A classification framework for lung tissue categorization

Adrien Depeursinge[a], Jimison Iavindrasana[a], Asmâa Hidki[a], Gilles Cohen[a], Antoine Geissbuhler[a], Alexandra Platon[b], Pierre–Alexandre Poletti[b] and Henning Müller[a,c]

[a]University & Hospitals of Geneva, Service of Medical Informatics, 24, rue Micheli–du–Crest, CH–1211 Geneva 14, Switzerland;
[b]University & Hospitals of Geneva, Service of Emergency Radiology, 24, rue Micheli–du–Crest, CH–1211 Geneva 14, Switzerland;
[c]Business Information Systems, University of Applied Sciences Sierre, Switzerland;

ABSTRACT

We compare five common classifier families in their ability to categorize six lung tissue patterns in high–resolution computed tomography (HRCT) images of patients affected with interstitial lung diseases (ILD) but also normal tissue. The evaluated classifiers are *Naive Bayes*, *k–Nearest Neighbor* (*k*–NN), *J48 decision trees*, *Multi–Layer Perceptron* (MLP) and *Support Vector Machines* (SVM). The dataset used contains 843 regions of interest (ROI) of healthy and five pathologic lung tissue patterns identified by two radiologists at the University Hospitals of Geneva. Correlation of the feature space composed of 39 texture attributes is studied. A grid search for optimal parameters is carried out for each classifier family. Two complementary metrics are used to characterize the performances of classification. Those are based on McNemar's statistical tests and global accuracy. SVM reached best values for each metric and allowed a mean correct prediction rate of 87.9% with high class–specific precision on testing sets of 423 ROIs.

Keywords: quantitative image analysis, feature extraction, texture analysis, chest high–resolution CT, supervised learning, support vector machines.

1. INTRODUCTION

Interpreting high–resolution computed tomography (HRCT) images of the chest showing patterns associated with interstitial lung diseases (ILDs) is time–consuming and requires experience. ILDs are a heterogeneous group of around 150 illnesses of which many forms are rare and thus many radiologists have little experience. The diagnosis of ILDs is often established through the collaborations of the clinicians, radiologists and pathologists. Images play an important role and patients may not require surgical lung biopsy when the clinical and radiographic (HRCT) impression is consistent with a safe diagnosis of ILD[1]. The first imaging examination used is the chest radiograph because of its low cost and weak radiation dose. When the chest x–ray does not carry enough elements to finalize the diagnosis, HRCT is used to provide an accurate assessment of lung tissue patterns[2]. Computerized HRCT analysis can provide quick and precious information for emergency radiologists and other non–chest specialists[3,4]. Whereas the radiologists' ability to interpret HRCT data is likely to change based on the domain–specific experience, human factors and time of the day, computerized classification of lung tissue patterns is 100% reproducible. The computer–aided detection (CAD) system should be used as first reader in order to improve radiologists productivity and reduce reading fatigue[5,6]. One approach for building image–based computerized diagnostic aid for ILDs is to imitate the radiologists' human vision system. This latter can be schematized into two main parts:

- the eyes, which act as captors and aims at extracting relevant features from the observed scene[7],
- the visual cortex that takes decisions based on the pre–processed information provided by the eyes as input, as well as the knowledge and experience of the radiologist as information processor.

Further author information: (Send correspondence to Adrien Depeursinge)
Adrien Depeursinge: E–mail: adrien.depeursinge@sim.hcuge.ch, Telephone: +41 (0)22 372–8875

In pattern recognition, these two tasks can be respectively identified as feature extraction and supervised machine learning. The feature extraction part is based on texture properties and grey–level analysis (grey–level histograms) along with complementary analysis of spacial variations in the image through discrete wavelet frames with a quincunx subsampling scheme[8–10]. The two are described in Section 3.1. Texture properties have shown having high importance for medical image analysis in CADe systems[11]. In this paper, the supervised machine learning part is studied.

1.1. Supervised learning

Since the outputs of the CADe are the detected classes of lung tissue patterns, the machine learning task involved is a classification task. Once the feature space is built, algorithms have to be used to detect and create boundaries among the several classes of lung tissue patterns. This process is called supervised learning.

In order to classify unknown regions of interest (ROIs) of lung tissue, a model has to be built from known labeled data through the training phase. The training is challenging as it is partly based on experience of the radiologists. The goal is to find the functions $\mathcal{F}$ which modelize best the boundaries among the distinct classes of lung tissue patterns represented in the feature space. The best functions are those that achieve classification of a test set with the lowest error rate. The test set is composed of labeled ROIs, which have not been used to train the classifier. It simulates future unknown instances and thus allows to measure the generalization performance. Indeed, the objective is to minimize the error rate on the training set while at the same time avoiding overfitting of the training instances. Several approaches are available to implement $\mathcal{F}$. Three general approaches including five classifier families are studied in this paper:

- Learning by density estimation with *Naive Bayes* and k–Nearest Neighbor (k–NN) classifiers.
- Recursive partitioning of the feature space with $J48$ decision trees.
- Nonlinear numerical approaches with the Multi–Layer Perceptron (MLP) and kernel Support Vector Machines (SVM).

