International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:8, 2008

ANN-Based Classification of Indirect ImmunoFluorescence Images

P. Soda, and G.Iannello

***Abstract*—**In this paper we address the issue of classifying the fluorescent intensity of a sample in Indirect Immuno-Fluorescence (IIF). Since IIF is a subjective, semi-quantitative test in its very nature, we discuss a strategy to reliably label the image data set by using the diagnoses performed by different physicians. Then, we discuss image pre-processing, feature extraction and selection. Finally, we propose two ANN-based classifiers that can separate intrinsically dubious samples and whose error tolerance can be flexibly set. Measured performance shows error rates less than 1%, which candidates the method to be used in daily medical practice either to perform pre-selection of cases to be examined, or to act as a second reader.

***Keywords*—**Artificial neural networks, computer aided diagnosis, image classification, indirect immuno-fluorescence, pattern recognition.

I. Introduction

CONNECTIVE tissue diseases (CTD) are autoimmune disorders of unknown aetiology characterized by a chronic inflammatory process involving connective tissues. A common marker of CTD, although it occurs at a variable rate in the different forms, is the presence of serum antinuclear autoantibodies (ANA) [1]. The recommended method for ANA testing is indirect immunofluorescence (IIF) [2], [3]. In IIF a serum sample is tested with a substrate containing a specific antigen. Fluorochrome conjugated anti human immunoglobulin antibodies reveal the antigen antibody reaction, and the slide is examined at fluorescence microscope.

The readings in IIF are subjected to interobserver variability that limits the reproducibility of the method. To date, the highest level of automation in IIF tests is the preparation of slides with robotic devices performing dilution, dispensation and washing operations [4], [5]. The development of a system that can offer a support to physician decision is therefore an evident medical demand [3].

In this paper we focus on the development of a system that would be able to classify the fluorescent intensity of IIF samples. The system should be usable in practice, capable of both performing a pre-selection of the cases to be examined and serving as a second reader. Hence, the human expert is assumed to intervene when the system cannot cast a reliable result. In order to achieve this goal, the false-positive and false-negative rate should be as low as possible. Indeed, the former leads to non-necessary analysis, whereas the latter leads to a worse scenario, where there is a possible disease but the test indicates that the patient is healthy.

The paper is organized as follows. After reasoning the IIF diagnostic procedure in section II, in section III we present the state of the art and the motivations. In section IV we describe the image acquisition, focusing on image annotation to get a reliable data set. Section V then discusses the image analysis procedure, section VI describes the adopted classification rule and section VII presents the result of the proposed approach. Finally in section VIII, we conclude the paper.

II. IIF Diagnostic Procedure

Current guidelines for appropriate IIF tests recommend the use of tumour cell line (HEp-2) substrate [2], [3] with the 1:80 titer. In IIF diagnosis, the physician looks the sample at the fluorescence microscope, reporting both the fluorescent intensity and the staining pattern description. Since technical problems can affect test sensitivity and specificity, positive and negative controls are used [6]. The positive control allows the physician to check the correctness of the preparation process; the negative one represents the auto-fluorescence level of the slide under examination. To classify the images, physicians should execute patient serum progressive dilution, until the fluorescent intensity disappears (end-point dilution). However this practice is very expensive in time and cost, because the analysis of a single patient requires more than a well. Hence, physicians use one fixed dilution, typically 1:80, and evaluate the fluorescent intensity with respect to the negative control. The guidelines described by the Centers for Diseases Control (Atlanta, USA) are used to classify image fluorescent intensity, which identify two classes (positive and negative) and four subgroups in the positive class (Table I). Specifically, the physician classifies a sample as positive if it is more fluorescent than the negative control, negative otherwise.

Using HEp-2 cells as a substrate, the resulting sample may reveal different patterns of immuno-fluorescent staining that are relevant to diagnostic purposes [2], [3]. As an example, Figure 1 shows two images of HEp-2 cells, belonging to different subgroups of the positive class.

