Specificity improvement of a CAD system for multiparametric MR prostate cancer using texture features and artificial neural networks

Specificity improvement of a CAD System for multiparametric MR Prostate Cancer using texture features and Artificial Neural Networks.

Author's names:

V. Giannini[1], S. Rosati[2], D. Regge[1,3] and G. Balestra[2]

Author's affiliation:

[1] Department of Radiology at the Candiolo Cancer Institute – FPO, IRCCS, Strada Provinciale 142 km 3.95, 10060 Candiolo, Italy.
[2] Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
[3] Department of Surgical Sciences, University of Torino, A.O.U. Città della Salute e della Scienza, Via Genova 3, 10126 Torino, Italy.

Corresponding author:
Valentina Giannini
Radiology Dept. Candiolo Cancer Institute – FPO, IRCCS
Strada Provinciale 142 Km 3,95
10060 Candiolo, Torino, Italy
Tel: +39 (0)11 9933327
fax: +39 (0)11 9933527
valentina.giannini@ircc.it

ACKNOWLEDGMENT

This work was funded by Fondazione Piemontese per la Ricerca sul Cancro FPRC-onlus, grant Pro-Cure, 5 per Mille 2009 Ministero della Salute.

Abstract

Prostate cancer (PCa) is the most common cancer afflicting men in USA. Multiparametric Magnetic Resonance imaging is recently emerging as a powerful tool for PCa diagnosis, but its analysis and interpretation is time-consuming and affected by the radiologist experience. Computer aided detection (CAD) systems have been developed to overcome this limitation and to support radiologists in the PCa diagnosis. Although several studies proposed CAD systems with very high performances in terms of sensitivity, the analysis of false positive (FP) areas is usually not clearly presented. The aim of this study is to improve the performance of a CAD system in term of reduction of FPs findings, without affecting the sensitivity. To this scope, we developed a classifier composed by 3 Artificial Neural Networks (ANN) able to distinguish between malignant and healthy areas through a voting strategy. In this method, we exploit the role of the Gray Level Co-occurrence Matrix, the Gray Level Difference Method and Gray Level Run Length Method Matrix in differentiating tumoural from healthy tissues. We first extract 64 textural features from T2-weighted (T2w) images and the apparent diffusion coefficient (ADC) maps, then we discretized them to reduce the data variability. A features selection method, based on the correlation matrix, is finally applied to remove redundant variables, that are those highly correlated with others. The remaining set of features is fed into the three ANNs and a post-processing step is applied to remove very small areas.

Results applied on a dataset of 58 patients showed a significant decrease of FPs (20 vs 12; p-value<0.0001) and an increase of the precision of PCa segmentation (0.62 vs 0.71; p-value<0.0001). Having less FPs is helpful to increase the performance of CAD systems in terms of specificity and to decrease the reporting time of radiologists. Moreover, having more precise PCa segmentation areas could be useful if a step of PCa characterization will be added to the CAD system.

Keywords: texture features, artificial neural networks, CAD system, prostate cancer, multiparametric MRI.

1. Introduction

Prostate cancer (PCa) is the most common cancer expected to occur in men in United States in 2016, accounting for 1 in 5 new diagnoses [1]. Statistically, the number of new diagnosed cases was estimated to be 180890, with no less than 26120 deaths [1]. Unfortunately, the current clinical practice to diagnose PCa, which includes PSA blood examination followed by transrectal ultrasound (TRUS) biopsy, is characterized by low sensitivity and specificity in detecting clinical significant lesions [2,3]. To overcome these limitations, multiparametric (mp)-Magnetic Resonance (MR) imaging is increasingly used to diagnose PCa as it has been demonstrated to

improve sensitivity and specificity over PSA and TRUS [3-5]. However, mp-MR imaging has not yet progressed to a first line modality because it is a labor-intensive examination and has a steep learning curve. Indeed, interpretation requires experienced radiologists capable of analyzing data extracted from the different MR sequences [2,6,7].

Recently, computer aided detection (CAD) systems have been presented to support the radiologist by indicating suspicious regions and reducing oversight, perception errors [8], and reporting time [9]. Sensitivity reported by these CADs ranges from 60% to 100% [2], however only few studies, who performed a lesion segmentation step, computed the free-response receiver operating characteristic analysis to provide the number of false positives (FP) candidates at a given level of sensitivity. Among them the best results were reported by Litjens et al. [6] which obtained 10 FPs with a sensitivity of 89%. Conversely, since most studies provide only a voxel-based segmentation step, results are usually presented by computing the receiver operating characteristic (ROC) curves on some malignant and benign/healthy regions of interest (ROIs). However, ROC curves are not able to evaluate neither the number of FPs within the prostate nor the precision of the segmentation of the lesion. Indeed, high value of area under the ROC curve could be reached even if the lesion is oversegmented or if there are many FP voxels outside the malignant/healthy ROI used for the analysis. Having a low number of FPs and more precise lesion segmentation may help to develop CAD systems able not only to detect PCas but also to provide the characterization of PCa aggressiveness, which represents the key to personalized treatment options, minimizing the risk of overtreatment [10].

