COMPUTER-AIDED DIAGNOSTIC SYSTEM FOR PROSTATE CANCER DETECTION AND CHARACTERIZATION COMBINING LEARNED DICTIONARIES AND SUPERVISED CLASSIFICATION

Jerome Lehaire[1,2], Rémi Flamary[3], Olivier Rouvière[1], Carole Lartizien[2]

[1]INSERM, U1032, LabTau, Lyon, F-69003, France; Université de Lyon, Lyon, F-69003, France
[2]Université de Lyon, CREATIS; CNRS UMR5220; INSERM U1044; INSA-Lyon; Université Lyon 1, France
[3]Laboratoire Lagrange, UMR 7293, Université de Nice Sophia-Antipolis, CNRS, Observatoire de la Côte d'Azur, Nice, France

ABSTRACT

This paper aims at presenting results of a computer-aided diagnostic (CAD) system for voxel based detection and characterization of prostate cancer in the peripheral zone based on multiparametric magnetic resonance (mp-MR) imaging. We propose an original scheme with the combination of a feature extraction step based on a sparse dictionary learning (DL) method and a supervised classification in order to discriminate normal {N}, normal but suspect {NS} tissues as well as different classes of cancer tissue whose aggressiveness is characterized by the Gleason score ranging from 6 {GL6} to 9 {GL9}. We compare the classification performance of two supervised methods, the linear support vector machine (SVM) and the logistic regression (LR) classifiers in a binary classification task. Classification performances were evaluated over an mp-MR image database of 35 patients where each voxel was labeled, based on a ground truth, by an expert radiologist. Results show that the proposed method in addition to being explicable thanks to the sparse representation of the voxels compares well (AUC>0.8) with recent state-of-the-art performances. Preliminary visual analysis of example patient cancer probability maps indicate that cancer probabilities tend to increase as a function of the Gleason score.

Index Terms— CAD, Prostate cancer, MRI, Dictionary learning, SVM, Logistic regression

1. INTRODUCTION

Prostate cancer is the most frequent cancer and the second cause of mortality in France. The actual gold standard diagnostic method is the echo-guided biopsy which is mostly randomly conducted because of the lack of sensitivity and specificity of ultrasound imaging. Radiologists are therefore exploring the feasibility of multiparametric magnetic resonance (mp-MR) imaging combining various MR sequences to target biopsies towards suspicious areas or ultimately to allow the non invasive active staging and follow-up of patients. Nevertheless, the interpretation of mp-MR images is complex since cancer lesions may generate conflicting signatures on the different sequences. To leverage this problem, computer-aided diagnostic (CAD) systems recently demonstrated promising results in assisting radiologists in the diagnostic phase. By extracting features from a single or multiple MR sequences followed by a supervised classification step, the proposed systems generate probability scores of malignancy either at a voxel level (CADe) thus providing probability maps in the whole prostate [1,2,3,4,5], or at the level of a region of interest (CADx) outlined by the radiologist [6,7].

Recently, we achieved promising results with CADx system based on the combination of a series of 110 statistical, structural and functional features extracted from three MR sequences (T2, Apparent Diffusion Coefficient and Dynamic Contrast Enhanced) and a SVM classifier [7,8]. We now aim at going one step further by proposing an original CADe scheme that outputs a probability score correlated with the Gleason score (GL) characterizing the aggressiveness of cancer lesions. The challenge we address is thus to design a system that can discriminate among the different types of prostate cancer lesions. This study is a preliminary step toward that goal.