This work was supported by "Regione Lazio" under the Programme "DOCUP 2000/2006—Sottomisura II.5.2—Progetto ITINERIS", by MIUR under the PRIN Project "Automatic analysis in Immunofluorescence for the diagnosis of autoimmune diseases: image classification and management of images and clinical data", and by DAS s.r.l of Palombara Sabina (www.dasitaly.com).

P. Soda and G. Iannello are with the Faculty of Engineering, Università CAMPUS Bio-Medico di Roma, Via Longoni 83, 00155, Roma, Italy. (corresponding author: p.soda@unicampus.it).

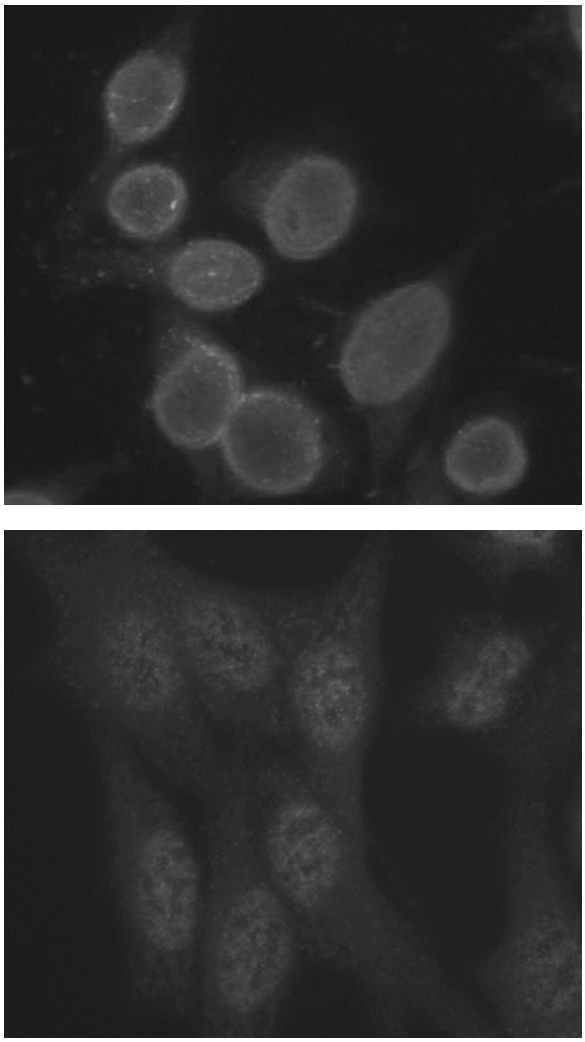

Fig. 1 Example of images that constitute the data set. In the top image, labeled with four plus, the positivity is given by the fluorescent staining of the whole cell body, whereas in the bottom image, labeled with three plus, the positivity is given by the fluorescent dots inside the cell body

TABLE I
FLUORESCENT INTENSITY CLASSIFICATION GUIDELINES

Subgroup	Description
++++	Bright green fluorescence
+++	Apple green fluorescence
++	Positive fluorescence clearly observable
+	Fluorescence level which allows clearly discrimination from background
0	Negative

III. STATE OF ART AND MOTIVATIONS

IIF is the recommended method for ANA testing [2], [3] and 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 improved diagnostic capabilities, and by the growing awareness of this clinical problem in general medicine. Currently, a higher number of health care structures need laboratories to perform these tests, but the major disadvantages of IIF method are:

- the low level of standardization;
- the lack of automated solutions;
- the inter-observer variability, which limits the reproducibility of IIF readings;
- the lack of resources and adequately trained personnel [3].

Up to now, the physician uses only his/her skills to classify the slide, without some piece of quantitative information.

In other medical contexts, Computer Aided Diagnosis system (CAD) has proven definitely effective [7], [8]. Hence, in the field of ImmunoFluorescence analysis, a CAD would attain three major objectives:

- the possibility of performing a pre-selection of the cases to be examined, both allowing the physicians to concentrate his/her attention only on relevant cases and saving time;
- the possibility of serving as a second reader, thus augmenting the physician capabilities in order to reduce mistakes;
- the possibility of working as a tool for training and education of medical personnel.