The analysis of the voxel-wise probabilities maps obtained with a previously developed CAD systems [7] and the maps of semi-quantitative and quantitative parameters available in the clinical practice prompted us to investigate the role of the spatial variations of the signal intensity in an image, i.e., texture features, to overcome the limits of some voxel-based parameters of being intrinsically noisy and sparsely distributed across the prostate region. Textures features are used in medical image analysis to quantify image properties such as homogeneity and contrast. In particular, the texture analysis method based on gray level co-occurrence matrix (GLCM) [11] has gained wide applications in medical image analysis for its ability to characterize the spatial dependence of gray-levels using the second-order statistics. Rosati et al. [12] proved that the analysis of carotid ultrasound images by means of textural features based on GLCM contributes to identify high cardiovascular risk subjects and it is able to capture the progressive presence of pathological conditions in vessel wall layers. Recently, it has been demonstrated that some texture features are correlated with Gleason Score (GS) [13].

In this paper we present an improvement of the methodology reported in [14], able to reduce the number of FPs and to provide a more precise lesion segmentation of a previously developed CAD system [2] by using some GLCM texture features computed on both the ADC maps and the T2w images.

2. *Materials And Methods*

2.1 Patients

The dataset is composed of 58 individuals that complied with the following inclusion criteria: (a) biopsy-proven prostate adenocarcinoma, (b) mp-MR imaging examination between April 2010 and November 2012, including axial T2-weighted (T2w), diffusion weighted (DW), and Dynamic Contrast Enhanced (DCE) MR sequences, (c) radical prostatectomy within 3 months of MR imaging, and (d) a clinically significant peripheral zone (PZ) lesion (tumor volume≥0.5 cc) at the whole-mount histopathologic analysis. The local Ethics Committee approved the study and participants into the study signed informed consent forms. This study was in accordance with the Helsinki Declaration.

2.2 MR Image Acquisition and reference standard

Images were acquired with a 1.5 T scanner (Signa Excite HD, GE Healthcare, Milwakee, Illinois, USA) using a four-channel phased-array coil combined with an endorectal coil (Medrad, Indianola, Pa). Axial T2w images were obtained using the following protocol: slice-thickness, 3 mm; FOV, 16x16 cm; NEX, 2; acquisition matrix, 384x288; reconstruction matrix, 512x512; TR/TE ratio, 3020/85. DW imaging was obtained using axial Echo-Planar Imaging sequences as follows: slice-thickness, 3 mm; FOV, 16x16 cm; acquisition matrix, 128x128; reconstruction matrix, 256x256; NEX, 6; TR/TE, 7000/101; b-values, 0 and 1000 s/mm 2. Finally, a 13 s time resolution DCE study was performed, with an axial 3D Spoiled Gradient echo (SPGR) sequence using the following parameters: TR/TE/FA, 3.6/1.3/20°; FOV, 20x20 cm; slice thickness, 3 mm; acquisition matrix, 224x192; reconstruction matrix, 512x512. Scanning started simultaneously with the intravenous injection of 0.1 mmol/kg gadobutrol (Gadovist, Bayer Schering, Berlin, Germany) through a peripheral line at 0.7 ml/s, using a power injector (Medrad Spectris, Maastricht, The Netherlands), followed by infusion of 20 cc normal saline at same rate. Twenty-six contrast-enhanced frames were obtained. The average time to complete the whole MR exam, including two additional T2w scans in the sagittal and coronal plane, was 40 minutes. Overall imaging parameters satisfied the minimal scanning requirements [3].

All patients underwent prostatectomy within 3 months of mp-MR exam. The prostate specimen was step-sectioned at 3 mm intervals perpendicular to the long axis (apical-basal) of the gland, thus reproducing the inclination of axial T2w images, as previously detailed [7]. An experienced radiologist (with an experience of more than 500 prostate mp-MRI studies interpreted per year for 6 years) in consensus with the pathologist, established the reference standard for PCa on T2w images drawing freehand ROIs on cancer foci, following the outlines drawn by the pathologist on digital images of the pathologic slices.

2.3 CAD system for PCa detection

The CAD system consists of multiple sequential steps [7]. First, DW images are upsampled to match the in-slice resolution of the T2w, then all dataset are automatically aligned so that features, derived from all the MR sequences and referring to the same pixel or group of pixels, may be compared and studied. In this study, the algorithm developed by Giannini et al. [15] is used to correct for both patient movements and image distortions. Once all datasets are aligned, four quantitative features are extracted from each voxel belonging to all the MR sequences. These features are: a) the ADC value derived from the DW images, b) the normalized T2w signal intensity, and c) two parameters extracted from the DCE sequence (a_0 and r). The latter were extracted fitting the normalized time-intensity curve using the Phenomenological Universalities (PUN) approach (equation 1) [7,16,17], that was demonstrated capable to reproduce all curve types one can obtain in a DCE-MRI session. The PUN equation is characterized by three fitting parameters, a_0, β and r, and the model implemented is reported in equation (1)

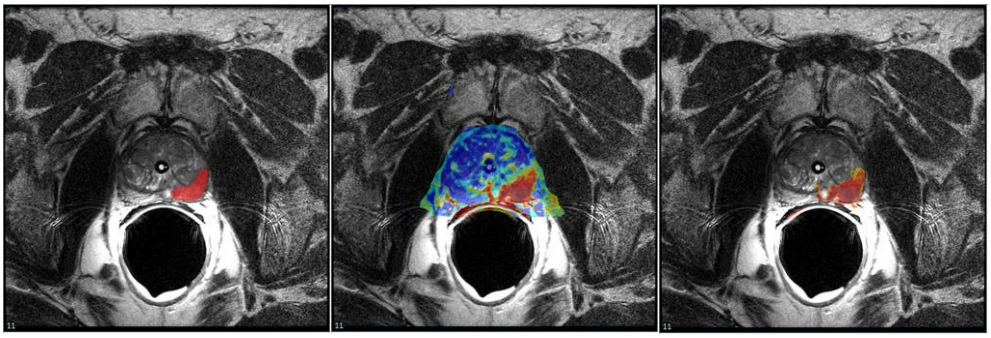
$$y_{PUN}(t) = exp\left[rt + \frac{1}{\beta}(a_0 - r)(\exp(\beta t) - 1)\right] \quad (1)$$