This problematic shares similar characteristics with that of hyperspectral imaging (HSI) [9]. We indeed hypothesize that the voxel MR signature is the resulting effect of the linear and non linear combination of different types of organic material (blood, cancer and normal tissue etc) and we aim at characterizing the voxel element, ie either quantifying the fractions of each component or classifying the voxel in the class of the dominant one. Similar factors make the unmixing and classification of the HSI and MR-spectral signature task complex: the high dimensionality and size of the data, the linear and nonlinear spectral mixing, and the low number of labeled training data. Following the unmixing formalism, we consider a voxel $\mathbf{x_i} \in \mathbb{R}^f$ as a linear combination of a finite number of basis elements that constitute a dictionary $\mathbf{D} \in \mathbb{R}^{f \times K}$

$$\hat{\mathbf{x_i}} = \sum_{k=1}^{K} \alpha_i^k \boldsymbol{d_k} = \mathbf{D}.\boldsymbol{\alpha_i} \qquad (1)$$

where K is the size of the dictionary, f the number of features, $\boldsymbol{d_k}$ the approximation elements of $\mathbf{D}$ and $\boldsymbol{\alpha_i} = [\alpha_i^1, \alpha_i^2, ..., \alpha_i^K]$ are the decomposition coefficients for voxel $\boldsymbol{x_i}$. The objective of unmixing is to both estimate the series of $\boldsymbol{d_k}$ and $\boldsymbol{\alpha_k}$ elements. The interest of such a representation is first to characterize the content of the image at a voxel level and second to reduce dimensionality by considering the series of $\boldsymbol{\alpha_k}$ as features for classification purpose in the context of small training samples. Among the different unmixing algorithm, those based on simple linear models such as VCA [10] postulate that the elements $\boldsymbol{d_k}$ are 'pure' components referred to as endmembers. As for HSI, this assumption is likely to be violated with mp-MR features because of nonlinear effects introduced during the MR acquisition and image processing steps that generate the features and the absence of 'pure' voxels in the MR images. Sparse coding unmixing models have been shown as a promising alternative to the linear methods for HSI [9]. They assume that the voxel is a linear combination of a few basis elements of a larger dictionary and learn these elements from the data so that they can encode nonlinear variations (e.g. different elements of the dictionaries can indeed represent the same tissue).

In this paper, we propose and evaluate different CAD schemes for mp-MR prostate screening that combine a feature extraction step based on a sparse representation of the data and a classification step performed by two classifiers, the linear support vector machine (SVM) and the logistic regression (LR) classifiers. Performance analysis of these schemes is presented based on a series of 35 annotated patients.

2. SUPERVISED CLASSIFICATION WITH LEARNED DICTIONARIES

2.1. Feature extraction

A voxel $\mathbf{x_i}$ can be represented by $\mathbf{x_i} = [x_i^1, x_i^2, ..., x_i^f] \in \mathbb{R}^f$ where *f* represents the number of features for a given voxel. Each voxel $\mathbf{x_i}$ was given a label y_i by an expert radiologist, where y_i encodes normal {N}, normal but suspicious {NS} and cancer tissues with Gleason scores ranging from 6 to 9.

2.1. Dictionary learning

Following the general linear model of Eq. 1, the sparse dictionary learning (DL) methods are based on the optimization of a cost function that both attempt to find the representation of the dictionary D that best describes the data while promoting a sparse representation. This leads to the following joint optimization problem:

$$\min_{\mathbf{D}\in\mathcal{C},\boldsymbol{\alpha_i}\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x_i} - \mathbf{D}\boldsymbol{\alpha_i}\|_2^2 + \lambda\|\boldsymbol{\alpha_i}\|_1$$

where n is the number of voxels, λ is a regularization parameter for coefficients sparsity and C is a convex set for the dictionary elements. In our numerical experiments, we force the dictionary elements to be of unitary euclidean norm. We chose the online dictionary learning scheme proposed by Mairal et al [12] to solve the above problem. This method was indeed proven to converge and scale up to large scale data such as the one in our application (the dimension of each feature vector representing a voxel is of size f = 70, and the total number n of processed voxels is about 455 000). The stochastic class of the dictionary learning method, which process one sample at a time during T iterations, implies to define the number of randomly chosen voxels at each iteration of the method. According to default setups defined in [12], this parameter, defined as the mini-batch extension was set to 512 and the number of iteration T to 100.

2.2. Classification

In this paper, we consider a binary classification task so that $y_i \in \{-1; 1\}$ and two classifiers were evaluated, the SVM which has shown good performances for binary classification tasks of prostate mp-MR imaging [7] and the LR which directly outputs probability.

