Pattern Anal Applic (2009) 12:215–226
DOI 10.1007/s10044-008-0116-z

THEORETICAL ADVANCES

A multiple expert system for classifying fluorescent intensity in antinuclear autoantibodies analysis

Paolo Soda · Giulio Iannello · Mario Vento

Received: 19 December 2006 / Accepted: 25 February 2008 / Published online: 22 April 2008

Abstract At the present, Indirect Immunofluorescence (IIF) is the recommended method for the detection of antinuclear autoantibodies (ANA). IIF diagnosis requires both the estimation of the fluorescent intensity and the description of the staining pattern, but resources and adequately trained personnel are not always available for these tasks. In this respect, an evident medical demand is the development of computer-aided diagnosis (CAD) tools that can offer a support to physician decision. In this paper we first propose a strategy to reliably label the image data set by using the diagnoses performed by different physicians, and then we present a system to classify the fluorescent intensity. Such a system adopts a multiple expert system architecture (MES), based on the classifier selection paradigm. Two different selection rules are presented and, given the application domain, the convenience of using one of them is analyzed. Different sets of operating points are determined, making the recognition system suited to application in daily practice and in a wide spectrum of scenarios. The measured performance on an annotated database of IIF images shows a low overall miss rate (<1.5%, 0.00% of false negative).

P. Soda (✉) · G. Iannello
Facoltà di Ingegneria,
Università Campus Bio-Medico di Roma, Rome, Italy
e-mail: p.soda@unicampus.it

G. Iannello
e-mail: g.iannello@unicampus.it

M. Vento
Dipartimento di Ingegneria dell'Informazione ed Ingegneria Elettrica, Università di Salerno, Salerno, Italy
e-mail: mvento@unisa.it

Keywords Indirect immunofluorescence · Computer-aided diagnosis · Pattern recognition · Multiple expert systems · Reject option

1 Originality and contribution

Indirect immunofluorescence (IIF), a technique that detects the presence of antinuclear antibodies (ANA) in patient serum, is nowadays the recommended method for the diagnosis of autoimmune diseases. Currently, the highest level of automation in IIF tests is the preparation of slides with robotic devices performing dilution, dispensation and washing operations. In this paper, we present a system that classifies the fluorescence intensity of IIF images, that is, it automatically discriminates positive, negative and intermediate tests. The system may be the basis for developing a computer-aided diagnosis (CAD) system, which may support the physician's decision and overcome some limitations of the current methods (e.g., the photo-bleaching effect, the interobserver variability, low level of standardization). The recognition system is based on a multiple expert system (MES) paradigm and employs a classifier selection approach. Two different selection rules are presented and the convenience of using either of them in different application domains is analyzed. Different sets of operating points are determined, making the recognition system suited for application in daily practice and in a wide spectrum of scenarios. The measured performance on an annotated database of IIF images shows a low overall miss rate (<1.5%, 0.00% of false negative). The specialists can test the tool through a web-based interface at http://slideimaging.unicampus.it.

2 Introduction

Connective tissue diseases (CTD) are autoimmune disorders characterized by a chronic inflammatory process involving connective tissues. Detection of ANA is a common marker in patients with suspected CTD. ANA directed against a variety of nuclear antigens have been detected in the serum of patients with many rheumatic and non-rheumatic diseases [1]. The recommended method for ANA testing is Immunofluorescence microscopy, particularly the IIF [2, 3]. In IIF, a serum sample is tested with a substrate containing a specific antigen, and the antigen antibody reaction is revealed by fluorochrome conjugated anti-human immunoglobulin antibodies through examination with fluorescence microscope.

In autoimmune diseases, the availability of accurately performed and correctly reported laboratory determinations is crucial for the clinicians. The relevance of the issue is emphasized by the increase in the incidence of autoimmune diseases observed over the last years, partly attributable to both improved diagnostic capabilities and growing awareness of this clinical problem in general medicine. A growing number of health-care structures need laboratories to perform these tests, but the major disadvantages of the IIF method are:

- the lack of resources and adequately trained personnel [3];
- the low level of standardization [4];
- the photobleaching effect, which bleaches significantly in a few seconds biological tissues stained with fluorescent dyes [5];
- interobserver variability, which limits the reproducibility of IIF readings [6];
- the lack of automatized procedures.

To date, the highest level of automation in IIF tests is the preparation of slides with robotic devices performing dilution, dispensation and washing operations [7, 8]. Being able to automatically determine the presence of autoantibodies in IIF would enable easier, faster and more reliable tests. Hence, an evident medical demand is the development of a CAD system, which may support the physician's decision and overcome the limitations of the current methods. In response to this medical demand, some recent works on both the automated HEp-2 pattern description [9–11] and the fluorescent intensity classification [12, 13] may be found in the literature.

In this paper, we propose a strategy to reliably label the image data set by using the diagnoses performed by different physicians, and present a system to classify the fluorescent intensity. With respect to [12, 13], we adopt different features, different system architecture and different classifiers, improving the management of samples that are intrinsically hard to classify (e.g., samples that are borderline between different classes) and developing a more flexible recognition system that should fit to different working scenarios. The system proposed here is based on an MES paradigm and uses a classifier selection approach; two different selection rules are introduced and experimentally evaluated. Starting from the widely accepted result that an MES approach generally produces a better performance than those obtained by individual experts, the rationale is inspired by the results coming out from the feature selection phase: the relatively small set of stable and effective features obtained for each class enforced the evidence that the classification could be reliably faced by introducing one specialized expert for each class that the system should recognize. By performing a convenience analysis on the selection rules, we determine three different sets of operating points that allow applying the classifier to the main working scenarios of a CAD system. This feature makes such innovative CAD suited for application in daily practice.

The paper is organized as follows. After presenting the state of art and motivations in Sect. 2, in Sect. 3 we describe the peculiarities of the application domain. Section 4 presents features extraction and selection, Sect. 5 describes the MES system architecture, whereas Sect. 6 introduces two rules to select the expert, which is more likely to be correct for each input sample. Section 7 reports the experimental results and in Sect. 8 we conclude the paper.

3 Background

3.1 Application context

Humans are limited in their ability to detect and diagnose disease during image interpretation due to their non-systematic search patterns and to the presence of noise. In addition, the vast amount of image data that is generated by some imaging devices makes the detection of potential disease a burdensome task and may cause oversight errors. Another problem is that similar characteristics of some abnormal and normal structure may cause interpretational errors. Developments in computer vision and artificial intelligence in medical image interpretation have shown that CAD system can pursue four major objectives: (1) performing a pre-selection of the cases to be examined, enabling the physician to focus his/her attention only on relevant cases, making it easier to carry out mass screening campaigns, (2) serving as a second reader, thus augmenting the physician capabilities and reducing errors, (3) aiding the physician while he/she carries out the diagnosis, (4) working as a tool for training and education of specialized medical personnel [14–18].