in which a_0 controls the steepness of the curve at t = 0 and together with β, which is related to the time the knee of the curve is reached, it primarily affects the growth rate of the curve in its first part; r determines the behavior of the second part of the curve (r > 0 if the signal keeps increasing, r < 0 there is there is a wash-out phase). The PUN approach has been demonstrated useful to provide information about tumor's vascularization without making any assumption on the tissue physiology and overcoming the limitations of the Tofts pharmacokinetic model, which requires the conversion of MR signal intensities into contrast agent concentration [16]. Parameters a0 and r are chosen because they were the most discriminative and less correlated in detecting malignant voxels [7].

These features are fed into a SVM classifier which uses the radial basis kernel and parameter C and γ equal to 2 and 0.002, respectively and each voxel is classified using the leave-one-patient out approach, as previously described [7] (Fig. 1b).

The SVM classifier computes the probability that each voxel belongs to the malignant class, thus providing a map representing the likelihood of malignancy for the whole prostate (figure 1b). Finally, a step is applied to extract 3D candidates highly suspected to be cancers. Therefore, we discarded all voxels having ADC values < 400 or >1600 mm^2/s [4] and probability to be malignant < 60%.

The schematic representation of the CAD system for PCa detection is depicted in the left part of Fig. 2.



a) b) c)

Fig. 1 a) T2w image of a patient with a prostatectomy Gleason Score =4+3 PCa of volume=6.25 cc. The red ROI is the tumor drawn by the radiology using the histological data as reference standard. b) Voxel-wise probability map of the correctly segmented tumor provided by the SVM classifier; c) output of the candidate segmentation step. Probability ($p \geq 0.8$ in red, $0.7 \leq p < 0.8$ in orange, $0.6 \leq p < 0.7$ in yellow, $0.5 \leq p < 0.6$ in green, $0.4 \leq p < 0.5$ in cyan, and $p < 0.4$ in blue.

2.4 FP reduction system

Starting from the voxel-wise probability maps obtained from the CAD system, we further process the malignancy probability maps using a system based on Artificial Neural Networks (ANNs) able to reduce the number of FP voxels and consequently improve specificity in PCa detection. In this phase each slice in the prostate volume is processed independently as a 2D image, considering each voxel as a pixel. The procedure for the construction of the FP reduction system and its application for the FP reduction is described below and schematically represented in the right part of Fig. 2.

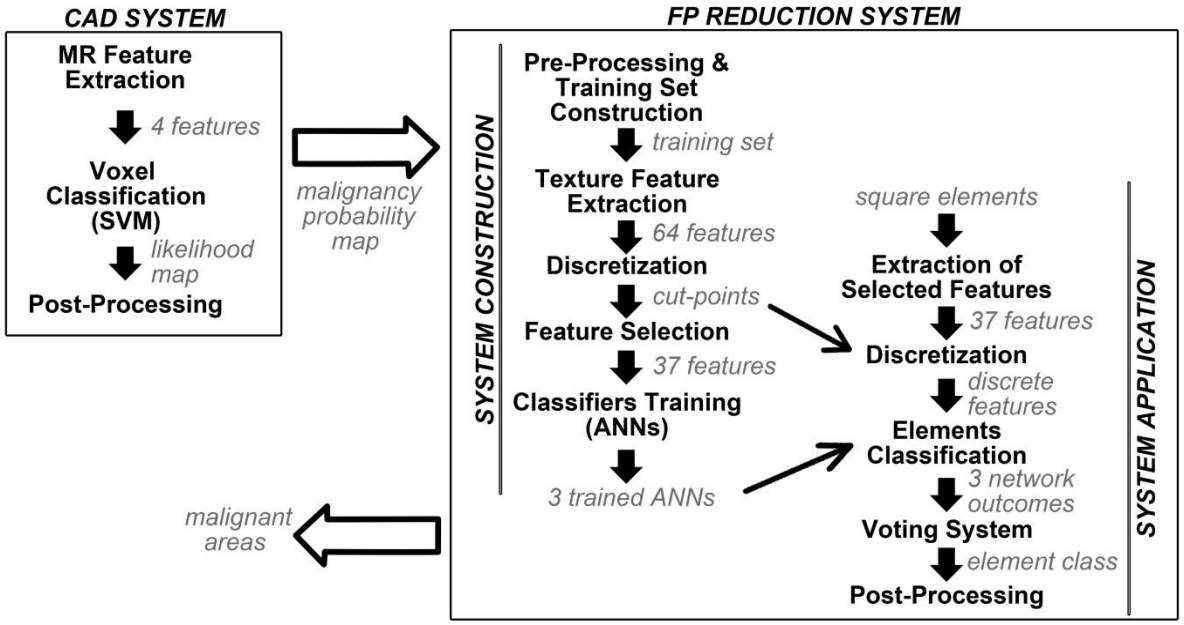


Fig. 2. Schematic representation of the CAD system for PCa detection and the FP reduction system.

