Università Campus Bio-Medico di Roma
School of Engineering

PhD Course in Biomedical Engineering
(XX - 2004/2007)

Computer-Aided Diagnosis in Antinuclear Autoantibodies Analysis

Paolo Soda

Computer-Aided Diagnosis in Antinuclear Autoantibodies Analysis

A thesis presented by
Paolo Soda
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biomedical Engineering

Università Campus Bio-Medico di Roma
School of Engineering

Coordinator	Supervisor
Prof. Saverio Cristina	Prof. Giulio Iannello

December 2007

Contents

List of Figures

List of Tables

Ringraziamenti

E' difficile in poche righe ricordare tutte le persone che, a vario titolo, mi hanno accompagnato in questo percorso triennale. Desidero innanzitutto ringraziare Giulio Iannello, il mio supervisore, per avermi sempre seguito, sia professionalmente che umanamente, appoggiandomi ben oltre la sfera lavorativa. Dario Malosti, con i suoi preziosi consigli e la sua infinita disponibilità, è sempre stato un importante punto di riferimento.

Un sentito ringraziamento è rivolto alla Das s.r.l. ed al suo Presidente, Mario Orsini, perché dopo la tesi di laurea specialistica ha mostrato fiducia nei nostri confronti ed ha finanziato la mia borsa di dottorato.

Desidero inoltre ringraziare Antonella Afeltra ed il suo gruppo per aver fornito un impagabile supporto clinico. Tra di loro, un grazie particolare va ad Amelia Rigon per tutte le ore trascorse insieme davanti al microscopio ad acquisire e refertare le immagini necessarie allo svolgimento di questo lavoro.

Sarebbe troppo lungo, ma doveroso, ringraziare tutte le persone con cui ho collaborato in questi anni. Tra di loro voglio ricordare Luca Malosti, Massimo Rossi e Roberto Setola per il loro apporto; Carlo Sansone e Mario Vento che hanno contribuito ad alcuni dei risultati presentati in questa tesi. Sono riconoscente inoltre ai componenti del Laboratorio di Sistemi di Elaborazione e Bioingegneria Informatica dell'Università Campus Bio-Medico, ed in particolare a Roberto Valenti, "colonna portante" dell'organizzazione tecnica del laboratorio stesso, così come al personale (sia di ricerca che non) del Campus e della Das s.r.l., per aver dimostrato sempre disponibilita' nei miei confronti. Vorrei ringraziare anche i miei colleghi dottorandi la cui compagnia ha reso più gradevoli questi anni.

Sono stati altresí importanti i ragazzi che, in qualità di correlatore, ho seguito nello svolgimento delle loro tesi di laurea: Paolo Piro, Valerio Cusimano, Marco Figura, Beniamino Savo, Flavia Giovannetti De Sanctis, Francesca Pizzorni Ferrarese, Ilaria Renna, Stefano Pizzini, Marianna Meo, Francesco Spoto, Michele Ciani e Pierfrancesco D'Annibale. Insieme a loro infatti ho trascorso momenti piacevoli ed ho sempre imparato qualcosa di nuovo.

"Last, but not least" è il ringraziamento speciale a Giulia per essermi stata sempre molto vicina, per aver mischiato le nostre passioni e avermi fatto osservare il mondo anche da un'angolazione differente.

Un grazie di cuore anche ai miei genitori, Francesco ed Annamaria, a mia sorsella Gaia, ed ai miei amici più cari che mi hanno sempre sostenuto.

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 fluorescence intensity and the description of 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 work we present a CAD system suited to application in the IIF field. To this aim, we discuss different issues that should be considered to introduce viable solutions. Firstly, we evaluate the reliability of automatically acquired digital images of IIF slides for diagnostic purposes, in order to pursue a high image quality without artefacts and reduce inter-observer variability. Second, we report our experience in looking for effective autofocus functions that cope with the peculiarities of these images. We propose to use two functions that greatly improve the performance with respect to functions commonly proposed in the literature. The first one is based on image histogram, whereas the second is a popular autofocus function properly modified to compensate the photobleaching effect. Effectiveness of the proposed functions has been assessed on real images both quantitatively and qualitatively, confirming that they allow obtaining high quality images, in most cases better than those manually acquired. Third, we present a system to classify IIF images. It is based on a cascade of two steps: the first classifies the fluorescence intensity, whereas the second recognizes the staining pattern of positive wells. On the one hand, the fluorescence intensity recognition system adopts a decomposition approach, referred to as Multy Dichotomies System or MDS in short. It decomposes the classification polychotomy into a series of dichotomies on the basis of the one-per-class method. Two different aggregation rules are presented and, given the application domain, the convenience of using one of them is analyzed. In this framework, we introduce also a novel parameter that measures the reliability of the final classification provided by the MDS. This feature is used to introduce a reject option that allows reducing the error rate and determining different sets of operating points, 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). On the other hand, the staining pattern of positive wells is classified on the strength of the recognition of their cells. The core is a MDS based on the one-per-class approach devised to label the pattern on single cells. It employs a hybrid approach since each composing binary module is constituted by an ensemble of classifiers combined by a fusion rule. The error reject analysis permits to achieve system flexibility and to determine various operating points. The approach has been evaluated on 37 wells, for a total of 573 cells. The measured performance shows a low overall error rate ($2.7\% \div 5.8\%$), which is below the observed intra-laboratory variability. The analysis of perspective performance attained combining the fluorescence intensity and the staining pattern classifications shows that the use of CAD system in IIF has the potential for lowering the method variability, for increasing the standardization level and for reducing the specialists workload.

Sommario

L'Immunofluorescenza Indiretta (IIF) è oggigiorno il metodo di riferimento per determinare la presenza o meno degli anticorpi antinucleo (ANA). La loro diagnosi richiede la valutazione dell'intensità di fluorescenza e la descrizione del pattern mostrato dal substrato. Però, risorse e personale adeguatamente preparato per questi compiti non sempre sono disponibili. A tal riguardo, un'evidente esigenza è lo sviluppo di strumenti di Diagnosi Assistita al Calcolatore (CAD), che possano supportare lo specialista nel processo decisionale. In questo lavoro viene presentato un sistema CAD specifico il settore IIF e si discutono i differenti aspetti che concorrono alla realizzazione di uno strumento che sia realmente utilizzabile. In primo luogo, è stato valutato l'impatto dell'introduzione delle immagini digitali nello specifico settore al fine di ridurre la variabilità fra le diagnosi di differenti specialisti. In secondo luogo, si discute una specifica procedura di autofocus che sia adatta alle caratteristiche delle immagini in esame, al fine di sviluppare un sistema automatico di acquisizione delle stesse. A tal riguardo, si propongono due funzioni che migliorano le prestazioni rispetto ad altre presentate in letteratura. La prima si basa sull'istogramma dell'immagine, mentre le seconda è una nota funzione di autofocus adeguatamente modificata per compensare l'effetto del decadimento luminoso. Le prestazioni, valutate sia qualitativamente che quantitativamente, confermano l'efficacia del metodo. In terzo luogo, viene presentato un sistema per la classificazione delle immagini IIF basato sulla combinazione in cascata di un elemento che classifica l'intensità di fluorescenza e di un altro che riconosce il pattern.Da un lato, il sistema che classifica l'intensità di fluorescenza si basa su un paradigma di decomposizione, denominato Multy Dichotomies System o, brevemente, MDS. Tale metodo decompone il problema di classificazione distribuito su più classi in una serie di dicotomie, sulla base di un approccio noto come uno-per-classe. Vengono presentate due differenti regole di aggregazione e si analizza la convenienza di utilizzare una delle due. Si introduce inoltre un nuovo stimatore dell'affidabilità della classificazione del MDS. Tale caratteristica è poi utilizzata per introdurre un'opzione di rigetto che permette di ridurre la percentuale degli errori e di determinare un insieme di punti di lavoro del sistema. Il sistema di riconoscimento è così in grado di adattarsi a differenti scenari applicativi. Le prestazioni misurate su un insieme di immagini annotate presentano un basso valore dell'errore ($< 1.5\%, 0.00\%$ di falsi negativi). D'altro lato, il pattern dei pozzetti positivi è determinato sulla base del riconoscimento del pattern delle cellule che lo costituiscono. Il sistema che etichetta le singole cellule utilizza un MDS basato sull'approccio uno-per-classe e integra al suo interno un metodo ibrido, perché ogni classificatore binario è a sua volta composto da un insieme di esperti combinati secondo una regola di fusione. L'applicazione di un'opzione di rigetto permette di ottenere flessibilità del sistema rispetto a differenti condizioni di funzionamento. Le prestazioni, valutate su 37 pozzetti (per un totale di 573 cellule) evidenziano una percentuale d'errore ($2.7\% \div 5.8\%$) che risulta inferiore alla variabilità intra-laboratorio. L'analisi combinata delle prestazioni dei due sistemi di classificazione mostra che l'utilizzo di un CAD in IIF permetterebbe di diminuire la variabilità della metodica, di aumentare il livello di standardizzazione e di ridurre il carico di lavoro dello specialista.

Chapter 1

Introduction

Connective tissue diseases (CTD) are autoimmune disorders characterized by a chronic inflammatory process involving different organs. Antinuclear Antibodies (ANA) directed against a variety of nuclear antigens are detectable in the serum of patients with many rheumatic and non-rheumatic diseases [57]. The usefulness of ANAs tests depends on the clinical situation. If the clinical history and physical examination reveal symptoms or signs suggestive of CTD, a positive ANAs test contributes to the diagnosis. In addition, as many CTD have common clinical manifestations, the laboratory may play a fundamental role in formulating the correct diagnosis.

The recommended method for ANA testing is the Indirect Immunofluorescence (IIF) [8, 53, 95, 100]. 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 the examination at the fluorescence microscope.

There are more than 30 nuclear antigen-antibody (Ab-Ag) specificities that have been identified [95, 105]. Often the specificity Ag-Ab is associated with a specific staining pattern in IIF, which may have diagnostic value in differentiating between types of CTD.

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 improved diagnostic capabilities, and to the growing awareness of this clinical problem in general medicine. Concurrently, a higher number of health care structures need laboratories to perform these tests, but the major disadvantages of IIF method are:

- the lack of resources and adequately trained personnel [71];
- the low level of standardization [82];
- interobserver variability which limits the reproducibility of IIF readings [83];
- the photobleaching effect, which bleaches significantly in a few seconds biological tissues stained with fluorescence dyes [96];
- 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 [4, 17]. Although this greatly helps in speeding up the routine part of the tests and in improving the standardization level, it does not affect most of the above problems.

These observations suggest that the development of a Computer-Aided Diagnosis (CAD) system supporting the IIF diagnostic procedure would be beneficial in many respects. Being able to determine the presence of autoantibodies in IIF automatically would enable easier and faster test execution and result reporting, increase test repeatability, and lower costs. In response to this medical demand, some recent works addresses both the topics of image acquisition [42] and image classification [40, 41, 76, 81, 85, 90, 92, 93], respectively.

In this work we discuss different issues that should be considered in a systematic way to provide effective and viable CAD solution in this biomedical field.

To this aim, we review both the specific biomedical background and the literature on CAD solution for IIF application. Hence, we investigate the main limitations of the current methods and focus our efforts on the main research topics that are still opened. The different issues that we present can be grouped into three categories: assessment, acquisition and classification of IIF images. Indeed, we first discuss the introduction of digital images in IIF practice and full validation of their use both in manual and assisted diagnosis. In this respect our goals are twofold. On the one hand we would assess that the use of digital images neither introduces artefacts, nor leads to losses of useful information that significantly change the results of IIF tests. To this end, we propose a strategy to reliably label the sample image data set by using the diagnoses performed by different physicians. On the other hand we would provide an evaluation of the improvement that could be expected by using automatically acquired digital images in place of direct inspection of IIF samples at the fluorescence microscope.

Second, we deem that the ability to automatically and reliably acquire IIF images is a basic milestone and, therefore, we report our experience in the development of a system for automatic acquisition of IIF images. Since the conventional autofocus functions are not specifically suited to IIF images, they are limited by the photobleaching effect. In this respect, we determine an effective autofocus function that copes with it.

Third, we present a system to classify the fluorescence intensity. With respect to previous works reported in the literature [92, 93], 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 scenario. The system proposed here is based on the decomposition approach, which reduces the polychotomy, i.e. the multiclass learning problem, in less complex subtasks. Two different rules that provide the final classification on the basis of dichotomizers outputs are introduced and experimentally evaluated. 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 per each class that the system should recognize. Performing a convenience anal-

ysis on the reconstruction rules, we determine three different sets of operating points that allow applying the classifier to the main working scenarios of a CAD system, thus making it suited for application in daily practice.

Fourth, we present a recognition system that supports the classification of the staining pattern of the whole well. The recognition of the whole well staining pattern is carried out on the basis of single cells classification. The latter task employs a decomposition approach where each dichotomizer is composed of an ensemble of classifiers. The rationale starts from the widely accepted result that a multiple expert system paradigm approach generally produces better performance than those obtained by individuals composing experts. Such a configuration has been successfully tested and a reject option has been developed to adapt the system to different working scenarios. With respect to [90], we use a more sophisticated classification architecture, a different criterion to determine the well pattern and we introduce a novel reliability estimator tailored to the decomposition approach. Furthermore, our work differs from [41], [81] and [85] for two main reasons. First, they aim only at classifying the pattern of individual cells. Second, their datasets differ from ours since we use images acquired from the real patients sera diluted at 1:80, which therefore exhibits positive fluorescence intensity at various grading. Indeed, in [41] the authors employed only sera of positive controls, whereas in [81] and [85] the authors used a different data set, which is constituted by samples diluted at 1:160 and also containing cells that were negative, i.e. they did not exhibit a detectable fluorescence intensity. Finally, our system, differently from the others, may vary its operating point, which it makes it suited to different scenarios.

The analysis of perspective performance attained combining the fluorescence intensity and the staining pattern classifications shows that the use of CAD system in IIF has the potential for lowering the method variability, for increasing the standardization level and for reducing the specialists' workload. Indeed, in the most conservative and liberal setups the CAD achieves error rates (3.1% and 10.3%, respectively) that are comparable to the observed intra-laboratory variability.

Chapter 2

Background and Motivations

This chapter presents the background and motivations of the field on which this thesis focuses. It is organized as follows: in sections 2.1 and 2.2 we describe the IIF test, from its preparation to its interpretation, with particular reference to the international guidelines. Section 2.3 details the aims of the project, whereas section 2.4 reports the related works corresponding to the different fields covered by this work.

2.1 Domain Application

The rheumatic diseases are characterized by the presence of one or more autoantibodies that react with components of the nucleus, cytoplasm or surface of cells. They are disorders characterized by a chronic inflammatory process involving different organs. Exempla of autoimmune rheumatic diseases, which vary with the type of autoantibodies and the extent and severity of lesions in the various organ systems, are [8, 100]:

- systemic and discoid lupus erythematosus, named as SLE and DLE, respectively;
- mixed connective tissue disease (CTD);
- Sjogren's syndrome;
- systemic sclerosis and CREST (Calcinosis, Raynaud's phenomenon, Esophageal dysfunction, Sclerodactyly and Telangiectasia);
- rheumatoid arthritis;
- dermatomyositis and polymyositis;
- other connective tissue disease syndromes, including tumors, drug reactions and infectious diseases.

Over the past twenty years, there has been a progressive characterization of the immunochemical and molecular nature of various auto-antigens. An increased number of antigen-antibody systems associated with specific diseases have been identified. The terms "autoantibodies to nuclear antigens" or "antinuclear antibodies", commonly referred to as ANAs, have gained widespread use as generic descriptions of a group of autoantibodies. Several features of ANAs and their relationship to the rheumatic diseases have been reported in the literature [30, 53, 57, 83]. Therefore, autoantibodies directed against intracellular antigens and autoimmune rheumatic diseases constitute a major interest of physicians working in the field of clinical immunology (eg, rheumatologists, immunologists, dermatologists) and laboratory medicine [46].

ANAs have been used as diagnostic marker, since the presence of diagnostic criteria in the serum autoantibody profile of the patient represents an important requisite for the clinical diagnosis, as centralized by the progressive review of the cardinal criteria for the diagnosis of the principal autoimmune rheumatic diseases. Indeed, test for ANAs have a very high diagnostic sensitivity[1], such that it is considered the best test for the screening of the diseases reported above [8, 95, 100].

The antinuclear antibodies are associated with many immunologic disorders; however they are the essential hallmark of the systemic rheumatic diseases. Indeed, ANAs directed against a variety of nuclear antigens are detectable in the serum of patients with many rheumatic and non-rheumatic diseases [53]. Their significance is as follows:

- useful for screening and diagnostic evaluation of systemic rheumatic diseases; if the clinical history and physical examination reveal symptoms or signs suggestive of disease, a positive ANA test contributes to the diagnosis;
- ANAs positive is a classification criteria for some diseases (e.g. SLE);
- some of these diseases have distinct profile of ANA. Significant changes in the levels of certain specific ANA, such as antibodies to a-RNP, are useful in following the causes of the diseases ad their responses to therapy;
- ANAs can be useful as experimental reagents in the isolation of nuclear antigens, especially nonhistone antigens or basic studies in cell biology.

Indirect immunofluorescence (IIF) is the most commonly used technique for ANA determination because it is highly sensitive with good specificity[2], easier to perform and less expensive than other methods [8, 100].

[1]The sensitivity of a test can be described as the proportion of true positives it detects of all the positives. Therefore, it is the ratio between the true positive rate and the sum over true positive and false negative rates.

[2]The specificity of a test can be described as the proportion of true negatives it detects of all the negatives. It is thus a measure of how accurately it identifies negatives. Therefore, it is the ratio between the true negative rate and the sum over true negative and false positive rates.

Table 2.1: Recommendations for ANAs tests based on IIF.

• Test for ANAs only in patients with symptoms of autoimmune rheumatic disease (ARD) because weak ANA reactivity may be present in many nonrheumatic patients and even in healthy control subjects.
• To diagnose ARD, screen for ANAs using IIF on HEp-2 cells, and specify immunohistochemical pattern (nuclear, cytoplasmic, mitotic) and quantity (titer, concentration).
• To evaluate the positivity to ANAs execute patient serum dilution (from 1:40 up to 1:160) or use a fixed titer (1:80) and scores the fluorescence intensity semi-quantitatively (from 0 up to 4+).
• Do not use ANAs titer or concentration to monitor ARD.
• Use enzyme-linked immunosorbent assay (ELISA)-ANA screening test only when the procedure has shown good clinical and analytic correlation with the IIF method.
• Test for antinuclear specific antibody only in IIF ANAs positive patients or IIF-ANA negative patients who have clear symptoms of ARD.

2.2 Requirements for the Immunofluorescence ANA Test

In the last years, we have observed a proliferation of new methods and analytic systems in clinical immunology. It has involved a constantly increasing expenditure of economic resources for the assay of autoantibodies, estimated in hundreds of millions of dollars throughout the world [89]; this phenomenon thus has demanded the adoption of measures of standardization and verification of the quality of the methods, as well as the introduction and widespread use of guidelines for the diagnosis and monitoring of autoimmune diseases, such those proposed by the Centers for Disease Control and Prevention, Atlanta, Georgia (CDC) [8].

In this respect, some recent works present guidelines for the laboratory use of autoantibody tests in the diagnosis and monitoring of autoimmune rheumatic diseases [53, 95, 100]. Following [100], we discuss some recommendations for ANA tests, which are summarized in Table 2.1. We aim at clarifying the strength and weakness of such a test, thus providing the basis to present the motivations of our work.

2.2.1 Guidelines for the Laboratory Diagnosis of Inflammatory Connective Tissue Diseases

Many literature reports describe the standardization of methods and analytic procedures for the detection of autoantibodies [8, 53]. Here we do not detail the different guidelines, but we present some general considerations that help comprehending the key-points of IIF test.

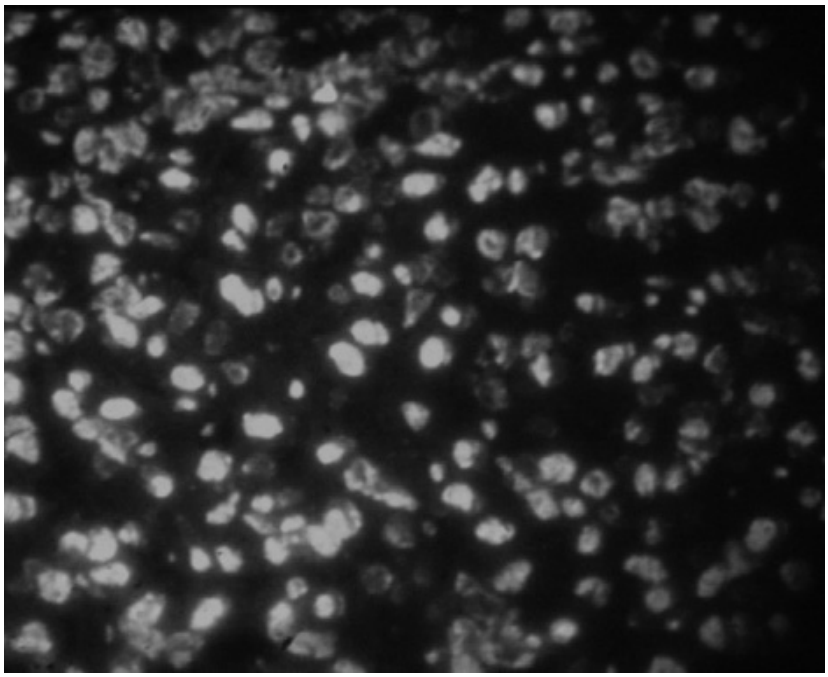

Figure 2.1: Example of the murine liver and kidney substrate.

First, the prevalence of autoimmune rheumatic diseases in the population is not high. With the exception of rheumatoid arthritis, which affects about 1% of adults, the other rheumatic diseases on the whole do not affect more than 0.5% of the population [48]. Autoantibodies in low titer without clinical significance in healthy subjects or individuals with diseases other than rheumatic diseases are a relatively frequent finding in the population (approximately in 25% of cases). This finding does not necessarily indicate the existence of an autoimmune rheumatic disease [30]. Therefore, the guidelines suggest performing ANA tests only in patients with clinical symptoms of autoimmune rheumatic disease.

Second, different substrates may be used in the IIF technique. They are:

- murine liver and kidney (figure 2.1);
- epithelial cell lines obtained from human laryngeal carcinoma (figure 2.2).

The latter substrate, named as HEp-2, has many advantages over the other tissues: *(a)* good visualization of all cellular structures, *(b)* the presence of a homogeneous, monolayer cell population, *(c)* ability to express antigens present in all phases of the cell cycle, *(d)* possibility to identify antibodies restricted to human nuclear antigens [8]. The use of HEp-2 substrate has increased the technique's sensitivity while maintaining a high specificity for detecting a wide spectrum of autoantibodies directed against intracellular antigens.

The third key-point of IIF test concerns the evaluation of the observed fluorescence intensity. To this aim, specialists should execute patient serum progressive dilution, until the fluorescence intensity disappears (end-point dilution). The recommended initial dilution of the sample is 1:40 and the guidelines suggest

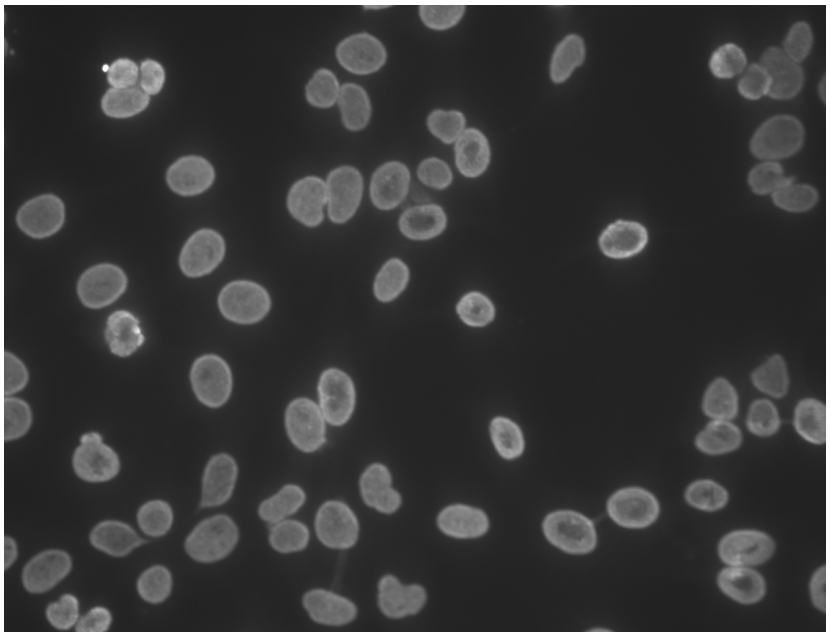

Figure 2.2: Example of the epithelial cell lines obtained from human laryngeal carcinoma (HEp-2) substrate.

progressing it up to the 1:160 titer. These two dilutions are considered decision-making levels. Indeed, titers less than 1:40 should be considered as negative, titers equal to or more than 1:40 and less than 1:160 should be considered positive at low titer (in the absence of specific symptoms, further diagnostic study is not advised, but the patient should be clinically monitored). Titers equal to or higher than 1:160 should be considered positive, and patients should undergo further diagnostic study because they are probably affected by an autoimmune disease. It is worth noting that each laboratory should verify the consistency of these cutoff levels by titrating conjugates positive sera of known titer to establish the correct optimal working dilution required in the procedure.

However this practice is very expensive in time and cost, because the analysis of a single patient requires the preparation and observation of more than a well per each serum [100].

Hence, the guidelines suggest grading the fluorescence intensity semi-quantitatively using one fixed dilution, typically 1:80. The scoring ranges from 0 up to 4+ relative to the intensity of a negative and a positive (4+) control as follows [8]:

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

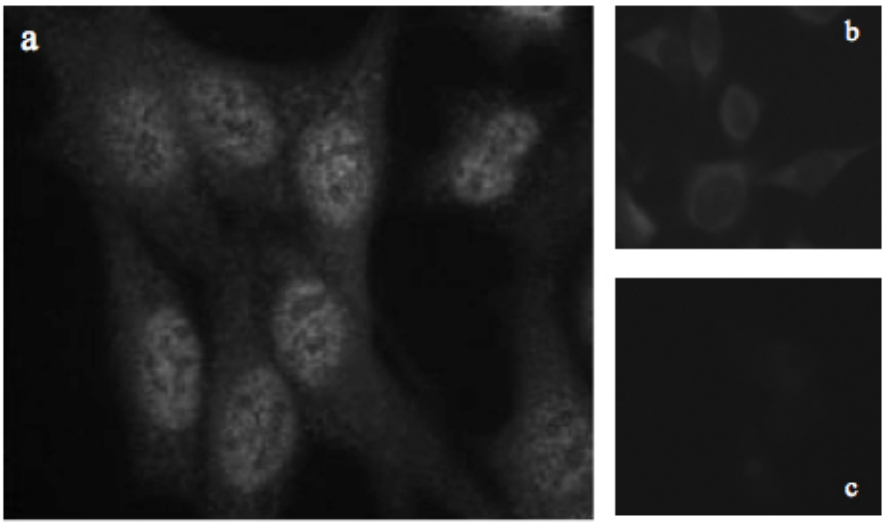

Figure 2.3: Examples of IIF images and diagnosis complexity. On the left is reported a sample (a), whose fluorescence intensity is given by the dots inside the cells, whereas on the right two different negative controls are shown (b and c). The same sample is labelled as 2+ with respect to b, whereas it is labelled as 4+ with respect to c.

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, figure 2.3 shows the images of a sample and two different negative controls, referred to as a, b and c, respectively. Note that the same sample, whose fluorescence intensity is given by the fluorescence dots inside the cells, can be labelled as 2+ with respect to the more fluorescence negative control (b), whereas it is labelled as 4+ with respect to the less fluorescence negative control (c).