Therefore, each of the previous working scenarios has its requirements and the CAD is expected to apply to them. In general, a classification system makes or does not make a decision on all input samples. In the former case, the CAD acts as a zero-reject system, whereas in the latter one or some samples are rejected. Note that the introduction of a reject option aims to reject the highest possible percentage of samples that would otherwise be misclassified. However, it introduces a side effect whereby some samples that otherwise would have been correctly classified are rejected. The relationship between the error rate and the reject rate is represented as an error–reject trade-off curve, which can be used to set the desired operating point of the classification system. This curve is monotonically non-increasing, since rejecting more patterns either reduces the error rate or keeps it the same.

Based on these considerations, for each working scenario the CAD behavior can be further characterized.

In case (1), the CAD carries out mass screening campaigns. In this respect, two opposite situations may occur. In the first one (referred to as α_1) the CAD acts as a full-automated system that labels all input samples, that is, it acts as a zero-reject system. In the second one (referred to as α_2), the physician must perform the pre-selection on cases rejected by the CAD. Therefore, the recognition system approaches a zero-error classifier whatever the reject rate (although in a real application, it is almost impossible to not have misclassification). Indeed, with reference to a theoretical error–reject curve, more the error approaches zero, greater is the reject rate (at limit 100%).

Case α_1 on the one hand allows carrying out many tests, since the CAD classifies all input samples thus increasing the throughput. On the other hand, the physician a priori knows that some samples will be misclassified, since the error rate (false positive and false negative) of a given recognition system should be evaluated. In this respect, it is important to keep the false negative rate as low as possible.

In case α_2, fewer even though more accurate tests are performed, since the CAD approaches a zero-error system and rejects doubtful samples.

This paper does not address the issue of proposing when and which one of the two situations should be preferred, but we would like to remark the following observation. Regarding only the error rate, case α_2 should appear preferable to case α_1, since more of the true ill patients are identified and treated. However, more tests in the pre-selection phase can be executed in case α_1 than in α_2. Therefore, some potential sick patients could not be processed in case α_2, unless additional work was performed by physicians to screen rejected samples. Hence, a dichotomy arises between performing more tests with a known error and performing less tests with less misclassifications, but excluding some people.

In case (2), referred to as β, the CAD serves as a second reader, supplying an opinion to the physician. Now, the CAD acts as a zero-reject system, providing also a reliability measure of its decision.

In case (3), referred to as γ, the recognition system aids the physician during the diagnosis, performing as a zero-reject system.

Finally in case (4), referred to as δ, the CAD acts differently on the basis of both training purpose and people skills.

It is worth noting that between these two extreme performances (i.e., the zero-error and the zero-reject) several intermediate operating points can be set on the basis of the error–reject curve.

3.2 Related work

IIF is the recommended method for ANA determination, but up to now, the physicians rarely made use of quantitative information. The development of a CAD in this field would improve medical practice, achieving the advantages mentioned above.

Since IIF diagnosis requires the classification of both fluorescent intensity and staining pattern, some previous works on these topics are reported in [9–13]. In [9] and [10], the authors present some results on mining staining pattern of fluorescent cells. The used image data set consists of 321 fluorescent samples with clear patterns, diluted at 1:160. Hence, the samples can be ascribed either to the 4+ class (see CDC criteria, in Sect. 3) or to the negative case. Each image is segmented to locate the cells, and then 132 texture-based features are computed. These features are given to a decision tree induction program to find out the most relevant subset and to construct the classification knowledge. These systems exhibit an error rate of 25.6% [9] and 16.9% [10], acting only as a zero-reject system without providing a reliability measure of final classification. With reference to the previous observations on working scenarios, it is therefore evident that such systems can operate only in case α_1, β, γ and δ, respectively.

With regard to fluorescent intensity classification, a system based on a multi-layer perceptrons (MLP) and a radial basis network has been proposed in [12, 13]. That system, which makes use of features inspired by medical practice shows low error rates (false positive plus false negative) up to 1%, but it uses a reject option and it does not produce a result in about 50% of cases. It uses two features related to the mean of the fluorescent intensity among the cells of the image. The small set of features is chosen with reference to the number of samples in the data set in order to avoid the course of dimensionality. The critical point of this approach is the cell segmentation algorithm, since it does not get to deal effectively with the

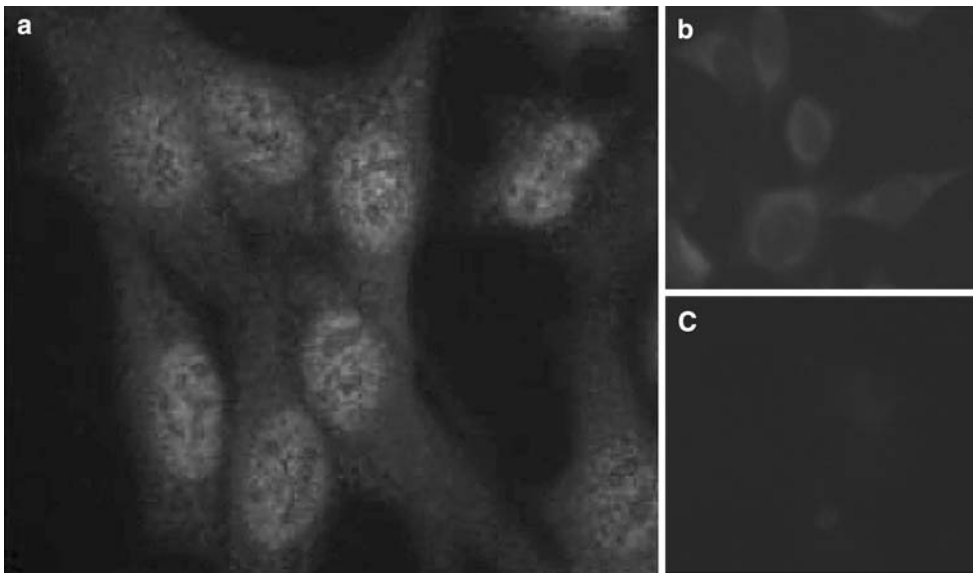

Fig. 1 Examples of IIF images and diagnosis complexity. On the *left* is reported a sample (**a**), whose fluorescent intensity is given by dots inside the cells, whereas on the *right*, two different negative controls are shown (**b**, **c**). The same sample is labeled as 2+ with respect to (**b**), whereas it is labeled as 4+ with respect to (**c**)

great difference in appearance between the fluorescent images and the very low contrast of the negative samples (Fig. 1). Moreover, the adopted classification rule does not allow a flexible management of samples that are intrinsically hard to classify. This justifies the high reject rate required to obtain low error rates, making the CAD suited for application in case α_2.