Recently, in the literature some papers proposed CAD systems to automate the HEp-2 pattern classification [9], [10]. The used image data set is made up of fluorescent images with clear patterns at dilution higher than 1:160. The system presented classifies the fluorescent pattern and exhibits an error rate of 25.6% [9] and 16.9% [10]. Note that our approach here is differently focused, since we aim to assist the fluorescent intensity evaluation (i.e. classification according to table I) using the fixed dilution of 1:80, as recommended in the guidelines [2], [3].

IV. DATA SET

A. Image Acquisition

Since, to our knowledge, there are not reference databases of IIF images publicly available, we populated a database of 540 annotated IIF images. In this respect, we use slides of HEp-2 substrate, at the fixed dilution of 1:80. A physician takes images of slides with an acquisition unit consisting of the fluorescence microscope by Leica, coupled with a 50 W mercury vapour lamp and with a digital camera. The last one has a monochrome CCD, with squared pixels of equal side to

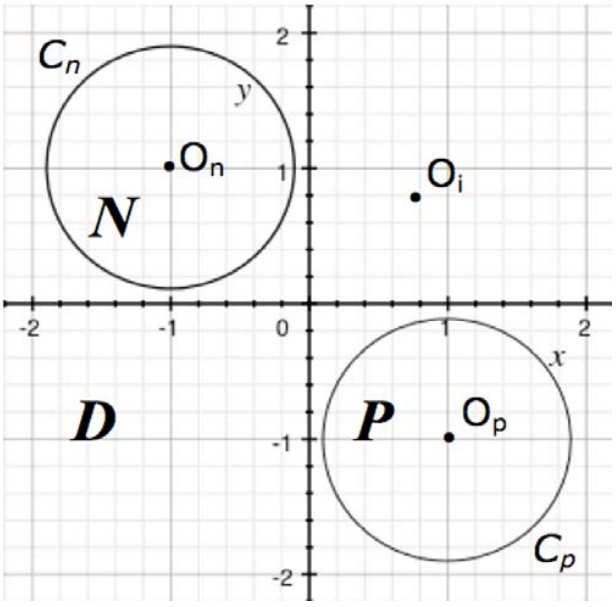

Fig. 2 Example of the classification rule used to vary the error tolerance of the classification system. The point O_i represents the *i*th output vector, points O_p and O_n represent ideal positive and negative classification, respectively. C_p and C_n are the two circles, centred at O_p and O_n, used to partition the plane in three regions

6.45 µm. The microscope objective is a 40-fold magnification and the medium is the air. The exposure time of slides to incident light is 0.4 s. The images have a resolution of 1024x1344 pixels, a colour-depth of 8 bits and they are stored in TIFF format. These images are stored in the database together with their diagnoses, as described in the following subsection.

B. Image Annotation

In the applications of pattern recognition, it is important to reliably label a data set with its true category. In the supervised classification approach, in which the input pattern is identified as a member of a predefined class, such a labelling is crucial both in the training and in the test phase.

Hence, one of the first steps in the development of a classification system is to get the ground truth. In IIF application, the ground truth is made by labelled images both with fluorescence intensity and staining pattern classification.

Since IIF is a subjective, semi-quantitative test, in [10] an objective independent method (e.g. ELISA, which permits verification of autoantibodies entities) is used to assess the human expert diagnosis on staining patterns. However, a correlation upon the 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, in order to label the data set samples and getting the ground truth for this specific application, we made use of the physician classification. Furthermore, to improve the reliability of the data set, two different physicians independently diagnosed each sample.

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 chose the most widely used one: the Cohen's kappa [11]. Its estimate, kappa (k), is expressible as a function of observed frequencies. Although the true parameter value varies from a lower bound of -1 to an upper bound of 1, the usual region of interest is $k>0$. In the literature, the following guidelines for interpreting kappa values are used: $0<k<0.2$ 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, and $0.8<k<1$ implies almost perfect agreement [12].

