques in both instances and variables dimensions to overcome this problem. In this work, we propose an instance variable selection approach based on the random forest ensemble method named IVsel. Its principle is ranking the instances and variables in a learning process based on the ensemble margin and the importance variable measure of Random Forest algorithm.

To evaluate our proposed approach, the pixel-based classification of white blood cells WBC (nucleus and cytoplasm) in cytological images was performed. Results of IVsel show that our method reaches the same results obtained by evolutionary algorithm CHC, PBIL and IFS_CoCo with higher reduction rate and lower execution time. These results were also verified on the standardized datasets from UCI and ASU.

The perspectives are innumerable whether in the fundamental side of our automatic segmentation approach or in the reduction process. Now, we are working on the identification of relevant features, in the case of images that need color, spatial and texture characterization such as mammographic images or ultrasound images of placenta. As future work, we consider to adapt IVsel for real-time segmentation of medical videos.

References

1. Azmi R, Norozi N, Anbiaee R, Salehi L, Amirzadi A (2011) Impst: a new interactive self-training approach to segmentation suspicious lesions in breast MRI. J Med Signals Sens 1(2):138–148
2. Baluja S (1994) Population-based incremental learning: a method for integrating genetic search based function optimization and competitive learning. Technical Report, CMU-CS-94-163, Computer Science Department, Carnegie Mellon University
3. Baluja S (1995) An empirical comparison of seven iterative and evolutionary function optimization heuristics. Technical report, School of Computer Science Carnegie Mellon University
4. Baluja S, Caruana R (1995) Removing the genetics from the standard genetic algorithm. Technical report, School of Computer Science Carnegie Mellon University
5. Bechar ME, Settouti N, Barra V, Chikh MA (2017) Semi-supervised superpixel classification for medical images segmentation: application to detection of glaucoma disease. Multidimens Syst Signal Process. https://doi.org/10.1007/s11045-017-0483-y