2.2.1. Support Vector Machine classifier (SVM)

The linear SVM separates the data by an hyperplane of equation $f(\boldsymbol{x}) = \boldsymbol{w}^t\boldsymbol{x} + b$ which maximizes the distance (margin) between the data of the two classes [11,14]. The parameters $\boldsymbol{w}$ and b result from the resolution of the following optimization problem

$$\begin{array}{c} \underset{\boldsymbol{w},b,\xi}{argmin}\ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ y_i(\boldsymbol{w_i^t}\boldsymbol{x_i} + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{array}$$

where ξ_i are slack variables corresponding to the distance to the margin of possibly misclassified samples $\mathbf{x_i}$ and C weights the empirical classification error. Note that we chose linear SVM that can be efficiently learned on large scale datasets [15].

2.2.2. Multinomial logistic regression classifier (MLR)

Multinomial logistic regression is a supervised learning method that also leads to a linear decision function but simultaneously provides probability estimates. In multiclass classification, MLR estimates one linear model w_i and b_i per class. The predicted conditional probability for class l is of the form:

$$p(y_i = l|\boldsymbol{x_i}) = \frac{e^{\boldsymbol{w_l}.\boldsymbol{x_i}+b_l}}{\sum_{j=1}^{n}(e^{\boldsymbol{w_j}.\boldsymbol{x_i}+b_j})}$$

The linear parameters $\boldsymbol{w_i}$ and b_i are estimated by maximizing the likelihood of the parameters given the training examples with a quadratic regularization term modeling a Gaussian prior knowledge on the model. The full optimization problem is not given in this work due to

lack of space but we refer the reader to classical statistical learning references such as [13]. The decision is then performed by selecting the class maximizing the conditional probability. Note that in this work we only used the binary classification variant of MLR known as logistic regression LR.

3. EXPERIMENTS AND RESULTS

3.1. Database description

The experimental database consisted of 35 patients who underwent mp-MR imaging following the protocol described in [7]. Each tumor or suspicious tissue was outlined by an expert radiologist over the three MR sequences. The nature of the tumor as well as its GL score were confirmed by an anatomo-pathologist. The total number of voxels and their percentage belonging to each of the four classes {N; NS; GS6; GS>6} are respectively {358 929 (79%); 31747 (7%); 10835 (2%); 54225 (12%)}.

3.2. Experiments

Eight different CAD schemes were evaluated based on the combination of the SVM and LR classifiers with four types of feature vectors. The first type referred to as *raw_feat* corresponds to the 110 features extracted from the 3 MR sequences [7]. The second input referred to as *alpha_VCA* are decomposition coefficients (abundances) obtained by the VCA linear model [10]. Two types of sparse coefficients were then considered. The first ones *alpha_DL* result from the learning of a global dictionary over the whole data set based on the software implementation of Mairal et al [12] while the second ones, *alpha_ssDL*, are generated from the concatenation of four dictionaries, each learned separately on one of the four classes {N, NS, GS6, GS>6} with the same algorithm.

The linear SVM and LR classifiers are used as binary classifiers where the discrimination task is to separate aggressive cancers (GS>6) from the rest (N, NS, and non aggressive cancer GS6). A leave-one-patient out cross-validation strategy (LOPO) is employed to evaluate the classification performance based on the area under the ROC curve (AUC).

Optimal methods and classification parameters were obtained by maximizing the area under the curve using the same LOPO scheme. A feature selection step was first performed based on a wrapper method described in [7]. Each classifier was first trained on all patients and all 110 features; the features were then ranked by descending order of their corresponding SVM and LR weights. The optimal number of features was selected from this ranked list based on a feature forward selection (FFS) strategy following the LOPO scheme, resulting in f=70 selected features for each classifier. The optimal number of coefficients p for *alpha_VCA* input was selected by varying this parameter between 4 and 30. Regarding the first dictionary learning method *alpha_DL*, the size K_{DL} of the dictionary was varied between 5 and 30. For the second dictionary learning method *alpha_ssDL*, the number of dictionaries K_{ssDL} was set to 1 for the {NS; GS6; GS>6} classes and 3 for the normal class {N} to account for the majority of samples from this class (79%). Then we multiplied simultaneously, from 1 to 5, the number of dictionary per class. We justified the choice of K_{ssDL}=3 by the majority of examples coming from the N class. Optimal values of C and λ for the SVM and LR classifiers respectively, were selected in the range $[10^{-5}; 10^{5}]$. This led to the following optimal parameters: f=70, p=6, K_{DL}=30 and K_{ssDL}=[9 3 3 3] for SVM and f =70, p=6, K_{DL}=20 and K_{ssDL}=[3 1 1 1] for LR.