All previous observations show that a CAD should be able to work in different situations to be effectively used by the medical community. The analysis of the literature in the field of ANA detection reveals that a CAD with this feature has not been developed yet. Therefore, we present an innovative recognition system devoted to classify the fluorescent intensity, which can work in several situations and that improves the present performance.

4 Domain application

IIF diagnosis consists of fluorescent intensity and staining pattern classification. With regard to the classification of fluorescent intensity, the guidelines suggest scoring it semi-quantitatively and independently by two physician experts of IIF. The scoring ranges from 0 up to 4+ relative to the intensity of a negative and a positive (4+) control, following the guidelines established by the Centers for Disease Control and Prevention, Atlanta, Georgia (CDC) [19]:

- 4+ brilliant green (maximal fluorescence);
- 3+ less brilliant green fluorescence;
- 2+ defined pattern, but diminished fluorescence;
- 1+ very subdued fluorescence;
- 0 negative.

Since technical problems can affect test sensitivity and specificity, the same guidelines suggest using both positive and negative controls. The former allows the physician to check the correctness of the preparation process; the latter represents the auto-fluorescence level of the slide under examination. Therefore, the physician has to compare the sample with the corresponding positive and negative control. This comparison is a very problematic task, and it affects the reliability of sample diagnosis. For instance, Fig. 1 shows a sample and two different negative controls, referred to as *a*, *b* and *c*, respectively. Note that the same sample, whose fluorescent intensity is given by the fluorescent dots inside the cells, can be labeled as 2+ with respect to the more fluorescent negative control (*b*), whereas it is labeled as 4+ with respect to the less fluorescent negative control (*c*).

Since IIF is a subjective, semi-quantitative method, in [10] an objective independent criteria (e.g., ELISA, which permits verification of autoantibodies entities) is used to assess the human expert diagnosis on staining patterns. However, a correlation upon positivity and negativity cannot be established between IIF and ELISA tests (e.g., a sample that is negative at IIF should be positive at ELISA, and vice versa). Furthermore, even if a correlation between IIF patterns and autoantibodies entities has been established [2], the same autoantibodies may be found in different patterns making the correspondence not univocal. Hence, in the general case, ELISA cannot be taken as a golden standard for IIF classification.

For all these reasons, we made use of the physician's classification, although image annotation by human experts suffers from variability; indeed, physicians report trouble in interpreting the images, since they are relative in nature. Another significant motivation is that both borderline and

negative samples exhibit low contrast. Furthermore, physicians act differently when the same sample is presented to them: some are more conservative and others more liberal, depending mostly on their skills and background. Hence, we deem that an image annotation procedure performed only by one physician is useless. In this respect, two different physicians independently and blindly diagnose each sample to improve the reliability of the data set.

The variability between physician's classifications has been statistically evaluated, pointing out that it is quite difficult to get substantial or perfect agreement, when each sample can be labeled into one of the previous five subgroups. Data analysis suggests the following class revision: the samples are classified into three classes, named *negative*, *intermediate* and *positive* (for detailed description and motivation see Sect. 7.1). Briefly, a sample is assigned to the negative class if both physicians classify it as negative, whereas it is labeled positive if both physicians mark it with two pluses or more. Finally, a sample is assigned to the intermediate class when some disagreement happens or when both physicians mark it as 1+. In the medical practice, this case usually corresponds to weak positive patients, who should re-execute the tests within 6 months to check disease development.

It is worth noting that, on the one hand, in the physicians' opinion these three classes maintain the clinical significance of the IIF test and, on the other hand, this class revision gets a more robust ground truth. Therefore, such a revised classification protocol has been adopted by our hospital and by the proposed CAD system. As a consequence, it is used in the following to manage the data given as input to the classifiers.

5 Feature extraction and selection

The choice of a suitable set of features is crucial for the performance of the classification system. Initially, we compute a set of statistical features, based on first and second-order gray-level histograms. The rationale lies in the meaning of these histograms: the former describes gray-level distributions, whereas the latter generally provides a good representation of the overall nature of the texture. For the definition of these features see [20]. Preliminary tests were performed on this set. Specifically, such a feature set was extracted from both the segmented images, that is, the features were computed on the cells' area only, and the whole image. In this initial phase, the results of the discriminant analysis suggest that features computed on the whole image have better separation capability than features extracted from the segmented cells. In our opinion, the whole image contains as much information as the segmented cells since:

- the background may be considered uniformly dark and its contribution to the statistical features is negligible;
- the cells of the same image have similar texture; hence all of them contribute the same to the extracted features;
- artifacts possibly due to the limitations of the segmentation algorithm (see Sect. 2) are avoided.

For all these motivations, the system described in the following is based on features extracted from the whole image.

To further increase the separation capability of these features, we combine the features computed on each sample with the same features extracted from the corresponding positive and negative controls. Indeed, the classification guidelines require comparing each sample with the corresponding positive and negative controls, as explained in Sect. 3. The combinations use both linear and non-linear strategies. Specifically, in Table 1, the first row corresponds to the value of ith feature of sample x (denoted by F_x^i and referred to as *absolute feature*), whereas the other entries correspond to four different combinations with the positive and negative control. Applying this strategy, we compute 95 features, 19 for each mode reported in Table 1.

Discriminant analysis shows that all the above features have limited discriminant strength over three classes (i.e., positive, intermediate and negative), but different feature subsets discriminate very well each class from the other two. For the sake of completeness, notice that the search of the best discriminant subsets has been carried out, first by a sequential forward selection and then it has been refined by an exhaustive search, taking into account the dimensionality of the data set and of the feature space.