When the physicians diagnosed the samples following the CDC guidelines, the measured kappa is 0.46 ± 0.13 ($p<0.05$). Since this value implies moderate agreement, we concluded that labelling the sample in five subgroups was hard and not completely reliable. Indeed, the disagreement between physicians was twofold. In one case, physicians assigned the sample to different classes (i.e. one to positive, the other to negative). In the other case, physicians disagreed about the subgroups to which a positive sample has to be assigned, i.e. physicians labelled it with a different number of plus. At a deeper examination, it appeared that physicians always agreed each other when the sample was marked either with two plus or more, or when it was definitely negative.

This observation suggested choosing a classification of data samples into three classes (i.e. negative, positive and dubious). A sample was assigned to the negative class if both physicians classify it as negative, whereas it was labelled positive if both physicians mark it with two pluses or more. Finally, a sample was assigned to the dubious class when either of the two types of disagreement described above happens or when both physicians mark it with one plus. Adopting this classification rule, the measured Cohen's kappa was 0.62 ± 0.13, implying substantial agreement.

Note that, according to this approach, the original classification problem on five classes is simplified in a classification problem on three classes. While the motivation for this class revision is the ability to get a more robust ground truth, it is worth noting that in the physicians' opinion these three classes maintain the clinical significance of the IIF test. Hence, such a classification was used in the following to manage input data to the classifiers.

V. Image Analysis

A. Image Segmentation

Each image of the data set was pre-elaborated in order to improve the contrast; then morphological filters, such as erosion and dilation have been applied to remove noise.

Using Otsu's algorithm [13], automatic thresholding was performed to locate the cells. Then, using other morphological operations, such as filling and connection analysis, a binary mask for cutting out the cells from the image was obtained. Cells connected with the image border have been suppressed.

The most and the least fluorescent cells have not been considered for further analysis, because physicians reports them as damaged, i.e. cells corrupted during slide production process. To remove overlapping cells we computed the following circularity measure:

$$circularity = \frac{4 \cdot \pi \cdot (cell\ Area)}{(cell\ Perimeter)^2} \qquad (1)$$

Then, based on a simple heuristic, we removed the cells for which this parameter is less than 0.5. After these operations, the image was properly segmented and it contained only isolated cells.

B. Feature Extraction and Selection

From these segmented images, we extracted a set of features, which are chosen considering the physician expertise. They are mostly related to measures of fluorescent intensity. Also features of positive and negative controls have been considered.

To reduce the dimensionality of the feature space, we perform the Principal Component Analysis on the extracted features. The principal components that exhibit the largest variation, and that we used as input to the classifiers, were the following:

$$RG_{sample/pos\,ctrl} = \frac{G_{medium\ sample}}{G_{medium\ pos\ ctrl}} \qquad (2)$$

$$RG_{sample/neg\,ctrl} = \frac{G_{medium\ sample}}{G_{medium\ neg\ ctrl}} \qquad (3)$$

where $G_{medium\ sample}$, $G_{medium\ pos\ ctrl}$, and $G_{medium\ neg\ ctrl}$, are the mean of the fluorescent intensity over all cells of the sample, of the positive control and of the negative control, respectively.

VI. Classification Rule

Following other approaches to similar problems, we investigated several classifiers belonging to the family of Artificial Neural Network architectures [14].

After some preliminary tests, we decided to use ANNs whose output measures the probability of belonging to a class (*measurement* classifiers), in place of classifiers whose output is a discrete label. Indeed, these classifiers naturally yield a numeric value that represents the degree to which an instance is a member of a class [15].