3.3. Results

Table 1 reports the values of the AUC and Acc metrics derived from the binary classification task performed by each of the 8 CAD schemes. These metrics were averaged over the 35 patients. All CAD schemes performed similarly according to the AUC metrics with a mean value of 0.78+/-0.1. The LR classifier allowed achieving higher accuracy performance (Mean Acc value of 0.84) than the SVM classifier (Mean Acc value of 0.72). This comes from the validation scheme that maximizes the AUC of the method. Note that a similar accuracy can be obtained from the SVM classifier by adjusting the bias b.

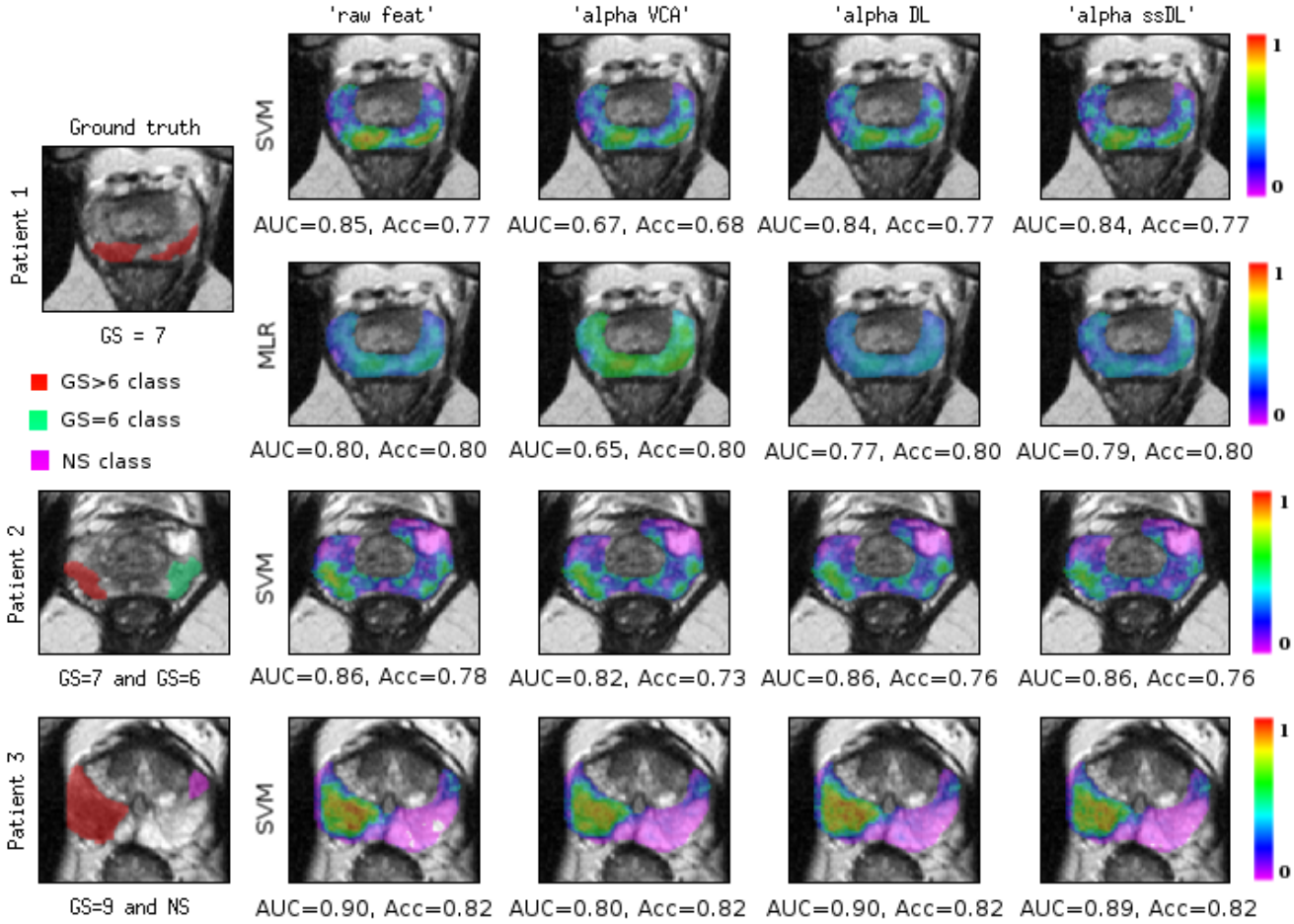
Table 1. Performances achieved for binary classification task, averaged over 35 patients