These observations suggest adopting a multi-expert approach [21–24] based on three classifiers, each one specialized in recognizing one of the three input classes. Specifically, the three experts are:

- Positive Expert (PE): classifier specialized on the classification of positive sample;

Table 1 Combination mode of the feature selection procedure

Description	Formula	Mode #
Absolute feature	F_x^i	1
Combination with the positive control	$F_x^i - F^i{}_{\text{pos ctrl of } x}$	2
	$\frac{F_x^i}{F^i_{\text{pos ctrl of x}}}$	3
Combination with the negative control	$F_x^i - F^i{}_{\text{neg ctrl of } x}$	4
	$\frac{F_x^i}{F^i_{\text{neg ctrl of x}}}$	5

F_x^i represents the value of ith feature of sample x, $F^i{}_{\text{pos ctrl of } x}$ and $F^i{}_{\text{neg ctrl of } x}$ represents the value of ith feature of positive and negative control of x, respectively

Table 2 Selected features for each expert

Positive expert (PE)	Intermediate expert (IE)	Negative expert (NE)
Kurtosis of H_1 using mode # 2	Skewness of H_1 using mode # 1	Kurtosis of H_1 using mode # 3
Autocorrelation of H_2 using mode # 2	Kurtosis of H_1 using mode # 1	Energy of H_1 using mode # 3
		Entropy of H_1 using mode # 3
Covariance of H_2 using mode # 2	Inverse of H_2 using mode # 1	Covariance of H_2 using mode # 3
		Energy of H_1 using mode # 5

H_1 and H_2 represent the first and second-order gray-level histogram, respectively. *Skewness* and *kurtosis* are the third and fourth moment of the histogram, respectively. *Inverse* stands for the inverse difference moment, that is, a measure of local homogeneity. For the description of features mode combination (#) see Table 1

– Negative Expert (NE): classifier specialized on the classification of negative sample;
– Intermediate Expert (IE): classifier specialized on the classification of sample belonging to the intermediate class.

Each classifier uses its own representation of the input pattern integrating physically different types of measurements. The different representations used, corresponding to different feature sets, are reported in Table 2.

6 System architecture

In the literature, it has been observed that different features and pattern recognition systems complement one other in classification performance [21–24]. The idea is that the recognition performance attainable combining set of experts should be improved by taking advantage of the strengths of the single experts, without being affected by their weakness.

Based on these considerations and motivated by the previous reasons (see Sect. 4), we adopt the system architecture shown in Fig. 2. The resulting MES aggregates three different experts, each one specialized in recognizing one of three input classes (i.e., positive, negative, intermediate). Each expert is a nearest neighbor (NN) classifier.

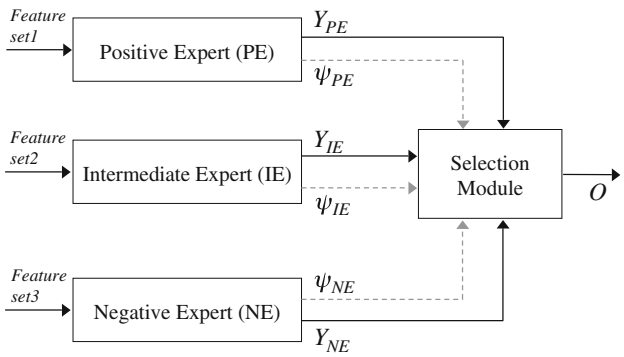

Fig. 2 The architecture of the proposed system. To obtain the decision O of the MES, the decisions Y_{PE}, Y_{IE}, Y_{NE} of the component experts are selected according to a rule, which can take or not take into account the reliability parameters ψ_{PE}, ψ_{IE}, ψ_{NE}, evaluated on the basis of the expert output vectors. In such a way, different reliability values can be associated with each classification act of an expert

To further validate such a design choice, we also explore some other solutions. First, we test a single classifier architecture and, second, we try out an MLP with a hidden layer of ten neurons as classifier of each specialized expert, that is, PE, NE and IE. The corresponding results, reported in Sect. 7.2, confirm that the MES constituted by NN classifiers outperforms the other solutions. Moreover, it is worth observing that the classification system has to be integrated with a reject option to operate in the different working scenarios presented in Sect. 2.1. In this respect, the NN classifiers can be effectively employed since the paradigm used for the reject option has been presently validated on them [25].

In the general case, the rule adopted in the selection module depends on the assumptions about classifier dependencies, the type of classifier outputs, the aggregation strategy (global or local) and the aggregation procedure (a function, a neural network, an algorithm), etc. [23]. In the literature, there are two types of combinations: classifier fusion and classifier selection. In the classifier fusion algorithms, individual classifiers are applied in parallel over the whole feature space. In the classifier selection scheme, each classifier is an expert in some local area of the feature space. In the latter case, the classifiers should be considered complementary rather than competitive [22, 23], and the algorithm attempts to predict which expert is most likely to be correct for a given sample. One classifier as in [26], or more than one as in [24, 27], can be nominated to make the decision.

Since classifiers of the proposed MES are specialized, we decide to adopt a classifier selection approach in the *selection module*. Moreover, since in the literature it has been observed that the evaluation of the classification reliability should be useful for solving complex pattern recognition tasks [21, 22, 28, 29], we decide to use this parameter from individual classifier to select the final output (dotted arrows in Fig. 2).

In the next two sections we propose two selection rules and present their results.

7 The classifier selection scheme

As pointed out previously, to combine the outputs of the classifiers we adopt a selection approach among complementary experts. The issue requires evaluating which single classifier is most likely to be correct for any given sample. In our case we have three specialized classifiers, each one with a binary output, and three desired output classes: in Table 3 all possible combinations of the classifiers' outputs are reported. In the following subsections, two classifier selection rules are proposed.

7.1 Binary combination

A first selection scheme for our MES architecture is a binary combination of the experts' outputs. Let us denote $O(x)$ for the MES output and $Y_k(x)$ for the output of the kth classifier on sample x. Furthermore, let C_k be the class on which the kth classifier is specialized. Note that in our system, k should be PE, NE or IE. According to Table 3, the possible combinations can be grouped into three categories: (1) those for which only one expert k classifies the sample in its class C_k (labeled as a), (2) those for which more experts classify the sample in its own class (labeled as b or c), (3) those for which none expert classifies the sample in its class (labeled as d).

According to these considerations, the following selection rule is adopted. In case (1) as a final output is chosen the class of the expert whose output is 1, since all the classifiers agree in their decision. In case (2) the sample is rejected since two or more experts indicate that the sample belongs to their own class. In case (3) the sample is rejected since none of the experts indicate that the sample belongs to its class. Note that this choice is conservative since no decision is taken in cases (2) and (3), when there is ambiguity in the outputs of the three experts. It is worth noting also that this approach does not require any reliability estimation.