In practice, the choice of a classifier family is a difficult problem and it is often based on the classifier happening to be available, or best known to the user[12,13].

1.2. Classifier families

1.2.1. Naive Bayes

The Naive Bayes classifier is based on a probability model and assigns the class, which has the maximum estimated posterior probability to the feature vector extracted from the ROI. The posterior probability $P(c_i|\vec{v})$ of a class c_i given a feature vector $\vec{v}$ is determined using Bayes' theorem:

$$P(c_i|\vec{v}) = \frac{P(\vec{v}|c_i)P(c_i)}{P(\vec{v})} \tag{1}$$

This method is optimal when features are orthogonal, but in reality it works well without this assumption. The simplicity of the straightforward method allows good performance with small training sets[14]. Indeed, by building probabilistic models, it is robust to outliers (feature vectors that are not representative of the class to which they belong) and it creates soft decision boundaries which has the effect to avoid overtraining. However, the arbitrary choice of the distribution model for estimating the probabilities $P(x)$ along with the lack of flexibility of the decision boundaries results in limited performance for complex multiclass configurations.

1.2.2. k–NN

The k–nearest neighbor classifier cuts out hyperspheres in the space of instances by assigning the majority class of the k nearest instances according to a defined metric (e.g. Euclidean distance)[15] It is asymptotically optimal and its straightforward implementation allows rapid tests for example for evaluating features. However, several shortcomings inhere in this method. It is very sensitive to the curse of the dimensionality. Indeed, increasing the dimensionality has the effect to sparse the feature space and local homogeneous regions that represent the prototypes of the diverse classes are spread out. The classification performance strongly depends upon the used metric[14]. Moreover, a small value of k results in chaotic boundaries and makes the method very sensible to outliers.

1.2.3. $J48$ decision trees

The $J48$ decision trees algorithm divides the feature space successively by choosing primarily features with highest information gain[16]. $J48$ is an implementation of the $C4.5$ algorithm. In medicine, it is in correspondence to the approach used by clinicians to establish a diagnostic by answering successive questions. This is nevertheless not fully true while radiologists interpret HRCT images. This method is robust to noisy features because only those with high information gain are used. However it is sensitive to variability of data. The structure of the tree is likely to change completely when a new instance is added to the training set. Another drawback is its incapability to detect interactions between features as it treats those separately. This results in decision boundaries that are orthogonal to dimensions which is not accurate for highly nonlinear problems. Two main parameters which influences the generalization performance require optimization:

- $N_{instances}$: the minimum number of instances per leaf, which determines the size of the tree.
- $C_{pruning}$: the feature confidence factor used for pruning the tree, which consists in removing branches that are deemed to provide little or no gain in statistical accuracy of the model.

1.2.4. Multi–layer perceptron

The multi–layer perceptron (MLP) is inspired by the human nervous system, where information is processed through interconnected neurons[17]. The MLP is a feed–forward neural network, which means that the information propagates from input to output. The inputs are fed with values of each feature and the outputs are providing the class value. With one layer of neurons, the output is a weighted linear combination of the inputs. This network is called the linear perceptron. By adding an extra layer of neurons with nonlinear activation functions (the hidden layer), a nonlinear mapping between the input and output is possible[18]. The training phase consists in iterative optimization of the weights connecting the neurons by minimizing the mean squared error rate of classification. The learning rate R_{learn} which controls the adjustments of the weights during the training phase must be chosen as a trade–off between error on the training set and overtraining. Another critical parameter is the number of units N_{hidden} of the hidden layer. Indeed the MLP is subject to overfitting and requires optimal choice of the parameters for regularization. The MLP can create models with arbitrary complexity by drawing unlimited decision boundaries. It is also robust to noisy features as those will obtain low weight value after training.

1.2.5. Kernel support vector machines

Kernel support vector machines (SVMs) implicitly map input feature vectors $\vec{v_i}$ to a higher dimensional space by using the kernel function $K(\vec{v_i}, \vec{v_j}) = \langle \phi(\vec{v_i}), \phi(\vec{v_j}) \rangle$. For example, the Gaussian kernel is defined by:

$$K(\vec{v_i}, \vec{v_j}) = e^{\frac{-\|\vec{v_i} - \vec{v_j}\|^2}{2\sigma}} \tag{2}$$

with σ being the width of the Gaussian to determine. In the transformed space, a maximal separating hyperplane is built considering a two–class problem. Two parallel hyperplanes are constructed symmetrically on each side of the hyperplane that separates the data. The goal is to maximize the distance between the two external hyperplanes, called the margin[19,20]. An assumption is made that the larger the margin is the better the generalization error of the classifier will be. Indeed, SVMs were developed according to the *Structural Risk*

Minimization principle which seeks to minimize an upper bound of the generalization error, while most of the classifiers aims at minimizing the empirical risk, the error on the training set[21]. The SVM algorithm aims at finding a decision function $f(\vec{v})$, which minimizes the functional:

$$\min C \sum_{i}^{N} \max\left(0, 1 - y_i f(\vec{v}_i)\right)^2 + \|f\|_K \tag{3}$$

where N is the total number of feature vectors, $\|f\|_K$ is a norm in a *Reproducing Kernel Hilbert Space* $\mathcal{H}$ defined by the positive definite function K, which means that the functionals f are bounded. y_i is the label of $\vec{v_i}$ with $y_i \in \{-1; 1\}$ (2–class problem). The parameter C determines the cost attributed to errors and requires optimization. For the multiclass configuration, several SVM models are built using one versus one combinations. Finally, the majority class is attributed.