2.4.1 System Construction

a) Map Pre-Processing and Training set Construction

For each slice in the prostate volume, we divided the suspicious regions remaining from CAD system into small squares made of 5-by-5 voxels (25 voxels). This size has been chosen considering that smaller elements, i.e., 3x3, could be sensitive to image's noise, and bigger elements could lay outside the tumors, thus affecting texture analysis. Indeed, since PCas could have different size and shape, it is important to use an element size that can guarantee to have at least one element inside each tumor. Each square area is considered as a single element for all phases of system construction and application. Following this procedure for all patients, we obtained a dataset composed of 33839 elements. A total of 4558 elements fell inside the manual mask drawn by the radiologist on the malignant regions (using the histopathology as reference standard): for 3766 of them the overlap with the mask was higher than 50% (malignant elements), while 792 elements had an overlap with the tumor less than 50% . All elements having no overlap with the manual mask were considered healthy (29281 elements).

As it has been proved that using strongly imbalanced dataset can negatively affect the following steps of feature extraction, discretization, feature selection and classifier training [18], we build a smaller dataset to be used as training set for the system construction phases. However, as the aim of this procedure is to improve the identification of benign voxels and consequently reduce the number of FPs, we decide to maintain in our training set a slight predominance of benign elements. In particular, we consider all malignant elements with overlap greater

than 50% (3766 elements) and a double number of healthy elements randomly selected (7532 elements), for a total of 11298 squares included in the training set.

b) Feature Extraction

The second step of the construction of the FP reduction system consists in the identification and extraction of a set of features able to characterize each element. In particular, a set of 32 texture features is computed for each square element included in the training set, considering both the ADC maps and the T2w images, for a total of 64 features.

The 22 features proposed by Haralick et al. [11] are extracted from the GLCM, that is a matrix counting the number of occurrences for which a pixel with a gray level *i* occurs at distance *Δx* and *Δy* from another pixel with grey level *j*. The GLCM for a generic image I is obtained according to equation 2:

$$GLCM_{\Delta x,\Delta y}(i,j) = \sum_{p=1}^{m}\sum_{q=1}^{n}\begin{Bmatrix} 1\ if\ I(p,q) = i\ \&\ I(p+\Delta x, q+\Delta y) = j \\ 0\ otherwise \end{Bmatrix} \quad (2)$$

where *m* x *n* is the image size, that in our case corresponds with the slice size, and $\delta = (\Delta x, \Delta y)$ is the displacement. Since for this specific application the analyzed slices didn't show any preferential texture direction, we decide to calculate the GLCM in the four main directions 0°, 45°, 90°, 135° and to average the matrices in order to preserve all information. Starting from the mean GLCM, the following features are calculated: autocorrelation [19], contrast [11,19,20], correlation1 [20], correlation2 [11,19], cluster prominence [11], cluster shade [11], dissimilarity [11,20], energy [11,19,20], entropy [11,19,20], homogeneity1 [20], homogeneity2 [11], maximum probability [11], sum of squares [19], , sum average [19], sum variance [19], sum entropy [19], difference variance [19], difference entropy [19], information measure of correlation1 [19], information measure of correlation2 [19], inverse difference normalized [20], inverse difference moment normalized [20],

In addition to these features, other 5 are extracted from the *Gray Level Difference Method* [21] (GLDM) calculated as in equation 3:

$$GLDM_{\delta}(k) = \sum_i \sum_j GLCM_{\Delta x,\Delta y}(i,j) \quad (3)$$

where $k=|i-j|$, with $k=0,1,\ldots,n-1$, *n* is the number of gray levels and $\delta = (\Delta x, \Delta y)$ is the displacement. Similarly to the GLCM, the GLDM is calculated in the four main directions 0°, 45°, 90°, 135° and the resulting vectors are averaged. The following 5 features are obtained from the averaged GLDM vector: contrast [21], angular second moment [21], entropy [21], mean [21], inverse difference moment [21].

Finally, the *Gray Level Run Length Method* [21] (GLRLM) is used to obtain the last 5 features. $GLRLM_\theta$ is a matrix in which each element *(i,j)* counts the number of occurrences of the *j* adjacent elements with gray level *i* calculated in direction θ. Also in this case, we average the GLRLM evaluated in directions 0°, 45°, 90° and 135° and we used the this matrix for the computation of the last 5 variables: short run emphasis [21], long run emphasis [21], gray level distribution [21], run length distribution [21], run percentage [21].

c) Discretization

In the third step the training set characterized by the 64 features is discretized using the *ChiMerge algorithm* [22]. This phase is introduced in our pipeline in order to reduce the data variability and improve the classification performance [23].

ChiMerge algorithm is a supervised and bottom-up discretization method that acts variable by variable using χ^2 statistic to test if the class of all elements in two adjacent intervals is independent from the values that the variable assumes for those elements. ChiMerge algorithm iteratively proceeds evaluating the $\chi 2$ value for every pair of adjacent intervals and merging those intervals with the minimum value of $\chi 2$. The algorithm stops when the $\chi 2$ value for all adjacent intervals is greater than the χ^2 value for a give significance level and a number of degrees of freedom equal to the total number of classes - 1. For this system, we fixed the significance level to 0.95 and the degree of freedom to 1, as we had two classes – malignant and non-tumoural.

d) Feature Selection

In order to reduce the number of features to be managed in the following classification step, a Feature Selection (FS) is performed on the discretized training set, based on the correlation among features. The idea is that features with high correlation contain the same information, so they can be considered as redundant. For this purpose, we compute the correlation between each couple of features: if the correlation is higher than 0.9 we randomly discard one of the two features.

e) Classifier Construction

The last phase of the system construction procedure consists in the definition and training of a classifier able to correctly identify the class of each square element. In particular, we implemented a classifier based on three similar feed-forward ANNs [24], combined using the majority voting in order to reduce errors due to a single classifier and to improve the classification accuracy [25, 26]

The structure of three ANNs is composed of one hidden layer with 19 neurons, an input layer with a number of neurons equal to the number of selected features, and one neuron in the output layer returning the class which

the element belongs to (tumor vs non-tumor). The neuron activation functions are set to hyperbolic tangent sigmoid function for the hidden layer and to logarithmic function for the output layer. Bayesian back-propagation is chosen as the learning algorithm and the mean squared error is used as performance function. The initial values of interconnection weights are set randomly.