Table 3 Combination of experts output

Expert output			Case label
Positive expert	Negative expert	Intermediate expert	
0	0	0	*d*
0	0	1	*a*
0	1	0	*a*
0	1	1	*b*
1	0	0	*a*
1	0	1	*b*
1	1	0	*b*
1	1	1	*c*

Meaning of symbols: 1(0) the sample is assigned (is not classified) to the class; *a* the sample is assigned to just one class; *b* the sample is assigned to two classes; *c* the sample is assigned to three classes; *d* the sample is assigned to none class

7.2 Zero-reject selection rule

Alternatively, a zero-reject strategy that chooses an output in any of the eight cases reported in Table 3 may be introduced. Once more, the eight cases can be reduced to four main alternatives, referred to in the following as a, b, c and d, respectively. In cases labeled as a, only one expert votes 1, in cases labeled as b, two experts vote 1 and in cases labeled as c or d, all the classifiers vote the same.

Let us then denote $\psi_k(x)$ for the reliability parameter of the kth classifier when it classifies the sample x. Since all the classifiers agree in their decision in case a, we choose as before the class of the classifier whose output is 1 as a final output. Conversely, in cases b, c and d, the final decision is performed looking at the accuracy of single expert's classifications.

More specifically, in case b, two experts vote for their own class, whereas the third one indicates that x does not belong to its own class. To solve the dichotomy between the two conflicting experts we look at the reliability of their classification and choose the more reliable one. Formally:

$$O(x) = C_k, \quad \text{where } k = \arg \max_{i:Y_i(x)=1} (\psi_i(x)) \tag{1}$$

In case c, all experts classify x as belonging to the class they specialize in. Since the three classifications are now in competition, the bigger the $\psi_k(x)$, the less is the misclassification risk by the kth expert. This evidence suggests using again the selection rule (1).

In the case d, all experts classify x as belonging to another class than the one they specialize in. In this case, the bigger the reliability parameter $\psi_k(x)$, the less is the probability that x belongs to C_k, and the bigger the probability that it belongs to the other classes. These observations suggest selecting the following selection rule:

$$O(x) = C_k, \quad \text{where } k = \arg \min_{i:Y_i(x)=0} (\psi_i(x)) \tag{2}$$

In other words, we first find out which classifier has the minimum reliability and then we choose the class associated with this classifier as a final output.

7.3 Reliability parameter

The approach described above for deriving a zero-reject classifier from our MES requires the introduction of a reliability parameter that evaluates the accuracy of the classification performed by each expert.

A reliability parameter should permit distinguishing between the two reasons causing unreliable classifications [28]: (a) either the sample is significantly different from those presented in the training set, that is, in the feature space the sample point is far from those associated with any class, (b) the sample point lies in the region where two or more classes overlap. To distinguish between these situations we introduce two reliability parameters, named ψ_a and ψ_b, which correspond to the two previous cases, respectively. Note that these values vary in the interval [0,1], where the more the parameter approaches one, the more reliable is the classification. Based on these definitions, the parameter providing an inclusive measure of the classification reliability can be defined as follows:

$$\psi = \min(\psi_a, \psi_b) \tag{3}$$

This form is conservative since it considers a classification unreliable as soon as one of the two alternatives causing unreliable classifications happens.

The definition of both the parameters ψ_a and ψ_b relies on the particular classifier architecture adopted. Following [28], the samples belonging to the training set are divided into two sets: the reference set and the test set. The former is used to perform the classification of the unknown pattern x, that is, it plays the role of training set for the NN classifier, whereas the latter provides further information needed to evaluate the ψ_a parameter. More specifically, the two reliability estimators are defined as:

$$\begin{aligned} \psi_a &= \max\left(1 - \frac{O_{\min}}{O_{\max}}, 0\right) \\ \psi_b &= 1 - \frac{O_{\min}}{O_{\min 2}} \end{aligned} \tag{4}$$

where: $O_{\min}$ is the distance between x and the nearest sample of the reference set, that is, the sample determining the class $Y(x)$, $O_{\max}$ is the highest among the values of $O_{\min}$ obtained from all samples of class $Y(x)$ belonging to the test set, and $O_{\min 2}$ is the distance between x and the nearest sample in the reference set belonging to a class other than $Y(x)$. For further information, see [28].

8 Experimental results

Since, to our knowledge, there are not reference databases of IIF images publicly available, several slides of HEp-2 cells were read with a fluorescence microscope in order to populate an image database.

The slides we use are diluted at 1:80. One of the two physicians, randomly chosen, takes digital images of slides with an acquisition unit consisting of a fluorescence microscope, coupled with a 50 W mercury vapor lamp and with a digital camera. The last one has a monochrome CCD, with squared pixels of equal side to 6.45 μm. The microscope objective has a 40-fold magnification and the medium is air. The exposure time of slides to incident light is 0.4 s. The images have a resolution of 1,024 × 1,344 pixels, a color-depth of 8 bits and they are stored in TIFF format.

Up to now, the image data set consists of 600 images, stored in the database together with the ground truth.

8.1 Ground truth

In IIF, the ground truth is made by labeled images both with fluorescent intensity and staining pattern classification. As motivated in Sect. 3, we made use of two physician's classifications to get it.

Clearly, such an approach relies upon the agreement between multiple readers. In other words, its reliability depends on the degree of agreement between physicians. In the literature, many non-equivalent measures of agreement have been proposed. We choose the most widely used one: the Cohen's *kappa* [30]. Its estimate, *kappa* (k), can be expressed as a function of observed frequencies. Although the true parameter value may vary from -1 to 1, the usual region of interest is $k > 0$. In the literature, the following guidelines for interpreting kappa values are used [31]:

- $0 < k < 0.2$ implies slight agreement;
- $0.2 < k < 0.4$ implies fair agreement;
- $0.4 < k < 0.6$ implies moderate agreement;
- $0.6 < k < 0.8$ implies substantial agreement;
- $0.8 < k < 1$ implies almost perfect agreement.