The fourth key point of IIF test refers to the staining pattern that can be observed from positive samples. Indeed, when the IIF HEp-2 slides are examined at the fluorescence microscope, they may reveal different patterns of immunofluorescence staining that are relevant to diagnostic purposes. These patterns can originate from the *nuclear*, *mitotic* or *cytoplasmic* domains of the cells. To date, more than thirty different patterns could be identified, which are given by upwards of one hundred different autoantibodies [95, 105].

Among the nuclear patterns, the most frequent consists of *homogeneous* fluorescence, which is specific to DNA and histones, *peripheral* fluorescence, which is specific to deoxyribonucleoprotein and Lamin-B receptor, and *speckled* fluorescence, which is associated to RNP, Sm, Ro/SSA, La/SSB and Scl-70. The relatively frequent patterns include the *centromeric*, specific to CENP-A, CENP-B and CENP-C, the *nucleolar*, specific to PM/Scl, nucleolin, fibrillarin,

RNA polymerase I, and the *diffuse grainy*, associated to topoisomerase I or Scl70, whereas the rare patterns include the *nuclear dots*, specific to coilin, Sp-100, and the *nuclear membrane* pattern, specific to lamins A, lamins B, lamins C, glycoprotein 210.

Cytoplasmic patterns are relatively less frequent and consist of *speckled fluorescence*, specific to tRNA synthetase, *mitochondrial* fluorescence, associated to proteins of the pyruvate-dehydrogenase complex, *ribosomal* fluorescence, specific to ribosomal ribonucleoprotein, *cytoskeletal filament* fluorescence, *Golgi apparatus* and *lysosomes* fluorescence.

With reference to the rare mitotic patterns, these include fluorescence of the *spindle*, specific to tubulin, *centrosomes*, specific to enolase, *poles* or *NuMa*, specific to nuclear matrix protein,*midbody*, and the *chromosomal protein CENP-F*, which is associated mainly with cancer and hepatitis B or C and not with systemic autoimmune disease. Table 2.2 reports the previous associations between the possible fluorescence patterns and the diseases.

Among such a comprehensive prospectus of staining patterns, the CDC standardization comittee suggests reporting the following most relevant cases [8]:

- *Homogeneous*: characterized by a diffuse staining of the interphase nuclei and staining of the chromatin of mitotic cells;
- *Peripheral nuclear* or *Rim*: characterized by a solid staining, primarily around the outer region of the nucleus, with weaker staining toward the center of the nucleus;
- *Speckled*: characterized by a fine or coarse granular nuclear staining of the interphase cell nuclei;
- *Nucleolar*: characterized by a large coarse speckled staining within the nucleus, less than six in number per cell;
- *Cytoplasmic*: characterized by a cytoplasmic staining.

The fifth key-point of IIF technique concerns the use of other methods to detect the autoantibodies. Although IIF is a subjective, semi-quantitative method that presents a high analytical variability for the fluorescence intensity as well as for the staining pattern, in [85] an objective independent criteria (ELISA) is used to assess the specialist diagnosis on staining patterns. However the ELISA technique, which reveals different immunoglobulin classes and is more sensitive than the IIF technique, may also detect antibodies with low avidity having uncertain clinical significance. On the one hand, positive results require subsequent confirmation by IIF, which has a higher specificity [100]. On the other hand, 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 [53], 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 and the literature accords that ELISA tests cannot yet substitute for the IIF procedure, which could be still considered the reference method for ANAs detection.

Table 2.2: Association between HEp-2 fluorescence staining pattern and systemic autoimmune disease. They are reported from the most frequent (nuclear) to the most rare (mitotic).

Pattern	**Specificity**
Nuclear	
Homogeneous	DNA
	histones
Peripheral	deoxyribonucleoprotein
	Lamin-B
Speckled	RNP
	Sm
	Ro/SSA
	Ro/SSB
	La/SSB
	Scl-70
Centromeric	CENP-A
	CENP-B
	CENP-C
Nucleolar	PM/Scl
	nucleolin
	fibrillarin
	RNA polymerase I
Diffuse grainy	topoisomerase I
	Scl-70
Nuclear dots	coilin
	Sp-100
Nuclear membrane	lamins A
	lamins B
	lamins C
	glycoprotein 210
Cytoplasmic	
Speckled	tRNA synthetase
Mitochondrial	proteins of the
	pyruvate-dehydrogenase complex
Ribosomal	ribosomal ribonucleoprotein
Cytoskeletal filament	
Golgi apparatus	
Lysosomes	
Mitotic	
Spindle	tubulin
Centrosomes	enolase
NuMa or poles	nuclear matrix protein
Chromosomal protein	CENP-F

Finally, note that the evaluation of fluorescence intensity and staining pattern should be executed independently by two physicians experts of IIF to improve the reliability of the decision, as the literature recommends [8, 53, 100].

In summary, in the initial diagnostic phase in patients with clinical symptoms that raise suspicion for autoimmune rheumatic disease, the first test is the detection of ANAs by IIF; the pattern of nuclear or cytoplasmic fluorescence determines the subsequent test, represented by the search for autoantibodies directed against one or more specific intracellular autoantigens.

2.2.2 Principles of the IIF Test

The IIF test utilizes the indirect fluorescence antibody techniques first described by Coons and Kaplan [12]. The following steps are performed when using such a method:

1. mounted tissue sections of tissue culture cells, which are fixed on microscope slides, serve as the source of the nuclear antigens. Tissue culture cells are recommended as the most reproducibile source of substrate tissue;
2. titered test serum is added, and incubation follows. Known ANA positive and ANA negative control sera are also added to separate substrate vials or slides;
3. excess serum protein is washed off;
4. the detecting reagent, fluorescein-conjugated antihuman polyvalent immunoglobulin, is added, followed by incubation;
5. excess conjugate is washed off;
6. the slides are observed for the presence of specific nuclear immunofluorescence.

The ability to perform the IIF at room temperature is convenient.

2.2.3 IIF Limitations

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 the IIF tests, but the major disadvantages of the method are:

1. the lack of resources and adequately trained personnel [71];
2. the low level of standardization [82];
3. interobserver variability which limits the reproducibility of IIF readings [83];
4. the photobleaching effect, which bleaches significantly in a few seconds biological tissues stained with fluorescence dyes [96];

5. the lack of automatized procedures.

In the following, we discuss in details such limitations.

1. The autoimmune diseases have increased their incidence in the last years and, therefore, the health care structures need to carry out more tests. However, performing more tests implies more workload, that may lead to a less accurate diagnosis. Indeed, on the one hand, the specialists may allocate less time to diagnose each IIF image to increase the throughput. On the other hand, the health center may assign the diagnosis task to personnel who can be not adequately trained to analyze the many qualitative or semi-quantitative IIF features.

2-3. The proliferation of new methods and analytic systems in immunology observed in the last years has demanded the development of both standardization and verification measures of the quality of the methods. To this aim, recent works have proposed guidelines for the IIF test [8, 53, 95, 100], as summarized in section 2.2.1. However, up-to-date researches demonstrate that ANAs detection in routine practice is far from being standardized [82, 83]. Indeed, the former paper reported a six-year consecutive survey that involves between 606 to 687 laboratories together with six university reference laboratories experienced in performing tests in autoimmunity. The authors observed variability in IIF methodological procedure, such as the use of inappropriate microscope magnifications for reading slides or nonconventional cutoff dilution of serum. Concerning ANAs measurement, the rate of agreement on positive and negative samples ranged from 92.7% up to 99.5%. The latter paper, i.e. [83], involved sixteen industrial and two university laboratories that determinate the anti-nuclear ANAs in 11 sera from patients with clinically diagnosed systemic rheumatic disease, using IIF and other methodologies[3]. The authors found that the agreement for positive-negative results and for fluorescence patterns description was 92.6% and 76.0%, respectively. Furthermore, the intra-laboratory variability was measured equal to 7.4%. They observed that this agreement variability, which is consistent with past studies [28], may be also due to the difference in the subjective interpretation of the HEp-2 preparations. Such results prove that interobserver variability limits the reproducibility of IIF readings.

4. The indirect immunofluorescence, as the name means, uses fluorescence intensity to extract information about the local concentrations of autoantibodies that are previously labelled with fluorescence probes. The photobleaching effect is a dynamic process in which fluorochrome molecules undergo photo-induced chemical destruction upon exposure to excitation light, thus lose their ability to generate fluorescence. It can be described as the cumulative effect of fluorophore loss from each excitation-emission cycle over time. The bleached molecules can no longer participate in the excitation-emission cycle [96, 97]. The rapid decrease of fluorescence is due to the interaction between dyes (D-D *dye to dye*) and also by the interaction between dye and oxygen molecules (D-O *dye to oxygen*). The

[3]The other methods tested are immunoenzymatic assay, counterimmunolectrophoresis, immuno and western blotting.

D-D reaction probability is a function of the intermolecular distance between the dye molecules. In order to consider the two reactions, it has been proven that the total amount of fluorescence light shows an exponential decay in time [21], thus altering the observed intensity of the sample under examination.

5. 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, 17]. Although this greatly helps in speeding up the routine part of the tests and in improving the standardization level, it does not affect most of the above problems.

2.3 Project Description

The previous sections have presented both the relevance of the IIF method and its limitations. Furthermore, the literature accords that it can be considered the reference method for ANAs detection at the present time [8, 53, 82, 83, 95, 100].

However, 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.

Automation may offer a solution to the growing demand of diagnostic tests for systemic autoimmune diseases. Indeed, it has been introduced in other areas of medicine in general, e.g. radiography, mammography or hematology [10, 102, 103, 112], and laboratory medicine in particular, e.g. the analysis of tumor cells [9, 107], skin lesions [20, 26], bone marrow [54, 74], brain [99], lymphocites [77], proving its usefulness. 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 Computer-Aided Diagnosis (CAD) system, which may support physician's decision and overcome current method limitations.

Figure 2.4 shows a typical architecture of a CAD system devised to analyze microscope images. The user controls and interacts with the system through a Graphical User Interface (GUI), which is located at a workstation. The workstation controls a full-motorized microscope, whose parameters, such as focusing, movement stage, objective, illumination and magnification can be flexibly set. Moreover, the microscope is equipped with a digital camera to acquire one or more images of the sample under examination. The digital images are transferred to the workstation, and then they are automatically post-processed, stored and shared on the web. It is well known that such systems play an important role since they not only support the specialist in the image analysis task, but they also provide a set of useful functionalities that help in speeding up the routine part of the work.

Besides providing traditional image post-processing tools, e.g. noise reduction, filtering and image enhancement, the main functionality of a CAD regards the automatic classification of the images.

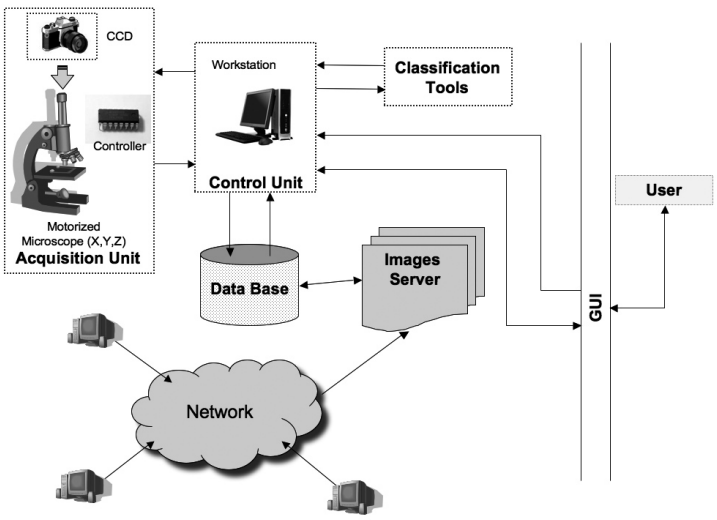

Figure 2.4: Typical architecture of a computer-aided diagnosis (CAD) system.

The analysis of the literature in the field of ANAs detection reveals that a comprehensive CAD system in IIF is not available yet, as discussed in the following section. Therefore, this work aims at developing a system that addresses the limitations reported above. It should improve both data management and standardization level, with particular reference to image acquisition and classification. th Figure 2.5 shows the schematic of the system that we present in this work, whereas figure 2.6 depicts the flow chart of the main operations. The former figure details the data that are transferred from one functional block to the others. The control unit of the CAD, e.g. a workstation, interacts with the user via a GUI, which allows introducing metadata and clinical remarks in the system. The central unit transmits the signals that control the acquisition unit. The acquired images are then transferred to the workstation. The central unit provides images and the related metadata to the classification unit. Such a block extracts a set of features, classifies the fluorescence intensity and the staining pattern, and returns the predicted labels. Furthermore, the control unit stores or retrieves data from the repository. Figure 2.6 shows that the user can acquire or retrieve patient data. Then, they can be managed or processed using a set of computer based tools.

With reference to figure 2.4, this work focuses on the acquisition and classification units, since their development presents the main research challenge in indirect immunofluorescence. However, we deem that the first issue that should be considered in a systematic way to provide effective and viable CAD solutions is the introduction of digital images in IIF practice and full validation of their use both in manual and assisted diagnosis. Hence, this last topic is discussed in chapter 3, whereas chapters 4 and 5 present the acquisition and classification of IIF images, respectively.

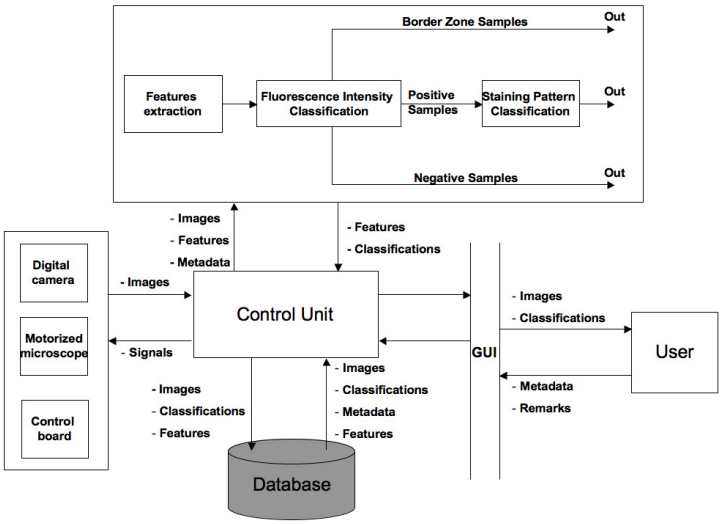

Figure 2.5: Schematic of the CAD system that we present in this work. The figure reports the data that are transferred from one functional block the others.

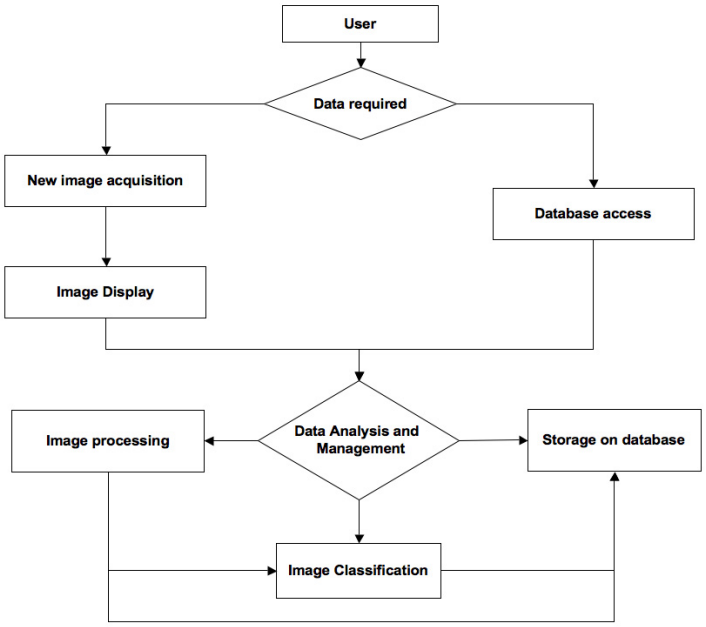

Figure 2.6: Flow chart of the main operations.

2.4 Related Work

2.4.1 Image Assessment and Acquisition

Previous works in the field of IIF image assessment and acquisition are rare. Indeed, on the one hand, only [42] discusses the issue of IIF image quality evaluation to optimize automatic image acquisition. On the other hand, many papers study and evaluate the autofocus functions that are suited to be used in general microscope applications [6, 29, 32, 35, 87, 110].

In [42] the authors developed a method to determine image parameters allowing an objective appraisal of quality of image data as well as a separation of object and background. They calculated both shape and texture features on IIF images. Area, perimeter and shape factor of the objects in the images belong to the former type of parameters, whereas image variance, contrast, correlation, minimum object grey value (MinGW), maximum object grey value (MaxGW) and range (MaxGW−MinGW) belong to the latter type of features. The results showed that the values of variance, correlation and contrast features significantly vary between corrected exposed and overexposed images. The authors therefore proposed to evaluate these parameters to adjust camera setting for automatic image acquisition, allowing optimized image data without conditional subjective misalignment of camera system.

With reference to the literature on autofocus algorithms, many functions have been proposed and compared [6, 29, 32, 35, 87, 110]. They are based on previous knowledge about the differences in information content in focused and unfocused images. The basic assumption behind most of these functions is that defocused image results from the convolution of the image with a certain point-spread function [35], which usually produces a decrease in the high frequencies of the image. This result can also be seen on the assumption that well-focused images contain more information and detail (edges) than unfocused ones [115]. According to a general classification reported in the literature [29, 35, 87, 115], the focus functions can be divided into five groups: functions based on (1) image differentiation, (2) depth of peaks and valleys, (3) image contrast, (4) histogram and (5) correlation measurements. For further details on such functions, the interested reader may refer to [6, 29, 32, 35, 87, 110].

2.4.2 Image Classification

Up to now, during the inspection of micrographs[4], the physicians use only qualitative or semi-quantitative information, thus limiting method reproducibility. In order to overcome these drawbacks, the recent researches have demonstrated that supporting more the specialist by moving from a structural approach to a semantic interpretation of micrographs is a promising trend [7, 9, 20, 26, 44, 54, 69, 70, 74, 75, 77, 78, 88, 94, 99, 102, 107, 109, 112, 113]. The final goal is the classification of the whole image as well as regions or objects detected by the segmentation[5] procedures.

To this aim, pattern recognition techniques are employed, such as those

[4] In the field of microscope analysis, its images are also referred to as micrographs.

[5] Classically, image segmentation is defined as the partitioning of an image into not overlapping, constituent regions that are homogeneous with respect to some characteristic, such as intensity or texture.

Figure 2.7: Example of a theoretical error-reject trade-off curve.

based on supervised or unsupervised approaches. Note that, in spite of several years of research and development in the field of pattern recognition, the general problem of recognizing complex patterns with arbitrary scale, orientation and location remains unsolved. Therefore these systems have to be tailored to the specific application domain.

The ultimate performance measure of a general pattern recognition system is the overall error rate (P_e). In general, a classification system makes or not a decision on all input samples. In the former case, the CAD acts as a zero-reject system, whereas in the latter one some samples are rejected. Indeed in many applications, such as the medical ones, the designers can implement a *reject option* that aims at rejecting the highest possible percentage of samples that would otherwise be misclassified (i.e. misclassified without a reject option). It is worth noting that, even when used adequately, this criterion introduces a side effect whereby some samples that otherwise would have been correctly classified are rejected. Invoking the reject option reduces the error rate; the larger the reject rate, the smaller the error rate. This relationship is represented as an error-reject trade-off curve that can be used to set the desired operating point of the classifier. It is monotonically non-increasing, since rejecting more patterns either reduces the error rate or keeps it the same (figure 2.7). As example, a good choice for the reject rate is based on the costs associated with reject and incorrect decisions [19, 34, 86].

Taxonomy of Classification Systems in Medicine

Developments in computer vision and artificial intelligence in medical image interpretation have shown that the classification module of a CAD system can pursue four major objectives [10, 18, 77, 102, 103]:

(i) performing a pre-selection of the cases to be examined, enabling the physician to focus his/her attention only on relevant cases, making easier to

carry out mass screening campaigns;

(ii) serving as a second reader, thus augmenting the physician capabilities and reducing errors;

(iii) aiding the physician while he/she carries out the diagnosis;

(iv) working as a tool for training and education of specialized medical personnel.

Therefore, each of the previous working scenarios has its requirements and the CAD it is expected to apply to them. Based both on these observations and on the considerations reported above, for each working scenario the CAD behaviour can be further characterized (Table 2.3).

In case (i), 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, i.e. it acts as 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 to 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 to a zero-error system and rejects doubtful samples.

This section 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 α_2. Therefore some potential sick patients could not be processed in case α_2, unless additional work is 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 (ii), 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 (iii), referred to as γ, the recognition system aids the physician during the diagnosis, performing as a zero-reject system.

Finally in case (iv), 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.

Table 2.3: Description and requirement of a CAD system for different working scenarios.

Case	Working scenario description	CAD requirement
α_1	Mass screening campaigns	Zero-reject system
α_2	Samples pre-selection	Reject system
β	Second reader	Zero-reject system and reliability measure
γ	Aiding expert in the diagnosis	Zero-reject system
δ	Training and education of specialists	Depends on the purpose

Classification of IIF Images

IIF is the recommended method for ANAs determination but up to now, the physicians rarely made use of quantitative information (see sections 2.2 and 2.3). The development of a CAD in such field would improve the medical practice, achieving the advantages mentioned above.

Previous works on the classification of both fluorescence intensity and staining pattern are reported in [40, 41, 76, 81, 85, 90, 92, 93]. They focused on such topics since they represent the two sides of IIF diagnosis.

With regard to fluorescence intensity classification, a system based on a Multi-Layer Perceptrons and a Radial Basis Network has been proposed in [92, 93]. That system, which made use of features inspired to medical practice, showed low error rates (false positive plus false negative) up to 1%, but it used a reject option and it did not cast a result in about 50% of cases. It used two features related to the mean of the fluorescence intensity among the cells of the image. The small set of features was chosen with reference to the number of samples in the data set in order to avoid the course of dimensionality[6] [3, 101]. The critical point of this approach was the cell segmentation algorithm, since it did not get to deal effectively with the great difference in appearance between the fluorescence images and the very low contrast of the negative samples (figure 2.3). Moreover, the adopted classification rule did not allow a flexible management of samples that are intrinsically hard to classify. This justified the high reject rate required to obtain low error rates, making the CAD suited for application in case α_2.

With regard to staining pattern classification, different works have been proposed in the literature [40, 41, 76, 81, 85, 90]. The approaches presented in [81] and [85] are similar since both used ELISA test to construct the ground truth[7]. However, such a choice introduces the limitations and the drawbacks presented in section 2.2.1, since ELISA cannot be taken as a golden standard for IIF classification [8, 100]. In these two works, the data set consists of 321

[6]The course of dimensionality is a phenomenon where the added features may actually degrade the performance of a classifier if the number of training samples that are used to design the classifier is small relative to the number of features [49]. It is also known as *peaking phenomenon*.

[7]The term *ground truth* refers to a labelled data set which is used for testing and training purposes in the application of supervised pattern recognition techniques.

and 1041 cells, respectively. The samples were distributed not only among the five staining pattern classes reported in section 2.2.1, but also a *negative class* was introduced. The latter had been included because the authors aimed also at discriminating between positive and negative samples. The cells were localized applying global thresholding by the Otsu's algorithm [80]. The algorithm well located the cells with their cytoplasmic structure, but not the nuclear envelope itself. To overcome such limitations, the authors used morphological filters and the overlapping cells were eliminated by a simple heuristic based on compactness evaluation [81, 85]. After image segmentation, more than one hundred texture-based features were computed on segmented cells, and then they were given to a decision tree induction algorithm to find out the most relevant subset and to construct the classification knowledge. These systems exhibited an error rate of 16.9% [85] and 25.6% [81]. acting only as a zero-reject systems without providing a reliability measure of final classification. With reference to Table 2.3, such systems can operate only in case α_1, γ and δ, respectively.

In [40], the authors emphasized the relevance of developing a fast, objective, safe and economical automatic analysis in the IIF field.

Recent results have been published in [41, 90]. In the former paper, the authors presented results on three experiments to show the robustness and the reliability of the image capturing process proposed [41]. In the first experiment, several positive and negative control sera were scanned and image quality (sharpness and brightness) of every scene was checked. The results showed that 25% of acquired images were evaluated as unsuitable for further uses. Most of the errors were due to inhomogeneous cell distributions, procedure artefacts, such as destroyed cells on the covered slide. In the second experiment, influences on the main scanning time were examined, recognizing technical and biological factors which influence the complete analysis time of every well of the serum. In this respect, the authors proposed a high-transmission filter and a high-numerical aperture objective to optimize the scanning system. The third experiments regarded the recognition of the fluorescence pattern of single cells into one of the following groups: homogeneous, nucleolar, speckled, rim, nuclear dot, centromere and mitotic associated (spindle). To this aim, the authors segmented the cells using a histogram-based mixture model threshold algorithm, which models the background intensity, in conjunction with the watershed transform [104]. Then the segmented objects were described with boundary, regional, topological, and texture/surface features. The set of attributes was then reduced to 86 descriptors that are given to different recognition algorithms, e.g. decision tree, Nearest Neighbour, Random Forest, Naive Bayes or Radial Basis Network, to find out the most performing classifier. The experiments showed that the lowest error rate ($P_e = 3.1\%$) were achieved using a logistic model tree. It is a classifier that builds classification trees with logistic regression functions at the leaves [66]. However, it is worth noting that in this paper only sera of positive controls were used. They are reference sera that exhibit an evident fluorescence intensity and staining pattern and they do not represent the real cases that occur in the daily practice. Hence, a direct comparison of these results with the other presented either in [81, 85] or [90] is not possible.

In [90] we presented a recognition system devised to label the pattern of HEp-2 wells diluted at 1:80 on the strength of its cells classification. We employed an approach based on multiple classifiers (commonly referred to as Multi-Experts System or MES in short) and introduced a fixed reject that aims at lowering

the misclassification rate. The error rate (P_e) and the reject rate (P_r) of that system were 13.1% and 42.6% for individual cells, respectively. To determine the staining pattern of the whole well, two different procedures based on relative and absolute majority of cells distribution among the different staining classes, i.e. homogeneous, rim, speckled, nucleolar and no pattern, were proposed. In the first rule, the class with the relative majority of cells sets the pattern of the well. If the two most populated classes have the same cardinality, the recognition system does not cast a decision, i.e. the well is rejected. In the second rule, the well is labelled as the class with more than half of samples (absolute majority). If no class wins the absolute majority of cells, the system does not cast a decision. Using the former rule, $P_e = 12.9\%$ and $P_r = 16.6\%$, whereas adopting the latter rule $P_e = 8.3\%$ and $P_r = 29.5\%$, respectively.

Finally, a different approach has been proposed in [76], where the authors evaluated an automatic fluorescence image analyzer (Image Titer, Tripath Imaging, Burlington NC), which measures the titer of the sample by means of only one image per subject. Indeed, the algorithm created images of samples diluted stepwise by extrapolation from a slide at one dilution point. Therefore, it did not required the staining of a series of diluted samples as did the manual method. The method was tested on 132 serum samples prepared using HEp-2 substrate diluted at 1:40. Different people, i.e. a well-trained, a newly trained, and an inexperienced person, reported both the staining pattern and the titer using the manual method and the Image Titer system. In relation to the staining pattern determination, the results showed that the people reported the same ANAs staining pattern for all the samples tested by the Image Titer as by the manual method. With reference to the dilution evaluation, the authors regarded as coincident the two titers provided by the specialists, i.e. estimated looking at images of virtual and real dilution, if they differ within ± 1 dilution. They observed that in the he 93.9% of sera, the value of the titer obtained using the Image Titer coincided with the titer obtained by the manual method with sufficient accuracy, i.e. within ± 1 dilution. The others 6.1% were out of range, i.e. they were 2 dilutions higher or lower than by the manual method. Those samples belonged both to homogeneous and to speckled pattern classes, respectively. It is worth noting that such an analysis did not supply an estimation of the error rate, as usual in the pattern recognition field. Furthermore, although the system provided virtual dilutions of the serum samples, it did not support their classifications which was therefore carried out by specialists themselves. On these motivations, we deem that the system cannot be considered a "traditional" classification system. Hence, a direct comparison of its performance with respect to the others is not possible.