TABLE II
Confusion Matrix for a Three Inputs-Three Outputs Classifiers

		Input Class (true class)		
		p	n	d
Output class	P	True Positive (TP)	False Positive (FP)	False Positive (FPd)
	N	False Negative (FN)	True Negative (TN)	False Negative (FNd)
	D	Dubious Positive (DP)	Dubious Negative (DN)	True Dubious (TD)

All considered classifiers have two output neurons, associated to the positive and negative classes, respectively. When a sample belonging to the positive class is presented to the network, the output neurons should ideally assume the values (1,-1), whereas when the presented sample is negative, the outputs should be (-1,1). Dubious samples should lie somewhere between these two extremes.

Indeed, since the selected classifiers work at the measurement level, the output vector corresponding to generic sample *i* measures how much *i* belongs to positive or negative classes, and it is attributed to a class according to a given rule. The simplest rule is the Winner-Takes-All, which attributes a sample to the class whose output neuron has the biggest value.

To discuss the proposed classification rule, let us consider the *xy* plane in figure 2, where the *x* coordinate represents the output of first neuron and the *y* coordinate represents the output of second neuron. The point marked O_i represents the generic *i*th output vector, whereas points marked O_p and O_n represent the ideal positive and negative classifications, respectively. The smaller is the distance between *i*th output point and point O_p or O_n, the greater is the accuracy of each single classification act of an expert.

Exploiting the presence of a third class (i.e. dubious), we therefore propose a classification rule that allows making the experts more or less conservative. To pursue this goal, we make use of two circles, with the same radius φ and centre in O_p and O_n, respectively. The φ parameter could range in the interval $[0, \sqrt{2}]$ to avoid overlapping.

Three zones, named P, N and D can be distinguished and three different corresponding cases can occur: (i) if the point O_i is inside the circle C_p, the sample is assigned to the positive class, (ii) if the point O_i is inside the circle C_n, the sample is assigned to the negative class, (iii) if the point O_i is outside both circles it is classified as dubious. As the radius value

increases, the circle areas increase as well, and the classification system is less accurate in distinguishing between positive and negative classes. In other words, the classifier becomes less conservative.

According to this classification rule and to the observations reported in the fourth section, we excluded from the training set the samples that are intrinsically dubious (i.e. samples labelled as 'd' in table II), and use only positive and negative samples to train the selected classifiers. All types of samples are instead present in the test set. Hence, the confusion matrix used in the test phase is the one reported in Table II.

To evaluate the overall classification capabilities of the approach, we selected two classifiers based on the Multi-Layer Perceptrons (MLPs) and the Radial Basis Network (RBF) architecture, since they exhibited the best performance in the current application. In the following, to evaluate the error tolerance of the experts we used a set of fourteen radii of the circles identifying the positive and negative classes, ranging in the interval $[0, \sqrt{2}]$ and regularly spaced.

VII. Results

To estimate the error rate we adopted a leave-one-out approach [16]. We therefore divided the sample set in k folds, with k = 8; the rates reported in the following are the mean of k tests.

We investigated several Multi-Layer Perceptrons (MLPs) and Radial Basis Networks (RBF) classifiers, varying both the number of hidden layers and the number of neurons for layer.

In all tests we perform, the number of hidden layers and the number of neurons per layer did not exceed three and thirty, respectively. These thresholds were chosen taking into account both the number of samples in the database and the feature space dimensionality.

To measure the ability of these classifiers to recognize negative samples, avoiding erroneous recognition of positive samples as negative, we realized the ROC curve. To globally compare the performance of each expert with respect the others, we measure the area under the respective ROC curve [17]. In the ideal case this measures 1; in real situations, the more the area approaches 1, the better is the classification system. Hence, we selected the two classifiers that exhibited the biggest value of area under ROC curve. They were:

- MLPs network trained with the backpropagation algorithm, with 6 hidden neurons and 2 output neurons. The transfer function at each layer is the hyperbolic tangent sigmoid and linear function, respectively;
- RBF network, with *spread* parameter of 6 (such a parameter is related to the selectivity of the neuron [18]).

The two experts are named 6-2 MLPs and RBF 6, respectively.