6. Benazzouz M, Baghli I, Chikh MA (2013) Microscopic image segmentation based on pixel classification and dimensionality reduction. Int J Imaging Syst Technol 23(1):22–28
7. Boukir S, Guo L, Chehata N (2013) Classification of remote sensing data using margin-based ensemble methods. In: 2013 IEEE international conference on image processing, pp 2602–2606. https://doi.org/10.1109/ICIP.2013.6738536
8. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140. https://doi.org/10.1023/A:1018054314350
9. Breiman L (2001) Random forests. Mach Learn 45:5–32
10. Cano J, Herrera F, Lozano M (2003) Using evolutionary algorithms as instance selection for data reduction in KDD: an experimental study. IEEE Trans Evolut Comput 7:561–575
11. Chen ZY, Lin WC, Ke SW, Tsai CF (2015) Evolutionary feature and instance selection for traffic sign recognition. Comput Ind 74:201–211. https://doi.org/10.1016/j.compind.2015.08.007
12. Cicconet M, Hochbaum DR, Richmond D, Sabatini BL (2017) Bots for software-assisted analysis of image-based transcriptomics. bioRxiv 5:4. https://doi.org/10.1101/172296
13. do Carmo RAF, de Freitas FG, de Souza JT (2010) Empowering simultaneous feature and instance selection in classification problems through the adaptation of two selection algorithms. In: Proceedings of the 2010 9th international conference on machine learning and applications
14. Derrac J, Garcia S, Herrera F (2010) IFs-CoCo: instance and feature selection based on cooperative coevolution with nearest neighbor rule. Pattern Recognit 49:2082–2105
15. Derrac J, Triguero I, Garcia S, Herrera F (2012) Integrating instance selection, instance weighting, and feature weighting for nearest neighbor classifiers by coevolutionary algorithms. IEEE Trans Syst Man Cybern 42:1383–1397
16. Drimbarean A, Whelan P (2001) Experiments in colour texture analysis. Pattern Recognit Lett 22(10):1161–1167. https://doi.org/10.1016/S0167-8655(01)00058-7
17. Ebner M (2007) Color constancy. Wiley, London
18. Eshelman L (1991) The CHC adaptive search algorithm: how to have safe search when engaging in nontraditional genetic recombination. Morgan Kaufmann, Los Altos, pp 265–283
19. Gao C, Wang L, Xiao Y, Zhao Q, Meng D (2018) Infrared small-dim target detection based on markov random field guided noise modeling. Pattern Recognit 76:463–475. https://doi.org/10.1016/j.patcog.2017.11.016
20. García-Pedrajas N, Romero del Castillo J, Ortiz-Boyer D (2010) A cooperative coevolutionary algorithm for instance selection for instance-based learning. Mach Learn 78:381–420
21. Garcia-Pedrajas N, de Haro-Garcia A, Pérez-Rodriguez J (2014) A scalable memetic algorithm for simultaneous instance and feature selection. Evolut Comput 22(1):1–45. https://doi.org/10.1162/EVCO_a_00102**(PMID: 23544367)**
22. Guo L, Boukir S (2014) Ensemble margin framework for image classification. In: 2014 IEEE international conference on image processing (ICIP), pp 4231–4235. https://doi.org/10.1109/ICIP.2014.7025859
23. Gupta V, Bhavsar A (2017) Random forest-based feature importance for hep-2 cell image classification. In: Valdés Hernández M, González-Castro V (eds) Medical image understanding and analysis. Springer International Publishing, Cham, pp 922–934
24. Hamidzadeh J, Monsefi R, Yazdi HS (2016) Large symmetric margin instance selection algorithm. Int J Mach Learn Cybern 7:25–45
25. Hoehfeld M, Rudolph G (1997) Towards a theory of population based incremental learning. In: Proceedings of the IEEE conference on evolutionary computation
26. Ishibuchi H, Nakashima T, Nii M (2001) Genetic-algorithm-based instance and feature selection, chap. 6. Springer, Dordrecht, pp 95–112
27. Kim JH, Park YS, Ahn SH, Kim SK (2014) A feature-based small target detection system. In: Park JJJH, Adeli H, Park N, Woungang I (eds) Mobile, ubiquitous, and intelligent computing. Springer, Berlin, pp 541–548
28. Kursa MB (2014) Robustness of random forest-based gene selection methods. BMC Bioinform 15(1):8. https://doi.org/10.1186/1471-2105-15-8
29. Laszlo L, Szidonia L, Simina E, Mircea Florin V (2017) Random forest feature selection approach for image segmentation. https://doi.org/10.1117/12.2268694
30. Lefkovits L, Lefkovits S, Vaida MF, Emerich S, Maluţan R (2017) Comparison of classifiers for brain tumor segmentation. In: Vlad S, Roman NM (eds) International conference on advancements of medicine and health care through technology; 12th–15th Oct 2016, Cluj-Napoca, Romania. Springer International Publishing, Cham, pp 195–200
31. Li H, Tan Y, Li Y, Tian J (2014) Image layering based small infrared target detection method. Electron Lett 50:42–44
32. Li Y, Zhang Y (2018) Robust infrared small target detection using local steering kernel reconstruction. Pattern Recognit 77(C):113–125. https://doi.org/10.1016/j.patcog.2017.12.012
33. Lim YW, Lee SU (1990) On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques. Pattern Recognit 23(9):935–952
34. Liu Y, Zhao H (2017) Variable importance-weighted random forests. Quant Biol 5(4):338–351. https://doi.org/10.1007/s40484-017-0121-6
35. Lizarraga-Morales RA, Sanchez-Yanez RE, Ayala-Ramirez V, Patlan-Rosales AJ (2014) Improving a rough set theory-based segmentation approach using adaptable threshold selection and perceptual color spaces. J Electron Imaging 23(1):013024–013024
36. Martinez W, Gray JB (2014) The role of margins in boosting and ensemble performance. Wiley Interdiscip Rev Comput Stat 6(2):124–131. https://doi.org/10.1002/wics.1292
37. Matale SM, Banait SS (2017) A review on instance and feature selection in big data environment. Int J Adv Res Innov Ideas Educ 3(2):519–523
38. Mellor A, Boukir S, Haywood A, Jones S (2015) Using ensemble margin to explore issues of training data imbalance and mislabeling on large area land cover classification. In: 2014 IEEE international conference on image processing, ICIP 2014, pp 5067–5071. https://doi.org/10.1109/ICIP.2014.7026026
39. Newman D, Hettich S, Blake C, Merz C (1998) UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html. Retrieved 21 May 2019
40. Nguyen TT, Zhao H, Huang JZ, Nguyen TT, Li MJ (2015) A new feature sampling method in random forests for predicting high-dimensional data. In: Cao T, Lim EP, Zhou ZH, Ho TB, Cheung D, Motoda H (eds) Advances in knowledge discovery and data mining. Springer International Publishing, Cham, pp 459–470
41. Ohta YI, Kanade T, Sakai T (1980) Color information for region segmentation. Comput Graph Image Process 13(3):222–241
42. Paschos G (2001) Perceptually uniform color spaces for color texture analysis: an empirical evaluation. IEEE Trans Image Process 10(6):932–937. https://doi.org/10.1109/83.923289
43. Phung SL, Bouzerdoum A, Chai D (2005) Skin segmentation using color pixel classification: analysis and comparison. IEEE Trans Pattern Anal Mach Intell 27(1):148–154
44. Potter MA, De Jong K (2000) Cooperative coevolution: an architecture for evolving coadapted subcomponents. Evolut Comput 8:1–29
45. Pérez-Rodríguez J, Arroyo-Peña AG, García-Pedrajas N (2015) Simultaneous instance and feature selection and weighting using evolutionary computation: proposal and study. Appl Soft Comput 37:416–443. https://doi.org/10.1016/j.asoc.2015.07.046

