A self-tuned graph-based framework for localization and grading prostate cancer lesions: An initial evaluation based on multiparametric magnetic resonance imaging

Weifu Chen 𝑎,𝑏, Mingquan Lin 𝑏, Eli Gibson 𝑐, Matthew Bastian-Jordan 𝑑, Derek W. Cool 𝑑, Zahra Kassam 𝑑, Huageng Liang 𝑓, Guocan Feng 𝑎, Aaron D. Ward 𝑒, Bernard Chiu 𝑏,*

𝑎 School of Mathematics, Sun Yat-sen University, Guangzhou, China
𝑏 Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China
𝑐 Biomedical Engineering, University of Western Ontario, London, Ontario, Canada
𝑑 Department of Medical Imaging, University of Western Ontario, London, Ontario, Canada
𝑒 Lawson Health Research Institute, London, Ontario, Canada
𝑓 Department of Urology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, China

ARTICLE INFO

Keywords:
Prostate cancer
Multiparametric MRI (mpMRI)
Prostate Imaging and Reporting Data System (PI-RADS)
Laplacian regularized regression model
Multiple kernel learning

ABSTRACT

Multiparametric magnetic resonance imaging (mpMRI) has been established as the state-of-the-art examination for the detection and localization of prostate cancer lesions. Prostate Imaging-Reporting and Data System (PI-RADS) has been established as a scheme to standardize the reporting of mpMRI findings. Although lesion delineation and PI-RADS ratings could be performed manually, human delineation and ratings are subjective and time-consuming. In this article, we developed and validated a self-tuned graph-based model for PI-RADS rating prediction. 34 features were obtained at the pixel level from T2-weighted (T2W), apparent diffusion coefficient (ADC) and dynamic contrast enhanced (DCE) images, from which PI-RADS scores were predicted. Two major innovations were involved in this self-tuned graph-based model. First, graph-based approaches are sensitive to the choice of the edge weight. The proposed model tuned the edge weights automatically based on the structure of the data, thereby obviating empirical edge weight selection. Second, the feature weights were tuned automatically to give heavier weights to features important for PI-RADS rating estimation. The proposed framework was evaluated for its lesion localization performance in mpMRI datasets of 12 patients. In the evaluation, the PI-RADS score distribution map generated by the algorithm and from the observers’ ratings were binarized by thresholds of 3 and 4. The sensitivity, specificity and accuracy obtained in these two threshold settings ranged from 65 to 77%, 86 to 93% and 85 to 88% respectively, which are comparable to results obtained in previous studies in which non-clinical T2 maps were available. The proposed algorithm took 10s to estimate the PI-RADS score distribution in an axial image. The efficiency achievable suggests that this technique can be developed into a prostate MR analysis system suitable for clinical use after a thorough validation involving more patients.

1. Introduction

Prostate cancer is the most common non-skin cancer in the United States with an estimated of 220,800 new cases in 2015 [1]. In Hong Kong, prostate cancer was the third most common cancer in men and accounted for 11.3% of all new cancer cases [2]. Fortunately, more than 90% of all prostate cancers are diagnosed at the localized stage and five-year survival rate is almost 100% for men diagnosed with localized cancer [3]. Hence, it is important for men with elevated risk to be periodically screened. The first-line screening tests include Digital Rectal Examination (DRE) and serum Prostate Specific Antigen (PSA) tests. If the DRE or PSA result is suspicious for cancer, transrectal-ultrasound-guided (TRUS-guided) biopsy is performed. Since prostate cancer lesions are difficult to be seen in ultrasound, TRUS-guided biopsy is not a procedure that targets suspicious lesions but a systematic technique that samples prostate regions in which tumours occur most frequently [4]. As a result, TRUS-guided biopsy missed 20 – 35% of detectable lesion in the first biopsy [5–8]. To increase the cancer
detection yield, repeated biopsies are required, leading to increased anxiety pain and morbidity for patients. Thus, sensitive image-based tools allowing for precise lesion localization are required in the development of targeted sampling strategies.

The widespread use of PSA screening since the early 90’s has led to a higher detection rate of localized and less aggressive tumours [9]. The development of focal therapies, such as cryotherapy and high-intensity focused ultrasound, has provided options for localized tumours to be treated with a lower risk of morbidity. Delineation of tumours is required for the administration of these therapies in order to minimize damage to the surrounding healthy tissues and organs. In addition to tumour localization, risk assessment is also important to identify suitable candidates for focal therapies.

Multiparametric MRI (mpMRI) combines anatomic and functional imaging techniques and has been shown to have high sensitivity and specificity in cancer localization [10–12]. Consensus guidelines have been established for the use of mpMRI [13], which recommended the combination of T2-weighted (T2W) images with at least two functional MRI techniques, typically the dynamic contrast enhanced (DCE) and diffusion-weighted (DW) MRI. T2W MR imaging is the most widely used MR sequence for anatomy visualization. It has high tissue contrast and spatial resolution for visualization of zonal anatomy and tumours, which typically appears as homogeneous low-intensity regions in the peripheral and transition zones [13][Fig. 1 (a)]. However, the specificity of T2W imaging is limited since benign abnormalities, such as post-biopsy membranes of malignant lesions. The apparent diffusion coefficients (ADC) characterizing the amount of diffusion are calculated from multiple DW images, and are typically displayed as a parametric map with foci are clinically significant when combined with DCE-MRI [11,19].