Input	Classifier	AUC	Acc
raw_feat	SVM	0.79+/-0.11	0.72+/-0.10
	LR	0.78+/-0.11	0.83+/-0.11
alpha_VCA	SVM	0.77+/-0.11	0.71+/-0.09
	LR	0.77+/-0.11	0.84+/-0.11
alpha_DL	SVM	0.79+/-0.11	0.70+/-0.12
	LR	0.78+/-0.11	0.84+/-0.11
alpha_ssDL	SVM	0.78+/-0.11	0.71+/-0.10
	LR	0.78+/-0.11	0.83+/-0.11

A sign test between all CAD schemes based on the AUC values derived for each of the 35 patients was performed. For the SVM classifier, all feature sets were shown to produce similar performance except for the *raw_feat* set which was shown to outperform the *alpha_ssDL* feature set (p=0.02). For the LR classifier, the *raw_feat* and *the alpha_DL* feature sets were both shown to outperform the *alpha_VCA* features (p=0.04) but all other paired comparisons concluded on similar performance of the different feature sets. There was no statistical difference between the performance of the two classification methods.

Fig 1. illustrates examples of the predicted probability maps (column 2 to 5) for two patients, obtained with the different CAD schemes and superimposed on the T2-weighted transverse slice.

Fig. 1. Example of predicted probability maps for three patients, four types of input: *raw_feat*, *alpha_VCA*, *alpha_DL*, *alpha_SSDL* and the SVM and LR classifiers.



The first example is a patient with two aggressive cancers (GS=7 for the two lesions). The Acc and AUC values are displayed under each image for this patient. Comparison of the first and second lines indicates that the SVM probability maps better highlight the two lesions than the LR classifier. This result is also correlated with the AUC metric. The Acc metric does not correctly capture the visual performance; It is indeed likely to be more sensitive to the very unbalanced class sample size, i.e. strongly impacted by the majority of normal (N) voxels (79%). The four features sets allow achieving similar probability maps with a higher contrast achieved for the *alpha_ssDL*. The second example shows probability maps obtained with the SVM classifier for a patient with one aggressive lesion (GS=7) and a non-aggressive lesion (GS=6). The two lesions are well depicted by the four features sets with the highest contrast achieved with the raw features. The mean probability achieved for the GS=7 lesion is slightly higher than that of the GS=6 lesion thus suggesting that the probability values might be correlated with the GS score. Same comment applies to the third patient images that clearly indicate that the very aggressive lesion (GS=9) is scored as highly suspicious (yellow-red area, proba>0.85) while the NS lesion appears with a much lower cancer probability (blue area; proba<0.5) but still higher than that of the normal tissue (pink area, proba~0).