46. Ramirez-Cruz JF, Fuentes O, V AA, L GB (2006) Instance selection and feature weighting using evolutionary algorithms. In: Proceedings of the 15th international conference on computing (CIC'06)
47. Ros F, Harba R, Pintore M (2012) Fast dual selection using genetic algorithms for large data sets. In: 12th international conference on intelligent systems design and applications (ISDA)
48. Saidi M, Bechar MEA, Settouti N, Chikh MA (2017) Instances selection algorithm by ensemble margin. J Exp Theor Artif Intell. https://doi.org/10.1080/0952813X.2017.1409283
49. Saidi M, El Amine Bechar M, Settouti N, Chikh MA (2016) Application of pixel selection in pixel-based classification for automatic white blood cell segmentation. In: Proceedings of the Mediterranean conference on pattern recognition and artificial intelligence, MedPRAI-2016. ACM, New York, pp 31–38. https://doi.org/10.1145/3038884.3038890
50. Sakinah S, Ahmad S, Pedrycz W (2011) Feature and instance selection via cooperative PSO. IEEE
51. Saraswat M, Arya KV (2014) Feature selection and classification of leukocytes using random forest. Med Biol Eng Comput 52(12):1041–1052. https://doi.org/10.1007/s11517-014-1200-8
52. Schapire R, Freund F (2012) Boosting: foundations and algorithms. The MIT Press, Cambridge
53. Serra J (1986) Introduction to mathematical morphology. Comput Vis Graph Image Process 35(3):283–305. https://doi.org/10.1016/0734-189X(86)90002-2
54. Settouti N, El Habib Daho M, Bechar MEA, Lazouni MA, Chikh MA (2018) Semi-automated method for the glaucoma monitoring. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-63754-9_11
55. Sirikulviriya N, Sinthupinyo S (2011) Integration of rules from a random forest. In: International conference on information and electronics engineering IPCSIT, vol 6. IACSIT Press, Singapore
56. Soltaninejad M, Zhang L, Lambrou T, Allinson NM, Ye X (2017) Multimodal MRI brain tumor segmentation using random forests with features learned from fully convolutional neural network. CoRR arXiv:abs/1704.08134. http://arxiv.org/abs/1704.08134
57. Teixeira de Souza J, Ferreira do Carmo RA, Lima De Campos GA (2008) A novel approach for integrating feature and instance selection. In: Proceedings of the 7th international conference on machine learning and cybernetics. Kunming
58. Tsai CF, Eberle W, Chu CY (2013) Genetic algorithms in feature and instance selection. Knowl-Based Syst 39:240–247
59. Vandenbroucke N, Macaire L, Postaire JG (2003) Color image segmentation by pixel classification in an adapted hybrid color space. Application to soccer image analysis. Comput Vis Image Underst 90(2):190–216. https://doi.org/10.1016/S1077-3142(03)00025-0
60. Villuendas-Rey Y, Caballero-Mota Y, Garcìa-Lorenzo M (2013) Intelligent feature and instance selection to improve nearest neighbor classifiers. Springer, Berlin
61. Wang H, Yang F, Zhang C, Ren M (2018) Infrared small target detection based on patch image model with local and global analysis. Int J Image Graph 18(01):1850002. https://doi.org/10.1142/S021946781850002X
62. Wang L, Gao Y, Shi F, Li G, Chen K, Tang Z, Xia J, Shen D (2016) Automated segmentation of CBCT image with prior-guided sequential random forest. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 9601 LNCS. Springer, Germany, pp 72–82. https://doi.org/10.1007/978-3-319-42016-5_7
63. Yang J, Yao D, Zhan X, Zhan X (2014) Predicting disease risks using feature selection based on random forest and support vector machine. In: Basu M, Pan Y, Wang J (eds) Bioinformatics research and applications. Springer International Publishing, Cham, pp 1–11
64. Zafarani R, Liu H (1998) Asu repository of social computing databases. http://socialcomputing.asu.edu/pages/datasets. Retrieved 21 May 2019
65. Zhang L, Chen C, Bu J, He X (2012) A unified feature and instance selection framework using optimum experimental design. IEEE Trans Image Process 21(5):2379–2388

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.