Each ANN is trained separately using datasets with a different number of elements. Specifically, starting from the discretized training set, one network is trained with the complete set of available elements (3766 tumoural elements and 7532 non-tumoural elements), the second ANN is trained using all malignant elements and an equal number of benign elements randomly extracted (3766 tumoural elements and 3766 non-tumoural elements), while for the training of the last network we use a number of benign elements that is the half of the malignant elements (3766 tumoural elements and 1883 non-tumoural elements). This procedure is employed because in a preliminary analysis we did not obtain satisfying results in terms of classification performance using only one network trained with both balanced and imbalanced training set.

Once the three ANNs are defined and trained, a majority voting system is implemented among them to obtain the final element class, that is the one obtaining at least two votes over the three ANNs outcomes. Having three nets and two possible classes, the element always results classified.

2.4.2 System application

Once the three ANNs are trained, the system for the FP reduction is applied for the classification of all square elements in the images remaining from the CAD system.

First of all, for each element only the selected texture features are calculated and discretized. The limits of the discretized features obtained for the training set are used for the discretization of the new values. The resulting discretized features vector is input in the three ANNs and the voting procedure is employed to obtain the final class that is assigned to all 25 voxels belonging to the considered element. Finally, when all 5-by-5 squares in the image are classified, a post-processing step is performed in order to remove small malignant ROI. In this phase, we test two different dimensions to be considered as clinically significant tumor, keeping only groups made of at least 25 or 50 adjacent voxels.

2.5 Statistical evaluation

Sensitivity of the CAD and the ANN outputs was computed. In order to compare our results to a validated system [7] we applied to the CAD output some heuristic criteria to reduce the number of FPs, and we compared the results of the proposed method with the performance of the previously developed system (*old CAD*). From the probability map, 3D connected regions with a size<100 voxels are discarded. This size represents the 60% of the

volume of the smallest clinical significant PCa, i.e. 0.5 cc [27], therefore it has been chosen in order to not discard tumors that might have been only partially segmented. Finally, for each of the remaining voxels the time-intensity curves from the DCE MR dataset are computed, and voxels having maximum uptake in the first minute < than 100% are discarded.

Sensitivity was defined as the number of correctly segmented tumor over the total number of tumors. A lesion was considered detected if the Dice overlap between the manually (M) segmented mask and the automatically segmented (S) mask was ≥ 0.05. Manual masks were obtained using the histopathological sections as reference standard, as previously described [7]. Volume and numbers of FPs, considered as 2D connected region, were also computed and compared for the two outputs.

As a secondary endpoint, we applied some metric to evaluate the overlap between the manually segmented ROIs and the output of both the CAD system and the ANN. Indeed, it is not only important to evaluate if the CAD is able to detect the PCa, but also how well it can segment it, i.e., in order to apply further steps for aggressiveness characterization. To this scope, we evaluated the Dice overlap, defined as the ratio between twice the intersection of M and S masks over the sum of M and S mask of the malignant ROIs, the precision, defined as the ratio between the intersection of M and S over S, which is a measure of exactness, and recall, defined as the ratio between the intersection of M and S over M, which is a measure of oversegmentation. Differences between the performances of the CAD and the ANN were compared using the Wilcoxon test, while the difference of FPs size were evaluated with Mann-Whitney test. Statistically significance was set at $p \leq 0.05$.

3. Results

The study population (58 patients) included 50 patients (86%) with one clinically significant PZ tumor focus, and 8 patients (14%) with two clinically significant PZ tumor foci, for a total of 66 clinically significant PZ PCas. Patients and lesions characteristics are summarized in table 1.

Table 1 Patients and lesions characteristics

Parameter	Value
No. of patients included in study	58
Patients median age [y] (1st-3rd quartile)	64 (60-70)
Median PSA at diagnosis [ng/ml] (1st-3rd quartile)	6.7 (5.1-8.4)
Median no. of days between biopsy and MR examination [d] (1st - 3rd quartile)	90 (60-111)
Median prostate volume [ml] (1st -3rd quartile)	43.6 (35.3- 60.5)
No. of lesions with tumour volume ≥ 0.5 ml	66
Median volume [ml] (1st-3rd quartile)	1.7 (1.0-2.6)
Distribution of pathologic Gleason scores [no. of patients]	
3+3	6 (9%)
3+4	33 (50%)
4+3	16 (24%)
4+4, 4+5	11 (17%)

The FS applied to the 64 texture features selected 37 variables, 18 texture features calculated from T2w images and 19 from ADC maps. The list of the selected variables is showed in table 2. As it emerges from the table, the same features are selected from T2w and ADC, except for the inverse difference moment normalized of GLCM that is selected only for the ADC. This can means that, after discretization, the most informative texture parameters remain the same for the two types of images, but at the same time they give complementary information.