In summary, SVMs allow training generalizable, nonlinear classifiers in high–dimensional spaces using a small training set. This is enabled through the selection of a subset of vectors (called the support vectors) which characterizes the true boundaries between the classes well.

1.3. Classifiers used for lung tissue categorization in HRCT data

A brief review of the recent techniques used for the categorization of lung tissue patterns in HRCT data are described in this section.

In[22] normal versus fibrotic patterns are classified using SVMs with grey–level histograms, co–occurrence and run–length matrices. No details about the choices of the parameters of the SVMs are communicated. The small dataset used (9 HRCT image series) leads to a biased classification task since training and testing using series from the same patient create instances artificially close together in the feature space.

Non–linear binning of grey–level values for co–occurrence matrices is proposed in[23] in order to qualify lung tissue fibrosis in HRCT data. A minimum Mahalanobis distance classifier is used. This extended Naive Bayes classifier relies on the assumption that the probability density functions of the classes are Gaussian leading to non–flexible decision boundaries.

Two classifiers along with two feature selection techniques are evaluated in[24] through their ability of detecting fibrosis in HRCT images using co–occurrence matrices. The two are Naive Bayes and $J48$ decision trees. The feature selection technique showed to improve classification accuracy whereas two classifier families achieved equivalent performance. Still, the dataset used for testing is fairly small and the classifiers may not be flexible enough for multiclass problems. Information about the localization of the lung tissue patterns within a lung atlas is integrated as an additional feature in[25] which allow a classification accuracy improvement.

In[26], six lung tissue patterns are classified using an adaptive multiple texture feature method. Correlated features are removed and a Bayesian classifier is used. The latter may not be accurate for classifying any type of lung tissue as it is sensitive to the choice of the probability density function of the features.

Optimization of the parameters of SVMs with Gaussian kernels using a gradient descent is carried out in[27]. Quincunx Wavelet frames are used as texture features. The dataset used is small containing 22 images for 4 lung pattern classes. The optimization of the cost C of the errors as well as the width σ of the Gaussian kernel is carried out for each 2–class combination. The use of an anisotropic Gaussian kernel is tested in[28], and did not lead to significant improvement of the classification accuracy.

In[10] grey–level histograms with discrete wavelet frame features were evaluated using a k–NN classifier. In this paper, we evaluate the ability of 5 optimized common classifier families to discriminate among 6 lung tissue patterns characterized by improved quincunx wavelet frames texture features.

2. METHOD(S)

The dataset used is part of an internal multimedia database of ILD cases[29,30] containing HRCT images created in the Talisman project. The slice thickness of the images is limited to $1mm$. Annotation of regions is performed by two radiologists. Around 100 clinical parameters related to the 15 most frequent ILDs are acquired with each

Table 1. Distribution of the ROIs per class of lung tissue pattern.

	healthy	emphysema	ground glass	fibrosis	micronodules	macronodules
# of ROIs	113	93	148	312	155	22
# of patients	11	6	14	28	5	5

case. A graphical user interface implemented in Java was developed in order to meet the needs of the radiologists for the various annotation tasks. It allows high–quality annotations in 3D HRCT data. 843 regions of interest (ROIs) from healthy and five pathologic lung tissue patterns commonly found in HRCT images of the chest are selected for training and testing the classifiers (see Table 1). The selected patterns are *healthy*, *emphysema*, *ground glass*, *fibrosis*, *micronodules* and *macronodules*. Distributions of the classes are highly imbalanced as the largest class *fibrosis* contains 312 ROIs and the smallest class *macronodules* only 22 ROIs. There is a mean of 140.5 ROIs per class.

Classifier implementations were taken from the open source Java library *Weka*[31,32]. The feature extraction and the optimization of the classifier parameters were implemented in Java. The implementation of quincunx wavelet frames in Java is courtesy of the *Biomedical Imaging Group (BIG)* led by Prof. Michael Unser at the *Swiss Federal Institute of Technology* of Lausanne (*EPFL*). *LIBSVM* library is used for the SVMs' *C*–support vector classification[33].

3. RESULTS

3.1. Texture features

The construction of the feature space is detailed in this section. It is composed of image texture features as the taxonomy used by radiologists to interpret patterns in HRCT images often relates to texture properties[10]. The two feature groups are grey–level histograms, air components and quincunx wavelet frame coefficients with *B–spline* wavelets.