Chapter 3

IIF Image Assessment

In the first chapter of this dissertation we have shown that the development of a Computer-Aided Diagnosis (CAD) system supporting the IIF diagnostic procedure would be beneficial in many respects. Being able to determine the presence of autoantibodies in IIF automatically would enable easier and faster test execution and result reporting, increase test repeatability and lower costs.

The first issue that should be considered in a systematic way to provide effective and viable CAD solutions is the introduction of digital images in IIF practice and full validation of their use both in manual and assisted diagnosis. In other words, we suggest that procedures to automatically acquire digital images of IIF slides should be defined and validated. Then these images can be used either to perform IIF diagnosis by specialists looking at them on a computer monitor in place of direct observation at the fluorescence microscope, or as input to a CAD system. Note that this achievement would hopefully provide significant advantages in a short-term over current practice, since it should greatly reduce the negative effects of both the photobleaching and the interobserver variability. Moreover, the availability of reliable digital images would enable the provisioning of a wide spectrum of useful features, ranging from easy repeatability of diagnosis over time to integration of IIF exams into Electronic Patient Records.

One previous work on the issue of IIF image acquisition can be found in [42], and commercial products for such a purpose have been recently released to the marked [79]. Nevertheless, to our knowledge, this is the first study about the validation of using digital images in IIF tests. Here we focus on the use of automatically acquired digital images for diagnostic purposes, i.e. if the physicians may reliably use digital IIF images in place of direct microscope observations to carry out the diagnosis. Specifically, we present data about diagnoses performed by experts looking at the fluorescence microscope and at digital images on the computer screen. Our goals are twofold. On the one hand we would assess that the use of digital images neither introduces artefacts, nor leads to losses of useful information that significantly change the results of IIF tests. On the other hand we would provide an evaluation of the improvement that could be expected by using automatically acquired digital images in place of direct inspection of IIF samples at the fluorescence microscope.

As the following results discussion demonstrates, specialists working at a computer monitor can profitably use digital images to perform diagnosis.

3.1 Materials and Methods

3.1.1 Slide Preparation, Acquisition and Diagnosis Procedure

The samples included in this study were obtained for diagnostic purposes and routine testing from consecutive outpatients and inpatients of the Campus Biomedico, University Hospital of Rome, Italy. Ninety-six sera (73 F; 23M) were screened for ANAs by IIF on HEp-2 cells (ATCC-CCL 23). Mean age was 49 years (73-15 ys). Clinical diagnoses are reported in Table 3.1.

Table 3.1: Clinical diagnoses of screened sera.

Diagnosis	**Num samples**	%
Healthy	44	45.8%
Connective tissue diseases (CTDs)		
Rheumatoid arthritis	4	4.2%
Sieronegative arthritis	4	4.2%
SLE	6	6.3%
Sjogren syndrome	9	9.2%
Undifferentiated connective tissue diseases	3	3.1%
Other CTDs	4	4.2%
Other autoimmune disorders		
PAPS	1	1.0%
Autoimmune thyroid disorders	6	6.3%
Autoimmune piastrinopenia	1	1.0%
Raynaud	4	4.2%
Autoimmune Gastrointestinal disorders (celiac disease, IBD)	4	4.2%
Viral hepatitis	6	6.3%
	Tot 96	

Sera were diluted 1:80 in 0.01M phosphatebuffered saline (PBS), pH 7.2 (50 microL of serum in 1950 microL of PBS). A positive and a negative reference controls were tested in each slide for quality control.

Pre-diluted sera were overlaid on fixed HEp-2 cells (The Binding Site Ltd, UK) for 30 min at room temperature. Slides were washed twice for 5 min each with PBS, overlaid with fluorescently labelled conjugate (FITC) (goat antihuman IgG, heavy and light chains; Binding site), and incubated for an additional 30 min. After a slide was washed twice, a cover slip was placed over the slide with mounting medium (The Binding Site Ltd, UK).

The slides were read with a Leica fluorescence microscope equipped with a 50 W mercury vapour lamp. The objective is a 40-fold magnification and the medium is the air. The fluorescence intensity was scored semi quantitatively from 1+ to 4+ relative to the intensity of a negative and a positive control (4+), by following the guidelines established by the Centers for Disease Control and Prevention, Atlanta, Georgia (CDC) [8]:

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

In order to simplify the staining pattern classification, the sera were classified in the following basic groups that are specific to the most relevant and recurrent ANAs:

- *Homogeneous*: characterized by a diffuse staining of the interphase nuclei and staining of the chromatin of mitotic cells;
- *Peripheral nuclear* or *Rim*: characterized by a solid staining, primarily around the outer region of the nucleus, with weaker staining toward the center of the nucleus;
- *Speckled*: characterized by a fine or coarse granular nuclear staining of the interphase cell nuclei;
- *Nucleolar*: characterized by a large coarse speckled staining within the nucleus, less than six in number per cell;
- *No pattern*: unclassifiable pattern.

Although just reported in section 2.2.1, we have repeated here the classes definitions for the sake of comprehension. We used the same staining pattern groups that are reported in other works [41, 81, 85]. Therefore, we do not consider the cytoplasmic pattern since it is not so much frequent as the others [100].

The images were blindly classified by two physicians, experts of IIF, working at the fluorescence microscope. The diagnoses (fluorescence intensity and staining pattern) were performed independently, at short temporal distance to minimize the effect of fluorescence decay. During the reading phase at the microscope, one of the two experts, randomly chosen, selects three different zones of the well under examination, based on their clinical significance to perform the diagnosis. These areas are then acquired with an acquisition unit consisting of the fluorescence microscope and a monochrome CCD camera, which has squared pixels of equal side to $6.45\mu m$. The exposure time of slides to incident light is 0.4 s. The images have a resolution of 1024x1344 pixels and a colour depth of 8 bits; they are stored in TIFF format (figure 3.1).

At a different time, the same two experts perform again the diagnosis procedure described above at a $19''$ flat monitor HP L1940, looking at the digital images previously acquired (with the corresponding positive and negative controls). Monitor settings are 1024x1280 pixels and refresh rate of 60 Hz.

At the monitor both the physicians examined the same regions of the well, whereas at the microscope they may observe the whole well. Motivation of this choice is that at the microscope it is not possible to manually acquire images of the whole well, at the decided resolution, even though we were developing an automated comprehensive system to acquire images of the whole well (see next chapter).

We performed some training session before starting the tests because physicians were not accustomed to look at IIF digital images.

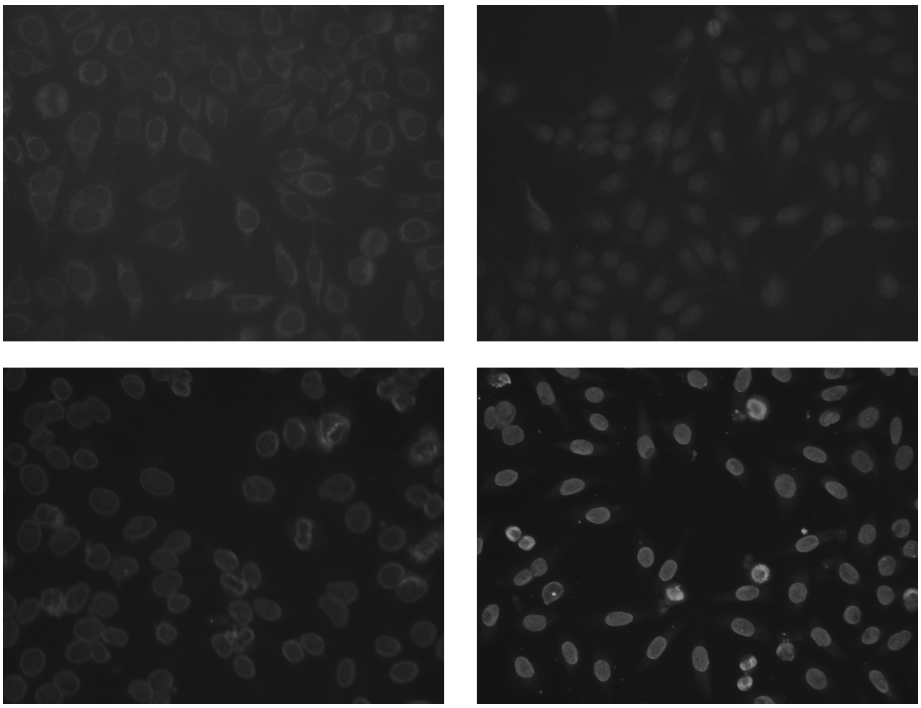

Figure 3.1: Examples of IIF images acquired using the equipment described in section 3.1.1.

3.1.2 Simplified CDC criteria

We decided to use a simplified version of CDC guidelines for fluorescence intensity in order to get a ground truth reliable enough to develop a CAD. On the one hand, this modified version should maintain the diagnostic meaning of IIF test, and, on the other hand, it should allow getting a well-founded data set.

We observed that the disagreement between physicians is twofold (see section 3.2.1). In one case, physicians assign the sample to different classes (i.e. one physician to positive, the other to negative). In the other case, physicians disagree about the subgroups to which a positive sample has to be assigned, i.e. physicians label it with a different number of plus. At a deeper examination, it appears that physicians always agree each other when the sample is marked either with two plus or more, or when it is definitely negative.

According to these observations, we decided to use a simplified classification of data samples into three classes (i.e. negative, positive and *border zone*). A sample is assigned to the negative class if both physicians classify it as negative, whereas it is labelled positive if both physicians mark it with two pluses or more. Finally, a sample is assigned to the border zone class when either of the two types of disagreement described above happens or when both physicians mark it with one plus.

Different from fluorescence intensity scale, the staining pattern classes are defined by well-distinguished features, as reported above. Therefore, we choose to consider only four main classes without modifying them.

3.1.3 Statistical analysis

The agreement between multiple ratings is indicative of the reliability of the single rating. Our agreement analysis regards both the classification of the fluorescence intensity and the description of the staining pattern. Since in all cases our data basically consist of two independent ratings per subject with respect to a dichotomous outcome, we decided to use the degree of agreement between ratings as our main indicator, adopting to this aim the Cohen's *kappa*, which is the most widely used index in the literature among the many non-equivalent proposed [11]. Its estimate, 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 [65]:

- $0 < k < 0.2$ 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.

3.2 Results

In order to validate the use of digital images in IIF diagnosis, we populated a database of IIF images.

We then analyzed the agreement between couple of diagnoses, reporting data on: (i) the agreement at the microscope and at the monitor for each expert and (ii) the agreement between experts for each procedure.

The first kind of data evaluate the reliability of using digital images in place of direct observation at the fluorescence microscope to carry out IIF diagnoses. The second kind of data evaluates advantages or disadvantages of using digital images in place of traditional instrumentation. The agreement analysis regards both the classification of fluorescence intensity and the description of staining pattern.

3.2.1 Fluorescence Intensity Evaluation

Tables 3.2 and 3.3 report the Cohen's kappa and the related confidence interval ($p < 0.05$). For each physician the agreement between traditional and digital supported diagnosis is different (Table 3.2). The computed Cohen's kappa suggests substantial agreement for Physician1 ($k = 0.66 \pm 0.12$) and almost substantial agreement for Physician2 ($k = 0.56 \pm 0.14$).

The agreement between pairs of experts remarkably depends on the method used to carry out the diagnosis. The measured kappa is almost 30% more at the monitor than at the microscope. Indeed, the measured kappa implies moderate agreement at the microscope ($k = 0.46 \pm 0.13$) and substantial agreement at the monitor ($k = 0.65 \pm 0.12$).

If we classify the fluorescence intensity in three classes instead of five, the measured kappa rises (Table 3.3). Specifically, with respect to the fluorescence

Table 3.2: Agreement on fluorescence intensity classification into 5 classes.

	kappa (k)	Confidence Interval
Physician1 looking at the microscope and at the monitor	0.66	0.12
Physician2 looking at the microscope and at the monitor	0.56	0.13
Between experts at the microscope	0.46	0.13
Between experts at the monitor	0.65	0.12

intensity classification in five classes the kappa increases of 15.85% and of 13.87% for Physician1 and Physician2, respectively. Now, the computed Cohen's kappa suggests almost perfect agreement for Physician1 ($k = 0.78 \pm 0.11$) and substantial agreement for Physician2 ($k = 0.66 \pm 0.13$).

Table 3.3: Agreement on fluorescence intensity classification into 3 classes, i.e. the simplified CDC criteria.

	kappa (k)	Confidence Interval
Physician1 looking at the microscope and at the monitor	0.78	0.11
Physician2 looking at the microscope and at the monitor	0.66	0.13
Between experts at the microscope	0.62	0.13
Between experts at the monitor	0.84	0.09

The measured kappa is nearly 26% more at the monitor than at the microscope, implying substantial agreement at the microscope ($k = 0.62 \pm 0.13$) and perfect agreement at the monitor ($k = 0.84 \pm 0.09$). Hence, with respect to the intensity classification in five classes, the kappa increases of 25.38% and of 22.60% working at the microscope and at the workstation monitor, respectively.

The observations that give rise to the simplified CDC criteria (section 3.1.2) can be deeper understood looking at figures 3.2 and 3.3. They report the percentage of agreement between the two physicians when they classify the fluorescence intensity of the samples into five subgroups, working at the microscope and at the workstation monitor, respectively.

The bigger the agreement between physicians' classification for each class, the brighter the grey level of the corresponding box in the figure. To better comprehend this symbolic representation, the grey levels are computed mapping the biggest agreement percentage to the biggest grey level of the image,

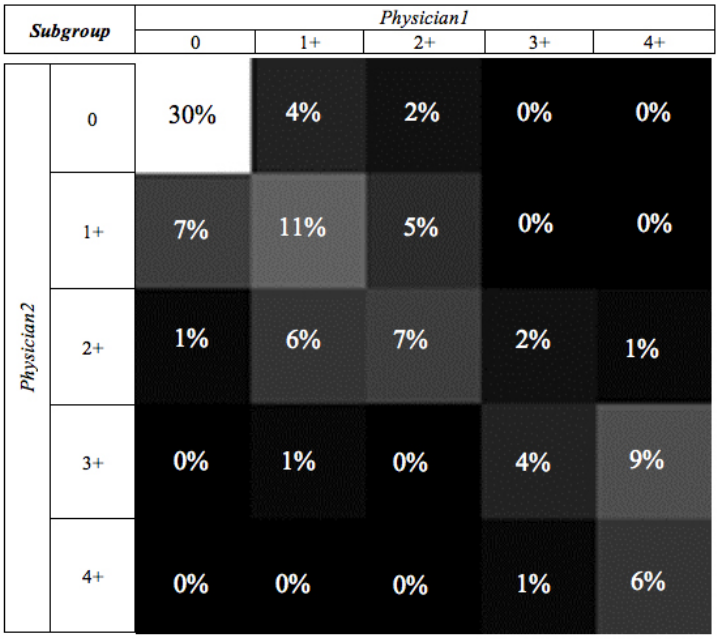

Figure 3.2: Grey level map of the agreement between physicians when they classify the samples into five subgroups at the fluorescence microscope.

i.e. the white. The other agreement percentages are mapped to a grey value proportional to their value. For the sake of comprehension, such percentages are reported upon the corresponding box. On the one hand, when the diagnosis is carried out at the fluorescence microscope, both physicians agree each other in the 58% of cases, i.e. the sum of the main diagonal in figure. Such a low agreement is obviously related to the low kappa value obtained labelling the sample into five subgroups ($k = 0.46 \pm 0.13$). On the other hand, when the diagnosis is performed at the workstation monitor, the agreement between the specialists rises up to 73%, in accordance with the corresponding increase of the kappa value ($k = 0.65 \pm 0.12$). Figure 3.4 represents the agreement between physicians' classification when they label the samples into the three classes at the fluorescence microscope. Now, at the fluorescence microscope 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 ± 0.13, implying substantial agreement, which is considered satisfactory to get a reliable ground truth. For the sake of completeness, figure 3.5 shows that the specialists agree in the 89% of cases working at the workstation monitor. Indeed in such a case the kappa is 0.84 ± 0.09.

3.2.2 Staining Pattern Evaluation

The observed frequencies of homogeneous, rim, speckled, nucleolar and unclassifiable pattern class were 21%, 1%, 29%, 1% and 48%, respectively. These data were computed averaging each class rate over the four readings on the same sample.

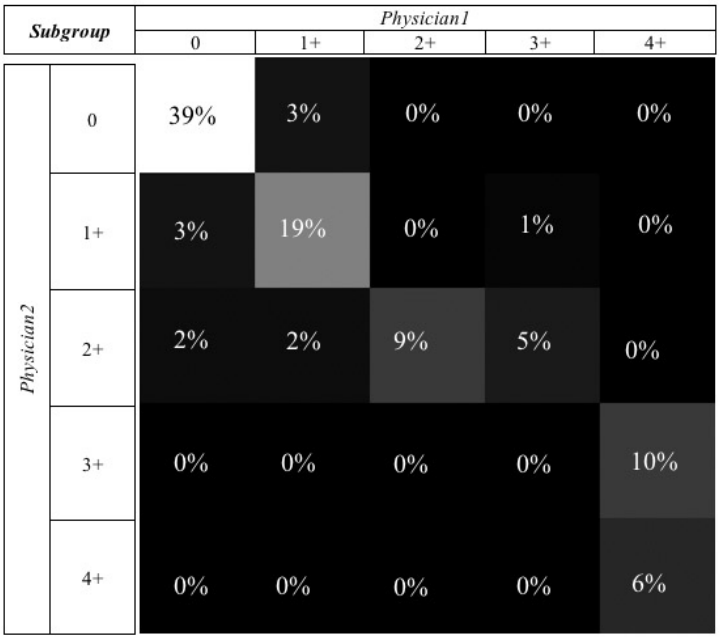

Figure 3.3: Grey level map of the agreement between physicians when they classify the samples into five subgroups at the workstation monitor.

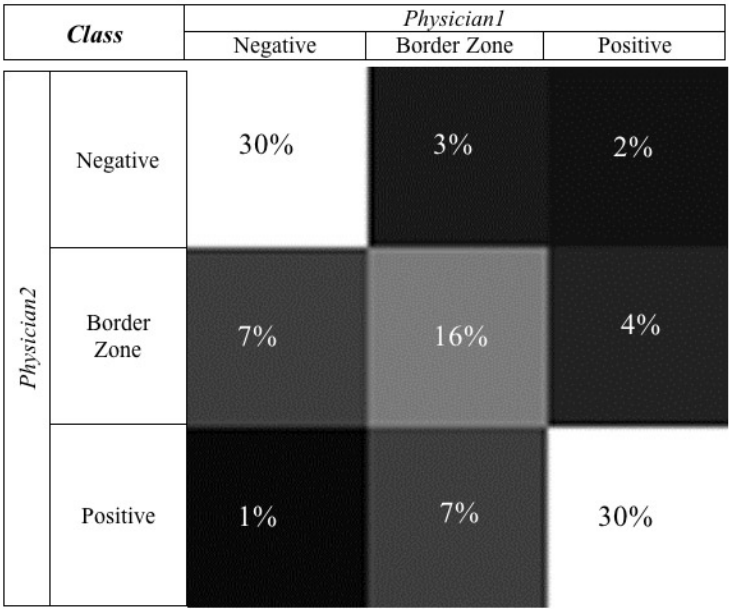

Figure 3.4: Grey level map of the agreement between physicians when they classify the samples into three classes at the fluorescence microscope.

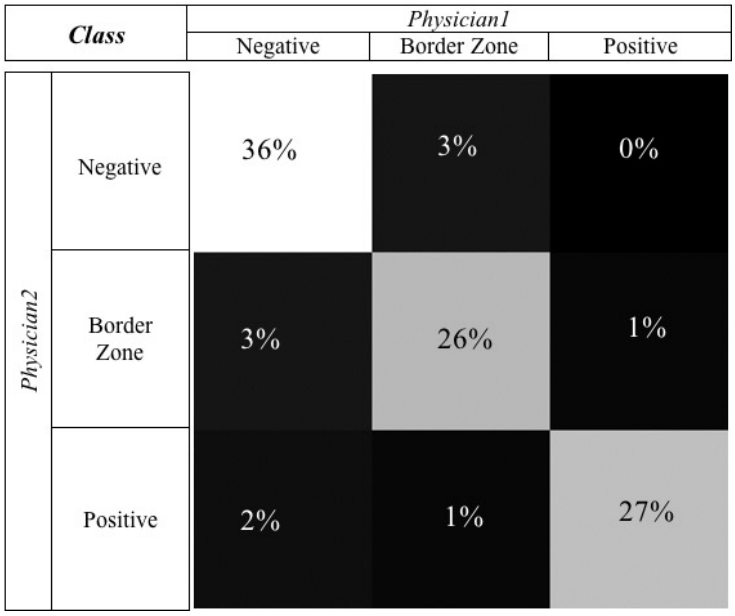

Figure 3.5: Grey level map of the agreement between physicians when they classify the samples into three classes at the workstation monitor.

An analysis similar to the intensity one has been performed for the staining pattern description (Table 3.4). Once more, the agreement between traditional and digital supported diagnosis is different for each physician, suggesting substantial and moderate agreement ($k = 0.66 \pm 0.13$ and $k = 0.56 \pm 0.14$), respectively.

Again, the agreement between pairs of experts remarkably depends on the method used to carry out the diagnosis. The measured kappa is nearly 10% more at the monitor than at the microscope. Furthermore, the Cohen's kappa implies substantial agreement both at the microscope ($k = 0.61 \pm 0.13$) and at the monitor ($k = 0.68 \pm 0.12$), respectively.

Table 3.4: Agreement on staining pattern classification.

	kappa (k)	Confidence Interval
Physician1 looking at the microscope and at the monitor	0.66	0.13
Physician2 looking at the microscope and at the monitor	0.56	0.14
Between experts at the microscope	0.61	0.13
Between experts at the monitor	0.68	0.12

3.3 Discussion

With regard to the fluorescence intensity evaluation for each physician, the measured kappa implies moderate and substantial agreement at the microscope and at the monitor, respectively (Table 3.2). At a further analysis, nearly the 85% of disagreements for both physicians occurs on samples that exhibit fluorescence intensity from classes from 0 up to 2+. The samples of such classes are intrinsically hard to classify, because they are on the borderline between positive and negative classes.

It is worth noting that the two diagnoses performed by each physician on the same sample should be considered independent one each other. These observations suggest that diagnosing fluorescence intensity of the sample using digital media is at least as much reliable as the classification carried out in the traditional way (i.e. the fluorescence microscope).

The measured kappa on the agreement between pairs of physicians shows moderate and substantial agreement at the microscope and at the monitor, respectively. These data suggest that performing the diagnosis by looking at digital images on a workstation screen allows the physicians to better concentrate on sample examination. Indeed digital images on the one hand avoid the photobleaching effect and, on the other hand, allow for a more careful intensity evaluation, by easily comparing the sample under examination with negative and positive controls.

Such an agreement, however, may not be satisfactory enough to develop a reliable CAD. In this respect, it is important to reliably label a data set with its true category. In the field of pattern recognition, a labelled data set is named ground truth. In IIF application, the ground truth is made by labelled images with both fluorescence intensity and staining pattern classification. IIF is a subjective test and no objective independent test could be used to assess the human expert diagnosis. Based on these observations we utilize two different and independent diagnoses for each sample to get ground truth. Clearly its reliability depends on the degree of agreement between physicians.

To overcome such a limitation, the original fluorescence intensity classification problem on five classes is simplified to a classification problem on three classes, as described in section 3.1.2. While the motivation for this class revision is the ability to get a better ground truth, it is worth noting that these three classes maintain the clinical significance of the test (i.e. positive, negative and border zone test). To evaluate the effectiveness of this simplified classification into three classes, we adopt the previous approach. Obviously, the agreement between the diagnoses of the same experts and between pairs of experts depends on the number of categories in which they should classify the fluorescence intensity of the examined sample. Indeed, using the classification in three classes, the Cohen's kappa suggests almost perfect and substantial agreement for Physician1 and for Physician2, respectively. On the other hand, with respect to the classification into five classes, the agreement at the microscope and at the monitor rise up from moderate and substantial to substantial and perfect, respectively. These data suggest that a classification in three classes leads to a more reliable ground truth, useful to develop a CAD supporting the classification of fluorescence intensity.

Concerning the staining pattern evaluation, the measured kappa for each physician suggests substantial and moderate agreement, respectively. Since the

diagnosis of the expert concerns both the fluorescence intensity and the staining pattern, the difference between experts is the same measured as the fluorescence intensity. Likewise to the grading of the fluorescence intensity, variability in the staining pattern classification is reported in the literature [16, 82]. Our data shows a reduction in this variability: it suggests that classifying the pattern using digital images is more effective than looking at the microscope.

The disagreements clustered both round the choices speckled vs. homogeneous and speckled vs. unclassifiable pattern; in all cases such disagreements ranged from 3% to 8% over the total number of sera. It is worth noting that most of them occur on samples whose fluorescence intensity was weakly positive, as usual in daily clinical practice.

Analysing the agreement between pairs of experts, the Cohen's kappa implies substantial agreement both at the microscope and at the monitor. Disagreements between experts mainly occur when the sample exhibits multiple patterns, or when one of the two experts reports an unclassifiable pattern. In this respect its worth noting that CAD systems have the potential for distinguishing not only the predominant pattern but also the minor one.

Different from fluorescence intensity grading, that should be considered a *continuous variable*, the classes of staining pattern are defined by well-distinguished features, or *well-separated* using a term of the pattern recognition field.

Our results suggest that observing the pattern by looking at digital images on a workstation monitor allows the physician to better concentrate on sample examination, e.g. to observe carefully fine details without take care of photobleaching effects.

As a final consideration, the physicians were initially not accustomed to diagnose the sample using the workstation monitor, while they were well skilled in carrying out the diagnosis at the microscope. Potentially, the results on digital image classification could remarkably improve as the expert "feeling" with this kind of diagnostic procedure increases.

Chapter 4

IIF Image Acquisition

In the previous chapter we have proven that the use of digital images on a workstation monitor allows getting a more reliable diagnosis, in the case of fluorescence intensity classification and staining pattern description. Furthermore, the intrinsic interobserver variability of IIF readings limits the reproducibility of the method and demands the development of a system to support the physician decision. In this respect, the ability to automatically and reliably acquire IIF images seems a basic milestone.

In this chapter we report our experience in the development of a system for automatic acquisition of IIF images, focusing our attention on the determination of an effective autofocus function. The difficulty here is the inadequacy of conventional autofocus functions to cope with photobleaching, a physical phenomenon affecting automatic acquisition of IIF images (see section 2.2.3). We therefore propose an autofocus procedure that overcomes the limitations of existing methods, and present a set of experiments on real images that confirm its effectiveness.

The chapter is organized as follows. In section 4.1 we discuss the limitations of existing autofocus functions and present our autofocus procedure that succeeds in dealing with the peculiarities of the IIF images acquisition process. In section 4.2 we discuss the performance of the proposed procedure and report experimental data confirming its effectiveness.

4.1 The Autofocus Algorithm

Focus algorithms proposed in literature [6, 29, 32, 35, 87, 110] are based on a criterion function applied to images of the same sample, acquired at different z axis positions. The autofocus functions give a value that indicates the degree of focusing of each image. The function maximum should correspond to the optimum focus position. In a typical focus function three different regions can be distinguished [6]: a near flat region, a sloped region and a quadratic region, which lies just around the function peak. The algorithm is therefore organized according three subsequent phases named *coarse*, *fine* and *refine*, respectively.

