COMPUTER AIDED STAGING OF LYMPHOMA PATIENTS WITH FDG PET/CT IMAGING BASED ON TEXTURAL INFORMATION

Carole Lartizien[1], Matthieu Rogez[1], Adeline Susset[1], Francesco Giammarile[2,3], Emilie Niaf[1], Fabien Ricard[2,3]

[1] Université de Lyon, CREATIS; CNRS UMR5220; Inserm U1044; INSA-Lyon; Univ. Lyon 1, France.
[2] Université de Lyon, EMR 3738 ; Université Lyon 1, France.
[3] Hospices Civils de Lyon, Département de Médecine Nucléaire ; Hôpital Lyon Sud, France.

ABSTRACT

We have designed a computer aided diagnosis (CADx) system to assess the presence of cancer in FDG PET/CT exams of lymphoma patients. Detection performances of the random decision forest (RDF) and support vector machine (SVM) classifiers were assessed based on a feature set including 115 PET and CT first order and textural parameters. An original feature selection method based on combining different filter methods was proposed. The evaluation database consisted of 156 lymphomatous (M for malignant), 158 physiologic (N for normal) and 32 inflammatory (NS for normal suspicious) regions of interest. An optimization study was performed for each classifier separately to select the best combination of parameters considering the two problems of discriminating the {M} and {NS+N} classes and the {M} and {NS} classes. Promising classification performance was achieved by the SVM combined with the 12 most discriminant features with AUC values of 0.97 and 0.91 for the first and second problem respectively.

Index Terms— Positron emission tomography (PET), classification, CAD, image texture analysis.

1. INTRODUCTION

The initial staging and response assessment of malignant lymphoma remain difficult and repetitive tasks for radiologists despite the development of diagnostic imaging modalities such as FDG PET/CT and contrast CT scans. The recommendations published in 2007 by the International Harmonization Project (IHP) imaging subcommittee in lymphoma [1] underlined that the contrast-enhanced diagnostic CT scan could be performed at initial staging using an integrated PET/CT system. The ongoing standardized evaluation protocol in lymphoma consists in selecting the 6 most significant targets based on their abnormal shape and size derived from the CT images, corresponding to FDG avid lesions on the PET images. The interim or end-of-therapy PET/CT scans are then designated positive or negative based on IHP criteria including the morphologic and metabolic responses of the target sites. One main limitation of this protocol is that the analysis does not always account for all abnormal regions. Indeed, lymphoma can have a spread presentation so that taking into account every pathological site would be overly time consuming for daily practice. Some cancer locations may thus be discarded from the follow-up global evaluation, potentially leading to diagnostic discordances based on the baseline targets choice. Another limitation is that radiologists choose and characterize the targets on the baseline PET/CT scan and then, manually find their corresponding locations on the follow-up PET/CT scan, which may be difficult, especially if the targets are partially or fully responding. Moreover, this follow-up analysis can occasionally be performed without visualizing the baseline scan, but only based on the recorded notes of the location, size and uptake values of the pre-defined targets. A recent study [2] showed that FDG PET/CT for early response assessment in lymphoma had a poor predictive value based on the IHP criteria and motivated the need to develop and validate more sophisticated prognostic criteria.

Radiologists may benefit from improved statistical image processing methods to help them facing the difficult and repetitive diagnostic cases. Computer aided detection (CAD) systems are well suited for that purpose. While such systems have been widely studied for mammography, CT and MRI, there have been only a few studies in PET and PET/CT oncology imaging [3-5]. We have recently designed a CADe system that performs guided localization and classification of small and low-contrast lesions in 3D PET imaging [6]. This CAD system was shown to significantly outperform human detection sensitivities, especially regarding the low-contrast lesions. The number of false detections per image, however, remains too high. This study aims at improving this performance by accounting for textural information that has been shown promising in

discriminating benign from malignant lesions in PET imaging [3]. We focus on prototyping a CADx system, where the radiologist interviews the system on pre-selected suspicious areas of interest, thus avoiding the challenging task of localization and classification of abnormalities on a per voxel basis.

2. DATABASE DESCRIPTION

2.1. Study population

Twenty-five patients (11 males and 14 females) aged 20 to 79 year-old at diagnosis were included. They had undergone enhanced CT and FDG PET in an integrated enhanced FDG PET/CT exam at baseline for aggressive lymphoma (B cell lymphoma or Hodgkin disease).