When the physicians diagnose the samples following the CDC guidelines, the measured kappa is 0.46 ± 0.13 ($p < 0.05$), corresponding to a moderate agreement. With reference to the *kappa* values, we believe that $k > 0.6$ corresponds to a reasonable agreement degree, and it should be considered satisfactory to get a reliable ground truth. Therefore, we deem that labeling the samples in five subgroups (i.e., four positive and one negative subgroups) is not completely reliable. Indeed, the disagreement between physicians is twofold motivated. In one case, physicians assign the sample to different classes (i.e., one to positive and the other to negative). In the other case, physicians disagree about the subgroups to which a positive sample has to be assigned, that is, physicians label it with a different number of plus. At a deeper examination, it appears that physicians always agree with each other when the sample is marked either with 2+ or more, or when it is definitely negative. These observations can be better understood by looking at Fig. 3. It reports the percentage of agreement between the two physicians when they classify the samples into five subgroups. The bigger the agreement between physicians' classification for each class, the brighter is the gray level of the corresponding box in the

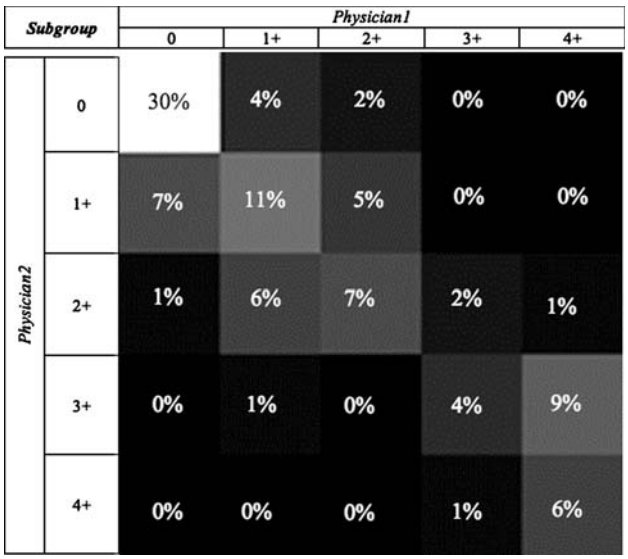

Fig. 3 Gray level map of the agreement between physicians when they classify the samples into five subgroups

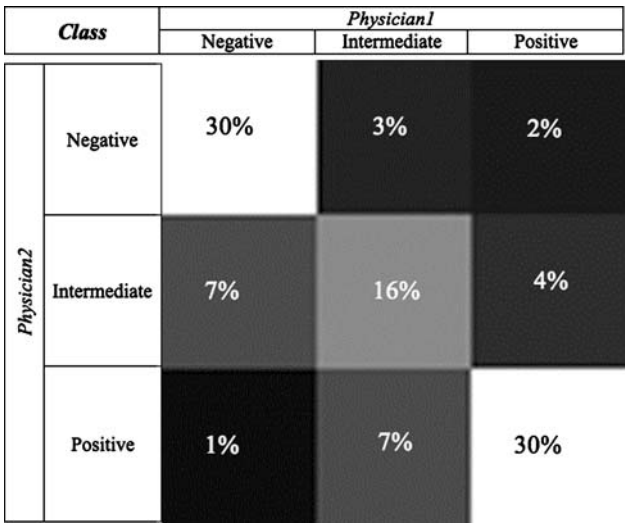

Fig. 4 Gray level map of the agreement between physicians when they classify the samples into three classes

figure. To better comprehend this symbolic representation, the gray levels are computed mapping the biggest agreement percentage to the biggest gray level of the image, that is, the white. The other agreement percentages are mapped to a gray value proportional to their value. For the sake of comprehension, such percentages are reported on the corresponding box. Note that both physicians agree with each other only in 58% of cases, that is, the sum of the main diagonal in the figure. Such a low agreement is obviously related to the low *kappa* value obtained by labeling the sample into five subgroups.

These observations suggest choosing a classification of data samples into three classes (i.e., negative, intermediate and positive). A sample is assigned to the negative class if both physicians classify it as negative, whereas it is labeled positive if both physicians mark it with two pluses or more. Finally, a sample is assigned to the intermediate class when either of the two types of disagreement described above happens or when both physicians mark it as 1+. Figure 4 represents the agreement between physicians' classification when they label the samples into these three classes. Now, the agreement percentage between the two physicians increases from 58 up to 76%. Consequently, adopting this classification rule, the measured Cohen's kappa is 0.62 $\pm$ 0.13, implying substantial agreement, which is considered satisfactory to get a reliable ground truth.

8.2 Recognition results

For testing the two introduced selection rules, that is, the binary combination and the zero-reject one, we have used the 600 images of the database. The a priori probability of positive, negative and intermediate class is 36.0, 32.5 and 31.5%, respectively.

The error rate has been evaluated according to a p-fold cross validation approach, dividing the sample set in eight-folds. The rates reported in the following are the mean of p tests. For each test, *1/p* part of the data set has been used as the validation set, another *1/p* as the test set, and the other parts as the reference set. Using classes reported in Table 4, the recognition rate is depicted in Tables 5 and 6, as relative and absolute values, respectively. Note that in the case of the zero-reject selection rule, the fourth row of Table 4 does not apply.

With respect to the binary selection rule, the overall miss rate is quite low. At a deeper analysis, the selection scheme does not exhibit false negative rate. Hence, the positive samples erroneously classified are assigned to the intermediate class, whereas intermediate samples wrongly recognized are assigned to the positive class. Furthermore, no negative samples are misclassified and occasionally they are rejected. The selection module rejects approximately 11% of samples, which is the counterpart we have to pay for such low error rates. Therefore, with reference to samples not rejected, the hit rate is 98.50%.

It is worth noting that in medical application, the two kinds of errors, that is, false positive and false negative, have very different relevance. Typically, the former kind of error can be tolerated to a larger extent since false positive leads to non-necessary analysis, whereas the false negative leads to a worse scenario, where there is a possible disease but the test indicates that the patient is healthy.

Turning attention to the zero-reject strategy based on reliability estimation of classification acts, we point out that the hit rate increases up from 87% to more than 94%.

Table 4 Output categories of the three inputs–three outputs classifier

		Input class		
		p	n	i
Output class	P	True positive (TP)	False positive (FPn)	False positive (FPi)
	N	False negative (FNp)	True negative (TN)	False negative (FNi)
	I	False intermediate (FIp)	False intermediate (FIn)	True intermediate (TI)
	R	Rejected (Rp)	Rejected (Rn)	Rejected (Ri)

Letters p, n, i and r stands for positive, negative, intermediate and rejected samples, respectively. Lower and upper case letters refers to input and output classes, respectively

Table 5 Relative performance of the recognition system, using the two selection rules

Class	Hit rate (recognition rate)		Reject rate		Miss rate		
	Binary selection (%)	Reliability-based selection (0-reject) (%)	Binary selection (%)	Reliability-based selection (0-reject)	Binary selection (%)		Reliability-based selection (0-reject) (%)
Positive samples	87.67	92.12	11.88	–	FIp	0.45	5.15
					FNp	0.00	2.73
Intermediate samples	85.03	92.24	11.53	–	FPi	3.43	6.61
					FNi	0.00	1.16
Negative samples	89.43	98.90	10.57	–	FPn	0.00	0.50
					FIn	0.00	0.60

Hence, some of the samples that are rejected by the previous approach are now correctly classified. Nevertheless, there are also samples previously rejected that are now misclassified, increasing the overall miss rate of the recognition system up to 5.67%. Note that, in this case, the performance on negative samples is still fine, since 99% of them are correctly recognized.