Since, for practical reasons, the autofocus procedure is usually started from a position moderately far from the focus, a robust autofocus function should show an evident, although limited, slope in regions far from the focus, and a

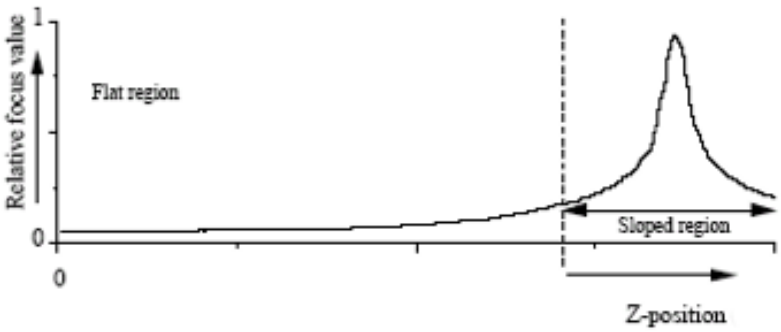

Figure 4.1: Different regions which can be distinguished in a focus function, near-flat and sloped. The figure has been taken from [6].

sharp peak at the focus position (figure 4.1).

Many functions have been proposed in the literature [6, 29, 32, 35, 87, 110]. Based on the assumption that unfocused images usually have slight differences between dark and bright objects [87], after some preliminary tests, we selected a focus function proposed in the literature based on image differentiation [6], and defined by the formula:

$$F_{\{1,-1\}} = \sum_x \sum_y (i[x,y] - i[x-1,y])^2 \tag{4.1}$$

with $i(x,y)$ the grey level of pixel (x,y). In the following it is referred to as $\{1,-1\}$ filter.

This filter, as well as most of the other filters proposed in the literature, assumes that the image roughly maintains its basic properties throughout the whole autofocus process. In particular, it assumes that the integral image intensity (i.e. the sum of all pixels) is constant. Unfortunately, integral intensity may change in fluorescence applications due to the photobleaching effect. It is a physical phenomenon that chemically alters the fluorescent molecules to a permanent non fluorescent state when they are exposed to a fluorescent lighting. Due to photobleaching, the total amount of fluorescent light from free fluorescent molecules shows an exponential decay in time [21, 96, 97]. Biological tissue stained with fluorescent dyes may bleach significantly in a few seconds. Hence, changes may be relevant if long exposure time of the slide to excitation light, or high lighting power are needed, as it happens in IIF applications. Figure 4.2 shows the decay of the image integral intensity when multiple images of the same slide are taken.

Since multiple images at regular steps are taken until the focus position is adjusted moving mechanically toward the slide, integral intensity changes when acquiring IIF images.

For this reason, the $\{1,-1\}$ filter does not perform well in the acquisition of IIF images in all the three phases mentioned above. Indeed, the local maximum corresponding to the focus position is not a global maximum (figures 4.3 and 4.4), which makes the location of focus position ineffective. Analogous behaviour has been observed also for other similar functions reported in the literature [29, 35, 87].

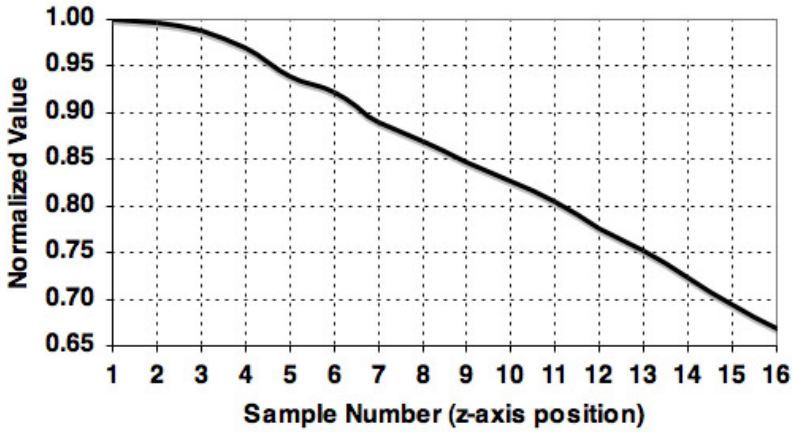

Figure 4.2: Image Intensity Integral. It is the sum of all pixel values in an image. Multiple images of the same area are taken at regular z axis position.

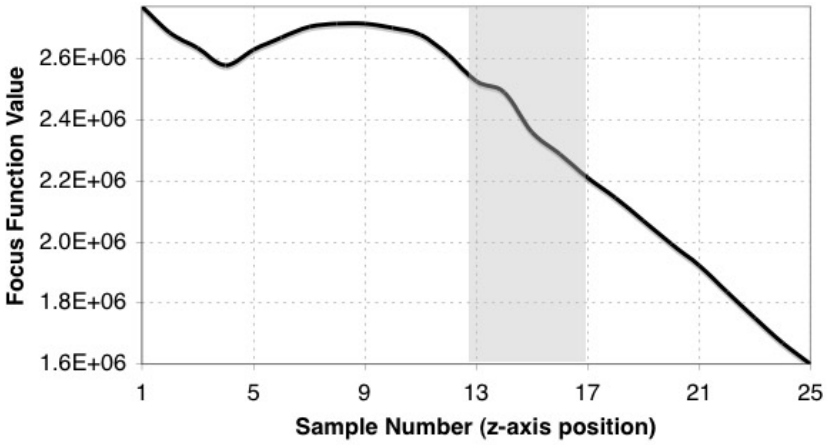

Figure 4.3: Shape of the $\{1, -1\}$ filter in the coarse phase. The shadowed area refers to the real focus position.

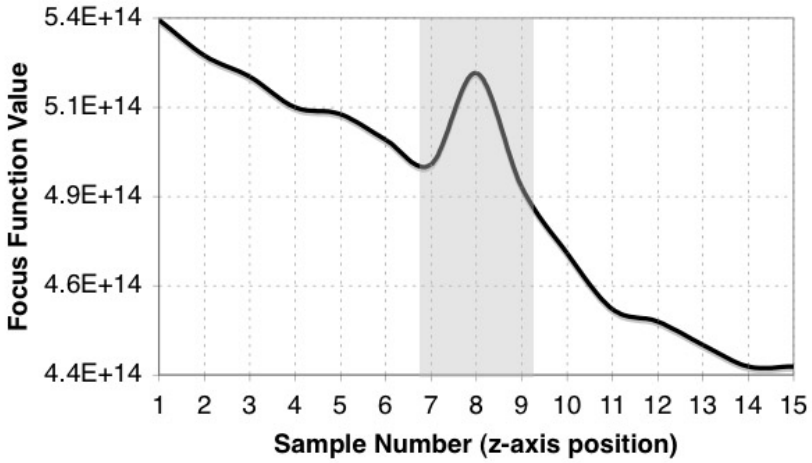

Figure 4.4: Shape of the $\{1, -1\}$ filter in the fine phase. The shadowed area refers to the real focus position.

To overcome such limitations, we decided both to compensate the photobleaching effect and vary the type of functions used in the three phases of the autofocus algorithm.

4.1.1 Photobleaching Compensation

The $\{1, -1\}$ filter is a high pass filter whose value is proportional to the energy of the image. If photobleaching is not negligible, the energy decreases during the autofocus procedure, making the filter ineffective. We then decided to compensate the fluorescence decay by normalizing the $\{1, -1\}$ filter with respect to the energy of the first image acquired. The advantage of this choice is that no direct estimation of fluorescence decay is needed. In the following, we will refer to this modified filter as the $\{1, -1\}$ *compensated filter*; it is defined by the formula:

$$F_{\{1,-1\}Cf} = (\frac{G_{FirstImage}}{G})^2 \cdot F_{\{1,-1\}} \tag{4.2}$$

where G is the integral image intensity.

This modified filter exhibits a very good performance. Figure 4.5 compares the shape of the original and modified filters. Since two of the properties of a useful focus functions are unimodality and reproducibility (i.e. only one maximum and a sharp top of the focus function, respectively) [35], it is apparent that the effects of photobleaching are well compensated by the modified filter.

4.1.2 Focus Function in the Coarse Phase

Figure 4.6 points out that when the $\{1, -1\}$ compensated filter is applied in the first phase, it does not succeed in clearly identifying the focus region. Indeed, when the z axis position of the image is far from the focus position, edges and borders are not observable, which makes the localization of the focus region difficult.

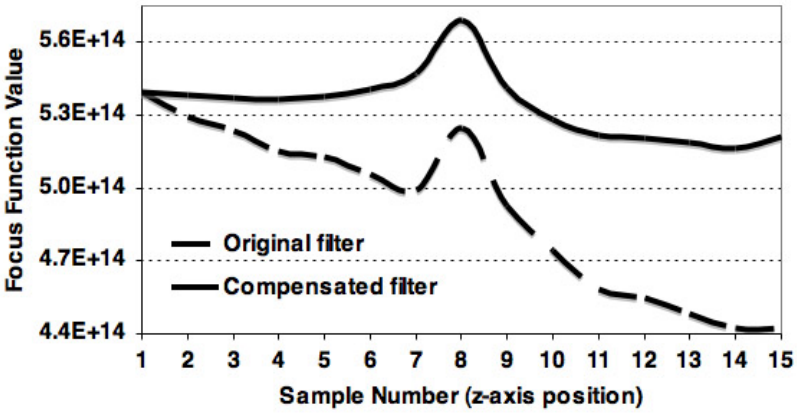

Figure 4.5: Shape of the $\{1, -1\}$ focus functions.

Based on the assumption that the range of grey levels of the image increases as the sample come into focus [29], we tried a focus function based on histogram range and found that in the coarse phase this function is characterized by a fairly sharp peak around the interval containing the focus position (figure 4.7). Denoted by H_i the number of pixels with intensity i, we used the following function:

$$F_{range} = max(i|H_i > 0) - min(i|H_i > 0) \tag{4.3}$$

It is worth noting that this histogram based function has been already considered in the literature, but it exhibits poor performance as a focus function in other applications [87].

4.1.3 Complete Procedure Definition

The last step in defining a complete autofocus procedure for our application was the determination of the steps to use in each of the three phases while moving along the z axis. Both to improve the overall efficiency of the procedure and to minimize the required number of images, after experimentation, we modified a little bit the three phase algorithm proposed in [6]. Whereas there $100\mu m$ is considered a wide range of z positions, robustness considerations related to the nature of our acquisition system motivated us to choose $500\mu m$ as the range where moving along the z axis. In order to cope with this wider range and reduce the number of acquired images, in the coarse phase z axis is stepped by two different steps of size $60\mu m$ and $20\mu m$. In the fine and refine phase, step sizes of $10\mu m$ and $1\mu m$, respectively, are chosen.

It is worth noting that the smallest step size of the complete procedure, i.e. $1\mu m$ applied in the refine phase, has been defined by determining the *depth-of-focus*. It is the distance Δz over which the image specimen can be expected to be observed without significant optical aberration. It can be derived from

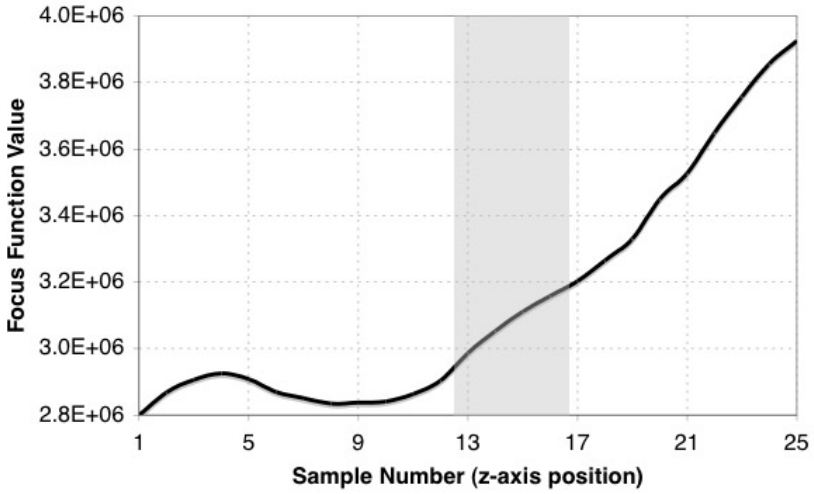

Figure 4.6: Shape of the $\{1, -1\}$ compensated filter in the coarse phase. The shadowed area refers to the real focus position.

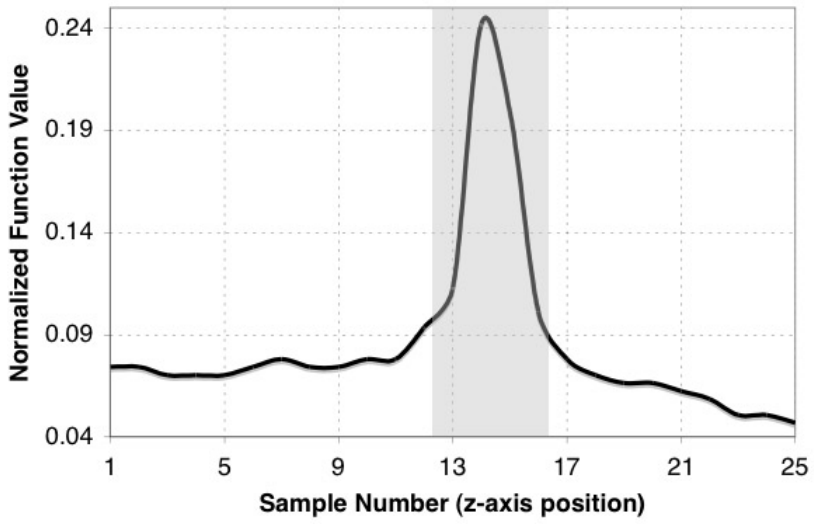

Figure 4.7: Shape of the histogram range focus function (coarse phase). The shadowed area refers to the real focus position.

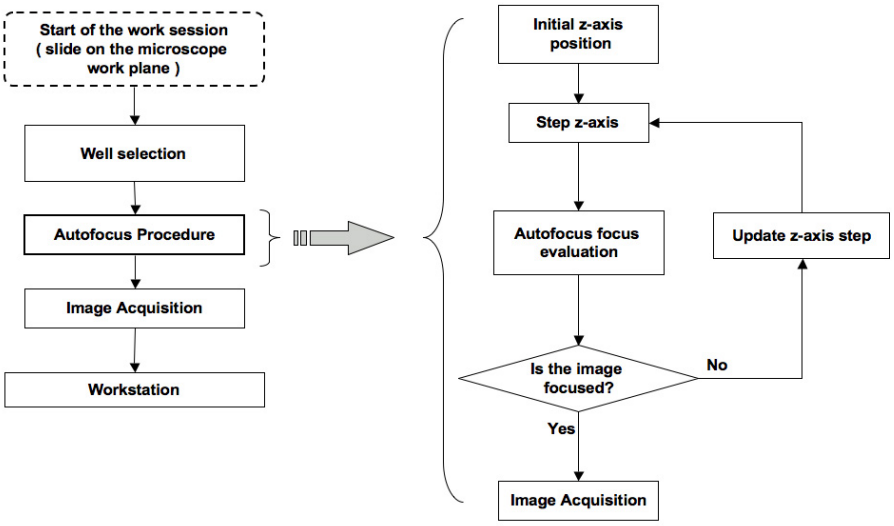

Figure 4.8: Flow chart of an acquisition session. The panel on the right details the autofocus procedure.

considerations of wave optics [84, 116, 117] and it is expressed as follows:

$$\Delta z = \frac{\lambda}{4\pi\left(1-\sqrt{1-\left(\frac{NA}{n}\right)^2}\right)} \tag{4.4}$$

where λ, NA and n are the wavelength of light, the numerical aperture of the lens and the index of refraction of the medium between the lens and the specimen, respectively. The numerical aperture measures the ability of the lens to collect light and is given by $n \cdot \sin(\theta)$, with θ the angle of acceptance of the microscope lens. In our tests $n = 1$ since the medium is the air, $\lambda = 500nm$ and $NA = 0.65$ for a 40x objective. Therefore, the application of equation 4.4 gives $\Delta z = 0.521\mu m$, thus confirming the usefulness of the refine phase step.

4.2 Performance Evaluation

To evaluate the effectiveness of the autofocus procedure described above, we manually applied the algorithm to acquire the images of 15 real IIF wells with the instruments described in section 3.1.1. In short, it is constituted by a fluorescence microscope equipped with a 40x objective, a monochrome CCD camera and a mercury vapour lamp. Figure 4.8 depicts the flow chart of an acquisition session. It starts when the user puts the slide on the microscope work plane and select the well to be examined. Starting from a default position, the z axis is stepped according to the procedure reported in section 4.1.3 until the focus position is determined. Then, the focused image is returned to the user.

To evaluate the algorithm performance we carry out both a quantitative and a qualitative evaluation.

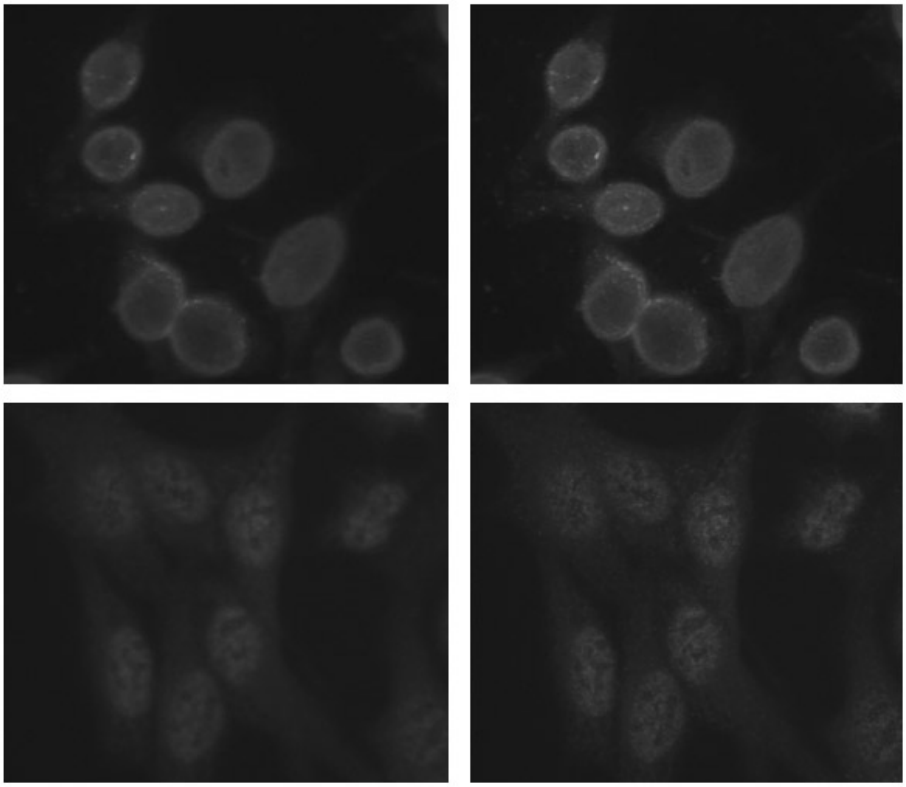

Figure 4.9: Examples of images used in the evaluation of autofocus algorithm performance. On the left there are manually focused images, whereas on the right there are images focused following our autofocus algorithm.

The quantitative evaluation concerns the total acquisition time of each image. In all tests we did, no more than 15 steps were needed to locate the focus position. Since each image requires 0.4 s of exposure time with our CCD camera, this means that about 6 s are needed to acquire the final image. Note that, at each step, the execution time of the autofocus algorithms can be largely overlapped to the next step, and in any case it is negligible with respect to the exposure time.

The qualitative evaluation concerns the subjective quality of the images obtained applying the proposed autofocus procedure. In order to carry out this evaluation, 31 testers have been selected and asked to look at 15 pairs of images on a computer monitor. Each pair consists of a manually focused image and an image acquired following our autofocus algorithm. In the following, the two images are referred to as *man-image* and *auto-image*, respectively. Figure 4.9 shows two examples of acquired images representative of the whole test set: man-images are reported on the left, with the corresponding auto-images on the right. Images on the top are characterized by the presence of circular cells with homogeneous background, whereas, images on the bottom present cell bodies with multiple fluorescent fine dots and cytoplasmic structures in the background. It is apparent that borders and details are much better defined in

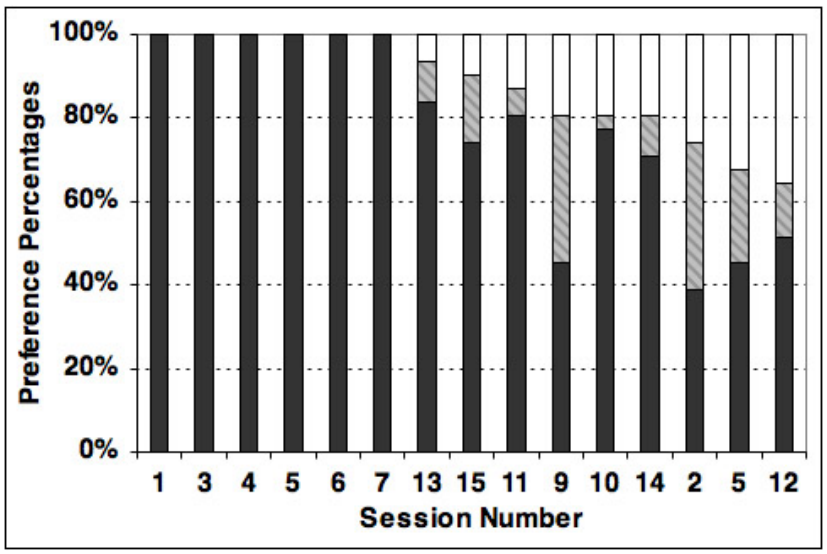

Figure 4.10: Percentages of preference for each pair of images: images focused with the proposed algorithm (black), images reported as equivalent (dashed), images manually focused (white).

auto-images.

For every pair of images, each physician selects on a subjective basis the best-focused one or, alternatively, reports that the images are equivalent. Using Student's t-statistic ($p < 0.05$), in mean each tester prefers auto-images in the $77.8\% \pm 4.6\%$ of cases, prefers the man-images in the $12.0\% \pm 3.6\%$, and reports the images as equivalent in the $10.3\% \pm 3.2\%$ of cases.

Preference percentages relative to the 15 pairs are reported in figure 4.10. We point out several issues: (i) the majority of testers never prefers images manually focused, (ii) in the 40% of sessions all testers prefer the image automatically focused, (iii) in the 80% of sessions more than half of testers prefer images automatically focused.

If we sum preferences of auto-images and equivalence between images, in the 80% of sessions the preferences for manually focused images is less than 20% and in no case such a percentage is larger than 40%.

Furthermore, there are two sessions (number 2 and 9) in which nearly one third of testers reports that the two images are equivalent. Looking carefully at these images we found out that in session 2 both the images are very well focused, whereas in session 9 both the images are poorly focused, probably because the well has more than one focus plane.

To deeper understand these observations, for each pair we computed the difference between the focus position of the two images (figure 4.11). Comparing preferences with differences in focus position, we notice that when preference is given to auto-image by all testers, there is actually a non-negligible difference in focus position between images. This seems an indication that the autofocus algorithm is effective in locating the right focus position in these cases. When preferences for man-images are more than 30% (sessions 5 and 12) there is again a non-negligible difference in focus position. Since also in this case a majority

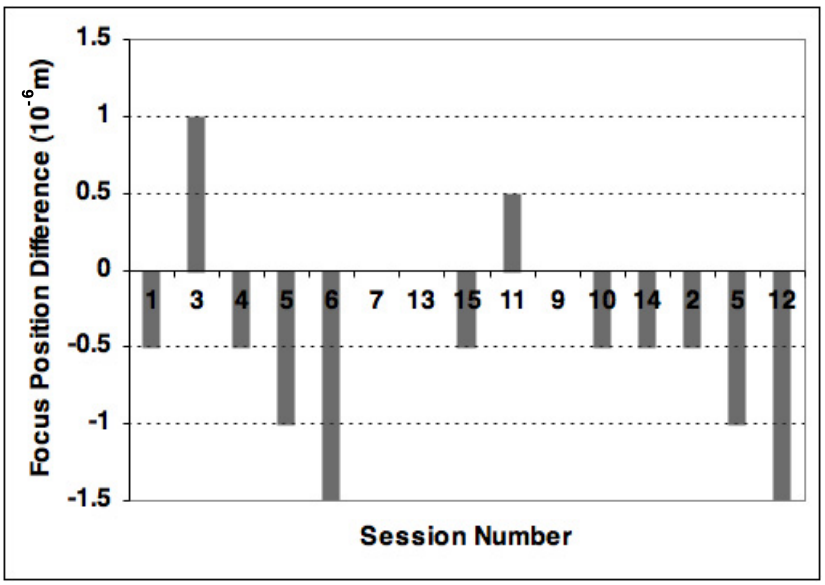

Figure 4.11: Differences between focus position computed by the algorithm and position manually chosen.

of the testers chose the auto-image, we argue that the problem does not concern the algorithm, but rather the presence of multiple focus planes in the wells.

From these observations, we deem that the proposed autofocus procedure seems effective enough for automatic acquisition of IIF images. We also note that such an algorithm is likely to substantially improve the overall acquisition process, since it provides a reliable mean to locate the focus position, irrespectively of phenomenon like photobleaching, which may seriously limit the precision of a human operator working at the microscope.

Chapter 5

IIF Images Classification

A typical CAD system is constituted by several blocks that control data acquisition and storage, interact with users and support the diagnosis 2.4. To this end, in the previous chapters we first focused on the assessment of the using digital images in place of direct observation at the fluorescence microscope to carry out the diagnosis and, second, we studied an image acquisition procedure. With reference to the aim of this work, the present chapter discusses the classification of both the fluorescence intensity and the staining pattern of IIF wells.

Figure 5.1 shows the flow-chart employed to classify each input sample. The recognition approach is based on a cascade of two steps: the first classifies the fluorescence intensity, whereas the second recognizes the staining pattern of positive wells, as the guidelines recommend (section 2.2.1). These two classification stages, although based on the same classification architecture, are discussed separately in the following.

5.1 Fluorescence Intensity Classification

In this section we propose a system devised to classify the fluorescence intensity according to the simplified CDC criteria 3.1.2. The recognition system belongs to the field of supervised statistical pattern recognition. The term supervised classification means that the input pattern is identified as a member of a predefined class defined by the system designer[1] [106]. In such a case, each input pattern is represented in terms of d features or measurements, and is viewed as a point in a d-dimensional space. The goal is to choose those features that allow pattern vectors belonging to different categories to occupy compact and disjoint regions in that feature space. Therefore, the choice of a suitable set of features is crucial for the performance of the classification system since the effectiveness of the representation space (feature set) is determined by how well patterns from different classes can be separated. The objective is to establish decision boundaries in the feature space that separate patterns belonging to different classes. In the statistical decision theoretic approach, the decision boundaries are determined by the probability distributions of the patterns belonging to each class,

[1]The other classification approach is named as unsupervised classification, in which the pattern is assigned to an unknown class. In this case, the classes are learned based on the similarity of patterns.