Full resolution (12–bit grey values) HRCT images are containing values in Hounsfield Units (H.U.) in the interval $[-1500; 1500]$. These values are corresponding univoquely to densities of the anatomic organs and thus allow the identification of lung tissue components. In order to take advantage of this, histograms of pixel values are computed over each ROI. Each bin value is integrated into the feature space and the optimal number was investigated in[10] where 40 bins constituted the best trade–off between classification accuracy and dimensionality of the feature space. 22 bins corresponding to pixel values in $[-1050; 600[$ were kept as the bins outside this interval were very sparsely populated. The air component value given by the number of pixels with value < -1000 H.U. is computed as an additional feature.

In order to study the spatial organization of the pixels quincunx wavelet frame (QWF) coefficients are extracted from the ROIs. Discrete wavelet frames have shown to perform well for texture analysis[8]. Compared to the wavelet transform, wavelet frames are redundant and offer more flexibility for image analysis: they enable translation–invariance by removing the subsampling part of the algorithm. A quincunx subsampling scheme is used in order to allow a finer scale progression compared to classical dyadic wavelet transform[9] (images are downsampled by a factor of $\sqrt{2}$ instead of 2 at each iteration). Moreover, its isotropy is suitable for analysis of axial images of the lung tissue as we made the assumption that no information is contained in directionality of patterns. The mean μ_i and variance σ_i of the coefficients of 8 iterations of QWF are computed over each ROI, resulting in 16 QWF features.

The feature space contains a total of 39 attributes that are normalized in order to give equivalent weight to each of them. The correlation of the values is showed in Figure 1.

3.2. Classifier family evaluation

The methodology utilized to compare the performance of each classifier family is described in this section. The full dataset (843 ROIs) is divided into two equal parts: 50% for training and 50% for testing. Training means

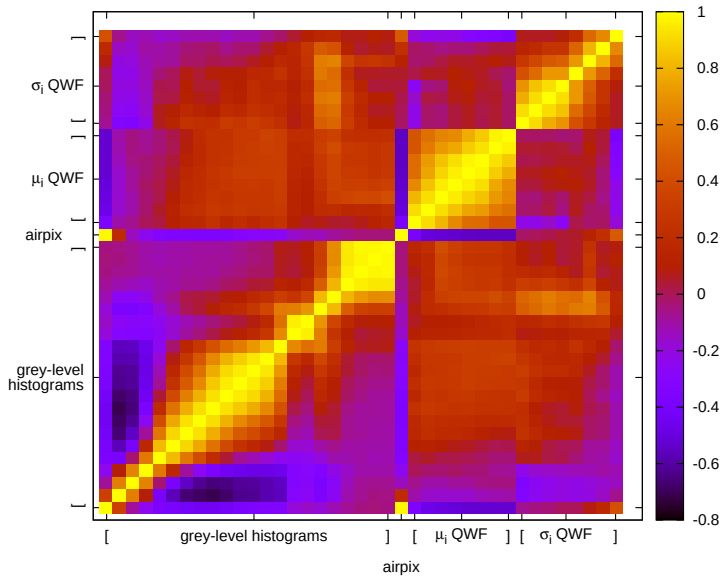

Figure 1. Correlation matrix of the feature space.

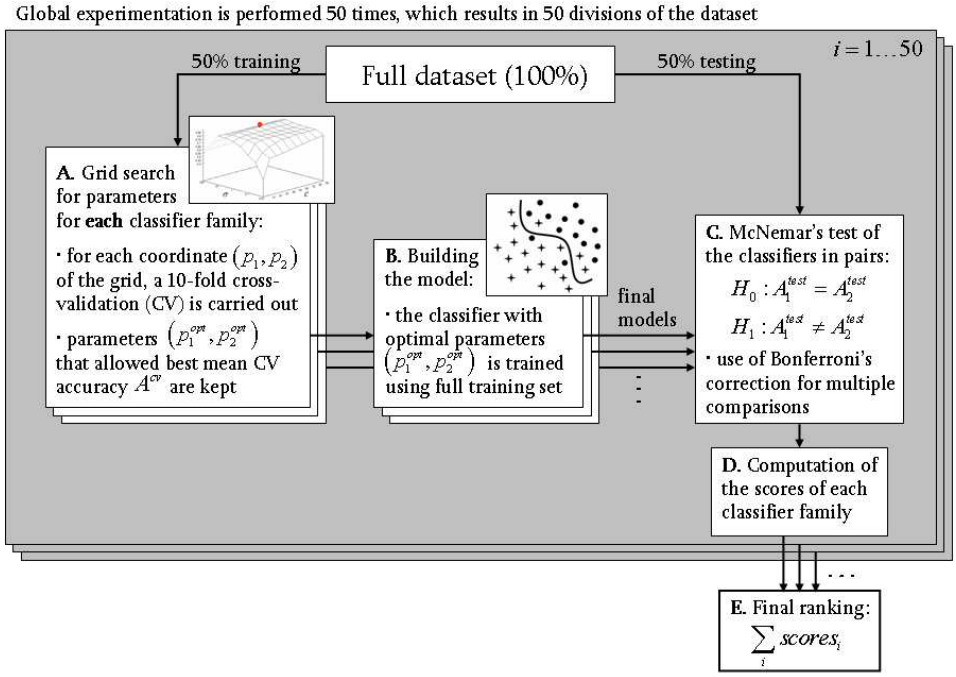

Figure 2. Methodology for benchmarking the classifiers.

both search for optimal parameters and creation of the model (i.e. adjustments of the decision boundary). The methodology is detailed in Figure 2 and in sections 3.2.1 and 3.2.2. .