Table 2 Features selected with the correlation

	T2w	ADC
GLCM	Autocorrelation	Autocorrelation
	Contrast	Contrast
	Correlation1	Correlation1
	Cluster Prominence	Cluster Prominence
	Cluster Shade	Cluster Shade
	Energy	Energy
	Entropy	Entropy
	Homogeneity1	Homogeneity1
	Maximum probability	Maximum probability
	Sum entropy	Sum entropy
	Difference entropy	Difference entropy
	Information measure of correlation1	Information measure of correlation1
	Information measure of correlation2	Information measure of correlation2
	Inverse difference normalized	Inverse difference normalized
		Inverse difference moment normalized
GLDM	Angular Second Moment	Angular Second Moment
GLRLM	Short Run Emphasis	Short Run Emphasis
	Long Run Emphasis	Long Run Emphasis
	Gray Level Distribution	Gray Level Distribution

The results obtained for each ANN in the FP reduction system are reported in table 3 in terms of percentage of elements correctly classified for the two classes of interest with respect to the dataset used for their training. Moreover, the same table shows the results of the voting procedure applied among the three networks outcomes for the classification of the whole training set.

Analyzing the classification results reported in table 3 it is evident the influence of class imbalance during the ANN training on the classification performances. In particular, the most represented class in the training set is

also the most correctly identified. In fact for ANN1 and ANN3 it is possible to reach very high percentages (>80%) of correct classification for the benign and malignant elements respectively. Conversely, for the remaining class the results are very poor, being lower (ANN1) or equal (ANN3) than the 50% of correct classified elements.

In the case of balanced classes (ANN2) the final classification performance is very similar for the two classes, achieving the 66.7% for the non-tumoural class and the 70.4% for the tumoural elements, even if it is not satisfactory for the purpose of this study.

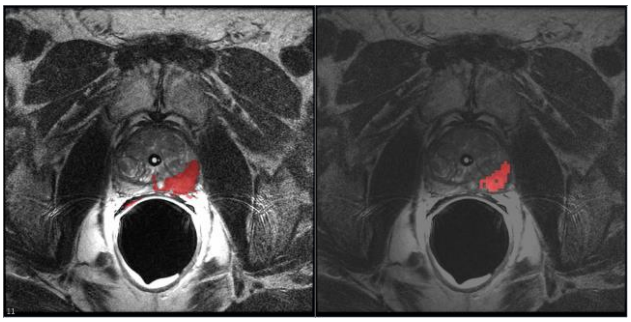
Using the voting procedure among the three classifiers (last row of table 3) it is possible to improve the classification for both classes with respect to the case with balanced dataset (ANN2). In this situation, in fact, the obtained results are similar to those reached for the most represented class in imbalance dataset: 77.1% of benign elements correctly classified by voting the outcomes with respect to 84.3% obtained with ANN1 and 81.4% of malignant elements correctly classified using voting procedure with respect to 85.3% realized with ANN3.

On the other hand, 18.6% of malignant elements are classified as benign. However this doesn't imply necessarily that the tumors are not identified, because more than a single 5-by-5 element is usually extracted from a single malignant area.

Table 3 Results of the 3 ANNs and the voting procedure of the FP reduction system for the classification of the training set. The two columns contain the percentage of correctly classified elements for the two classes.

	Malignant Elements (M)	**Benign Elements (B)**
ANN1 (3766 M / 7532 B)	48.8%	84.3%
ANN2 (3766 M / 3766 B)	70.4%	66.7%
ANN3 (3766 M / 1883 B)	85.3%	50.5%
VOTING (3766 M / 7532 B)	81.4%	77.1%

Figure 3 shows an example of the ROIs output of the CAD system (Fig.3a) and the output of the FP reduction method (Fig.3b).



a) b)

Fig. 3 a) ROIs segmented by the CAD system; b) ROIs segmented by the FP reduction system. All maps are superimposed to the T2w image.

Table 4 reports the results in terms of sensitivity and number and size of FPs obtained with the CAD and after the post-processing step. As it emerged from the table, a significant decrease of the number of FPs has been obtained when the connected region ≤ 50 pixels were discarded, while it remains the same when the ANNs used 25 pixels as minimum size limit. However, the median size of FPs strongly decreases in both conditions ($p<0.0001$). The median size of the FPs obtained with the size limit of 25 pixels is higher than those obtained when 50 pixels were used as minimum size, even if the number of FPs is higher. This behavior depends on the fact that the first method is able to detect smaller region as tumor, while for the second method is necessary that a region is > 50 to be kept as true positive. Therefore, using the limit of 25 pixels, we are able to not discard a higher number of small elements (i.e., < 50 pixels) than with the limit of 50 pixels. Conversely, all region > 50 pixels are kept by both limits. Five out of 6 FN lesions were the same for the two outputs; while one pGS=3+4 secondary lesion (size=1.67 cc) that was detected by the CAD was missed by the FP reduction method. Conversely, the ANNs were able to correctly detect a pGS=4+5 lesion (size=1 cc) that was missed by the CAD. Since the lesion missed by the ANNs was a secondary lesion in a patient in which the FP reduction method correctly detect the primary lesion, while the lesion missed by the CAD was the sole lesion in that patient, the ANN method obtained a higher per-patient sensitivity (95% vs 97%).

The other FNs were the following: 1 primary PCa with pGS=3+3 and size=1.66 cc, 1 primary PCa with pGS=3+4 and size=0.65 cc, 2 secondary PCas with pGS=3+4 and size of 0.55 cc and 1.55 cc respectively, and 1 secondary PCa having pGS=4+3 and size of 1.85 cc.

Besides, the FP reduction method obtains a high value of precision together with the decrease of recall, sign of a decrease of the oversegmentation.