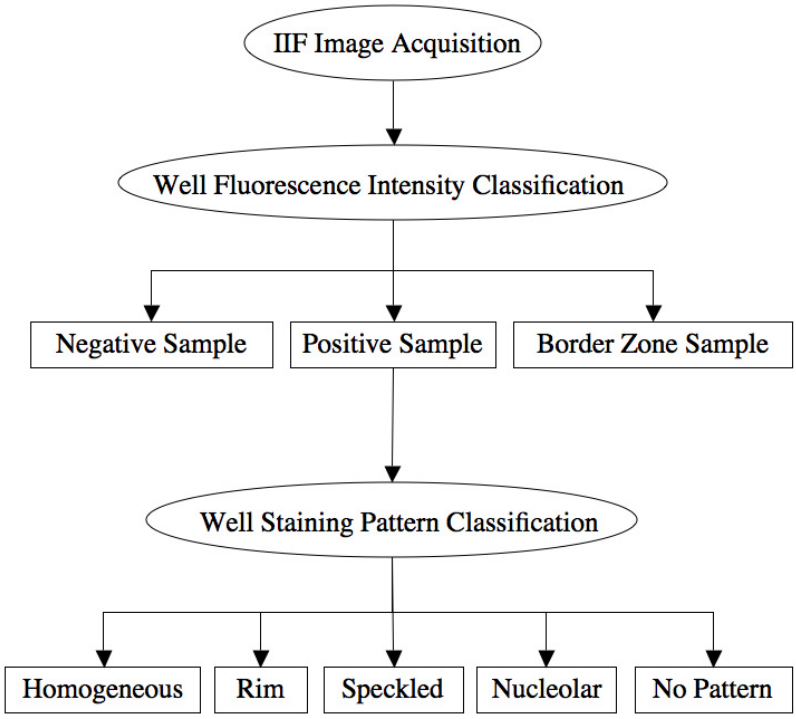

Figure 5.1: Flow-chart of the classification procedure. The approach is based on the cascade of two systems: the first classify the fluorescence intensity, whereas the second recognizes the staining pattern of positive wells.

which must either be specified or learned [22, 25, 49].

The background, the motivations and the literature on fluorescence intensity classification have been presented in chapter 2. In particular, with respect to [92, 93] 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 scenario.

Next section reports the features extraction and selection, whereas from section 5.1.2 up to section 5.1.5 we describe the architecture of the classification system. Section 5.1.6 reports the experimental results.

5.1.1 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 grey-level histograms. The rationale lies in the meaning of these histograms: the former describes grey-level distributions, whereas the latter generally provides a good representation of the overall nature of the texture. For the definition of these features see [23]. Preliminary tests were per-

Table 5.1: Combination mode of the feature selection procedure. F_x^i represents the value of ith feature of sample x, $F^i_{pos\ ctrl\ of\ x}$ and $F^i_{neg\ ctrl\ of\ x}$ represents the value of ith feature of positive and negative control of x, respectively.

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

formed on this set. Specifically, such a features set was extracted from both the segmented images, i.e. the features were computed on the cells' area only, and the whole image. For the description of the segmentation procedure see section 5.2.2. 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;
- artefacts possibly due to the limitations of the segmentation algorithm 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 presented in section 2.2.1. The combinations use both linear and non-linear strategies. Specifically, in Table 5.1, the first row corresponds to the value of ith feature of sample x (denoted as 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 per each mode reported in Table 5.1.

Table 5.2: Selected features for each expert. H_1 and H_2 represent the first and second order grey-level histogram, respectively. *Skewness* and *kurtosis* are the third and fourth moment of the histogram, respectively. *Inverse* stands for the inverse difference moment, i.e. a measure of local homogeneity. For the description of features mode combination (#) see Table 5.1.

Positive Expert (PE)	Border Zone Expert (BZM)	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

Discriminant analysis shows that all the above features have limited discriminant strength over three classes (i.e. positive, border zone 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 [50].

These observations suggest adopting an aggregation of experts rather than a single one. This resulting system is therefore based on three modules, each one specialized in recognizing one of the three input classes. Specifically, the three experts are:

- Positive Module (PM): module specialized on the classification of positive sample;
- Negative Module (NM): module specialized on the classification of negative sample;
- Border Zone Module (BZM): module specialized on the classification of sample belonging to the border zone 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 5.2.

5.1.2 System Architecture

Many supervised pattern recognition tasks can be cast as the problem of assigning elements to a finite set of classes or categories. Such tasks are referred to as binary learning, or dichotomies, when they aim at distinguishing instances

of two classes, whereas they are named multiclass learning, or polychotomies, if there are more categories.

There is a huge number of applications that require multiclass categorization. Some examples are text classification, object recognition and support to medical diagnosis, to name a few.

In the literature numerous learning algorithms have been devised for multiclass problems, such as neural networks or decision trees. However it exists a different approach that is based on the reduction of the multiclass task into multiple binary problems, referred to as *decomposition method*. The problem complexity is therefore reduced trough the decomposition of the polychotomy in less complex subtasks. The basic observation that supports such an approach is that in the literature most of the available algorithms, which handle classification problems, are best suited to learning binary function [24, 73]. Different dichotomizers, i.e. the discriminating functions that subdivide the input patterns in two separated classes, perform the corresponding recognition task. To provide the final classification, their outputs are combined according to a given rule, usually referred to as *aggregation* or *reconstruction rule*.

In the framework of decomposition methods for classification, the various methods proposed to-date can be traced back to the following three categories [1, 15, 24, 38, 51, 62, 72, 73].

The first one, called *one-per-class*, is based on a pool of binary learning functions, where each one separates a single class from all the others. The assignment of a new input to a certain class can be performed, for example, looking at the function that returns the highest activation [24, 72].

The second approach, commonly referred to as *distribuited output code*, assigns a unique codeword, i.e. a binary string, to each class. If we assume that the string has n bit, the recognition system is composed of n binary classification functions. Given an unknown pattern, the classifiers provide a n-bit string that is compared with the codeword to set the final decision. For example, the input sample is assigned to the class with the closest codeword, according to a distance measure, such as the Hamming one. In this framework, in [24] the authors proposed an approach, known as *error-correcting techniques* (ECOC), where they employed error-correcting codes as a distributed output representation. Their strategy was a decomposition method based on the coding theory that allowed obtaining a recognition system less sensitive to noise via the implementation of an error-recovering capability. Although the traditional measure of diversity between the codewords and the outputs of dichotomizers is the Hamming distance, other works proposed different measures. For example, Kuncheva in [62] presented a measure that accounted for the overall diversity in the ensemble of binary classifiers.

The last approach is called n^2 classifier. In this case the recognition system is composed of $(n^2 - n)/2$ base dichotomizers, where each one is specialized in discriminating respective pair of decision classes. Then, their predictions are aggregated to a final decision using a voting criterion. For example, in [51] the authors proposed a voting scheme adjusted by the credibilities of the base classifiers, which were calculated during the learning phase of the classification.

This short description of the methods so far shows that the recognition systems based on decomposition methods are constituted by an ensemble of binary discriminating functions. On this motivation, for brevity such systems are referred to as Multy Dichotomies System (MDS) in the following.

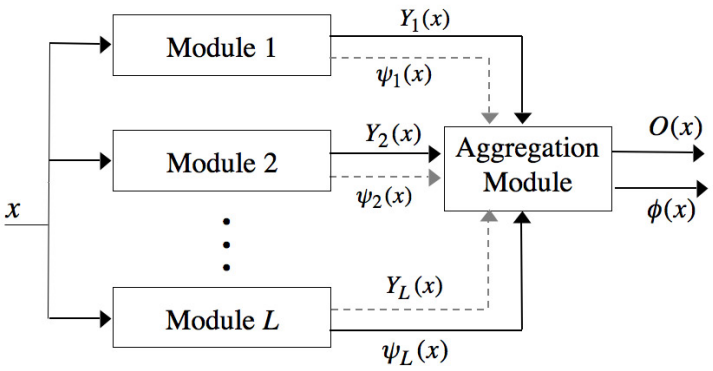

Figure 5.2: The system architecture. To obtain the decision $O(x)$ on the current input pattern x, the decisions $Y_1(x), Y_2(x), \cdots, Y_L(x)$ of the component modules are aggregated according to a rule which can take or not into account the reliability parameters $\psi_1(x), \psi_2(x), \cdots, \psi_L(x)$. Moreover, the overall system can provide a reliability measure $\phi(x)$ of the final decision.

Based on the results of the feature selection phase (section 5.1.1), we adopt here a MDS operating in the *one-per-class* paradigm. Indeed, the set of stable and effective features obtained for each class enforced the evidence that the classification could be reliably faced by introducing one specialized module per each class that the system should recognize. Given the number L of classes among which the input samples are distributed, the MDS is composed of L modules, each one being an expert in the separation of one input class from the others. Their predictions are aggregated to a final classification decision on the basis of a given rule. Figure 5.2 depicts the general architecture of the system, where the individual decisions are given to an *aggregation module* which identifies the module that is the most likely to be correct for any input sample.

The rules employed in the aggregation block evaluate which module is the most likely to be correct for any given sample. Since each module has a binary output, possible input combinations to the aggregation module can be grouped into three categories: (i) those for which only one module classifies the sample in its class, (ii) those for which more modules classify the sample in its own class, (iii) those for which none module classifies the sample in its class.

To deal with these cases, in the following we introduce two different strategies: the first, referred to as Binary-Aggregation (BA), is a conservative criterion that rejects the input samples when there is an ambiguity in the outputs of the composing modules. The second one, named as Reliability-based-Aggregation (RbA), is based on reliability estimation of each classification act and chooses an output in any of the possible combinations of modules' output. To formally present our system, let us introduce the following notations:

- x is the current input;
- $\Gamma = \{C_1, \cdots, C_i, \cdots, C_L\}$ is the set of classes;
- Y_i is the module specialized in discriminating the sample of class C_i, and

its output $Y_i(x)$ is 1 if $x \in C_i$, 0 otherwise. Formally:

$$Y_i(x) = \begin{cases} 1 & \text{if } x \in C_i \\ 0 & \text{if } x \in C_k,\ k \neq i \end{cases} \tag{5.1}$$

- $\psi_i(x)$ is the reliability parameter of the classification provided by Y_i on x;
- m is the number of modules that vote for their own class, i.e. the number of blocks for which $Y_i(x) = 1$:

$$m = \sum_{i=1}^{c} Y_i(x) = \begin{cases} 1, & \text{in case (i)} \\ [2,L], & \text{in case (ii)} \\ 0, & \text{in case (iii)} \end{cases} \tag{5.2}$$

- $O(x) \in \Gamma$ is the MDS final decision on x;
- j is the index of the selected class $O(x)$;
- $\phi(x)$ is the reliability parameter of the final classification $O(x)$;

Notice that we have just mentioned *module* and not *classifier* to emphasize that each dichotomy can be solved not only by a single expert, but also by an ensemble of classifiers. For further discussion see section 5.2.4.

It is worth observing that the modules, besides labelling each pattern, may supply other information typically related to the degree that the sample belongs to that class. In this respect, the various classification algorithms are divided into three categories, on the basis of the output information that they are able to provide [98]. The classifiers of *type 1* supply only the label of the presumed class and, therefore, they are also known as experts that work at the *abstract* level. *Type 2* classifiers work at the *rank* level, i.e. they rank all classes in a queue where the class at the top is the first choice. Learning functions of *type 3* operate at the *measurement* level, i.e. they attribute each class a value that measure the degree that the input sample belongs to that class. If a crisp label of the input pattern is needed, we can use the maximum membership rule that assigns x to the class for which the degree of support is maximum (ties are resolved arbitrarily). Although abstract classifiers provide a n-bit string that can be compared with the codewords, decision schemes that exploit information derived from the classifiers working at the measurement level permit us to define reconstruction rules that are potentially more effective. Furthermore, if the module is constituted by a multi-experts system, the information supplied by the single classifiers can be used to compute a measure similar to that provided by measurement classifiers.

Since measurement classifiers can provide more information with respect to the other two types, we assume that only measurement experts constitute our MDS. Therefore, the next two paragraphs presents two rules that, given the individual decision $Y_1(x), \cdots, Y_L(x)$ and, eventually, information directly derived from their outputs, set the final label.

The Binary-Aggregation (BA) Rule. With reference to possible input combinations to the aggregation module reported above, the BA rule is expressed as follows:

$$j = \begin{cases} i : Y_i(x) = 1, & \text{if } m = 1 \\ L + 1, & \text{otherwise} \end{cases} \tag{5.3}$$

where $L+1$ denotes the index associated to a rejection. When $m = 1$, as a final class is chosen the class related to the module whose output is 1. The rationale lies in the observation that in such case all modules agree in their decision, since one indicates that x belongs to its own class whereas the others suggest that it is not a member of the class on which they are specialized. Otherwise the sample is rejected since the modules disagree each other. Indeed, when $m = 0$ none expert indicates that the sample belongs to its class, whereas when $2 \leq m \leq L$, two or more experts reveal that the sample belongs to their own class.

It is worth noticing that such an aggregation method, that provides a fixed reject rate, represents a conservative choice since the sample is classified only when all the modules agree. Furthermore, it does not require any estimation of the recognition reliability and, hence, it is suited for classifiers that do not work at the measurement level.

The Reliability-based-Aggregation (RbA) Rule. Alternatively, a zero-reject strategy that chooses an output in any of the possible input combinations to the aggregation module may be introduced. Since in case (i) above all modules agree in their decision ($m = 1$), as a final output is chosen the class of the module whose output is 1, regardless of the reliabilities. Conversely, in cases (ii) and (iii) the final decision is performed looking at the classification reliability provided by each module. Indeed, we deem that the estimation of the reliability of each classification act is a viable method to employ the information directly derived from the classifiers output since it has demonstrated its convenience also in other field [13, 19, 37, 114].

More specifically, in case (ii), m modules vote for their own class, ($2 \leq m \leq L$), whereas the others ($L - m$) ones indicate that x does not belong to their own class. To solve the dichotomy between the m conflicting modules we look at the reliability of their classification and choose the more reliable one. In case (iii) $m = 0$, i.e. all modules classify x as belonging to another class than the one they are specialized in. In this case, the bigger is the reliability parameter $\psi_i(x)$, the less is the probability that x belongs to C_i, and the bigger is the probability that it belongs to another class. These observations suggest finding out which module has the minimum reliability and then choosing the class associated to it as a final output.

The RbA rule just presented is formally expressed by the following equation:

$$j = \begin{cases} \arg\min\left(\psi_i(x)\right) & \text{if } m = 0 \\ i : Y_i(x) = 1 & \text{if } m = 1 \\ \arg\max_{i:Y_i(x)=1}\left(\psi_i(x)\right) & \text{if } 2 \leq m \leq L \end{cases} \tag{5.4}$$

It is worth observing that ψ, the reliability of each module, depends on the input sample x, as presented below (section 5.1.4). Therefore, the aggregation rule too relies upon the quality of the current input. To our knowledge this choice constitutes the novelty of such a method since its application can not be found in the literature of decomposition methods. Indeed, previous works in this field used in the aggregation module the credibility values estimated during the learning phase of the classifiers, thus ignoring the features of the current input [1, 51]. On the other hand, the papers on decomposition methods that employed the information directly derived from the outputs of the base classifiers typically considered only the highest activation among the experts,

e.g. the maximum output from a pool of neural networks. However, this measure cannot be regarded as a reliability parameter, since it has been demonstrated that it should be computed considering not only the winner output neurons but also the losers [13].

Prior to proceeds in the other sections, let us qualitatively compare the RbA and the BA criterions. When all modules agree in their decisions, the two rules take the same decision. In the other cases, the RbA criterion classifies the samples that are rejected by the BA rule. Therefore the latter, providing a fixed-reject system, is less liberal than the former, which acts as a zero-reject classifier. When most of the samples rejected by the BA criterion are correctly labelled by the RbA rule, the number of correct classifications increases.

5.1.3 Fluorescence Intensity Recognition System Configuration

The MDS employed to recognize the fluorescence intensity is based on the architecture presented in the previous section and depicted in figure 5.2. The feature selection phase suggests employing three specialized modules tailored to recognize positive, negative and border zone samples, which are named as PM, NM and BZM, respectively (section 5.1.1). In the detail, each module is constitued by a Nearest Neighbour (NN) classifier.

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 a Multi-Layer Perceptrons (MLP) with an hidden layer of ten neurons as classifier of each specialized expert, i.e. PM, NM and BZM. The corresponding results, reported in section 5.1.6, confirm that the MDS 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 section 2.4.2. In this respect, the NN classifiers can be effectively employed since the paradigm used for the reject option has been presently validated on them [19].

5.1.4 Reliability Estimators

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

In general, the most common choice for evaluating the classification reliability is to use the confusion matrix, which reports for each entry (p, q) the percentage of samples of the class C_p assigned to the class C_q, or other measures that depend on the recognition performance achieved during the learning phase. For example, if an expert assigns the input sample to a certain class, a reliability proportional to the recognition rate achieved on the training set on that class is attributed to such a decision. The drawback of this approach is that all the patterns attributed to the same class have equal reliability, regardless of the quality of the sample. Indeed, the average performance on the learning set, although significant, does not necessarily reflect the actual reliability of each classification act. However, as discussed above, recent works have demonstrated that more effective solutions could be achieved by introducing parameters that estimate the accuracy of each single classification act of the system [13, 86, 114].

This is the approach adopted here, where the reliability estimators depend on the input sample x.

With reference to the RbA rule, it makes use of reliability parameters to select the specialized modules that contribute to set the final label of individual cells (figure 5.2). Furthermore, the MDS can provide the reliability measurement ϕ of the final decision taken by the aggregation module, which can be used to apply a reject option.

To this aim, we first define how to compute the reliability estimator of individual experts classification and then we present a way to compute ϕ. Note that all these quantities vary in the interval $[0, 1]$, and a value near 1 indicates a very reliable classification.

Reliability of the Individual Binary Classifiers The definition of parameters that compute the reliability of each classification act for measurement classifiers, i.e. the classifiers that are able to supply most information on their classification since they attribute each class a value representing the degree that the input sample belongs to that class, has been discussed in the literature [13, 19].

A reliability parameter should permit to distinguish between the two reasons causing unreliable classifications [13]: (a) either the sample is significantly different from those presented in the reference set, i.e. 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].Based on these definitions, the parameter providing an inclusive measure of the classification reliability can be defined as follows:

$$\psi = \min(\psi_a, \psi_b) \tag{5.5}$$

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. In the case of NN classifiers, following [13], the samples belonging to the training set are divided into two sets: the reference set and the training test set. The former is used to perform the classification of the unknown pattern x, i.e. 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:

$$\psi_a = \max\left(1 - \frac{O_{min}}{O_{max}}, 0\right) \tag{5.6}$$

$$\psi_b = 1 - \frac{O_{min}}{O_{min2}} \tag{5.7}$$

where: O_{min} is the distance between x and the nearest sample of the reference set, i.e. 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 training test set, and O_{min2} is the distance between x and the nearest sample in the reference set belonging to a class other than $Y(x)$.

In the case of MLP classifiers, which we test to further validate the choice of NN classifier in the fluorescence intensity classification, the reliability can be estimated as:

$$\psi = \min\left(O_{win}, O_{win} - O_{2win}\right) = O_{win} - O_{2win} = \psi_b \tag{5.8}$$

where: O_{win} is the output of the winner neuron, O_{2win} is the output of the neuron with the highest value after the winner. The interested reader may find further details in [13]. Note that such estimators have been useful also in other applications, e.g. in [19, 91].

Reliability ϕ of the RbA Aggregation To our knowledge, works that compute the reliability of the final classification on the basis of reliability estimation of each dichotomizer cannot be found. Indeed, some papers propose confidence estimators that vary according to different input patterns [24, 67], whereas others estimate the reliability of classification on the basis of the performance on the training set [51]. Indeed, the latter approach is a general method to compute the classification reliability and it is based on the confusion matrix estimated on the learning set. However, it introduces the limitations presented above.

The former approach, i.e. the computation of a confidence estimators that depend on the input patterns is therefore a better alternative. In this respect, in [67] the authors computed the reliability of one-per-class network as the difference in activity between the class with the highest activity and the class with the second-highest activity. If this difference is large, the chosen class is much better than the others. If the difference is small, the chosen class is nearly tied with another ones. For error-correcting output codes, in [24] the authors computed two Euclidean distance: the first between the outputs of the base classifiers and the nearest codeword; the second between the outputs vector and the second-nearest codeword. The confidence is then estimated as the difference between these two distances. If the difference is large, an algorithm can be quite confident of its classification decision. If the difference is small, the algorithm is not confident. On the one hand, it is worth noticing that such approaches for reliability computation are limited to using one expert per each binary module. On the other hand, they use only the highest activation provided by the classifier, and do not consider the quality of the sample, i.e. its position in the feature space. Indeed, in the literature related to methods for estimating the reliability of a decision taken by a classification system, this quantity has demonstrated its potential [13, 19, 37].

In this respect, the novel contribution of this section is the definition of an estimator that compute the reliability of the final decision provided by a Multi Dichotomies System operating in the one-per-class framework.

In order to present the rationale of such a definition, we consider the following analogy. Each specialized module of our MDS may be regarded as an output neuron of a Neural Network. More specifically, with reference to the possible input combinations to the aggregation module reported in section 5.1.2, case (i) corresponds to the ideal situation in which only one output neuron is activated, whereas in case (ii) more than one output neuron (at worst all) are activated. Finally, in case (iii) none neuron is activated. In such an analogy, the reliabilities ψ_i of the specialized modules play the role of the activation potential of neurons, which naturally leads to express the reliability ϕ of the aggregation as a function of the ψ_is.

Formally, let us introduce the auxiliaries quantities Θ_a and Θ_b defined as follows:

$$\Theta_a(x) = \begin{cases} 1 - \psi_j(x) & \text{if } m = 0 \\ \frac{\psi_j(x) + \sum_{i:Y_i(x)=0} \psi_i(x)}{L-m+1} & \text{if } 1 \leq m \leq L \end{cases} \tag{5.9}$$

$$\Theta_b(x) = \begin{cases} \frac{\sum_{i:(Y_i(x)=0 \wedge i \neq j)} \psi_i(x)}{L-1} & \text{if } m = 0 \\ 0 & \text{if } m = 1 \\ \frac{\sum_{i:(Y_i(x)=1 \wedge i \neq j)} \psi_i(x)}{m-1} & \text{if } 2 \leq m \leq L \end{cases} \tag{5.10}$$

where m and $(L-m)$ are the number of specialised experts that vote and do not vote for their own class, respectively. Furthermore, according to conventions reported in section 5.1.2, j stands for the index of the class selected by the aggregation rule.

If $m > 1$, Θ_a represents the average of the reliabilities of the $(L - m + 1)$ modules agreeing that x belongs to the selected class j. Indeed both the reliability ψ_j of the selected module and those of the blocks whose outputs are 0 contribute to Θ_a. Similarly, Θ_b indicates the mean of the reliabilities of the $(m - 1)$ specialized modules that vote for their own class $i \neq j$, i.e. a class different from the selected one.

If $m = 1$, both Θ_a and Θ_b are simply equal to ψ_j and 0, respectively, because only one module classifies the sample in its class, whereas the others indicate that x does not belong to their classes.

If $m = 0$, Θ_a is equal to $1 - \psi_j$ since the contribution of the selected module should be larger as long as its own reliability decreases. Conversely, $\Theta_b(x)$ is the average of the reliabilities of the $(L-1)$ modules whose outputs coherently indicate that the sample does not belong to the class on which they are specialized, since their contribution should be proportional to their own reliabilities.

To take into consideration all modules in the computation of final decision reliability, we define ϕ as a function of both Θ_a and Θ_b:

$$\phi(x) = max(\Theta_a(x) - \Theta_b(x), 0) \tag{5.11}$$

where the maximum with respect to 0 guarantees that $\phi \in [0, 1]$. Indeed, such a quantity gets bigger in proportion with the difference between the reliabilities of the modules that agree and disagree with the final decision, respectively. Note that Θ_a, Θ_b and ϕ have been indicated explicitly as functions of the input sample x to emphasize that they are computed for each classification act of the system.

5.1.5 Implementation Issues

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 that it is used for the classification of both fluorescence intensity and staining pattern.

To this aim, we used sera of consecutive outpatients and inpatients of the Campus Biomedico, University Hospital of Rome, Italy, collected as presented in section 3.1. Up to now, the image data set consists of 600 images, stored in the database together with the ground truth. They have a resolution of 1024x1344 pixels and colour-depth of 8 bits.

Fluorescence Intensity Ground Truth

In IIF, the ground truth is made by labelled images both with fluorescence intensity and staining pattern classification. As reported in section 3.1.1, the images where blindly classified by two physicians, expert of IIF, working at the microscope. Furthermore, the reliability of the ground truth so obtained depends upon the agreement between the readers. In other words, its reliability depends on the degree of agreement between physicians. In this respect, Table 3.2 reports the values of Cohen's kappa, i.e. the most widely used measures of agreement between multiple readers [11, 65], computed on the specialists classifications of patient sera.

Working at the fluorescence microscope, 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 (see section 3.1.3 for kappa value interpretation guidelines). With reference to such 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 labelling the samples in five subgroups, i.e. four positive (4+, 3+, 2+, 1+) and one negative (0) subgroups (section 2.2.1), is not completely reliable.

The considerations presented in sections 3.1.2 and 3.2.1 motivate us to choose a classification of data samples into three classes (i.e. negative, border zone and positive), which is referred to as simplified CDC criteria. Indeed, we observe that the disagreement between physicians is twofold motivated. On the one hand, physicians assign the sample to different classes (i.e. one to positive, the other to negative). In the other case, specialists disagree about the subgroups to which a positive sample has to be assigned, i.e. physicians label it with a different number of plus. At a deeper examination, it appears that physicians always agree each other when the sample is marked either with 2+ or more, or when it is definitely negative. Therefore, we decide to assign a sample to the negative class if both physicians classify it as negative, whereas it is labelled positive if both specialists mark it with two pluses or more. Finally, a sample is assigned to the border zone class when either of the two types of disagreement described above happens or when both physicians mark it as 1+. Such a simplified classification maintains the clinical significance of the tests. Adopting this classification rule, the measured Cohen's kappa is 0.62 ± 0.13, implying substantial agreement, which is considered satisfactory to get a reliable ground truth.

5.1.6 Recognition Results

For testing the two introduced aggregation rules, i.e. the binary combination and the zero-reject one, we have used the 600 images of the database, which have been collected and labelled as previously reported. The a priori probability of positive, negative and border zone class is 36.0%, 32.5% and 31.5%, respectively.

The error rate has been evaluated according to a k-fold cross validation approach [49, 58], dividing the sample set in 8 folds: The rates reported in the following are the mean of k tests. For each test, *1/k* part of the data set has been used as the test set, another *1/k* as the training test set, the other parts the reference set. Using classes reported in Table 5.3, the recognition rate is shown in Table 5.4 and 5.5, as relative and absolute values, respectively. Note that in case of the RbA rule, the fourth row of Table 5.3 does not apply.

Table 5.3: Output categories of the three Inputs-three Outputs Classifier. Letters p, n, b and r stands for positive, negative, border zone and rejected samples, respectively. Lower and upper case letters refers to input and output classes, respectively.

		Input class		
		p	n	b
Output class	P	True Positive (TP)	False Positive (FPn)	False Positive (FPb)
	N	False Negative (FNp)	True Negative (TN)	False Negative (FNb)
	B	False Border Zone (FBp)	False Border Zone (FBn)	True Border Zone (TB)
	R	Rejected (Rp)	Rejected (Rn)	Rejected (Rb)

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

It is worth noting that in medical application, the two kinds of errors, i.e. 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%. 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 the 99% of them are correctly recognized.

These results show that an approach based on the reliability evaluation of the modules decision 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 MDS approach such performance figures are compared with those achieved by the other explored solutions, i.e. a single classifier architecture and a MDS where each specialized expert is a MLP classifier (see section 5.1.2). In the first case, we train a single NN using the features reported

Table 5.4: Relative performance of the recognition system, using the two aggregation rules, i.e. the Binary-Aggregation (BA) and Reliability-based-Aggregation (RbA), respectively.