2.2. FDG PET/CT protocol

Enhanced FDG PET/CT exams were performed in daily practice conditions, meaning fasting patients with blood glucose level check prior to the intravenous injection of 5.2 MBq/kg of ^{18}F-FDG. PET/CT acquisition was performed approximately 60 minutes after the ^{18}F-FDG injection, using a Gemini PET/CT device (Gemini PET-CT camera Philips Medical Systems). Patient workup included a cervical–to–upper-thigh transmission contrast-enhanced CT scan with normal but shallow breathing (2 rows of detectors, 120kV, 100mAs, rotation: 0.5 s, slice: 3.2mm, increment: 2.5, pitch: 1.35, injection of 90-120mL of contrast agent containing 350 mg I/mL - Iobitridol, Xenetix 350- at a rate of 3mL/s). This was followed by a 3-dimensional PET scan (emission data, 3 minutes per bed position, 6-9 bed positions, bed length 18cm, overlap 9.6 cm, field of view 57.6cm). PET data was reconstructed with CT-based attenuation correction using an iterative algorithm (RAMLA).

2.3. Preprocessing and annotation

The PET images were linearly interpolated to the same voxel size as the CT images and normalized to SUV values to mimic the common clinical practice. Three-dimensional regions of interest (ROIs) were manually drawn on the PET images by a nuclear physician then reproduced and adjusted if necessary on the CT images to fit the anatomical contours. They were categorized as lymphomatous disease sites (nodal and extranodal sites), inflammatory non-lymphomatous sites and physiologic areas (lung, liver, spleen, bone, muscle, physiologic node, mammary gland) based on experts interpretation, patients follow-up, and biopsy results when available. A total of 156 lymphomatous ROIs ({M}), 158 physiologic ROIs ({N}) and 32 inflammatory non-lymphomatous ROIs ({NS}) were analyzed.

3. DESCRIPTION OF THE CAD SYSTEM

This paper focuses on a two-class classification problem (lesion-absent and lesion-present). The first step consists in extracting a vector V of some geometrical and textural features for each ROI. These features are the input data of a supervised classifier which outputs a malignancy score. An original feature selection protocol combining different filter methods was proposed to select the most discriminant features before the classification step.

3.1. Feature extraction

One hundred and fifteen features were extracted from the PET and CT images, including geometrical and textural information, as follows:

1) Geometrical features. Three metrics were computed for each ROI: the volume, the extent and solidity which are the ratio of the volume over the bounding box and convex hull respectively.

2) First order features. These global characteristics of the grey level frequency distribution include computation of the mean, standard deviation, skewness and kurtosis. We also included the PET SUV peak value.

3) Textural features. These second-order statistical measures consider the relationship between groups of two neighboring pixels in the image. Those used in this study were derived from the grey level co-occurrence matrix (GLC) [7], the grey level run-length matrix (GLRL) [8], the grey level intensity size zone matrix (GLISZ) [9] and the grey level difference matrix (GLD) [7]. As an example, the GLD matrix enables the computation of the coarseness, contrast and busyness which are thought to correlate well with human impression, while the GLC matrix is the basis for the computation of the Haralick coefficients such as energy, entropy and correlation. The GLC and GLRL matrices are parameterized by the angle θ and the distance d separating two pixels. In this study, they were computed in 3D considering a 26- neighborhood (d=1) and an average value over the 4 angular directions, θ =0°, 45°, 90° and 135°. The grey levels within each PET and CT images were linearly scaled to 16 levels prior the feature calculation in order to normalize among the different patients and reduce noise.

3.2. Feature selection

Six filter methods were implemented to rank the 115 features, including a distance metric, a Kruskal-Wallis non-parametric test, the area under the receiver operating curve (ROC), and three indices derived from the mutual information: the maximum relevancy (MR), the additive or multiplicative combination of the minimum redundancy (mR) and MR referred to as mRMRd and mRMRq respectively. The most discriminant features were those that were ranked two times among the 10 best ranked by each of the six selection methods.

3.3. Classification

Two supervised classifiers were considered in this study, the SVM and the RDF.

1) Support vector machine (SVM) [10]

The basic idea of the SVM classifier is to find an optimal hyperplane $w.V+b$ that maximizes the separation margin ($2/\|w\|$) between the two classes. The decision variable $\lambda_{SVM}(V)$ is then dependent on the signed distance separating V from the hyperplane. SVM allows solving nonlinear problems by mapping data into a higher dimensional space by means of a kernel function. They also account for nonseparable classes by allowing classification errors which are parameterized by a penalization parameter C. In this study, we chose the Radial Basis Function (RBF) as the kernel function which is parameterized by the standard deviation σ of the Gaussian kernel. The values of C and σ were derived based on the optimization study described in Section 4.1.