Table 4 Comparison of FP Reduction Method with CAD output. Two-tailed p-value obtained with Wilcoxon test was reported. To test differences between the median size of FP was used the Mann-Whitney test.

	Old CAD	**FP Reduction Method (25 pixels)**	**p-value**	**FP Reduction Method (50 pixels)**	**p-value**
No. of detected patients (per-patient sensitivity)	55/58 (95%)	56/58 (97%)		56/58 (97%)	
No. of detected PCas (per-lesion sensitivity)	60/66 (91%)	60/66 (91%)		60/66 (91%)	
Median no. of FP per patient (1st -3rd quartile)	20 (13-29)	20 (10-36)	0.6984	12 (5-22)	<0.0001
Median size of FP [voxels] (1st -3rd quartile)	194 (135-354)	75 (50-100)	<0.0001	100 (75-150)	<0.0001
Median number of FP voxels per patient (1st - 3rd quartile)	6127 (2768-10280)	1875 (850-3719)	<0.0001	1400 (762-3081)	<0.0001
Dice overlap	0.41 (0.29-0.54)	0.37 (0.18-0.48)	0.0118	0.36 (0.21-0.47)	0.0013
Precision	0.62 (0.43-0.70)	0.72 (0.60-0.82)	<0.0001	0.71 (0.63-0.81)	<0.0001
Recall	0.36 (0.20-0.50)	0.27 (0.10-0.37)	<0.0001	0.25 (0.12-0.35)	<0.0001

4. Discussions

In this study we presented a new method to reduce the number of FP voxels of a CAD system based on the textural information extracted from mp-MR images. Our preliminary results show that a subset of texture features contains suitable information to identify malignant and healthy voxels. Besides, since the output of the ANN classifiers could be multiplied to the output of the CAD system, this method is able to provide a probability map, in which the information of tumor heterogeneity are kept.

The ANNs obtained the same results of a previously developed CAD system in terms of per-lesion sensitivity, however it showed a higher per-patient sensitivity (97% vs 95%). This was due to the fact that the CAD system uses some heuristic criteria based on the signal-intensity curves of the DCE sequence. Indeed, it might happen either that a tumor is not vascularized and therefore it does not show contrast uptake or that it is connected to a large FP area and, consequently, its median signal-intensity curve is more correlated to the prostatic area. Conversely, the proposed method, using 5x5 pixels elements, is independent from both the precision of the segmentation obtained by the CAD system and the DCE sequences.

As an additional advantage, our method is able to strongly increase the precision of the segmentation which is very promising, since it means that most part of the segmented tumor lies inside the manual ROI, making possible to perform a subsequent aggressiveness analysis on those regions that are surely inside the tumor. Indeed, recently there is an increasing research in finding imaging biomarker able to not only detect PCa but also to characterize it in order to identify patients who can be managed without radical treatment, open the way to personalized treatment options, minimizing the risk of overtreatment [28].

Moreover, the proposed approach is completely user-independent, requiring no inputs or parameter setting, therefore it could be applied to different kinds of CADs and images. In fact, a change of the starting conditions will only produce a different subset of selected features, with no modifications in the general pipeline.

Finally, the proposed method is able to strongly decrease the number of FPs of the CAD output, making the subsequent analysis performed by the radiologist more robust and less time consuming. The median number of FP obtained by the method is 12 which might seem higher than other proposed CAD [6,7], however they considered 3D connected regions inside the prostate, while in our method we counted each 2D connected region in the image. Therefore, even if the prostate is segmented by the CAD, there are a consistent number of FPs laying outside the prostate, due to the oversegmentation of the prostate.

There are few limitations in this method. First, we tested the system using a dataset acquired with the same scanner and protocol and using the endorectal coil. We are planning to acquire more patients with different protocols and without the endorectal coil, however, since textural features measure spatial variations of the signal intensity, the proposed method does not necessitate post-processing strategies to correct differences in T2w signal intensity ranges across patients, avoiding a related potential source of errors.

Second, the ANNs and voting performances showed in table 3 were obtained considering the same set of elements for training and validation. This is usually performed to identify and compare the classifiers results in the best conditions, removing the variability introduced by cross-validation. Indeed, if the classification accuracy would not have obtained good results in this situation, it would be impossible to generalize the capability of the network, meaning its ability to correctly predict points not used during training, for classification purposes. However, even if all malignant elements were used both for training and validation, only a part of benign elements were included in the training set. Thus, observing the final performance of the system in reducing the number of FP voxels, meaning that a huge number of non-tumoural elements was correctly identified, we can conclude that the generalization capability for the benign class is very high and we also can suppose that similar results could be obtained for the tumoural voxels.

Moreover, the same training set was used for the discretization and feature selection phases, that can bias the final result. A sensitivity analysis should be conducted in order to understand the influence of the training set on these three phases of the FP reduction system construction, in which the information contained in the dataset is always used in supervised manner.

Finally, only a set of texture features based on GLCM was used in this work. However, other texture analyses, such as Fractal dimension texture analysis [29] and Laws' texture energy measures [30], could be employed in order to further improve the FP reduction system performances. Analogous considerations are valid for the classi-

fication of malignant and healthy elements, where other classifiers beyond ANNs could be tested in the wide field of supervised learning.

In conclusion, we presented a method able to consistently reduce the number of FP findings, without using heuristic criteria derived from the time-intensity curves, and to increase the precision of the segmentation of the PCa. If validated on a larger testing set, this system could be introduced in a CAD to provide more precise results, thus making the analysis less time consuming and that can be more easily interpreted.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin. 2016;66(1):7-30.