Class	Hit Rate (Recognition Rate)		Reject Rate		Miss Rate			
	Binary Aggregation	Reliability-based Aggregation (0-reject)	Binary Aggregation	Reliability-based Aggregation (0-reject)	Binary Aggregation		Reliability-based Aggregation (0-reject)	
Positive samples	87.67%	92.12%	11.88%	–	FBp	0.45%	FBp	5.15%
					FNp	0.00%	FNp	2.73%
Border Zone samples	85.03%	92.24%	11.53%	–	FPb	3.43%	FPb	6.61%
					FNb	0.00%	FNb	1.16%
Negative samples	89.43%	98.90%	10.57%	–	FPn	0.00%	FPn	0.50%
					FBn	0.00%	FBn	0.60%

Table 5.5: Absolute performance of the recognition system, using the two aggregation rules, i.e. the Binary-Aggregation (BA) and Reliability-based-Aggregation (RbA), respectively.

	Binary Aggregation	Reliability-based Aggregation (0-reject)
Hit Rate $(TP + TB + TN)$	87.34%	94.33%
Miss Rate $(FP + FB + FN)$	1.33%	5.67%
Reject Rate $(Rp + Rn + Rb)$	11.33%	-

in Table 5.2. The single classifier achieves a hit rate of 91.05%, which is less than the one attained by the MDS (94.33%). Furthermore, looking at the relative performance we note that the misclassification rates on positive, negative and border zone samples are 9.29% (*FBp*: 6.16%, *FNp*: 3.13%), 8.78% (*FPn*: 3.14%, *FBn*: 5.64%) and 8.76% (*FPb*: 3.31%, *FNb*: 5.45%), respectively. It is worth observing that these rates are higher than those reported in Table 5.4, showing that the proposed MDS improves the recognition performance attainable for all the three classes. In the second case, MLP classifiers replace the NN ones in each specialized modules, i.e. PM, NM and BZM. 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 MDS composed of NN classifiers (Table 5.5). Indeed, on the one hand using the binary aggregation rule the hit, miss and reject rates are 82.66%, 3.34% and 14.00%, respectively. On the other hand, employing the reliability-based aggregation the hit rate is 89.46%. These results show that NN classifiers outperform MLP ones in the proposed MDS.

It is worth noticing that the definition of the ϕ parameter, i.e. the reliability estimator of the final classification, permits to introduce an error-reject option, that aims at rejecting the samples which would otherwise be misclassified. It works as follows: given a sample x, if the value of the reliability $\phi(x)$ is greater than a threshold, the classification will be accepted, otherwise the sample will be rejected. The effect of its application can be evaluated having recourse to an error-reject curve, which plots, for all the threshold values, the error versus the reject rate. It should be a non-increasing curve, since as the error rate decreases, as long as a larger number of samples are rejected. In this respect, figure 5.3 reports the error-reject curve for the fluorescence intensity classification, where the point most on the right represents the zero-reject achieved employing the RbA criterion. The curve demonstrates that the use of such a rule in combination with the reliability estimation has the potential to adapt the system to different working scenario by means of varying the minimum reliability required to accept a classification, thus moving the operating point along the curve itself.

Finally, if we report on the error-reject plane a cross in correspondence with the performance attained applying the BA rule, we notice that the curve goes

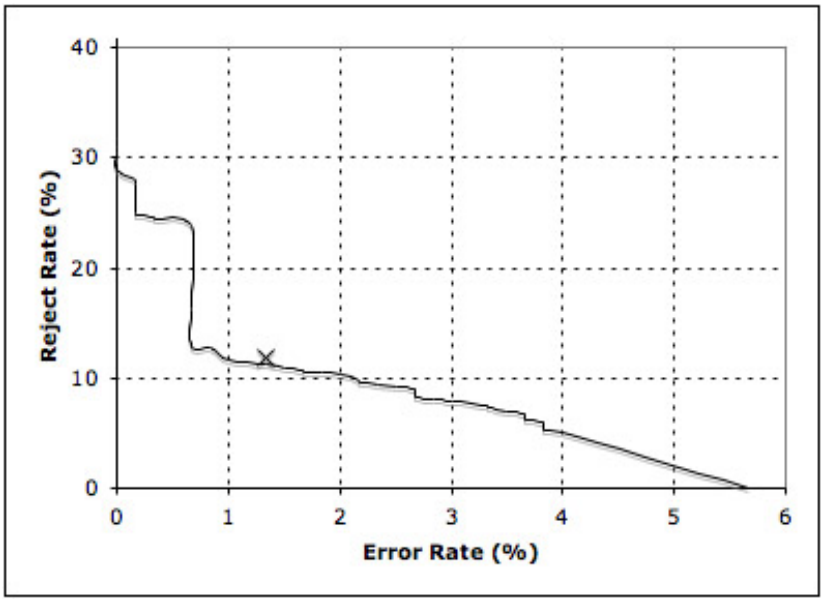

Figure 5.3: Error reject curve for the classification of the IIF fluorescence intensity.

below such a point (figure 5.3). This observation sustains the effectiveness of the reliability estimator ϕ and of the RbA rule, since they allow achieving better classification accuracy at the same reject rate and provide, in addition, the ability to flexibly set the operating point of the classification system.

Convenience Analysis

Now, given the two introduced aggregation rules, we are interested in understanding when it is preferable to use one strategy 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, 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 aggregation 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. Therefore, the overall cost is defined by:

$$C_{tot} = m\,C_m + r\,C_r \tag{5.12}$$

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

$$\overline{C_{tot}} = m + r\,\overline{C_r} \tag{5.13}$$

where the normalized global cost $\overline{C_{tot}}$ 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 aggregation rule performs better than the other, this last equation can be plotted in the $(\overline{C_r}, \overline{C_{tot}})$ plane (figure 5.4). With reference to this figure, each line represents the normalized global cost of the aggregation

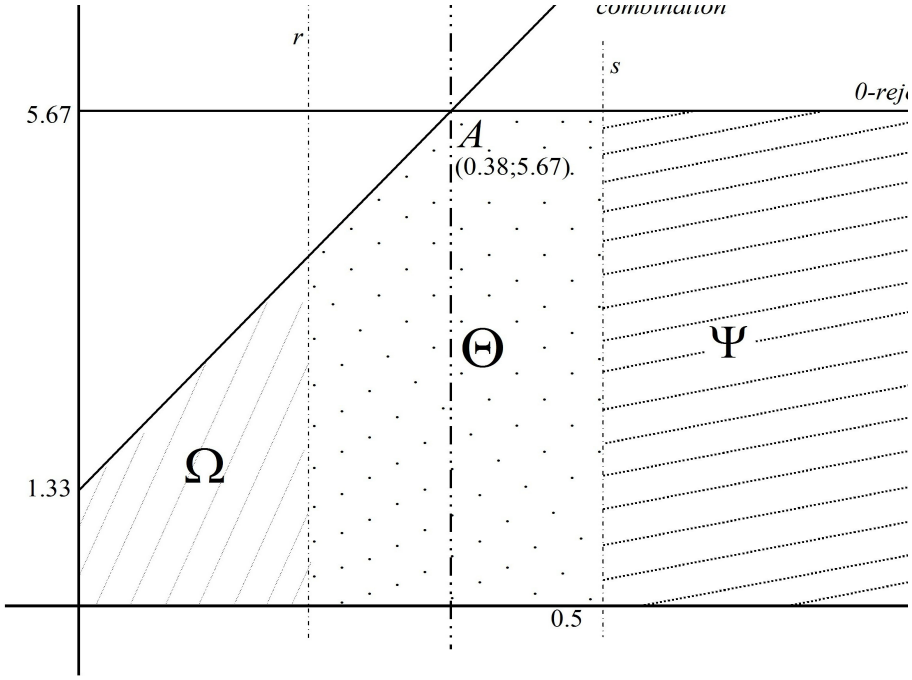

Figure 5.4: Convenience analysis of using one of the two aggregation rule (the binary or the zero-reject one) in a given application domain. The application domain is specified by the values of cost coefficients. A is the trade-off point between the two rules. Line r, s determine three different operating 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.

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 aggregation. In practical, the BA rule is preferable when the cost of a misclassification is less than 2.63 times the cost of a rejection.

In section 2.4.2, 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 to a zero-error system (remind that the binary aggregation rule does not exhibit false negative rate), although it shows a fixed reject rate. Therefore, for operating points located in such region the CAD is suited for application in case α_2.

For operating points in the shaded zone Ψ, the recognition system may adopt the aggregation 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.

5.2 Staining Pattern Classification

The two sides of IIF diagnosis concern both the fluorescence intensity classification and the staining pattern description. Having developed a system that addresses the former issue, we focus now on the latter one, i.e. on the recognition of the staining pattern. For the sake of comprehension, we report again the typical staining pattern classes:

- *Homogeneous*: characterized by a diffuse staining of the interphase nuclei and staining of the chromatin of mitotic cells;
- *Peripheral nuclear* or *Rim*: characterized by a solid staining, primarily around the outer region of the nucleus, with weaker staining toward the center of the nucleus;
- *Speckled*: characterized by a fine or coarse granular nuclear staining of the interphase cell nuclei;
- *Nucleolar*: characterized by a large coarse speckled staining within the nucleus, less than six in number per cell;
- *No pattern*: unclassifiable pattern.

Note that these classes are specific to the most relevant and recurrent ANAs. Figure 5.5 depicts four examples of these staining patterns. Notice that such images belong to strong positive sera, making the staining pattern easily distinguishable. No instance of the *no pattern* class is reported since it is quite impossible to find positive wells belonging to that class, whereas some cells without a classifiable pattern can occur in a given well.

This section presents a system that supports the classification of the staining pattern of the whole well at the 1:80 titer, i.e. the recommended one [8]. It belongs to the field of supervised statistical pattern recognition (see section 5.1). The background, the motivations and the related literature have been presented in chapter 2. In particular, with respect to [90] we use a more sophisticated classification architecture, a different criterion to determine the well pattern and we employ the MDS reliability estimator introduced in section 5.1.4. Furthermore, our work differs from [81] and [85] for three main reasons: first, such papers classified only the pattern of individual cells; second, they used samples diluted at 1:160 and, third, our data set contains only positive wells. With reference to the last point, notice that both the medical guidelines and the clinical practice require labelling the staining pattern only for positive samples. Finally, the approach presented here differs from [41] since they used only positive controls, whereas we use images acquired from sera of real patients, which therefore exhibits positive fluorescence intensity at various grading.

The rest of the chapter is organized as follows: from sections 5.2.1 up to section 5.2.5 we give an overall view of the proposed recognition system, first presenting in general the classification approach and then discussing each part of

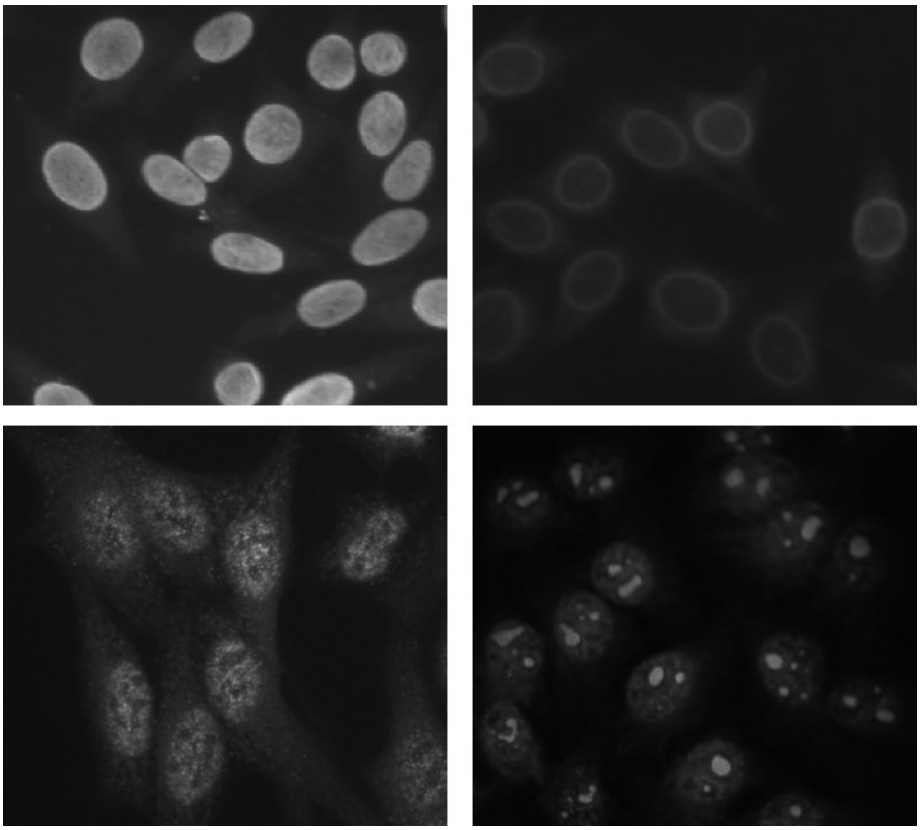

Figure 5.5: Examples of the homogenous, rim, nucleolar and speckled staining patterns (clockwise).

the system. Section 5.2.6 details the implementation issues, with particular reference to both the data set description and the feature extraction and selection phase. Section 5.2.7 contains the experimental results and the discussion. Finally, section 5.3 offers an overview of the results we get combining fluorescence intensity and staining pattern classifications.

5.2.1 Recognition Approach

To classify the well staining pattern into one of the basic groups reported above, we adopt the approach depicted in figure 5.6. First, we segment the image to locate the cells and we extract the features; second, we label the staining pattern of individual cells and, third, we classify the staining pattern of the whole well on the strength of the classification of its cells.

In our opinion, such an approach addresses some key-points of IIF staining pattern classification. Indeed, a recognition method based on the classification of individual cells is tolerant with respect to misclassifications, since the final label of the well is computed by using several pieces of information. If enough cells per well are available, it is reasonable that cells misclassification, if limited, does not affect the well pattern classification. Furthermore, this approach has the

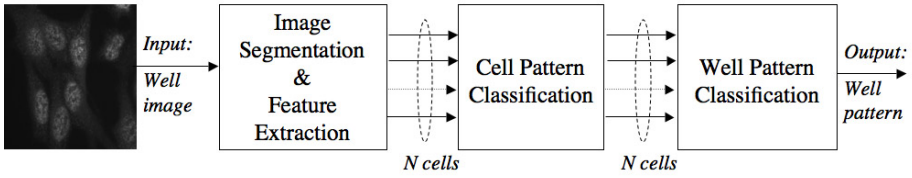

Figure 5.6: Representation of the proposed approach to classify the well staining pattern.

potential for detecting the occurrence of multiple patterns, i.e. the predominant and the minor ones. In the following subsections we focus on the first, second and third block of figure 5.6, respectively.

5.2.2 Cell Location

Each well image is pre-elaborated to improve the contrast and then morphological filters, such as erosion and dilation, have been applied to remove noise. Using Otsu's algorithm [80], automatic thresholding is performed to locate the cells. Next, the application of other morphological operations, such as filling and connection analysis, gives a binary mask for cutting out the cells from the image. 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 the slide production process. To remove overlapping cells we compute the following circularity measure:

$$\text{circularity} = \frac{4 \cdot \pi \cdot (\text{cell Area})}{(\text{cell Perimeter})^2} \tag{5.14}$$

Then, based on a simple heuristic, we remove the cells for which this parameter is less than 0.5. After these operations the images of the single cells are individually cropped to a rectangular region.

5.2.3 Feature Extraction and Selection

From an image processing point of view, we deem that texture analysis is the proper way to analyze the staining pattern. In this respect, for each sample, i.e. the cells, we compute a set of texture features related to statistical and spectral components.

The first kind of features is based on first and second order measures. First order features are related to the statistical properties of the intensity histogram. Examples are the moments up to the fourth order, the energy around the peak, the entropy, and so on. Second order statistics are considered since it has been shown that humans are sensitive to them [52]. The most used one is the grey level co-occurrence matrix, which estimates the spatial grey level dependence, computing the second-order joint conditional probability density function[2]. From a co-occurrence matrix a number of features that describe the

[2]The second-order joint conditional probability density function, $f(r, t|\delta, \varepsilon)$, is the probability of going from grey level r to grey level t, given the intersample spacing δ with direction ε.

coarseness or the finesse of the texture can be derived. The most used are: contrast, entropy, correlation, homogeneity, both angular and inverse difference moment up to the fifth order.

The second kind of features is related to spectral measures of texture. They are obtained analyzing the spectrum of the Fourier Transform (FT), which is useful for discriminating between periodic and non-periodic texture patterns. Indeed, the global texture patterns, easily distinguishable as concentration of high energy burst in the spectrum, are generally quite difficult to detect with spatial methods because of the local nature of these techniques. Since one way to extract features in the FT domain is to partition the FT space into bins, we compute the two commonly used ones, the angular and the radial [59].

Furthermore features related to the Wavelet Transform and Zernike moments have been computed. Indeed, the former have been used in many signal processing applications such as multiresolution signal processing, computer vision, image analysis and compression [68], whereas in some cases the latter have been demonstrated their superiority over regular moments and moment invariants [55]. In this work we employ both discontinuous (Haar) and continuous (Daubechies) wavelets, since they have proven their usefulness for the discrimination of textures in medical imaging [23, 39]. In particular, we compute features related to average, standard deviation, entropy and energy of wavelet coefficients. Although originally used in the description of optical aberration, the Zernike polynomials, on which the Zernike moments are based, have found application in pattern recognition [2, 55] to introduce rotation-invariant features. Here, we compute the magnitude of Zernike moments up to the 15^{th} degree.

For each cell, all features are computed from four different images:

- the original cell with the original background surrounding it, to consider the true image;
- a contrast enhanced version of the previous image, to emphasize the difference between uniform and variable patterns (e.g. the homogeneous and the speckled ones);
- the original image with the background set to zero, to minimize its contribution and maximize the one of cell border;
- a contrast enhanced version of the previous image, to emphasize its pixels variation.

In conclusion, the initial features set is constituted of 360 features. The search of the best discriminant subset 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 [50].

Discriminant analysis shows that all the above features have limited discriminant strength over all classes. Indeed, the best feature set reports maximum separation strength of approximately 62%.

On these motivations, and similarly to the experience on the classification of IIF fluorescence intensity reported in section 5.1.1, we take into consideration a recognition approach based on a MDS operating in the one-per-class framework. To this end, we perform discriminant analyses that aim at distinguishing the samples of one class from the others. The results show that the homogeneous

and the speckled classes are the most difficult to discriminate. Indeed, the best discrimination attained on them are 81.3% and 75.7%, respectively. The best feature subsets related to the rim class exhibits good discriminant strength (85.1%). Finally, nucleolar and artefact classes are well separated from the others (96.1% and 94.2%, respectively).

These results are expected: indeed, on the one hand the specialists report troubles in distinguish between homogeneous, rim and speckled classes, whereas on the other hand they report that the nucleolar samples are the easiest to be identified, since they exhibit one or more evident nucleoli, as discussed in section 3.3.

In conclusion, we deem that the set of stable and effective features obtained for each class enforced the choice of using a recognition approach based on the aggregation of binary classifiers.

5.2.4 Cell Recognition

The previous considerations and the preliminary results on the classification of the staining pattern of individual cells suggested us to use a combination of experts rather than a single one. Hence, similarly to the classification of the fluorescence intensity, we adopt a MDS that belongs to the one-per-class framework.

From a theoretical point of view, each module of the system can be constituted either by a single classifier or by employing a multiple experts scheme that aims at improving the distinguishing ability of the module, as introduced in section 5.1.2. However, to our knowledge, the dichotomizers typically adopt the former approach, i.e. they are composed of one classifier per specialized module. For example, for their experimental assessments the authors used a a decision tree and a multi layer perceptrons with one hidden layer both in [72] and [73], respectively. The same functions were employed by Dietterich and Bakiri for the evaluation of their proposal in [24], whereas Allwein et al. used a Support Vector Machine [1]. A viable alternative to using a single expert is the combination of classifiers outputs solving the same recognition task. The idea is that the classification performance attainable by their combination should be improved by taking advantage of the strength of the single classifiers, without being affected by their weakness [47, 56, 63, 98, 108]. Classifier selection and fusion are the two main combination strategies reported in the literature. The former presumes that each classifier has expertise in some local area of the feature space [33, 43, 45, 61, 108, 111]. For example, when an unknown pattern is submitted for classification, the more accurate classifier in the vicinity of the input is selected to label it [108]. The latter algorithms assume that the classifiers are applied in parallel and their outputs are combined to attain somehow a group of "consensus". It supposes that all classifiers are equally "skilled" and applied in parallel over the whole feature space, providing robustness by multiplying the number of observation channels, which are then combined in a data fusion block [5, 14, 36, 56, 63, 64]. Typical fusion techniques include weighted mean, voting, correlation, probability, etc.. Finally, a different approach is based on a mixture of selection and fusion (e.g. as in [60]).

In the case of the classification of the staining pattern of individual cell, we deem that the recognition performance achieved using one classifier per module are not satisfactory, as presented in section 5.2.7. Therefore, to improve these

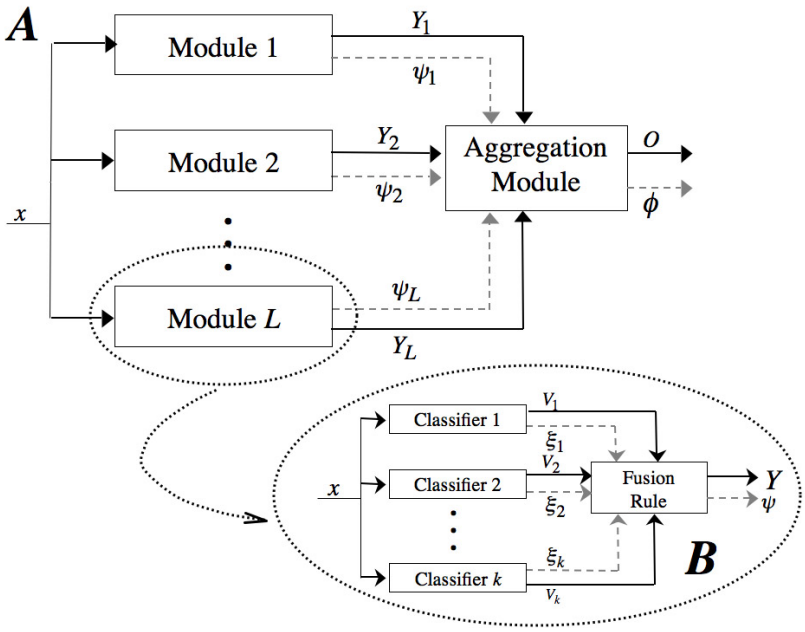

Figure 5.7: The architecture of the recognition system where the modules that operate in the one-per-class framework (panel ***A***) are composed of a fusion of experts (panel ***B***). $V(x)$, $Y(x)$ and $O(x)$ represents the outputs of the single classifier, of the binary module and of the overall recognition system, respectively. $\xi(x)$, $\psi(x)$ and $\phi(x)$ represents the classification reliability estimatiors of the single expert, of the binary modules and of the overall MDS, respectively.

performance, we implement the specialized modules with multiple binary classifiers combined by fusion (figure 5.7). In panel B it shows that the fusion of individual expert decisions provides the module, thus extending the general architecture depicted in figure 5.2. For the sake of brevity, such system is referred to as Hybrid-Classifier-Aggregation-Fusion (HCAF) MDS.

Among the many combination schemes proposed in the literature, as a fusion criterion internal to each specialized module we use the Weighted Voting (WV) rule, which is reported in the next paragraph.

Fusion Rule The Weighted Voting Rule (WV) is a voting technique in which the opinion of each expert about the class of the input pattern is weighted by a reliability parameter.

Since each expert deals with a binary learning task, we present here a simplified notation of the WV rule. Denoting as $V_k(x)$ and as $\xi_k(x)$ the output and the classification reliability of kth classifier on sample x, the weighted sum of the votes for each of the two classes is given by:

$$W_h(x) = \sum_{k:V_k(x)=h} \xi_k(x), \text{ with } h = \{0, 1\} \tag{5.15}$$

Therefore the fusion output, $Y(x)$, is defined by[3]:

$$Y(x) = \begin{cases} 1, & \text{if } W_1(x) > W_0(x) \\ 0, & \text{otherwise} \end{cases} \tag{5.16}$$

$$\tag{5.17}$$

Reliability of the Weighted Voting Classification When the modules of the overall recognition system are composed of a fusion of individual experts, the reliability ψ should be computed considering the accuracy of the classifications provided by the individual experts. For NN or MLP classifiers, i.e. the experts used here, the latter parameters can be computed applying equations 5.5, 5.7 and 5.8.

In the following we present a method to estimate the accuracy of each single classification act provided by the WV rule, thus overcoming the limitations introduced by the use of the confusion matrix discussed in section 5.1.4. To this aim we need to introduce the following two auxiliary quantities:

$$\pi_1(x) = \max\left(\{\xi_k(x) | k : V_k(x) = Y(x)\}\right) \tag{5.18}$$

$$\pi_2(x) = \max\left(\{\xi_k(x) | k : V_k(x) \neq Y(x)\} \cup \{0\}\right) \tag{5.19}$$

where $\pi_1(x)$ and $\pi_2(x)$ represent the maximum reliabilities of experts voting for the winning class and for other classes (0 if all the experts agree on the winner class), respectively. Given these definitions, the reliability of the WV rule can be evaluated according to a conservative choice, as follows:

$$\psi(x) = \min\left(\pi_1(x), \max\left(0, 1 - \frac{\pi_2(x)}{\pi_1(x)}\right)\right) \tag{5.20}$$

5.2.5 Well Recognition

The staining pattern of the whole well is determined on the basis of the classification of its cells. To this aim, different strategies can be applied. For example, the aggregation criterion presented in [90] allowed using only rules based on simple voting. Indeed in that paper we compared two strategies based on absolute and relative majority of cells distribution among the staining pattern classes (see section 2.4.2).

On the contrary, the aggregation rule adopted here permits to estimate the reliability ϕ of the staining pattern classification of individual cells. This choice allows us using a well recognition rule, named in the following as Weighted Sum (WS) rule, which is based on weighting the classifications of individual cells of the well. Formally, for each well we define WS_i as:

$$WS_i = \sum_X \phi(x) \cdot I_i(x) \tag{5.21}$$

where the summation is over the set X of cells that belong to the well under consideration and $I_i(x)$ denotes an indicator variable defined as follows:

$$I_i(x) = \begin{cases} 1 & \text{if the cell } x \text{ is classified to class } C_i \\ 0 & \text{otherwise} \end{cases} \tag{5.22}$$

$$\tag{5.23}$$

[3]In case of tie, i.e. if $W_1(x)$ is equal to $W_0(x)$, the output $Y(x)$ is set arbitrarily to zero. Note that it never occurred in all tests we performed.

The index of the final class of well staining pattern is $\upsilon = \arg\max_i(WS_i)$, i.e. the class for which WS_i is maximum.

5.2.6 Implementation Issues

Similarly to fluorescence intensity classification, for well staining pattern there are not reference databases of IIF images publicly available. Indeed, the other research groups involved in this field use private datasets [41, 81, 85].

To develop the system devised to classify the staining pattern, we randomly choose 37 images of positive wells from our IIF dataset, populated as presented in section 5.1.5. Note that we choose only positive images, i.e. images that can be ascribed either to 2+, 3+ or 4+ class, since the guidelines require to classify the staining pattern only for positive samples [8].

Staining Pattern Ground Truth

With reference to our classification approach (figure 5.6), we need the labels of both individual cells and whole well, respectively. To the first aim, i.e. to get the single cells ground truth, two specialists of IIF independently and blindly classify the pattern of each cell at a workstation monitor, since at the fluorescence microscope is not possible to observe one cell at a time.

More specifically, they classify about 15 cells per well, one at a time, which have been chosen at random from those segmented according to the procedure described in section 5.2.2. To this end, the classes reported at the beginning of section 5.2 do not cover all the possibilities. Indeed, on the one hand those classes represent a global pattern, i.e. the pattern of whole well that is given by the global observation of several cells. On the other hand, each cell could potentially show a staining pattern that is different from the well pattern.