2) Random decision forest (RDF)

The RDF classifier consists of a series of separate decision trees (DTs) whose outputs are combined based on a majority voting scheme [11]. DTs are robust, make no assumption about the data distribution and can handle data with incorrect labeling errors. Each decision tree of the RDF is built based on N training examples selected by bootstrap resampling of the original N training cases and considering a subset of k features randomly chosen from the K features (k< K). In this study, we set k=1+log2(K) and a maximal depth of 20. Three estimation criteria were considered to measure, at each node of the DT, how well a given feature separates the training examples according to their target classification: the information gain, the gain ratio and the Gini index. An optimization study (see Section 4.1) was conducted to select the best criterion as well as the optimal number of trees.

4. PERFORMANCE OF THE CAD SYSTEM

4.1. Optimization of the SVM and RDF intrinsic parameters

An exhaustive grid search was run so as to cover the space of C and σ for the SVM, and the space of the decision criterion and number N of trees for the RDF. For each couple of parameters (C, σ) or (criterion, N), a leave-one-out (LOO) cross-validation was performed using the area under the ROC curve (AUC) as the quality metric.

4.2. ROC-based performance evaluation

Estimation of the classifiers performance was done using a LOO cross-validation approach, which avoids to train and test on the same data and is known to produce a pessimistic metric of the general population performance. ROC curves were plotted for the SVM and RDF classifiers combined with the whole feature set and the most discriminant ones. Two problems were addressed: the first approach tried to separate the {N, NS} class from the {M} class, the second, to separate the {NS} class from the {M} class. Performance was evaluated based on the AUC metric.

5. RESULTS

5.1. Feature selection

Results of the feature selection described in Section 3.2 are reported in Table 1. Twelve features were extracted for each discrimination problem. For the problem of discriminating {NS, N} from {M}, 7 PET and 5 CT characteristics were selected, including 8 textural parameters such as the PET coarseness as well as the standard clinical SUV metrics. Nine PET and 3 CT characteristics were selected for the more complex problem of discriminating {NS} from {M}, including 5 texture parameters such as the PET coarseness and the clinical SUV metrics.

Table 1
List of the most discriminant features*

{NS, N} vs {M}	{NS } vs {M}
SUV max	PET kurtosis
SUV peak	PET cluster prominence
PET coarseness	CT RL LRLGLE
CT ISZ RP	SUV min
SUV mean	SUV mean
SUV standard deviation	SUV peak
PET contrast	PET sum entropy
CT ISZ LRHGLE	CT mean
CT contrast	SUV max
CT dissimilarity	SUV standard deviation
CT difference variance	CT ISZ LRLGLE
PET RL RP	PET ISZ LRHGLE

*RP: run percent, LRHGLE: long run high grey level emphasis, LRLGLE: long run low grey level emphasis. These second-order statistical parameters are derived from the GLRL or GLISZ matrices.

5.2. ROC analysis

The optimization study showed that the highest SVM performance was achieved for C=32 and σ=0.125 when all features were considered and for C=8 and σ=2.3 accounting for the 12 most discriminant ones. For the RDF, the information gain associated to N=65 trees produced the best AUC without feature selection while the Gini index associated to N=25 trees was the best combination when accounting for the 12 most discriminant features.

Figure 1 shows the detection performance achieved by the SVM and RDF classifiers based on the whole set of 115 features (no selection) or on the 12 most discriminant ones (selection). Both classifiers achieve very high detection performance with AUC values ranging between 0.94 and 0.97. Feature selection does not allow improving classification performance for this task.

Figure 2 shows similar results for the more complex task of discriminating the {NS} and {M} cases. As expected, performance is decreased as compared with that achieved for the first problem with AUC values ranging

from 0.81 for the RDF without selection to 0.91 for the SVM classifier with feature selection. Feature selection has a significant impact on the performance of both classifiers unlike in the first detection problem. This might be explained by the smaller size of the sample cases in the second problem (32 samples for the {NS} class and 156 samples for the {M} class). It is indeed likely that accounting for the whole feature set increases the risk of overfitting the training data, thus leading to lower performance estimation with the LOO cross-validation strategy.