[2] Lemaître G, Martí R, Freixenet J, et al. Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review. Comput Biol Med. 2015;60:8-31.

[3] Russo F, Regge D, Armando E et al. Detection of prostate cancer index lesions with multiparametric magnetic resonance imaging (mp-MRI) using whole-mount histological sections as the reference standard. BJU Int 2015. 118(1):84-94

[4] Arumainayagam N, Ahmed HU, Moore CM et al. Multiparametric MR imaging for detection of clinically significant prostate cancer: a validation cohort study with transperineal template prostate mapping as the reference standard. Radiology 2013;268(3):761-9.

[5] Bratan F, Niaf E, Melodelima C et al. Influence of imaging and histological factors on prostate cancer detection and localisation on multiparametric MRI: a prospective study. Eur Radiol 2013;23(7):2019-29.

[6] Litjens G, Debats O, Barentsz J, et al. Computer-aided detection of prostate cancer in MRI. IEEE Trans Med Imaging. 2014;33:1083-1092.

[7] Giannini V, Mazzetti S, Vignati A, et al. A Fully Automatic Computer Aided Diagnosis System for Peripheral Zone Prostate Cancer Detection using multi-parametric Magnetic Resonance Imaging. Comput Med Imaging Graph. 2015;46:219-226

[8] Chan I, Wells W, Mulkern R V., et al. Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier. Med Phys. 2003;30:2390.

[9] Langer DL, van der Kwast TH, Evans AJ, et al. Prostate cancer detection with multi-parametric MRI: Logistic regression analysis of quantitative T2, diffusion-weighted imaging, and dynamic contrast-enhanced MRI. J Magn Reson Imaging. 2009;30:327–334.

[10] Donati OF, Mazaheri Y, Afaq A, et al. Prostate cancer aggressiveness: assessment with whole-lesion histogram analysis of the apparent diffusion coefficient. Radiology. 2014;71:143–52.

[11] Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Trans Syst Man Cybern. 1973;:610–621.

[12] Rosati S, Meiburger KM, Balestra G, Rajendra Acharya U. Carotid wall measurement and assessment based on pixel-based and local texture descriptors. Journal of Mechanics in Medicine and Biology 2016;16(1):1640006 (16 pages).

[13] Vignati A, Mazzetti S, Giannini V, et al. Texture features on T2-weighted magnetic resonance imaging: new potential biomarkers for prostate cancer aggressiveness. Phys Med Biol. 2015;60:2685–701.

[14] Giannini V, Rosati S, Regge D, Balestra G. Texture Features and Artificial Neural Networks: A Way to Improve the Specificity of a CAD System for Multiparametric MR Prostate Cancer. In: MEDICON 2016, pp. 296-301

[15] Giannini V, Vignati A, De Luca M, et al. A novel and fully automated registration method for prostate cancer detection using Multiparametric Magnetic Resonance Imaging. J Med Imaging Heal Informatics. 2015;5(6):1171-1182.

[16] Gliozzi AS, Mazzetti S, Delsanto PP, Regge D, Stasi M. Phenomenological universalities: a novel tool for the analysis of dynamic contrast enhancement in magnetic resonance imaging. Phys Med Biol 2011;7;56(3):573-86.

[17] Castorina P, Delsanto PP, et al. Classification scheme for phenomenological universalities in growth problems in physics and other sciences, Phys Rev Lett 2006;96:188701.

[18] Mazurowski M A, Habas P A, Zurada JM, et al. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Netw. 2008;21:427-436.

[19] Soh L-K, Tsatsoulis C Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. IEEE Trans Geosci Remote Sens,1999;37:780–795. doi: 10.1109/36.752194

[20] Clausi DA. An analysis of co-occurrence texture statistics as a function of grey level quantization. Can J Remote Sens. 2014;28:45-62.

[21] Conners RW, Harlow C. A theoretical comparison of texture algorithms. IEEE Trans Pattern Anal Mach Intell. 1980;2:204-222.

[22] Kerber R. Chimerge: Discretization of numeric attributes. Proc tenth Natl Conf Artif Intell. 1992;123-128.

[23] Rosati S, Balestra G, Giannini V, et al. ChiMerge discretization method: Impact on a computer aided diagnosis system for prostate cancer in MRI. In: 2015 IEEE Int. Symp. Med. Meas. Appl. Proc. IEEE, pp 297-302.
[24] Dalton J, Deshmane A. Artificial neural networks. IEEE Potentials. 1991; 10;33-36
[25] Kuncheva L I, Whitaker C J, Shipp C A, Duin R P W. Is independence good for combining classifiers?. In: Pattern Recognition, 2000. 15th International Conference on. pp. 168-171 vol.2.
[26] Lam L, Suen CY. Application of majority voting to pattern recognition: an analysis of its behaviour and performance. IEEE Trans Syst Man Cybern, 1997; 27(5):553-568.
[27] Stamey T a, Freiha FS, McNeal JE, et al. Localized prostate cancer. Relationship of tumor volume to clinical significance for treatment of prostate cancer. Cancer. 1993;71:933–938.
[28] Donati OF, Mazaheri Y, Afaq A, et al. Prostate cancer aggressiveness: assessment with whole-lesion histogram analysis of the apparent diffusion coefficient. Radiology 2014;271(1):143-52.
[29] Pentland AP. Fractal-based description of natural scenes. IEEE Trans Pattern Anal Mach Intell. 1984;6(6):661-74.
[30] Laws K I. Rapid Texture Identification. In: 24th annual technical symposium. International Society for Optics and Photonics, 1980. p. 376-381.