For manual labelling we therefore adopt the classes reported elsewhere [81]. The specialist can label each cell into one of the following groups (for definition of classes (i)–(iv) and (viii) see section 5.2):

(i) *homogeneous* (HO);

(ii) *peripheral nuclear* or *rim* (PN);

(iii) *speckled* (SP);

(iv) *nucleolar* (NU):

(v) *artefact* (AR), cell corrupted during the slide preparation process, identifiable with an irregular shape;

(vi) *positive mitosis*, the nonchromosome region of metaphase mitotic cells demonstrate staining:

(vii) *negative mitosis*, the nonchromosome region of metaphase mitotic cells is negative;

(viii) *no pattern* (NP).

Since the number of cells belonging from groups (vi)–(viii) in not statistically meaningful, they are not considered in the following.

The data set consists of 573 labelled cells, therefore subdivided: 23.9% HO, 21.8% PN, 37.0% SP, 8.2% NU and 9.1% AR.

Clearly, such an approach relies upon the agreement between multiple readers. In other words, the reliability of cells ground truth depends on the degree of agreement between the specialists. To evaluate it, we use the Cohen's kappa, similarly to the assessment of digital images for diagnostic purpose in IIF (chapter 3).

The statistical analysis shows a substantial agreement between the two readers ($k = 0.67 \pm 0.05, p < 0.05$) when they classify the staining pattern of cells into one of the above eight classes. We deem that this result is satisfactory to get a reliable ground truth.

To further assess its reliability, we observe that similar results were achieved in the labelling of the whole well staining pattern, as reported in section 3.2.2. There we observed that the readings between pairs of experts suggested substantial agreement when the diagnosis was performed both at the microscope ($k = 0.61 \pm 0.13, p < 0.05$) and at the monitor ($k = 0.68 \pm 0.12, p < 0.05$), respectively. It is apparent that the agreement value related to the labelling of individual cells is very close to that concerning the whole well classification, suggesting that the recognition approach proposed here is well founded.

We also note that both in well and individual cell staining pattern classification the main reason of disagreement between the readers clustered around the alternative speckled vs. homogeneous. Furthermore, for cell recognition a relevant disagreement may be found in the choice speckled vs. artefact patterns. Data analysis reveals that most of the controversies occur when the samples exhibit weak positive fluorescence.

Finally, to determine the ground truth related to staining pattern of the whole well we proceed as reported in 3.1.1. For the 37 wells that we used, the patterns distribution is: 24.3% homogeneous, 21.6% rim, 35.1% speckled and 18.9% nucleolar.

5.2.7 Recognition Results

Individual Cell Recognition

The HCAF system is a MDS constituted by five modules each one devised to discriminate one of the five input classes (i.e. HO, PN, SP, NU and AR) from the others. Each block is composed of a fusion of individual classifiers, such as k-Nearest Neighbour and Multi-Layer-Perceptrons with one or two hidden layers, combined by the WV algorithm. The recognition performance for the classification of individual cells has been evaluated according to a k-fold cross validation approach [49, 58], with $k = 8$.

With regard to the aggregation rules, Tables 5.6 and 5.7 report the confusion matrix computed averaging out the k tests for the BA (Binary-Aggregation) and RbA (Reliability-based-Aggregation) criterions, respectively. The former table shows that the classification accuracy of HO, PN and NU classes ranges from 51% to 60%, whereas the best and worst recognition performance are attained for cells of SP and AR classes, i.e. 75% and 29%, respectively. However, as introduced in section 5.1.2, such a rule introduces a fixed reject rate that aims

Table 5.6: Confusion matrix of HCAF classifier employing the Binary-Aggregation (BA) rule. The values are the mean of k-fold cross validation tests. Letter Rj stands for the reject class.

		Input class				
		HO	PN	SP	NU	AR
Output class	HO	60.6%	2.4%	1.9%	2.1%	11.5%
	PN	2.9%	52.8%	2.8%	4.3%	3.8%
	SP	6.6%	4.0%	75.9%	0.0%	9.6%
	NU	1.5%	0.8%	0.5%	51.1%	3.8%
	AR	2.9%	2.4%	0.5%	0.0%	28.8%
	Rj	25.5%	37.6%	18.4%	42.6%	42.3%

Table 5.7: Confusion matrix of HCAF classifier employing the Reliability-based-Aggregation (RbA) rule. The values are the mean of the k-fold cross validation tests.

		Input class				
		HO	PN	SP	NU	AR
Output class	HO	73.9%	5.6%	5.2%	8.5%	15.4%
	PN	10.0%	71.2%	3.8%	14.9%	13.5%
	SP	10.2%	12.8%	88.2%	0.0%	17.3%
	NU	1.5%	2.4%	0.5%	72.3%	9.6%
	AR	4.4%	8.0%	2.4%	4.3%	44.2%

at lowering the misclassifications. Indeed, the hit rate on the classified samples for HO, PN, SP, NU and AR classes is 81.3%, 84.6%, 93.0%, 89.0% and 50.1%, respectively. The latter table shows that the classification accuracy of HO, PN and NU classes ranges from 71% to 74%, whereas the best and worst recognition performance are attained for cells of SP and AR classes, i.e. 88% and 44%, respectively.

Whatever the aggregation rule, we deem that, on the one hand, misclassifications of HO, PN and SP samples are related to their similarities of staining pattern and texture. Indeed, the discrimination between such classes is a burdensome issue also for well-trained specialists. On the other hand, errors on NU and NP classes are related to the small cardinality of such sets. Moreover, the variability among AR samples is high, since such class contains those cells corrupted during the slide preparation that exhibit irregular shape and texture.

Turning our attention to the absolute performance of the HCAF MDS, when the BA rule is applied the 60.8% of cells are correctly classified whereas the 28.8% are rejected (Table 5.8). The experimental results achieved applying the RbA rule, which provides a 0-reject system, show that the MDS correctly label the 75.9% of samples.

When the system adopts the RbA criterion, it is possible to measure the

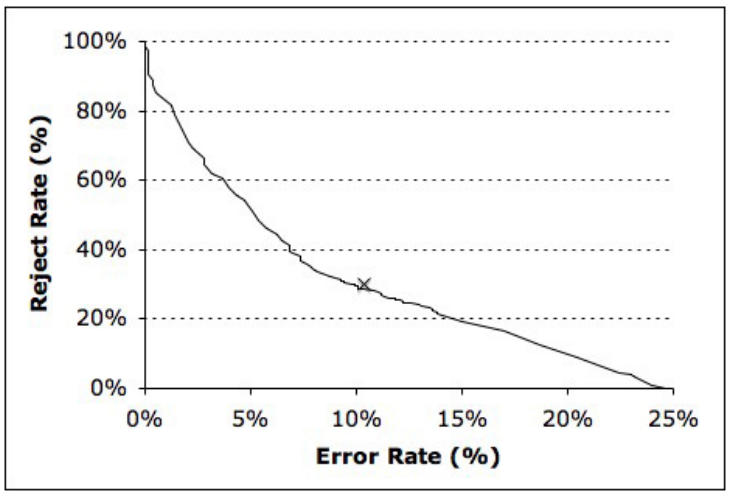

Figure 5.8: Error reject curve for the classification of the staining pattern of individual cells.

reliability of the cell classification using the estimator ϕ, which has been introduced in section 5.1.4. It permits to introduce an error-reject option that aims at rejecting the samples which would otherwise be misclassified. Similarly to section 5.1.6, it works as follows: given a sample x, if the value of the reliability $\phi(x)$ is greater than a threshold, the classification will be accepted, otherwise the sample will be rejected. The effect of its application can be evaluated having recourse to an error-reject curve, which plots, for all the threshold values, the error versus the reject rate. It should be a non-increasing curve, since as the error rate decreases, as long as a larger number of samples are rejected. With reference to individual cells staining pattern classification, figure 5.8 reports the estimated error-reject curve. Note that the point most on the right represents the zero-reject performance of the HCAF MDS, i.e. the recognition rate reported above. The curve demonstrates that the use of the RbA rule in combination with the reliability estimation has the potential to adapt the system to different working scenario by means of varying the minimum reliability required to accept a classification, thus moving the operating point along the curve itself.

Finally, if we report on the error-reject plane (figure 5.8) a cross in correspondence with the performance attained applying the BA rule to the HCAF system, we notice that the curve goes below this point. Such an observation sustains the effectiveness of the reliability estimator ϕ and of the RbA aggregation rule, since they allow achieving better classification accuracy at the same reject rate and provide, in addition, the ability to flexibly set the operating point.

Before presenting the results achieved in the recognition of whole well staining pattern, let us compare the results of the HCAF MDS in the discrimination of individual cells with those reported in [90]. In that paper the cells classifier is a MDS operating in the one-per-class framework where, however, each module is composed of a single classifier. That architecture is named as Single-Classifier-Aggregation (SCA). Furthermore, in [90] only the BA rule is applied. In order to provide a deep comparison between the two classification schemes,

Table 5.8: Absolute performance of both Hybrid-Classifier-Aggregation-Fusion (HCAF) and Single-Classifier-Aggregation (SCA) systems, respectively, using the two aggregation rules (BA and RbA).

	BA		RbA	
	SCA	*HCAF*	SCA	*HCAF*
Hit Rate	54.1%	60.8%	71.2%	75.9%
Miss Rate	11.0%	10.4%	28.8%	24.6%
Reject Rate	34.9%	28.8%	-	-

Table 5.9: Confusion matrix of SCA classifier employing the Binary-Aggregation (BA) rule. The values are the mean of k-fold cross validation tests. Letter Rj stands for the reject class.

		Input class				
		HO	PN	SP	NU	AR
Output class	HO	51.1%	2.4%	1.9%	6.4%	7.7%
	PN	2.9%	52.0%	1.4%	6.4%	5.8%
	SP	5.1%	3.2%	66.0%	2.1%	7.7%
	NU	0.0%	1.6%	1.4%	48.9%	5.8%
	AR	2.9%	3.2%	1.9%	0.0%	23.1%
	Rj	38.0%	37.6%	27.4%	36.2%	50.0%

i.e. SCA and HCAF, Tables 5.9 and 5.10 report the confusion matrices of the SCA system when the BA and RbA criterions are employed.

With reference to the BA rule, the comparison between Tables 5.9 and 5.6 shows that the HCAF system better recognizes than SCA classifier the sample of all classes. Indeed, adopting HCAF rather than SCA, the performance concerning the elements of HO, SP and AR classes raises of 15%, 13% and 19%, respectively. For element of classes PN and NU the improvement is less than 5%. Moreover, the reject rate (Rj) is always bigger using SCA than HCAF, except for the NU class. All these considerations are apparent looking at the absolute performance of the two approaches (second and third columns of Table 5.8): the improvement attained using the HCAF system rather than SCA concerns more both the hit and reject rate, and less the miss rate. Indeed, the hit rate increases of 11%, whereas the reject and the miss rate decrease of 27% and 6%, respectively.

Turning our attention to the RbA rule, the comparison between Tables 5.10 and 5.7 demonstrates that, once more, the HCAF better recognizes than SCA the sample of all classes. Indeed, the performance figures for elements of SP, NU and AR classes improve of 7%, 6% and 26%, respectively, using HCAF rather than SCA. For element of classes HO and SP the improvement is less than 5%. Looking at the absolute performance of the RbA rule, adopting the HCAF system instead of the SCA one, the hit rate increases of 7%, whereas the

Table 5.10: Confusion matrix of SCA classifier employing the Reliability-based-Aggregation (RbA) rule. The values are the mean of the k-fold cross validation tests.

		Input class				
		HO	PN	SP	NU	AR
Output class	HO	73.0%	10.4%	8.5%	12.8%	26.9%
	PN	7.3%	68.0%	6.1%	17.0%	13.5%
	SP	13.9%	10.4%	82.1%	0.0%	13.5%
	NU	0.7%	5.6%	1.9%	68.1%	13.5%
	AR	5.1%	5.6%	1.4%	2.1%	32.7%

miss rate decreases of 19%.

In conclusion, the results clearly show that the recognition accuracy improves adopting a hybrid approach, where an ensemble of classifiers composes each binary module. Indeed, the combination by fusion rather than the use of a single classifier gets better discrimination capability at the stage of both individual module performance and overall classification system.

In summary, we observe that the overall performance of the presented cells classifier outperforms that reported in [90]. Furthermore, a direct comparison of this results with respect to [41], [81] and [85] is not possible, since their recognition tasks differ from ours. Indeed, in [41] the authors employed only sera of positive controls, whereas in [81] and [85] the authors used a different data set, which is not only constituted by samples diluted at 1:160, but also containing cells that were negative, i.e. they did not exhibit a detectable fluorescence intensity. However our system, differently from the others, may vary its operating point, which it makes it suited to different scenarios.

Whole Well Recognition

Before presenting the performance achieved in the recognition of whole well staining pattern, notice that the established medical knowledge admits as well staining classes only those reported at the beginning of section 5.2 (i.e. homogeneous, rim, speckled, nucleolar and no pattern classes). On this basis, the cells that are labelled as AR are not considered to determine the well pattern.

In order to assess the performance of our proposal, we are interested not only in estimating the recognition rate, but also in comparing it with the one reported in [90]. Therefore, we initially adopt the same evaluation method reported in that paper, proceeding as follows. For all the wells, we randomly subdivide their cells into two equal partitions, and then each partition is first used as a training set and then as test set. We deem that this is a good balance between the need of keeping the training set representative as most as possible and having enough test cells per well to classify the staining pattern in accordance to the WS criterion. In the two trials, the system misclassified only one out of the 37 wells, attaining an hit rate equal to 97.3% and outperforming the one reported in [90] (see section 2.4.2).

However, a critical point of such an evaluation technique is that different

cells belonging to the same well are included in both the training and the test sets. Therefore, when an unknown cell is presented to the classifier, a similar one (in the sense of their features) is likely to be in the training set. In order to assess the system robustness in a working scenario in which input wells could contain cells different from those of the training set, we proceed similarly to a leave one out approach [31] working at the well level rather than at the cells one. This is the second estimation method. At each iteration one well (and therefore all its cells) constitutes the test set, while the others populate the training set. Therefore, the HCAF first learns from the training patterns and, second, it labels each cell of the test set; third, the reliability of cells classification is computed and, fourth, the error-reject curve is determined. Then, the application of the WS criterion determines the whole well staining pattern for each value of the threshold used to reject the individual cells. Notice that the rejected cells do not take part in the determination of well pattern. Figure 5.9 reports the hit rate of wells with respect to the threshold used for the application of the error-reject option at cell level. If all the input cells are accepted, the 82.4% of wells are correctly classified. The plot shows that it is convenient to reject the cells whose classification is not so much accurate, since the hit rate increases up to 85.3% moving along the curve. Notice that if the threshold further increases, i.e. most cells are rejected, the accuracy in the well classification decreases. It occurs at threshold values bigger than 0.7, when no enough cells are used to decide the well staining class. Indeed, although only cells with a very reliable classification are used, it can happen that these cells belong to a class different from the main well pattern.

If we look at the zero-reject performance achieved, the hit rate is still significantly greater than the one presented in [90]. Nevertheless, we deem that the error rate could be still too high to make the system usable in the medical practice. To overcome such a limitation, in an operating scenario we may apply the reject option to the decision taken by the WS criterion. To this end, we have to estimate the reliability of the decision provided by this rule, and then to compare it with respect to a threshold, similarly to what we did to reject individual cells. In this respect, it looks reasonable to adopt as a reliability estimator the quantity:

$$\rho = \frac{\max_i (WS_i)}{\sum_i WS_i} = \frac{WS_v}{\sum_i WS_i} \tag{5.24}$$

where v is the index of the final class of well staining pattern and i varies over the four classes homogeneous, rim, speckled and nucleolar (see the introduction). Indeed, the rationale of this choice is that the final classification is as much reliable as a larger number of cells are classified in the final class of the well. Applying such an option, with a threshold equal to 0.57 we get an error rate of 5.8%. Notice that this value is smaller than the estimated intra-laboratory variability, which it has been measured equal to 7.4% in [83]. The corresponding reject rate is 17.6% which looks fairly limited. This performance seems very good and makes the system usable in practice, especially as a second reader to support the specialists' decisions.

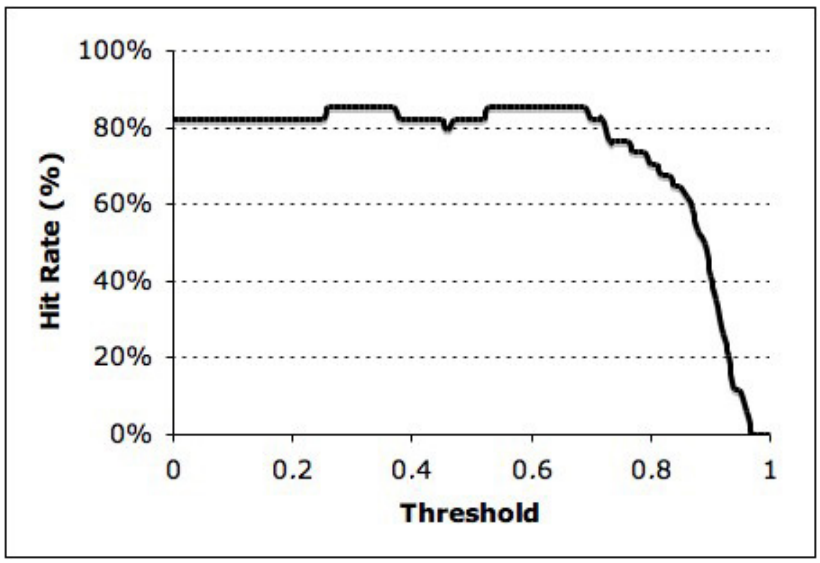

Figure 5.9: Plot of the hit rate achieved in the classification of whole well staining pattern on the basis of the threshold applied to reject the individual cells.

5.3 IIF Classification Overview

In the previous sections we reported the results on the classification of both IIF fluorescence intensity and staining pattern. They can be used to estimate the perspective recognition performance of the overall CAD system.

We have observed that the classification of IIF samples is based on both fluorescence intensity and staining pattern description. The former is reported for all the input samples, whereas the latter is detailed only for positive samples. Hence, the proposed recognition approach is based on a cascade of two steps: the first classifies the fluorescence intensity, whereas the second recognizes the staining pattern only for positive wells, as shown in figure 5.1.

On the basis of the recognition system configurations presented so far different setups can be used. Among all, in this section we focus on the two extreme available arrangements, which are referred to as liberal and conservative. On the one hand, a classifier may be thought as "liberal" when it makes positive classifications with weak evidence so it classify nearly all positives correctly, but it often has high false positive rates. On the other hand, it may be defined as "conservative" when it makes positive classifications only with strong evidence so it makes few false positive errors, but it often has low true positive rates as well [27]. In our case, the most liberal configuration is realized as follows. The fluorescence intensity classification system applies the RbA criterion.The staining pattern recognition system is based on the HCAF system in conjunction with the RbA rule to classify the pattern of individual cells. Then, it employs the WS criterion to label the staining pattern of the whole well.In particular, as system parameters we use those that provide the best well recognition rate (figure 5.9), i.e. 85.3%.

A conservative setup is carried out as follows. The fluorescence intensity

Table 5.11: Perspective performance of the overall CAD system in fluorescence intensity and staining pattern classifications when the most liberal setup is applied. The rates are computed on the basis of a priori probability distribution of IIF samples.

	Hit Rate	Miss Rate	Reject Rate
Positive Samples	78.6%	21.4%	0.0%
Negative Samples	98.9%	1.1%	0.0%
Border Zone Samples	92.3%	7.7%	0.0%
Total	89.5%	10.5%	0.0%

Table 5.12: Perspective performance of the overall CAD system in fluorescence intensity and staining pattern classifications when the most conservative setup is applied. The rates are computed on the basis of a priori probability distribution of IIF samples.

	Hit Rate	Miss Rate	Reject Rate
Positive Samples	67.2%	5.5%	27.3%
Negative Samples	89.4%	0.0%	10.6%
Border Zone Samples	85.0%	3.4%	11.6%
Total	80.0%	3.1%	16.9%

classification system employs the BA criterion. The system that recognizes the single cell staining pattern is based on the HCAF system in conjunction with the RbA rule. To label the staining pattern of the whole well, the WS criterion works with the reject option presented by equation 5.24, providing an hit, miss and reject rates of 76.6%, 5.8% and 17.6%, respectively.

The perspective results of the liberal and conservative setup are shown in Tables 5.11 and 5.12, respectively. They are computed on the basis of the a priori probabilities distribution of IIF samples (see section 5.1.6). In case of liberal configuration, the overall recognition rate is 90%, approximately, whereas in the conservative one it is 80%. Such a variation is essentially due to the introduction of reject options both at the stage of fluorescence intensity and staining pattern classification, respectively. Their use aims at lowering the misclassifications: indeed, the miss rate of the conservative configuration is one third of the corresponding one of the liberal setup, i.e. 3.1% vs. 10.5%. The side effect is that the 16.9% of samples are rejected. It is worth noting that the staining pattern classification influences only the recognition rate of positive samples. Therefore, the employment of a two stage recognition approach (figure 5.1) permits to achieve low false negative rate in both setups, as discussed for the results of the fluorescence intensity recognition (see section 5.1.6).

Besides the two configurations presented above, others should be used. However these two arrangements represent the most conservative and the most lib-

eral that can be set on the basis of the systems discussed in this work. The other setups present intermediate performance between such extrema.

The analysis of perspective performance so far shows that the use of CAD system in IIF has the potential for lowering the method variability, for increasing the standardization level and for reducing the specialists workload, e.g. if we assume that they diagnose only the rejected samples the reduction is more than 80%. Furthermore, the CAD system presented in this work can operate in all working scenarios described in Table 2.3, since it can acts both as a zero-reject and reject system as well as it can provide a reliability measurement of the decision.

Chapter 6

Conclusions

In this thesis, we have presented a CAD system tailored for application in the IIF field. This work has discussed the typical and different issues that regards the various functionalities of a CAD system, such as acquisition, processing, classification and management of data.

First, our results suggest that digital media is a reliable tool to help physicians in detecting autoantibodies in IIF. The data represent a first step to validate the use of digital images, thus offering an opportunity for standardizing and automatizing the detection of ANA by IIF. Indeed, the experiments show that performing the diagnosis by looking at digital images on a workstation monitor allows the specialists to better concentrate on sample examination. Furthermore, we have proposed a procedure that improves the reliability of the sample classification. It is based on a simplified version of the medical guidelines that, on the one hand, should maintain the diagnostic meaning of IIF test, and, on the other hand, allows getting a well-founded data set.

Second, we have proposed a complete autofocus procedure for automatic acquisition of IIF images. The procedure is effective in dealing with the peculiarities of these images and it can be easily implemented in a comprehensive acquisition system. Effectiveness of the autofocus procedure has been assessed on real images both quantitatively and qualitatively.

Since the diagnosis of IIF sample requires to estimate the fluorescence intensity and then, if the sample is positive, to determine the staining pattern, the third and fourth issue concerns the two recognition systems tailored to such aims. In order to develop the fluorescence intensity classification system, we initially addressed the key point of the feature extraction and selection, presenting three subsets of features that discriminate very well each class from the others. Then, we introduced a multi-dichotomy recognition system that aggregates three experts in the one-per-class framework. In this scenario, we have presented two aggregation rules and a novel parameter that evaluates the reliability of the final decision. Such an architecture has been experimentally evaluated on a IIF images dataset and exhibits very good performance, since the false positive and false negative rate approach zero in several cases.

Given an application domain, specified by the costs of a misclassification, a rejection and a right classification, we have performed a convenience analysis to find out where is convenient to use one of the two aggregation rules. Such an analysis reveals for what values of the cost coefficients one of the two rules

performs better than the other. Furthermore, the two criterions determine different regions in the convenience analysis plane, which can be complied with the different and peculiar objectives of a CAD system.

The fourth issue deals with the development of a system supporting the staining pattern classification of IIF slides diluted at 1:80 titer. To this end, we employ a two stages recognition approach: the first recognize the pattern of individual cells, whereas the second reconstruct the pattern of the whole well. This scheme provides a degree of redundancy that lowers the effect of cell misclassifications. The single cell classification system applies both a multi-dichotomy and a multi-experts technique. We have also discussed the results achieved using an error-reject option both at the level of single cell and whole well classification, respectively. The experimental evaluation shows good results in a field of laboratory medicine characterized by high variability. Furthermore, the application of an error-reject option allows varying the operating point of the system that therefore complies with the various objectives of a CAD system.

Finally, the analysis of perspective performance achieved combining the two recognition system, i.e. fluorescence intensity and staining pattern classifiers, shows that the use of CAD system in IIF has the potential for lowering the method variability, for increasing the standardization level and for reducing the specialists workload. Indeed, in the conservative configuration the miss rate is lower than the intra-laboratory variability. On these motivations, we deem that proposed system is suited for application in daily practice.

Appendix A

Software Organization

This appendix aims at describing the organization of the code that manages the classification tool. In this respect, chapter 5 have shown that both the fluorescence intensity and the staining pattern recognition systems are based on a Multi Dichotomies System (MDS) operating in the one-per-class framework. The system that recognizes the fluorescence intensity adopts a SCA architecture, i.e. each module is composed of a single classifier. The system that labels the staining pattern employs a HCAF scheme, i.e. each module is constituted by a Multi-Experts System (MES). In this case, the classifiers outputs are fused into a single decision applying the Weighted Fusion (WV) rule.

This appendix does not report the source code or the pseudo code, but provides its conceptual figures to present its course of operation.

The code, developed in Matlab©, is organized in several m-functions. The main one is named as `mds_classifier` (figure A.1). It first retrieves the data, i.e. the feature and the ground truth, and then acquires information about the setup of each binary module, such as classifier parameters, fusion rule, validation procedure etc.. This implies that the features have been previously extracted and selected. Next, the available samples are divided into training, test and training test set. The function `compute_index` returns k structures, one for each iteration of the k-fold cross validation method, which contains the indexes of the samples that constituted such sets. The ground truth, the features, the indexes and the classifiers parameters are given to the function `single_classifier` or `mes_classifier`, depending on the architecture of each binary module. Indeed, if the module employs a Single-Classifier-Aggregation (SCA) scheme the code calls the former function, whereas in case of use the Hybrid-Classifier-Aggregation-Fusion (HCAF) architecture the latter function is called. Both functions return the crisp labels, the module performance as well as the classification reliabilities. These data are then passed to the function `mds_aggregation` that applies the Binary-Aggregation (BA) or the Reliability-based-Aggregation (RbA) criterion to determine the final decision according to equations 5.3 and 5.4, respectively. It returns the final labels, the reliabilities (see formulas 5.10 and 5.11) and the measured performance of the overall system. Finally, it is possible to invoke the error-reject option, calling the function `error_reject_option`.