Table 2. Grid search for optimal parameters p_i^{opt}. The values for the number of hidden layer units N_{hidden} of the MLP are chosen as {none, number of classes, (number of attributes + number of classes)/2, number of attributes + number of classes}.

classifier family	parameters	ranges	step
Naive Bayes	-	-	-
k–NN	k	$[0; 100]$	linear
$J48$	$N_{instances}$, $C_{pruning}$	$[0; 5]$, $[0.02; 0.24]$	lin, lin
MLP	R_{learn}, N_{hidden}	$[10^{-10}; 10^{5}]$, $\{0, 6, 22, 45\}$	log, -
SVM	C, σ	$[1; 100]$, $[10^{-2}; 10^{2}]$	lin, log

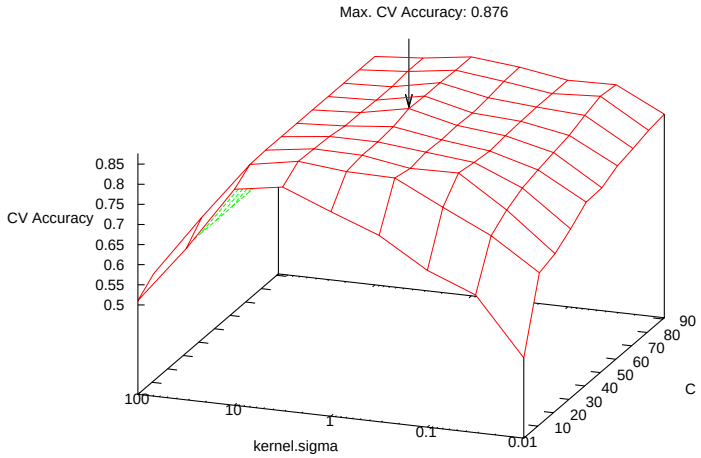

Figure 3. Grid search for SVM optimal parameters C and σ.

3.2.1. Grid search for optimal parameters

In order to determine the optimal parameters p_i, a grid search is performed for each classifier family. When required, exponential grid steps are used for coarse search. For every coordinate of the grid, a 10–fold cross–validation (CV) is carried out on the training set. Optimal parameters p_i^{opt} that allowed best mean CV accuracy A^{cv} are used to train the final model on the entire training set. Optimized parameters are detailed in Table 2. An example of grid search for best A^{cv} is shown in Figure 3 where the cost C and the σ value of the Gaussian kernel of the SVM are optimized. A preliminary coarse grid search is performed to locate regions of the space with high A^{cv} values.

3.2.2. Ranking

Instances of the test set are classified by each classifier family and McNemar's test is applied to the classifiers in pairs with the hypothesis:

$$H_0 : A_1^{test} = A_2^{test}$$
$$H_1 : A_1^{test} \neq A_2^{test}$$

with $A_{1,2}^{test}$ the testing accuracy of the classifiers 1,2 computed as the number of correctly classified instances divided by the total number of instances in the test set. Compared to other statistical tests for comparing supervised classification learning algorithms, McNemar's test showed to be the only test with acceptable type I error rate in[34]. Type I errors correspond to a false detection of difference in performance between two algorithms. Bonferroni's correction for multiple comparisons is used to adjust the threshold of the test. When H_0 is rejected

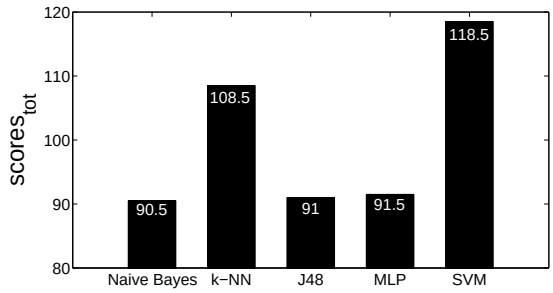

Figure 4. Final ranking based on the total of the scores of the classifiers. SVM reached the best score with 118.5.

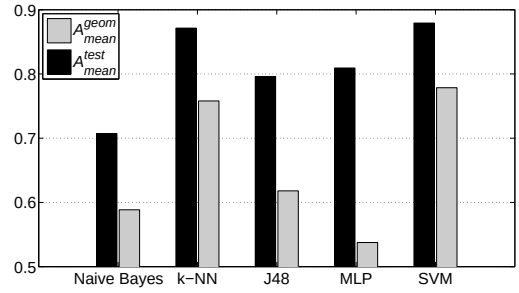

Figure 5. Overall mean testing accuracies A^{test}_{mean} and A^{geom}_{mean}. SVM reached best accuracies with $A^{test}_{mean} = 87.9\%$ and $A^{geom}_{mean} = 77.9\%$.