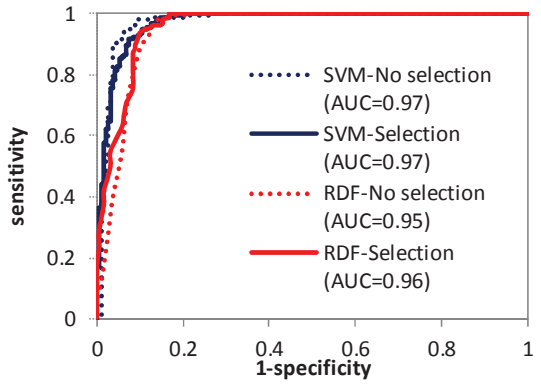


Fig 1. ROC curves for the {NS, N} vs {M} discrimination task

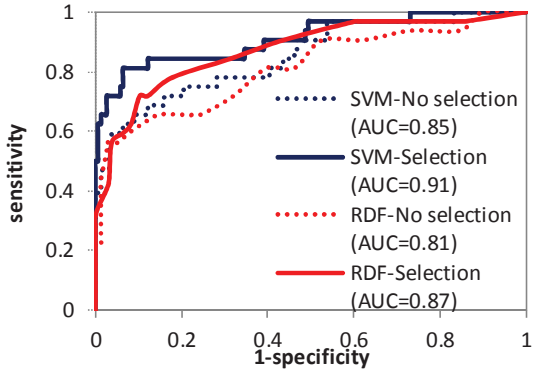


Fig 2. ROC curves for the {NS} vs {M} discrimination task

6. CONCLUSION

We have designed a computer aided diagnosis system to assess the presence of cancer in FDG PET/CT exams of lymphoma patients. The best performance was achieved by combining the SVM classifier with a series of first- and second-order (textural) statistical parameters extracted from the coregistered PET and contrast CT images, with AUC values of 0.97 and 0.91 for the tasks of discriminating the {NS, N} normal and suspicious ROIs from the {M} cancer ones and that of discriminating the {NS} suspicious ROIs from the {M} cancer regions, respectively. Perspectives include 1) comparing this CAD scheme with a classification system based on PET features only, 2) increase the PET/CT database to build separate training and testing datasets and 3) extending this ROI-based analysis to a voxel-based analysis, thus including the more challenging task of lesion detection and classification.

7. REFERENCES

[1] M. E. Juweid, S. Stroobants, O. S. Hoekstra, F. M. Mottaghy, M. Dietlein, A. Guermazi, G. A. Wiseman, L. Kostakoglu, K. Scheidhauer, A. Buck, R. Naumann, K. Spaepen, R. J. Hicks, W. A. Weber, S. N. Reske, M. Schwaiger, L. H. Schwartz, J. M. Zijlstra, B. A. Siegel, and B. D. Cheson, "Use of Positron Emission Tomography for Response Assessment of Lymphoma: Consensus of the Imaging Subcommittee of International Harmonization Project in Lymphoma," *J. Clinic. Oncol.,* vol. 25, pp. 571-578, February 10, 2007 2007.

[2] A. F. Cashen, F. Dehdashti, J. Luo, A. Homb, B. A. Siegel, and N. L. Bartlett, "18F-FDG PET/CT for Early Response Assessment in Diffuse Large B-Cell Lymphoma: Poor Predictive Value of International Harmonization Project Interpretation," *J. Nucl. Med.,* vol. 52, pp. 386-392, March 1, 2011.

[3] H. Yu, C. Caldwell, C. Mah, I. Poon, J. Balogh, R. MacKenzie, N. Khaouam, and R. Tirona, "Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images," *Int. J. Radiat. Oncol. Biol. Phys.,* vol. 75, pp. 618-625, 2009.

[4] Y. Nie, L. Q. Q. Li, F. Li, Y. Pu, D. Appelbaum, and K. O. Doi, "Integrating PET and CT information to improve diagnostic accuracy for lung nodules: a semiautomatic computer aided method," *J. Nucl. Med.,* vol. 47, pp. 1075-1080, 2006.

[5] G. V. Saradhi, G. Gopalakrishnan, A. S. Roy, R. Mullick, R. Manjeshwar, K. Thielemans, and U. Patil, "A framework for automated tumor detection in thoracic FDG PET images using texture-based features," in *International Symposium on Biomedical Imaging*, 2009, pp. 97-100.

[6] C. Lartizien, S. Tomei, and S. Marache-Francisco, "Improving performance of automated detection systems for 3D PET oncology imaging by use of resampled training images," *International Journal of Computer Assisted Radiology and Surgery,* vol. 4, pp. S186-193, 2009.

[7] R. Haralick, K. Shanmuggan, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 3, pp. 610-21, 1973.

[8] H. H. Loh, J. G. Leu, and R. C. Luo, "The analysis of natural textures using run length features," *IEEE Transactions on Industrial Electronics,* vol. 35, pp. 323-328, 1988.

[9] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, and J.-L. Mari, "Texture indexes and gray level size zone matrix. Application to cell nuclei classification.," in *Pattern Recognition and Information Processing (PRIP)*, Minsk, Belarus, 2009, pp. 140-145.

[10] V. Vapnik, *Statistical learning theory*. New York: Wiley, 1995.

[11] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Trans. on Pattern Anal. Machine Intel.,* vol. 20, pp. 832-44, 1998.