In the following performance analysis, we deemed FPd and FNd (i.e. dubious sample assignments to positive and negative classes, respectively) not as errors, since these samples are also intrinsically dubious for the physicians.

Figure 3 reports the performance of these classifiers for different threshold choices. Percentages for all curves are absolute, i.e. computed over the number of samples in the test

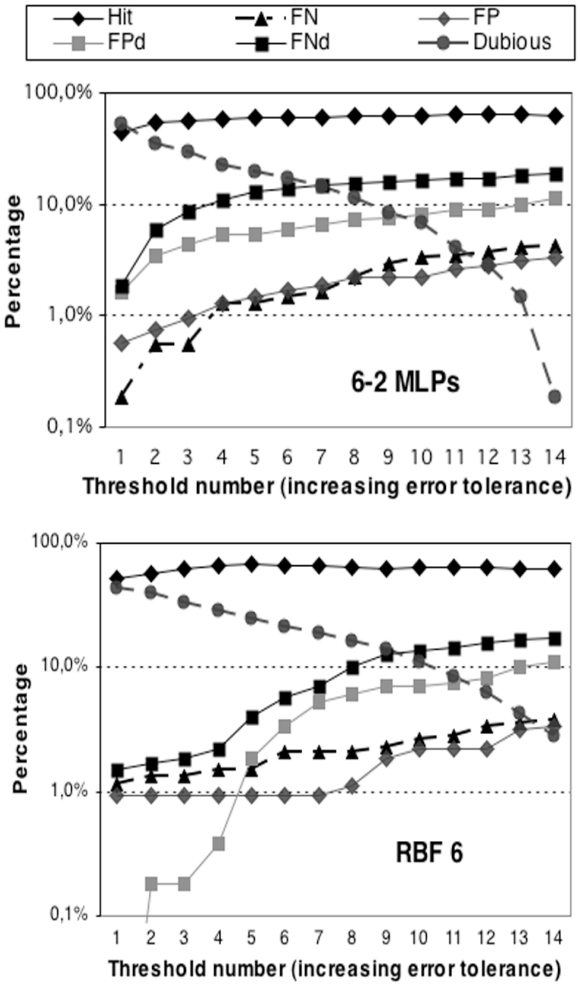

Fig. 3 Performances of the selected classifiers

set. Furthermore, they are in a logarithmic scale to better emphasize error variations, although this reduces the perception of hit variations. Note that in table II, the hit rate is the sum of TP, TN and TD, whereas the dubious rate is the sum of DP and DN.

Analysing the results obtained with the first threshold set (i.e. the most conservative), we note that: (i) all classifiers exhibit FP and FN rate approximately less than 1%, (ii) MLPs expert shows an overall error rate (FP plus FN) of 0.8%, whereas for the RBF 6 such a percentage is 1.9%, (iii) RBF 6 does not show FPd rate, (iv) RBF network exhibits a hit rate higher than MLPs expert (52.7% vs. 43.7%), (v) MLPs classifier shows a dubious rate higher than RBF expert (52.0% vs. 43.7%).

As stated previously, the focus of the paper is to develop a system that exhibit FP and FN rate as low as possible. In our system, up to the fifth threshold value, FP and FN rate are 1.3% and 1.5% for the MLPs expert, and 0.9% and 1.5% for the RBF network, respectively. At this same threshold, the hit

rate increases up to 59.7% and 67.1% for the MLPs and RBF classifier, respectively. At the same time, the dubious rate decreases to 19.2% and 24.5% for the MLPs and RBF expert, respectively. Finally, it is worth noting that, at the aforesaid threshold value, the RBF expert exhibits the maximum hit rate, whereas for the MLPs network it occurs at the twelfth threshold.

Based on these data, MLPs classifier seems complementary to Radial Basis Network. Indeed, up to the ninth threshold, on the one hand MLPs exhibits FP rate higher than FN rate, and on the other RBF expert shows the opposite attitude. In this respect, they should be the basis to realize a multi-expert classification system.