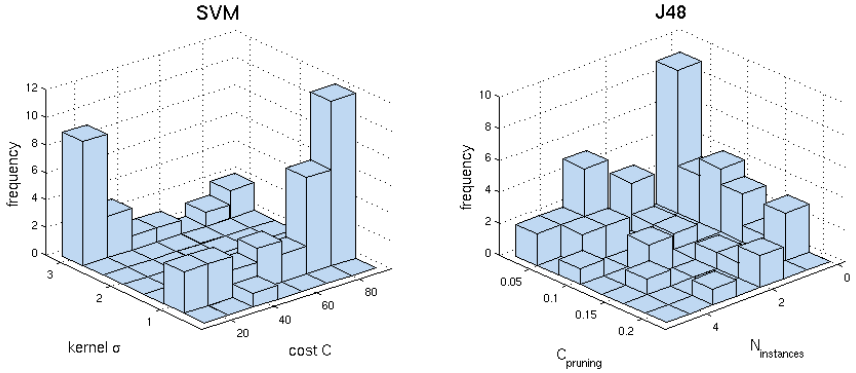

Figure 6. Bivariate histograms of the optimal parameters (p_1^{opt}, p_2^{opt}) for SVM and $J48$.

and A_1^{test} is greater than A_2^{test}, the score of the classifier 1 is incremented. When H_0 is accepted, 0.5 is added to the scores of both classifiers. The global experimentation is repeated 50 times and a final ranking based on the total of the scores is performed. As distribution of the classes are highly imbalanced, the geometric mean A^{geom} of each class–specific accuracy A^{c_i} on test set are computed for every classifier as follows:

$$A^{geom} = \sqrt[6]{\prod_{i=1}^{6} A^{c_i}} \tag{4}$$

A^{geom} gives the same importance to each class, even if the classes are imbalanced[35]. Final ranking, mean testing accuracies A^{test}_{mean} and mean geometric accuracies A^{geom}_{mean} are shown in Figures 4 and 5.

3.3. Stability

Optimal parameters of each classifier family were stored for every 50 : 50 division. In order to study the stability, histograms of the values of (p_1^{opt}, p_2^{opt}) are built for SVMs and $J48$ as shown in Figure 6.

4. INTERPRETATION

4.1. Feature space

In the correlation matrix (see Figure 1) of the feature space, three groups of features clearly appear as little correlated: the grey–level histograms, the mean μ_i of QWF and the variance σ_i of QWF. Bins of grey–level histograms are highly correlated in pairs, which is in accordance with the assumption that the density of the tissue extracted from the same ROI is roughly homogeneous. However, one can differentiate 2 subgroups: the 14^{th} first bins with values in $[-1050; 0]$ corresponding to various lung tissue patterns, the 15^{th} to 22^{nd} bins

with values in $]0;600]$ which corresponds to higher density tissue (i.e. vascular tissue). Features in this second subgroup are highly correlated due to sparsity. It is not surprising that the first bin is highly correlated with *airpix*. Means of QWF are anti–correlated with *airpix*, which is in accordance with the fact that regions with air are homogeneous (i.e. *emphysema* and interior of bronchus). Little correlation among the three groups of features suggests that the feature space contains little redundancy and is adapted to describe lung tissue texture.

4.2. Classifier performances

All scores are shown in Figure 4 resulting in strong variations among the classifiers. Moreover, the variations can be decreased by the use of Bonferroni's correction, which makes the tests more permissive (i.e. McNemar's tests rejects more easily H_0). Two classifier families reach scores out of the lot: k–NN and SVM. These performances are confirmed by their respective accuracies in Figure 5. Overall scores and mean testing accuracies A_{mean}^{test} show to be complementary metrics. For example, the MLP reaches high global accuracy of 81% with a low score of 91.5. Those discordances can be understood by looking at the mean geometric accuracies A_{mean}^{geom}. The latter is very low for the MLP with a value of 53.8% which indicates that the MLP has a very low class–specific accuracy, and thus a low precision for each class, which is not suitable for the characterization of lung tissue. Therefore, the SVM that reached best score and global accuracy is able to classify tissue of each class accurately, even from those that are little represented. Beyond the fact that the k–NN classifier reached a slightly lower score and global accuracy compared to SVM, one problem occurs with this classifier. The optimal number k of nearest neighbors for each of the 50 training/testing splits was 1. This strong tendency can be explained by the fact that for some classes, the number of patients is low and thus many ROIs are extracted from the same image series. Training and testing with images from the same image series can result in a biased classification as images are similar as they belong to the same patient. Two such instances are artificially close in the feature space and will facilitate the classification task while attributing the class of the closest neighbor, which probably belongs to the same image series. In that sense, the k–NN classifier carries out overfitting of the training instances which is not suitable for classifying ROIs from new ILD cases.