4. PERSPECTIVES AND CONCLUSION

This paper evaluates different CADe schemes for prostate cancer localization and characterization based on mp-MRI screening. The achieved AUC performances are shown to compare well with recent state-of-the-art performances [1,2,3,4,5]. The introduction of the sparse DL methods in the feature extraction step did not allow any performance gain for the specific binary classification task considered in this study ({NS;GS6;GS>6} versus {N}). Our ongoing research investigates multiclass and non linear detection tasks. We hypothesize that DL methods associated with non linear multiclass SVM may help achieving the challenging goal of deriving cancer probability maps correlated with the Gleason score.

5. REFERENCES

[1] Y. Artan, MA. Haider, DL. Langer, TH. Van der Kwast, AJ. Evans, Y. Yang, MN. Wernick, J. Trachtenberg, IS. Yetik "Prostate Cancer Localization With Multispectral MRI Using Cost-Sensitive Support Vector Machines and Conditional Random Field," IEEE Transactions on Image Processing, 19(9):2444–2455, September 2010.

[2] A. Madabhushi, J. Shi, M. Rosen, J.E. Tomaszewski, M.D. Feldman, "Comparing classification performance of feature ensembles: detecting prostate cancer from high resolution MRI," Computer Vision Methods in Medical Image Analysis (In Conjunction with ECCV vol 4241), Berlin: Springer, pp 25–36 , 2006.

[3] R. Lopes, A. Ayache, N. Makni, P. Puech, A. Villers, S. Mordon, N. Betrouni, "Prostate cancer characterization on mr images using fractal features," Medical Physics, 38: 83–95 , 2011.

[4] S. Viswanath, B.N. Bloch, M. Rosen, J. Chappelow, R. Toth, N. Rofsky, R. Lenkinski, E. Genega, A. Kalyanpur and A. Madabhushi, "Integrating structural and functional imaging for computer assisted detection of prostate cancer on multi-protocol," SPIE Medical Imaging, 7260, (Miami: FL) 72603I, 2009.

[5] D.L Langer, T.H van der Kwast, A.J Evans, J. Trachtenberg, B.C. Wilson, M.A Haider, "Prostate cancer detection with multiparametric MRI: logistic regression analysis of quantitative t2, diffusion-weighted imaging and dynamic contrast-enhanced MRI," Journal of Magnetic Resonance Imaging, 30: 327–34 , 2009.

[6] P.C Vos, T. Hambrock, J.O. Barentsz, H.J. Huisman, "Computer-assisted analysis of peripheral zone prostate lesions using T2-weighted and dynamic contrast enhanced T1-weighted MRI," Physics in Medicine and Biology, 55(6) :1719-1734, 2010.

[7] E.Niaf, O. Rouvière, F. Mège-Lechevallier, F. Bratan, C. Lartizien. "Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI," Physics in Medicine and Biology, 57: 3833–3851, 2012.

[8] E. Niaf, C. Lartizien, F. Bratan, L. Roche, M. Rabilloud, F. Mège-Lechevallier, O. Rouvière. "Prostate focal peripheral zone lesions characterization at multiparametric MR imaging: Influence of a computer-aided diagnosis system," Radiology, 721(3): 761-769, 2014.

[9] J.M Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Diu, P. Gader, J. Chanussot, "Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches," IEEE Journal of selected topics in applied Earth Observation and Remote Sensing, 5: 354-372, 2012.

[10] J.M. Nascimento, J.M. Bioucas-Dias, "Vertex Component Analysis: A Fast Algorithm to Unmix Hyperspectral Data," IEEE transaction on Geoscience and Remote Rensing, 43:898-910, 2005.

[11] B. Schölkopf and A. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond," Cambridge, MA: MIT Press, 2002.

[12] J. Mairal, F. Bach, J. Ponce, G. Sapiro, "Online Dictionary Learning for Sparse Coding," International Conference on Machine Learning, 2009.

[13] T. Hastie, J. Friedman and R. Tibshirani, Book: "The elements of statistical learning," Springer Series in Statistics, p.17, 2001.

[14] N. Cristianini, J. Shawe-Taylor, "An introduction to support vector machines and other kernel-based learning methods," Cambridge university press, 2000.

[15] C. Chang, C. Lin, "LIBSVM : a library for support vector machines". ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at www.csie.ntu.edu.tw/~cjlin/libsvm