These results show that an approach based on the reliability evaluation is well founded. In particular, the adopted reliability estimation integrates several pieces of information, considering not only the sample of the reference set that is nearest to the unknown sample, but also the nearest sample of a class different from the chosen one.

In order to validate the MES approach, such performance figures are compared with those achieved by the other explored solutions, that is, a single classifier architecture and an MES where each specialized expert is an MLP classifier (see Sect. 5). In the first case, we train a single NN using the features reported in Table 2. The single classifier achieves a hit rate of 91.05%, which is less than the one attained by the MES (94.33%). Furthermore, looking at the relative performance we note that the misclassification rates on positive, negative and intermediate samples are 9.29% (FIp: 6.16%, FNp: 3.13%), 8.78% (FPn: 3.14%, FIn: 5.64%) and 8.76% (FPi: 3.31%, FNi: 5.45%), respectively. It is worth observing that these rates are higher than those reported in Table 5, showing that the proposed MES improves the recognition performance attainable for all the three classes. In the second case, MLP classifiers replace the NN ones in each specialized experts, that is, PE, NE and IE. Such an architecture has been tested applying the two rules presented above: the achieved recognition performance is always worse than the one attained by the MES composed of NN classifiers (Table 6). Indeed, on the one hand, using the binary selection rule, the hit, miss and reject rates are 82.66, 3.34 and 14.00%, respectively. On the other hand, employing the reliability-based selection the hit rate is 89.46%. These results show that NN classifiers outperform MLP ones in the proposed MES.

Now, given the two introduced selection rules, we are interested in understanding when it is preferable to use one strategy with respect to the other. To this aim, let us introduce the cost of a misclassification (C_m), a rejection (C_r) and the gain of a right classification (C_h). Furthermore,

Table 6 Absolute performance of the recognition system using the two selection rules

	Binary selection (%)	Reliability-based selection (0-reject) (%)
Hit rate (TP + TI + TN)	87.34	94.33
Miss rate (FP + FI + FN)	1.33	5.67
Reject rate (Rp + Rn + Ri)	11.33	–

let us denote with m, r and h the miss rate, the reject rate and the hit rate, respectively.

The convenience of using one of the two selection rules can be evaluated by introducing a global cost function (C_{tot}) defined by the linear combination of the costs with the corresponding rate. However, since h is the 100's complement of both m and r, C_{tot} depends only on C_m, m, C_r and r, the overall cost is defined by:

$$C_{\text{tot}} = mC_m + rC_r \tag{5}$$

To further simplify the formula, we can normalize it with respect to C_m, obtaining:

$$\overline{C_{\text{tot}}} = m + r\overline{C_r} \tag{6}$$

where the normalized global cost $\overline{C_{\text{tot}}}$ is given by C_{tot}/C_m, whereas the normalized rejection cost $\overline{C_r}$ is given by C_r/C_m. To find out for which combination of cost coefficients one selection rule performs better than the other, this last equation can be plotted in the $(\overline{C_r}, \overline{C_{\text{tot}}})$ plane (Fig. 5). With reference to this figure, each line represents the normalized global cost of the selection rules, and the trade-off point A is given by their intersection. The data show that when the ratio between C_r and C_m is more than 0.38, it is more convenient to adopt the zero-reject strategy, whereas when this ratio decreases it is better to use the binary selection. In practice, the binary selection rule is preferable when the cost of a misclassification is less than 2.63 times the cost of a rejection.

In Sect. 2, we have discussed four different working scenarios of a CAD, which is therefore expected to apply to them. We note that different areas in the plot correspond to different operating points, making the proposed recognition system flexible enough to pursue the CAD major objectives. Indeed, in the shaded region, Ω the classifier keeps the error rate as low as possible, approaching a zero-error system (remind that the binary selection rule does not exhibit a false negative rate), although it shows a fixed reject rate. Therefore, for operating points located in such regions, the CAD is suited for application in case α_2.

For operating points in the shaded zone Ψ, the recognition system may adopt the selection rule based on reliability estimation, performing as a zero-reject system. Hence the CAD can carry out mass screening campaigns (case α_1), can serve as a second reader (case β) or can aid the physician (case γ).

For operating points in the dotted region Θ, the recognition system could perform intermediately between the two previous zones, depending on the objective.

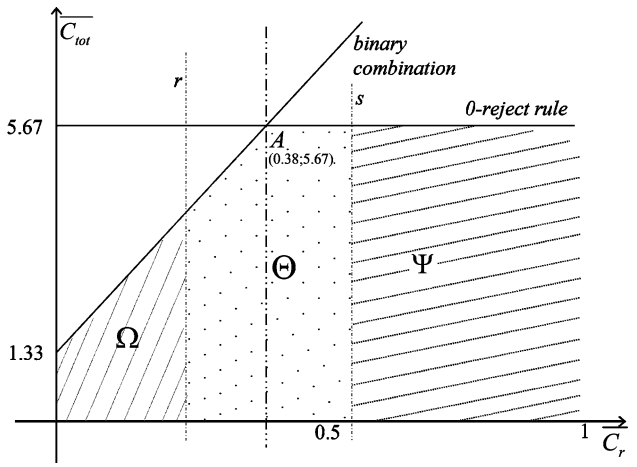

Fig. 5 Convenience analysis of using one of the two selection rule (the binary or the zero-reject one) in a given application domain. The application domain is specified by the values of cost coefficients. A is the trade-off point between the two rules. Line r, s determine three different operating regions Ω, Ψ and Θ. In the plot, we make an instance of possible values for these line equations. Note that line r and s has to be on the left and on the right side of A, respectively

9 Conclusions

In this paper we have proposed a system for the automatic classification of fluorescent intensity of IIF samples. The first issue we have dealt with concerns the procedure to reliably label the data set. Then, we addressed the key point of the feature extraction and selection, presenting three subsets of features that discriminate very well each class from the others. Therefore, we have presented a recognition system that aggregates three experts in an MES paradigm, using a classifier selection approach. In this framework, we have proposed two different selection rules, providing both a fixed-reject and a zero-reject system, respectively. The former one is based on the binary combination of the output of single classifiers, whereas the latter rule is based on the evaluation of the reliability of each recognition act of the classifiers. These rules have been experimentally evaluated, exhibiting very good performance, since the false positive and false negative rates approach zero in several cases.