Distributions of the optimal parameters (p_1^{opt}, p_2^{opt}) represented in Figure 6 shows distinct behavior for SVM and $J48$. Coupled parameters are more uniformly distributed for $J48$ compared to SVM: σ of the Gaussian kernel of SVM is characterized by a bimodal distribution. This means that two values of σ allow a convenient mapping of the feature space to higher dimensions for accurate separation of the classes. These values affect the optimal value of cost C. Indeed, the organization of the classes in the transformed space are fixed by the value of σ, which requires a corresponding readjustment of the optimal cost C. The most frequent pair of values of $J48$ occurs 9 times over 50, while the second most frequent pair occurs 5 times. For the SVM, the most frequent pair occurs 12 times over 50 while the second most frequent pair occurs 9 times. In that sense, the SVM classifier offers more stability. The stability has an important influence on the generalization performance: a classifier that frequently obtained identical pairs of optimal parameters has a high probability to be optimal for classifying new data.

5. CONCLUSIONS

In this paper, 5 common classifier families were tested to discriminate 6 classes of lung tissue patterns in HRCT data from healthy cases and cases affected with ILDs. Evaluation of the classifiers is based on a high–quality dataset taken from clinical routine. The classifiers were optimized in order to compare their best performance. The SVM classifier constitutes the best trade–off between the error rate on the training set and generalization, the ability to classify ROIs correctly from images of new patients. Since SVMs were designed to avoid overfitting of training samples, using them to classify medical images with much heterogeneity is adapted. The SVM classifier was able to correctly classify 87.9% of the instances into the 6 classes. Two metrics were used to characterize the performances of the classifiers: scores based on Mc Nemar's test along with global accuracy on the test set. The two metrics have shown to be complementary. The optimal classification algorithms were integrated into a software for classification of ROIs directly in three–dimensional DICOM images (Figure 7). The diagnostic aid tool is easy to integrate into the PACS having the same user interface and offers the possibility to add clinical data from the electronic patient record. The classifier belongs to the core of a computer–aided diagnosis system involved in the decision making process.

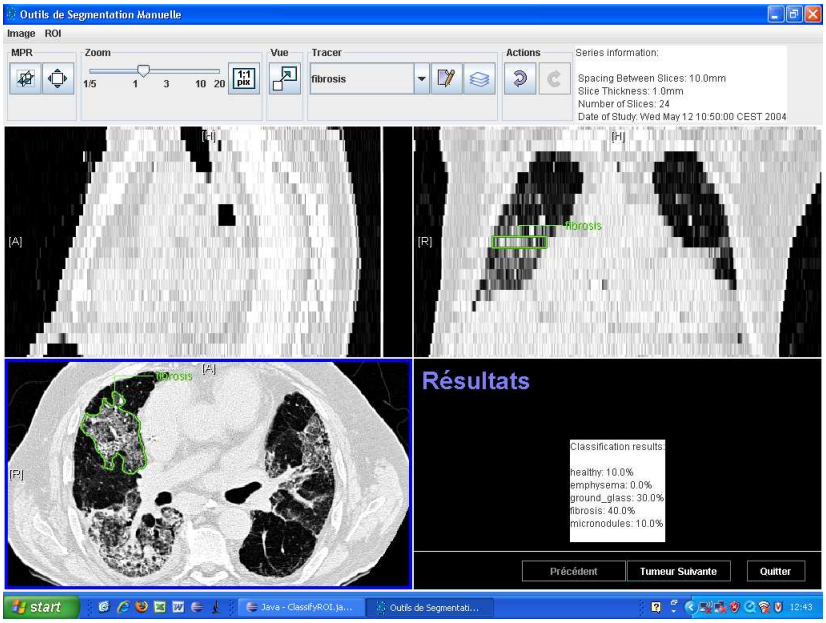

Figure 7. A screenshot of the DICOM viewer for the classification of image regions.

6. ACKNOWLEDGEMENTS

We thank Dr. Mélanie Hilario for her valuable comments on the methodology for benchmarking the classifiers. This work was supported by the Swiss National Science Foundation (FNS) with grant 200020–118638/1, the equalization fund of University and Hospitals of Geneva (grant 05–9–II) and the EU 6^{th} Framework Program in the context of the KnowARC project (IST 032691).

REFERENCES

1. K. R. Flaherty, T. E. King, J. Ganesh Raghu, J. P. Lynch III, T. V. Colby, W. D. Travis, B. H. Gross, E. A. Kazerooni, G. B. Toews, Q. Long, S. Murray, V. N. Lama, S. E. Gay, and F. J. Martinez, "Idiopathic interstitial pneumonia: What is the effect of a multidisciplinary approach to diagnosis?," *American Journal of Respiratory and Critical Care Medicine* **170**, pp. 904–910, July 2004.
2. P. Stark, "High resolution computed tomography of the lungs," *UpToDate* **September**, 2007.
3. C.-R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick, "ASSERT: A physician–in–the–loop content–based retrieval system for HRCT image databases," *Computer Vision and Image Understanding (special issue on content–based access for image and video libraries)* **75**, pp. 111–132, July/August 1999.
4. A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C.-R. Shyu, and A. Marchiori, "Automated storage and retrieval of thin–section CT images to assist diagnosis: System description and preliminary assessment," *Radiology* **228**, pp. 265–270, 2003.
5. R. M. Nishikawa, "Current status and future directions of computer-aided diagnosis in mammography," *Computerized Medical Imaging and Graphics* **31**, pp. 224–235, June 2007.
6. H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content–based image retrieval systems in medicine – clinical benefits and future directions," *International Journal of Medical Informatics* **73**, pp. 1–23, 2004.
7. I. Biedermann, "Recognition–by–components: A theory of human image understanding," *Psychological Review* **94 No 2**, pp. 115–147, 1987.
8. M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Transactions on Image Processing* **4**(11), pp. 1549–1560, 1995.