The description so far shows that the main function can call the `mes_classifier` or `single_classifier` functions on the basis of each binary module architec-

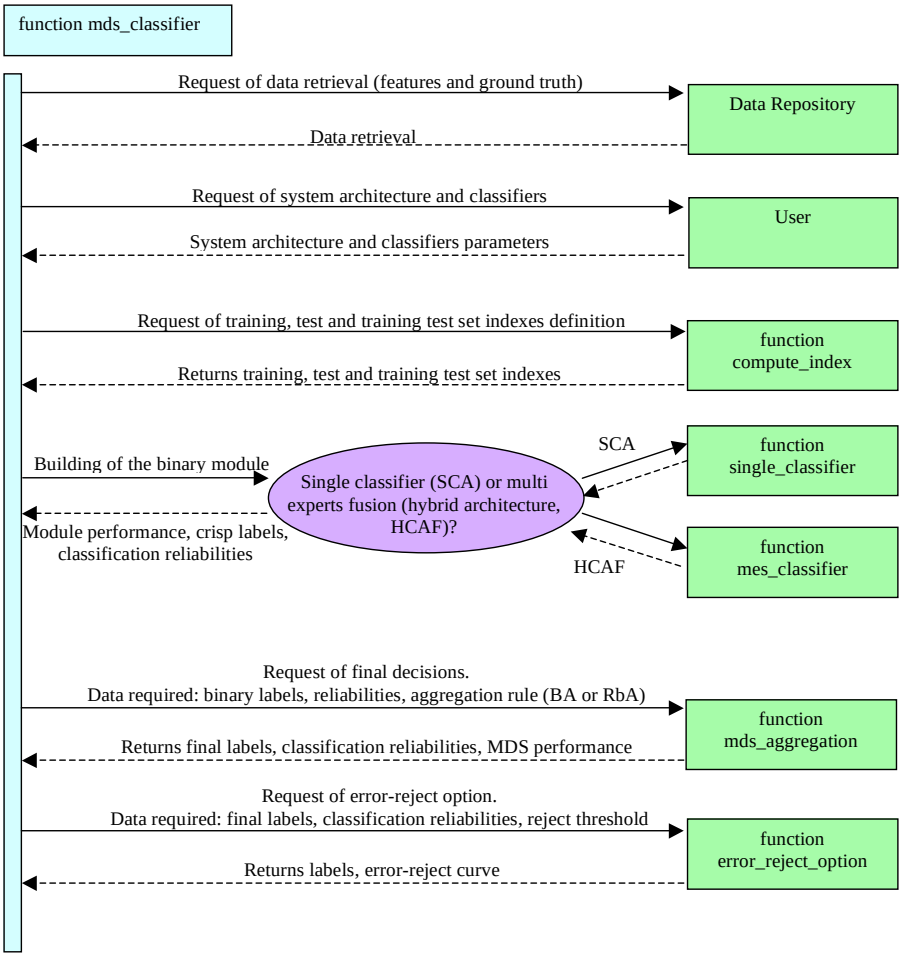

Figure A.1: Conceptual scheme of the function that realizes the Multi Dichotomies System (MDS).

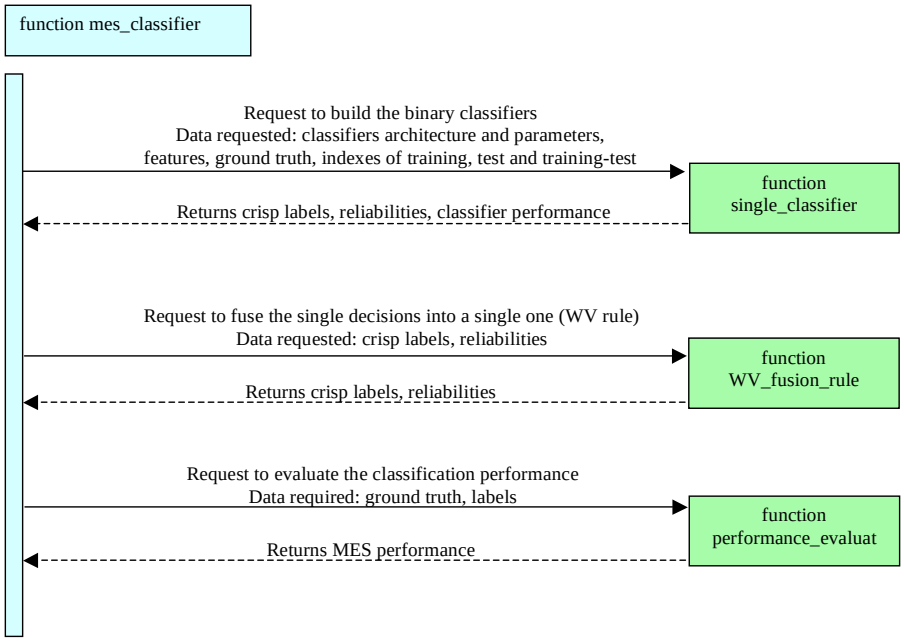

Figure A.2: Conceptual scheme of the function that realizes the Multi-Experts System (MES).

ture. In this respect, figures A.2 and A.3 depict their conceptual schemes. The former reports that the `mes_classifier` function passes data (i.e. ground truth, features, single classifier parameters, indexes of training, test and training test sets) to the `single_classifier` function in order to build the single expert that composes the Multi-Experts System (MES). The data returned by such a function are then given to `WV_fusion_rule` function that provides both the labels and the classification reliabilities, according to equations 5.17 and 5.20, respectively. Finally, the performance of the MES is evaluated.

Figure A.3 shows the conceptual scheme of function `single_classifier`, i.e. the function that implements the single classifiers. Initially, it normalizes the features in order to have zero mean and unit variances, applying the Matlab function `prestd`. Then, the samples label are dichotomized, i.e. the labels of the samples that belong to the class in which the module is specialized are set equal to 1, 0 otherwise. The function `make_classifier` builds, trains and tests the classifier specified by the corresponding parameters. This function than calls a different function for each possible classifier architecture. It returns the classifier outputs on training and test samples. Next, the function `crisp_label_comput` determines the crisp labels, whereas the function `reliability_label_comput` evaluates the reliability of each classification act according to formula 5.5. Finally, the function `performance_label_evaluat` measures the recognition performance of the classifier.

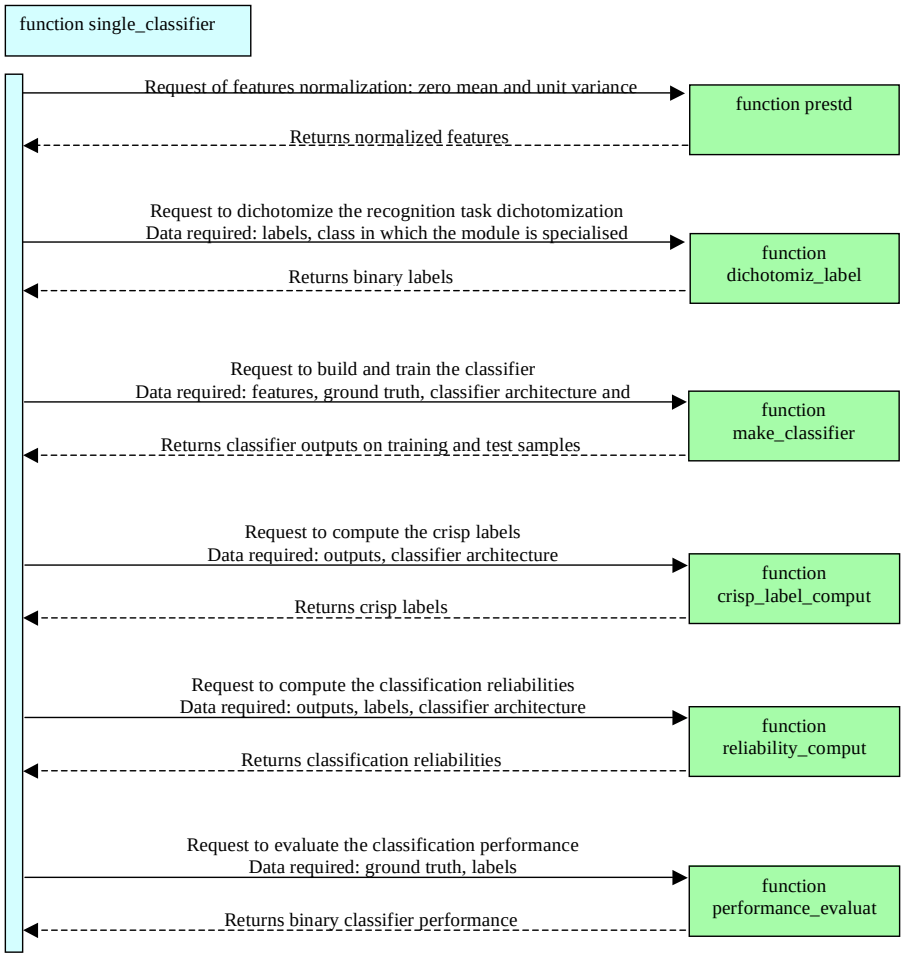

Figure A.3: Conceptual scheme of the function that realizes the single classifier.

Bibliography

[1] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2001.

[2] S. O. Belkasim, M. Shridhar, and M. Ahmadi. Pattern recognition with moment invariants: A comparative study and new results. *Pattern Recognition*, 24:1117–1138, 1991.

[3] S. K. Bhatia and J. S. Deogun. Conceptual clustering in information retrieval. *IEEE Transactions on Systems, Man, and Cybernetics*, 28(3):427–436, 1998.

[4] Bio-Rad Laboratories Inc. PhD System. http://www.bio-rad.com, 2004.

[5] I. Bloch. Information combination operators for data fusion: a comparativereview with classification. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 26(1):52–67, 1996.

[6] F. R. Boddeke, L. J. Van Vliet, H. Netten, and I. T. Young. Autofocusing in microscopy based on the OTF and sampling. *Bioimaging*, 2(4):193–203, 1994.

[7] M. V. Boland, M. K. Markey, and R. F. Murphy. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, 33(3):366–375, 1998.

[8] Center for Disease Control. Quality assurance for the indirect immunofluorescence test for autoantibodies to nuclear antigen (IF-ANA): approved guideline. *NCCLS I/LA2-A*, 16(11), December 1996.

[9] X. Chen, X. Zhou, and S. T. C. Wong. Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *Biomedical Engineering, IEEE Transactions on*, 53(4):762–766, 2006.

[10] H. D. Cheng, X. Cai, X. Chen, L. Hu, and X. Lou. Computer-aided detection and classification of microcalcifications in mammograms: a survey. *Pattern Recognition*, 36(12):2967–2991, 2003.

[11] J. Cohen. A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20:37–46, 1960.

[12] A. H. Coons and M. H. Kaplan. Localization of antigen in tissue cells ii. improvements in a method for the detection of antigen by means of fluorescent antibody. *Journal of Experimental Medicine*, 91(1):1–13, 1950.

[13] L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella, and M. Vento. Reliability parameters to improve combination strategies in multi-expert systems. *Pattern Analysis & Applications*, 2(3):205–214, August 1999.

[14] L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella, and M. Vento. A cascaded multiple expert system for verification. In *MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 330–339. Springer-Verlag, 2000.

[15] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2002.

[16] C. Dahle, T. Skogh, A. K. Aberg, A. Jalal, and P. Olcen. Methods of choice for diagnostic antinuclear antibody (ANA) screening. benefit of adding antigen-specific assays to immunofluorescence microscopy. *Journal of Autoimmunity*, 22(3):241–248, 2004.

[17] Das s.r.l. Service Manual AP16 IF Plus. Palombara Sabina (RI), March 2004.

[18] M. De Santo, F. Tortorella, M. Molinara, and M. Vento. Automatic classification of clustered microcalcifications by a multiple expert system. *Pattern Recognition*, 36:1467–1477, 2003.

[19] C. De Stefano, C. Sansone, and M. Vento. To reject or not to reject: that is the question: an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics–Part C*, 30(1):84–93, February 2000.

[20] O. Debeir, C. Decaestecker, J. L. Pasteels, I. Salmon, R. Kiss, and P. Van Ham. Computer-assisted analysis of epiluminescence microscopy images of pigmented skin lesions. *Cytometry*, 37(4):255–266, 1999.

[21] A. Delon, Y. Usson, J. Derouard, T. Biben, and C. Souchier. Photobleaching, mobility, and compartmentalisation: Inferences in fluorescence correlation spectroscopy. *Journal of Fluorescence*, 14(3):255–267, May 2004.

[22] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[23] A. P. Dhawan, Y. Chitre, C. Kaiser-Bonasso, and M. Moskowitz. Analysis of mammographic microcalcifications using gray-level image structure features. *IEEE Transactions on Medical Imaging*, 15(3):246–259, June 1996.

[24] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263, 1995.

[25] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2000.

[26] B. Fang, W. Hsu, and M. L. Lee. On the accurate counting of tumor cells. *Nanobioscience, IEEE Transactions on*, 2(2):94–103, 2003.

[27] T. Fawcett. ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 31, 2004.

[28] T. E. W. Feltkamp, F. Klein, and M. Janssens. Standardisation of the quantitative determination of antinuclear antibodies (ANAs) with a homogeneous pattern. *Annals of the Rheumatic Diseases*, 47(11):906–909, 1988.

[29] L. Firestone, K. Cook, K. Culp, N. Talsania, and K. Preston Jr. Comparison of autofocus methods for automated microscopy. *Cytometry*, 12(3):195–206, 1991.

[30] M. J. Fritzler. Immunofluorescent antinuclear antibody tests. *Manual of Clinical Laboratory Immunology,*, pages 733–739, 1986.

[31] K. Fukunaga and D. M. Hummels. Leave-one-out procedures for nonparametric error estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(4):421–423, April 1989.

[32] J. M. Geusebroek, F. Cornelissen, W. M. Arnold Smeulders, and H. Geerts. Robust autofocusing in microscopy. *Cytometry*, 39(1):1–9, 2000.

[33] G. Giacinto and F. Roli. Dynamic classifier selection based on multiple classifier behaviour. *Pattern Recognition*, 34(9):1879–1881, 2001.

[34] M. Golfarelli, D. Maio, and D. Maltoni. On the error-reject trade-off in biometric verification systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):786–796, 1997.

[35] F. C. A. Groen, I. T. Young, and G. Ligthart. A comparison of different focus functions for use in autofocus algorithms. *Cytometry*, 6(2):81–91, 1985.

[36] V. Gunes, M. Menard, P. Loonis, and S. Petit-Renaud. Combination, cooperation and selection of classifiers: a state of the art. *International Journal of Pattern Recognition and Artificial Intelligence*, 17(8):1303–1324, 2003.

[37] H. Hao, C. L. Liu, H. Sako, and T. Beijing. Confidence evaluation for combining diverse classifiers. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 760–765, 2003.

[38] T. Hastie and R. Tibshirani. Classification by pairwise coupling. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information & processing systems 10*, pages 507–513, Cambridge, MA, USA, 1998. MIT Press.

[39] P. Heinlein, J. Drexl, and W. Schneider. Integrated wavelets for enhancement of microcalcifications in digital mammography. *IEEE Transactions on Medical Imaging*, 22(3):402–413, 2003.

[40] R. Hiemann, N. Hilger, J. Michel, U. Anderer, M. Weigert, and U. Sack. Principles, methods and algorithms for automatic analysis of immunofluorescence patterns on HEp-2 cells. In *Autoimmunity Reviews*, page 86, 2006.

[41] R. Hiemann, N. Hilger, J. Michel, J. Nitscke, A. Bohm, U. Anderer, M. Weigert, and U. Sack. Automatic analysis of immunofluorescence patterns of HEp-2 cells. *Annals of the New York Academy of Sciences*, 1109(1):358–371, 2007.

[42] R. Hiemann, N. Hilger, U. Sack, and M. Weigert. Objective quality evaluation of fluorescence images to optimize automatic image acquisition. *Cytometry Part A*, 69:182–184, 2006.

[43] T. K. Ho, J. J. Hull, and S. N. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75, 1994.

[44] K. Huang and R. F. Murphy. Automated classification of subcellular patterns in multicell images without segmentation into single cells. In *Biomedical Imaging: Macro to Nano, 2004. IEEE International Symposium on*, pages 1139–1142, 2004.

[45] Y. S. Huang and C. Y. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):90–94, 1995.

[46] R. L. Humbel. Auto-immunité, auto-anticorps et maladie auto-immunes. *Autoanticorps et maladies autoimmunes*, pages 71–3, 1997.

[47] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[48] D. L. Jacobson, S. J. Gange, N. R. Rose, and N. M. H. Graham. Epidemiology and estimated population burden of selected autoimmune diseases in the united states. *Clinical Immunology and Immunopathology*, 84:223–243, 1997.

[49] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, January 2000.

[50] A. K. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

[51] J. Jelonek and J. Stefanowski. Experiments on solving multiclass learning problems by n^2 classifier. In *10th European Conference on Machine Learning*, pages 172–177. Springer-Verlag Lecture Notes in Artificial Intelligence, 1998.

[52] B. Julesz. Experiments in a visual perception of texture. *Scientific American*, 232:34–43, April 1975.

[53] A. Kavanaugh, R. Tomar, J. Reveille, D. H. Solomon, and H. A. Homburger. 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*, 124(1):71–81, 2000.

[54] J. M. Keller, P. D. Gader, S. Sohn, and C. W. Caldwell. Soft counting networks for bone marrow differentials. *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, 5, 2001.

[55] A. Khotanzad and Y. H. Hong. Invariant image recognition by zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):489–497, May 1990.

[56] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

[57] J. H. Klippel and P. A. Dieppe. *Rheumatology, 2nd edition*. Mosby International, March 1998.

[58] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–1145. Morgan Kaufmann, 1995.

[59] A. D. Kulkarni and P. Byars. Artificial neural network models for texture classification via the radon transform. In *SAC '92: Proceedings of the 1992 ACM/SIGAPP symposium on Applied computing*, pages 659–664, New York, NY, USA, 1992. ACM Press.

[60] L. I. Kuncheva. Switching between selection and fusion in combining classifiers: an experiment. *IEEE Transactions on Systems, Man and Cybernetics*, 32(2):146–156, 2002.

[61] L. I. Kuncheva. Switching between selection and fusion in combining classifiers: anexperiment. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 32(2):146–156, April 2002.

[62] L. I. Kuncheva. Using diversity measures for generating error-correcting output codes in classifier ensembles. *Patt. Recogn. Lett.*, 26(1):83–90, 2005.

[63] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34:299–314, 2001.

[64] L. Lam and C. Y. Suen. Optimal combinations of pattern classifiers. *Pattern Recognition Letters*, 16(9):945–954, 1995.

[65] J. R. Landis and G. G. Kock. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1997.

[66] N. Landwehr, M. Hall, and F. Eibe. Logistic model trees. *Machine Learning*, 59:161–205, 2005.

[67] Y.Ê LeCun, B.Ê Boser, J. S. Denker, D.ÊHenderson, R. E. Howard, W.ÊHubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[68] T. Li, Q. Li, S. Zhu, and M. Ogihara. A survey on wavelet applications in data mining. *ACM SIGKDD Explorations Newsletter*, 4(2):49–68, 2002.

[69] W. Lin, J. Xiao, and E. Micheli-Tzanakou. A computational intelligence system for cell classification. In *Information Technology Applications in Biomedicine, 1998. ITAB 98. Proceedings. 1998 IEEE International Conference on*, pages 105–109, 1998.

[70] X. Long, W. L. Cleveland, and Y. L. Yao. Automatic detection of unstained viable cells in bright field images using a support vector machine with an improved training procedure. *Computers in Biology and Medicine*, 36(4):339–362, 2006.

[71] R. Marcolongo, A. Ruffatti, and G. Morozzi. Presentazione linee guida del forum interdisciplinare per la ricerca sulle malattie autoimmuni (F.I.R.M.A.). *Reumatismo*, 55(2):9–21, 2003.

[72] F. Masulli and G. Valentini. Comparing decomposition methods for classication. In *KES'2000, Fourth International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies*, pages 788–791, 2000.

[73] E. Mayoraz and M. Moreira. On the decomposition of polychotomies into dichotomies. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 219–226, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[74] G. Méhes, T. Lörch, and P. F. Ambros. Quantitative analysis of disseminated tumor cells in the bone marrow by automated fluorescence image analysis. *Cytometry (Communications in Clinical Cytometry)*, 42:357–362, 2000.

[75] D. Milenovic, L. Stoiljkovic, and N. Stojanovic. Cell classification for diagnostic of reactive histocytichyperplasia using neural networks. In *Electrotechnical Conference, 1998. MELECON 98, 9th Mediterranean*, volume 2, 1998.

[76] T. Nakabayashi, T. Kumagai K. Yamauchi, M. Sugano, A. Kuramoto, K. Fujita, H. Hidaka, and M. Tozuka. Evaluation of the automatic fluorescent image analyzer, image titer, for quantitative analysis of antinuclear antibodies. *American Journal of Clinical Pathology*, 115(3):424–429, 2001.

[77] T. W. Nattkemper, H. J. Ritter, and W. Schubert. A neural classifier enabling high-throughput topological analysis of lymphocytes in tissue sections. *IEEE Transactions on Information Technology in Biomedicine*, 5(2):138–149, 2001.

[78] T. W. Nattkemper, T. Twellmann, H. Ritter, and W. Schubert. Human vs. machine: Evaluation of fluorescence micrographs. *Computers in Biology and Medicine*, 33(1):31–43, 2003.

[79] Kappa opto-electronics GmbH. Kameras DX4-40, 2006.

[80] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1970.

[81] P. Perner, H. Perner, and B. Muller. Mining knowledge for HEp-2 cell image classification. *Journal Artificial Intelligence in Medicine*, 26:161–173, 2002.

[82] B. N. Pham, S. Albarede, and P. Maisonneuve. Impact of external quality assessment on antinuclear antibody detection performance. *Lupus*, 14(2):113–119, February 2005.

[83] A. Piazza, F. Manoni, A. Ghirardello, D. Bassetti, D. Villalta, M. Pradella, and P. Rizzotti. Variability between methods to determine ANA, anti-dsDNA and anti-ENA autoantibodies: a collaborative study with the biomedical industry. *Journal of Immunological methods*, 219:99–107, August 1998.

[84] G. O. Reynolds, L. B. DeVelis, B. G. J. Parrent, and B. J. Thompson. *Physical optics notebook: tutorials in Fourier optics*. SPIE Optical Engineering Press, 1989.

[85] U. Sack, S. Knoechner, H. Warschkau, U. Pigla, F. Emmerich, and M. Kamprad. Computer-assisted classification of HEp-2 immunofluorescence patterns in autoimmune diagnostics. *Autoimmunity Reviews*, 2:298–304, 2003.

[86] C. Sansone, F. Tortorella, and M. Vento. A classification reliability driven reject rule for multi-expert systems. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(6):885–904, 2001.

[87] A. Santos, C. Ortiz De Solorzano, J. J. Vaquero, J. M. Pena, N. Malpica, and F. Del Pozo. Evaluation of autofocus functions in molecular cytogenetic analysis. *Journal of Microscopy*, 188(3):264–272, 1997.

[88] S. Seidenari, G. Pellacani, and P. Pepe. Digital videomicroscopy improves diagnostic accuracy for melanoma. *Journal of the American Academy of Dermatology*, 39:175–81, 1998.

[89] J. S. Smolen, G. Steiner, and E. M. Tan. Standards of care: the value and importance of standardization. *Arthritis & Rheumatism*, 40(3):410–2, 1997.

[90] P. Soda. Early experiences in the staining pattern classification of HEp-2 slides. In *Computer Based Medical Systems*, pages 219–224, Los Alamitos, CA, USA, 2007. IEEE Computer Society.

[91] P. Soda. Experiments on solving multiclass recognition tasks in the biological and medical domains. In *International Conference on Bio-inspired Systems and Signal Processing (Biosignals)*. In press, 2008.

[92] P. Soda and G. Iannello. Experiences in ANN-based classification of immunofluorescence images. *Enformatika Transactions on Engineering, Computing and Technology*, 14:252–257, August 2006.

[93] P. Soda and G. Iannello. A multi-expert system to classify fluorescent intensity in antinuclear autoantibodies testing. In *Computer Based Medical Systems*, pages 219–224, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

[94] P. Soda and G. Iannello. The relevance of computer-aided-diagnosis systems in microscopy applications to medicine and biology. *Encyclopaedia in Healthcare Information Systems - In press*, 2007.

[95] D. H. Solomon, A. J. Kavanaugh, and P. H. Schur. Evidence-based guidelines for the use of immunologic tests: Antinuclear antibody testing. *Arthritis Care & Research*, 47(4):434–444, 2002.

[96] L. Song, E. J. Hennink, I. T. Young, and H. J. Tanke. Photobleaching kinetics of fluorescein in quantitative fluorescence microscopy. *Biophysical Journal*, 68(6):2588–2600, June 1995.

[97] L. Song, C. A. Varma, J. W. Verhoeven, and H. J. Tanke. Influence of the triplet excited state on the photobleaching kinetics of fluorescein in microscopy. *Biophysical Journal*, 70(6):2959–2968, June 1996.

[98] C. Y. Suen and L. Lam. Multiple classifier combination methodologies for different output levels. In *MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 52–66, London, UK, 2000. Springer-Verlag.

[99] F. J. Theis, Z. Kohl, H. G. Kuhn, H. G. Stockmeier, and E. W. Lang. Automated counting of labelled cells in rodent brain section images. In *Proceedings BioMED 2004*, pages 209–212, 2004.

[100] R. Tozzoli, N. Bizzaro, E. Tonutti, D. Villalta, D. Bassetti, F. Manoni, A. Piazza, M. Pradella, and P. Rizzotti. Guidelines for the laboratory use of autoantibody tests in the diagnosis and monitoring of autoimmune rheumatic diseases. *American Journal of Clinical Pathology*, 117(2):316–324, 2002.

[101] G.V. Trunk. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(3):306–307, July 1979.

[102] O. Tuzel, L. Yang, P. Meer, and D. J. Foran. Classification of hematologic malignancies using texton signatures. *Pattern Analysis & Applications*, pages 1–14, 2007.

[103] B. Van Ginneken, B. M. Ter Haar Romeny, and M. A. Viergever. Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on Medical Imaging*, 20(12):1228–1241, December 2001.

[104] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991.

[105] C. A. Von Mühlen and E. M. Tan. Autoantibodies in the diagnosis of systemicrheumatic diseases. *Seminars in Arthritis and Rheumatism*, 24(5):323–358, 1995.

[106] S. Watanabe. *Pattern recognition: human and mechanical*. John Wiley & Sons, Inc. New York, NY, USA, 1985.

[107] B. Weyn, G. van de Wouwer, S. Kumar-Singh, A. van Daele, P. Scheunders, E. van Marck, and W. Jacob. Computer-assisted differential diagnosis of malignant mesothelioma based on syntactic structure analysis. *Cytometry*, 35(1):23–29, 1999.

[108] K. Woods, W. P. Kegelmeyer, and K. Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:405–410, April 1997.

[109] T. Würflinger, J. Stockhausen, D. Meyer-Ebrecht, and A. Böcking. Robust automatic coregistration, segmentation, and classification of cell nuclei in multimodal cytopathological microscopic images. *Computerized Medical Imaging and Graphics*, 28(1-2):87–98, 2004.

[110] W. Xiaoqiang, L. Xinmin, Z. Yiming, T. Deyuan, H. Xiaohai, and T. Qizhi. Autofocus methods for automated microscopy. *Proceedings of SPIE, the International Society for Optical Engineering*, 4224:114–117, 2000.

[111] L. Xu, A. Krzyzak, and C. Y. Suen. Method of combining multiple classifiers and their application to handwritten numeral recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435, 1992.

[112] S. Yamada, K. Komatsu, and T. Ema. Computer-aided diagnosis system for medical use, 1993. US Patent 5,235,510.

[113] R. Yamamura, T. Yamane, M. Hino, K. Ohta, H. Shibata, I. Tsuda, and N. Tatsumi. Possible automatic cell classification of bone marrow aspirate using the cell-dyn 4000® automatic blood cell analyzer. *Journal of Clinical Laboratory Analysis*, 16(2):86–90, 2002.

[114] X Ye, M. Cheriet, and C. Y. Suen. StrCombo: combination of string recognizers. *Pattern Recognition Letters*, 23(4):381–394, 2002.

[115] T. T. E. Yeo and S. H. Ong. Autofocusing for tissue microscopy. *Image and Vision Computing*, 11(10):629–639, 1993.

[116] I. T. Young. Quantitative microscopy. *Engineering in Medicine and Biology Magazine, IEEE*, 15(1):59–66, 1996.

[117] I. T. Young, R. Zagers, L. J. van Vliet, J. Mullikin, F. R. Boddeke, and H. Netten. Depth-of-focus in microscopy. In *Proceedings of the 8th Scandinavian Conference on Image Analysis (SCIA)*, volume 1, pages 493–498, 1993.