Finally, we have performed a convenience analysis of using one of the two selection rules in a given application domain, which can be specified by the costs of a misclassification, a rejection and a right classification. Such an analysis allows finding out for what values of the cost coefficients one of the two rules performs better than the other. In particular, the data show that, if the cost of a misclassification is nearly three times the cost of a rejection, the zero-reject selection rule should be used. Furthermore, the two selection rules determine different regions in the convenience analysis plane, which can be complied with the different and peculiar objectives of a CAD system.

We deem that such reasons make the proposed system suited for application in daily practice.

In the end, note that the HEp-2 substrate shows different patterns of fluorescent staining that are relevant to diagnostic purposes. Hence, we are already working on a CAD system that is also capable of supporting the physician in the classification of staining pattern. This work includes the generation of a labeled data set in order to attain the ground truth for training and testing purposes.

Acknowledgments We thank Antonella Afeltra, Amelia Rigon and Danila Zennaro for their collaboration in IIF images annotation and evaluation. We also thank Dario Malosti for his constant encouragement and support. This work has been funded by Das s.r.l of Palombara Sabina (http://www.dasitaly.com), by the "Regione Lazio" under the programme "DOCUP 2000/2006-Sottomisura II.5.2-Progetto ITINERIS".

References

1. Klippel JH, Dieppe PA (1998) Rheumatology, 2nd edn. Mosby International, St. Louis
2. Kavanaugh A, Tomar R, Reveille J, Solomon DH, Homburger HA (2000) Guidelines for clinical use of the antinuclear antibody test and tests for specific autoantibodies to nuclear antigens. Am Coll Pathol Arch Pathol Lab Med 124:71–81
3. Marcolongo R, Ruffatti A, Morozzi G (2003) Presentazione linee guida del forum interdisciplinare per la ricerca sulle malattie autoimmuni (F.I.R.M.A.). Reumatismo 55:9–21
4. Pham BN, Albarede S, Maisonneuve P (2005) Impact of external quality assessment on antinuclear antibody detection performance. Lupus 14:113–119
5. Song L, Hennink EJ, Young IT, Tanke HJ (1995) Photobleaching kinetics of fluorescein in quantitative fluorescence microscopy. Biophys J 68:2588–2600
6. Piazza A, Manoni F, Ghirardello A, Bassetti D, Villalta D, Pradella M, Rizzotti P (1998) Variability between methods to determine ANA, anti-ds DNA and anti-ENA autoantibodies: a collaborative study with the biomedical industry. J Immunol Methods 219:99–107
7. Das s.r.l. (2004) Service Manual AP16 IF Plus. Palombara Sabina (RI)
8. Bio-Rad Laboratories Inc. (2004) PhD System. http://www.bio-rad.com
9. Perner P, Perner H, Muller B (2002) Mining knowledge for hep-2 cell image classification. J Artif Intell Med 26:161–173
10. Sack U, Knoechner S, Warschkau H, Pigla U, Emmerich F, Kamprad M (2003) Computer-assisted classification of HEp-2 immunofluorescence patterns in autoimmune diagnostics. Autoimmun Rev 2:298–304
11. Hiemann R, Hilger N, Michel J, Anderer U, Weigert M, Sack U (2006) Principles, methods and algorithms for automatic analysis of immunofluorescence patterns on HEp-2 cells. In: 5th International Congress on Autoimmunity, Sorrento, Italy, Autoimmunity Review and other Journals 86
12. Soda P, Iannello G (2006) A multi-expert system to classify fluorescent intensity in antinuclear autoantibodies testing. In: Computer Based Medical Systems, Los Alamitos, CA, USA, IEEE Computer Society, pp 219–224
13. Soda P, Iannello G (2006) Experiences in ANN-based classification of immunofluorescence images. Enformatika Trans Eng Comput Technol 14:252–257
14. Nattkemper TW, Ritter HJ, Schubert W (2001) A neural classifier enabling high-throughput topological analysis of lymphocytes in tissue sections. IEEE Trans Inform Technol Biomed 5:138–149
15. Cheng HD, Cai X, Chen X, Hu L, Lou X (2003) Computer-aided detection and classification of microcalcifications in mammograms: a survey. Pattern Recognit 36:2967–2991
16. De Santo M, Tortorella F, Molinara M, Vento M (2003) Automatic classification of clustered microcalcifications by a multiple expert system. Pattern Recognit 36:1467–1477
17. Van Ginneken B, Ter Haar Romeny BM, Viergever MA (2001) Computer-aided diagnosis in chest radiography: a survey. IEEE Trans Med Imaging 20:1228–1241
18. Tuzel O, Yang L, Meer P, Foran DJ (2007) Classification of hematologic malignancies using texton signatures. Pattern Anal Appl 10:277–290
19. Center for Disease Control (1996) Quality assurance for the indirect immunofluorescence test for autoantibodies to nuclear antigen (IF-ANA): approved guideline. NCCLS I/LA2-A 16
20. Dhawan AP, Chitre Y, Kaiser-Bonasso C, Moskowitz M (1996) Analysis of mammographic microcalcifications using gray-level image structure features. IEEE Trans Med Imaging 15:246–259
21. Kittler J, Hatef M, Duin RPW, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20:226–239
22. Woods K, Kegelmeyer WP, Bowyer K (1997) Combination of multiple classifiers using local accuracy estimates. IEEE Trans Pattern Anal Mach Intell 19:405–410
23. Kuncheva LI, Bezdek JC, Duin RPW (2001) Decision template for multiple classifier fusion: an experimental comparison. Pattern Recognit 34:299–314
24. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Comput 3:79–87
25. Stefano CD, Sansone C, Vento M (2000) To reject or not to reject: that is the question : an answer in case of neural classifiers. IEEE Trans Syst Man Cybern C 30:84–93
26. Rastrigin LA, Erenstein RH (1982) Method of collective recognition. Energoizdat, Moscow
27. Alpaydin E, Jordan MI (1996) Local linear perceptrons for classification. IEEE Trans Neural Netw 7:788–792
28. Cordella LP, Foggia P, Sansone C, Tortorella F, Vento M (1999) Reliability parameters to improve combination strategies in multi-expert systems. Pattern Anal Appl 2:205–214
29. Cordella LP, Sansone C (2007) A multi-stage classification system for detecting intrusions in computer networks. Pattern Anal Appl 10:83–100
30. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46
31. Landis JR, Kock GG (1997) The measurement of observer agreement for categorical data. Biometrics 33:159–174