9. D. Van De Ville, T. Blu, and M. Unser, "Isotropic polyharmonic B-Splines: Scaling functions and wavelets," *IEEE Transactions on Image Processing* **14**, pp. 1798–1813, November 2005.
10. A. Depeursinge, D. Sage, A. Hidki, A. Platon, P.-A. Poletti, M. Unser, and H. Müller, "Lung tissue classification using wavelet frames," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, (Lyon, France), August 2007.
11. G. D. Tourassi, "Journey toward computer-aided diagnosis: Role of image texture analysis," *Radiology* **213**, pp. 317–320, July 1999.
12. 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)**, pp. 4–37, 2000.
13. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, August 2006.
14. C. van der Walt and E. Barnard, "Data characteristics that determine classifier performance," in *Proceedings of the Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa*, pp. 166–171, (Parys, South Africa), November 2006.
15. T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on* **13**(1), pp. 21–27, 1967.
16. J. R. Quinlan, "Induction of decision trees," *Machine Learning* **1**, pp. 81–106, March 1986.
17. C. M. Bishop, *Neural networks for pattern recognition*, Clarendon Press, Oxford, 1995.
18. A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: a tutorial," *Computer* **29**(3), pp. 31–44, 1996.
19. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery* **2**(2), pp. 121–167, 1998.
20. V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, November 1999.
21. G. Cohen, M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler, "Learning from imbalanced data in surveillance of nosocomial infection," *Artificial Intelligence in Medicine* **37**, pp. 7–18, May 2006.
22. J. J. Caban, J. Yao, N. A. Avila, J. R. Fontana, and V. C. Manganiello, "Texture–based computer–aided diagnosis system for lung fibrosis," in *SPIE Medical Imaging*, **6514**, (San Diego, CA, USA), February 2007.
23. V. Zavaletta, B. J. Bartholmai, and R. A. Robb, "Nonlinear histogram binning for quantitative analysis of lung tissue fibrosis in high–resolution CT data," in *SPIE Medical Imaging*, (San Diego, CA, USA), February 2007.
24. J. S. J. Wong and T. Zrimec, "Classification of lung disease pattern using seeded region growing," in *Australian Conference on Artificial Intelligence*, pp. 233–242, 2006.
25. T. Zrimec and J. Wong, "Improving computer aided disease detection using knowledge of disease appearance.," *Studies in Health Technology and Informatics* **129**, pp. 1324,1328, 2007.
26. R. Uppaluri, E. A. Hoffman, M. Sonka, P. G. Hartley, G. W. Hunninghake, and G. McLennan, "Computer recognition of regional lung disease patterns," *American Journal of Respiratory and Critical Care Medicine* **160**, pp. 648–654, August 1999.
27. A. Shamsheyeva and A. Sowmya, "Tuning kernel function parameters of support vector machines for segmentation of lung disease patterns in high-resolution computed tomography images," in *SPIE Medical Imaging*, **5370**, pp. 1548–1557, May 2004.
28. A. Shamsheyeva and A. Sowmya, "The anisotropic gaussian kernel for SVM classification of HRCT images of the lung," in *Proceedings of the 2004 Intelligent Sensors, Sensor Networks and Information Processing Conference*, pp. 439–444, December 2004.
29. A. Depeursinge, H. Müller, A. Hidki, P.-A. Poletti, T. Rochat, and A. Geissbuhler, "Building a library of annotated pulmonary CT cases for diagnostic aid," in *Swiss conference on medical informatics (SSIM 2006)*, (Basel, Switzerland), April 2006.
30. A. Depeursinge, H. Müller, A. Hidki, P.-A. Poletti, A. Platon, and A. Geissbuhler, "Image–based diagnostic aid for interstitial lung disease with secondary data integration," in *SPIE Medical Imaging*, (San Diego, CA, USA), February 2007.
31. I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Series in Data Management Sys, Morgan Kaufmann, second ed., June 2005.

32. E. Frank, M. A. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg, "Weka - a machine learning workbench for data mining," in *The Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, eds., pp. 1305–1314, Springer, 2005.
33. C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001.
34. T. G. Dietterich, "Approximate statistical test for comparing supervised classification learning algorithms," *Neural Computation* **10**(7), pp. 1895–1923, 1998.
35. M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proc. 14th International Conference on Machine Learning*, pp. 179–186, Morgan Kaufmann, 1997.