VIII. Conclusion

We have proposed a system for automatic classification of fluorescent intensity of IIF sample, without the use of information related to staining patterns. The procedure addresses the issue of realizing an expert that exhibits false-positive and false-negative rate as low as possible, making the system suited for application in daily practice. To pursue this objective we have proposed a classification rule that allows varying the working point of the system, i.e. it allows making the experts less or more conservative. Indeed the system may not cast a result occasionally, requesting that a human expert expresses the final classification of the sample.

Although in the medical application it is very hard to define a satisfactory error rate, the results are encouraging and we are currently engaged in populating a larger annotated database so as to improve the developed tools, particularly by using multi-expert system.

Finally, since the HEp-2 substrate shows different patterns of fluorescent staining that are relevant to diagnostic purposes, we are already working to a CAD system also capable to support the physician in the classification of staining pattern.

Acknowledgment

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 precious advices.

References

[1] J.H. Klippel, P.A. Dieppe, *Rheumatology*, 2nd edition, Mosby International, 1998.
[2] A. Kavanaugh, R. Tomar, J. Reveille, et al., "Guidelines for Clinical Use of the Antinuclear Antibody Test and Tests for Specific Autoantibodies to Nuclear Antigens", American College of Pathologists, Archives of Pathology and Laboratory Medicine, Vol. 124, No.1, 2000, pp. 71-81.
[3] R. Marcolongo et al., "Presentazione Linee Guida del Forum Interdisciplinare per la Ricerca sulle Malattie Autoimmuni (F.I.R.M.A.)", Reumatismo, Vol. 55, 2003, pp. 9-21.
[4] *Service Manual AP16 IF Plus*, Das s.r.l. , Palombara Sabina (RI), March 2004.
[5] *PhD System*, Bio-Rad Laboratories Inc., 2004, Available: http://www.bio-rad.com.
[6] D.H. Solomon, A. Kavanaugh, et al., "Evidence-Based Guidelines for the Use of Immunologic Tests: Antinuclear Antibody Testing", Arthritis & Rheumatism, Vol. 47, No. 4, 2002, pp. 434-444.
[7] A.K. Jain, R.P.W. Duin, J. Mao, "Statistical Pattern Recognition: A Review", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, January 2000, pp. 4-37.
[8] M. De Santo, F. Tortorella, M. Molinara, M. Vento, "Automatic Classification of Clustered Microcalcifications by a Multiple Expert System", Pattern Recognition, Vol. 36, 2003, pp. 1467-1477.
[9] P. Perner, H. Perner, B. Muller, "Mining Knowledge for HEp-2 Cell Image Classification", Journal Artificial Intelligence in Medicine, Vol. 26, 2002, pp. 161-173.
[10] U. Sack, S. Knoechner, H. Warschkau, U. Pigla, F. Emmerich, M. Kamprad, "Computer-Assisted Classification of HEp-2 Immunofluorescence Patterns in Autoimmune Diagnostics", Autoimmunity Reviews, Vol. 2, 2003, pp. 298-304.
[11] J. Cohen, "A coefficient of agreement for nominal scales", Education and Psychological Measurement, Vol. 20, 1960, pp. 37-46.
[12] J.R. Landis, G.G. Koch, "The measurement of observer agreement for categorical data", Biometrics, Vol.33, 1997, pp. 159-174.
[13] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, 1979, pp. 62-66.
[14] M. Egmont-Petersen, D. de Ridder, H. Handels, "Image Processing with Neural Networks - a review", Pattern Recognition, Vol. 35, 2002, pp. 2279–2301.
[15] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers", HP Laboratories, January 2003.
[16] K. Fukunaga, D.M. Hummels, "Leave-One-Out Procedures for Nonparametric Error Estimates", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 11, No. 4, April 1989, pp. 421-423.
[17] A.P. Bradley, "The Use of the Area Under the Roc Curve in the Evaluation of Machine Learning Algorithms", Pattern Recognition, Vol. 30, No. 7, 1997, pp. 1145-1159.
[18] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, NJ, USA, 1998.