Although prior studies have been performed to investigate the use of mpMRI for prostate lesion detection and localization [10,11,15,19–24], most studies involved visual identification of lesions from 6 to 30 coars regions in the prostate instead of pixel-accurate lesion delineation. The need for manual identification in these studies was time-consuming and added observer variability to the result. Since these studies defined prostate regions differently, there is a large variation in the results obtained across studies; for example, the sensitivity and specificity in lesion detection from T2W images ranged from 54 to 91% and 27–91% respectively [10,11,25]. Turkbey et al. [11] divided the prostate into 30 regions and compared two ways of quantifying sensitivity and specificity. In the first approach, known as the stringent approach, a lesion was deemed to be undetected if it is not detected in the exact region where the lesion was detected in the histological examination, even if it was detected in one of the neighbouring regions in mpMRI. In the second approach, known as the neighbouring approach, a lesion is deemed to be detected if it is detected in a neighbouring region. The sensitivity and specificity for lesion localization obtained in the neighbouring approach were much higher than in the stringent approach (sensitivity: 42% vs. 73%, specificity: 83% vs. 89% in T2W), suggesting that localization accuracy is sensitive to the detection criterion and the region size. To address these issues, pixel-wise binary cancer classification algorithms using mpMRI have been proposed [26–31], but these algorithms were not evaluated on mpMRI images acquired according to the clinical consensus guidelines [13,32]. In particular, pixel-by-pixel T2 maps were available as inputs to the algorithms proposed in Refs. [26–31]. Although quantitative T2 maps are superior to T2W imaging in that it is not affected by variabilities in TR and the bias field inhomogeneity due to the use of endorectal coils, repeated T2W acquisitions are required at tens of echo times, thereby substantially lengthening the acquisition time. Considering that the acquisition protocol specified in the consensus guidelines already take 30–45 min, the further lengthening of the acquisition time cannot be afforded in clinical practice. Furthermore, although binary classifiers discussed above can provide information on the location and size of cancer foci, the knowledge of how likely these foci are clinically significant will further optimize diagnosis and treatment planning. It has been demonstrated that ADC is correlated with tumour aggressiveness [20], but this information could not be conveyed by the results generated by binary classifiers described above. The study reported in Ref. [31] is a notable study that estimated the pixel-based malignancy probability using a logistic regression model. Although the model provided a pixel-based malignancy probability in the continuous range from 0 to 1, it was trained using binary classified data [i.e., malignant and benign regions of interest (ROIs)] and the intermediate probabilities were obtained by mathematically model fitting. Without concrete examples of lesions associated with intermediate cancer risk to establish the clinical meaning of the fitted malignancy probabilities and to train the model, it is unclear how the likelihood generated by the model should be interpreted clinically. In addition, it is not possible to validate the malignancy likelihood against expert observations without.

Fig. 1. Outlines of prostate lesions by two radiologists on T2W, ADC and DCE images. Each row shows the contour drawn by a radiologist. 7 DCE images were acquired sequentially (Fig. 2) and the one with the maximum enhancement is shown. The radiologists assigned a PI-RADS score to each lesion shown in the three sequences. The score is shown at the bottom of each image.
establishing a standard way to interpret the probability.

The difficulty in arriving at a proper clinical interpretation from the malignancy probability map points to a more general issue that is hindering a more routine of mpMRI in the clinical setting: the lack of a standardized clinical agreement or guideline to rate the likelihood of the presence of clinically significant cancers based on observations from mpMRI. To fill this gap, the European Society of Urogenital Radiology (ESUR) and the American College of Radiology (ACR) [13,32] developed the Prostate Imaging-Reporting and Data System (PI-RADS), a standardized 5-point scoring system used to grade prostate lesions according to the likelihood of the lesion being clinically significant. This paper proposes the first algorithm to predict the PI-RADS score distribution in the prostate based on the mpMR images acquired according to the consensus guidelines [13,32]. Apart from the clinical innovation, the proposed graph-based regression algorithm has two major technical contributions. First, existing graph-based approaches are sensitive to the choice of the similarity between data points (i.e., edge weight) [33]. The proposed algorithm automatically estimates the edge weights based on the structure of the data points. Second, each pixel is equipped with many features extracted from mpMRI. The proposed regression model has a feature selection capability that gives heavier weights to inputs features important for accurate PI-RADS score estimation. Optimizing feature and edge weights in the proposed framework involves solving a series of quadratic programming problems, which can be completed in 8s for each prostate image consisting of over 2000 pixels. To facilitate comparison with previously proposed binary classifiers, the predicted PI-RADS score distribution was first binarized to highlight cancerous tissues; the sensitivity, specificity and accuracy were then computed and compared with previous binary classifiers. In addition, adding a treatment margin is recommended to ensure that the entire lesion is covered in focal therapies [34], we also investigated how adding margins of different widths would improve the sensitivity in lesion delineation.

2. Methods

Fig. 3 shows a flowchart of the proposed cancerous lesion localization and assessment framework. Each step in the flowchart is explained in detail in the following sections.

2.1. Study subjects, mpMRI acquisition and surrogate ground truth PI-RADS score

Images were acquired for 13 patients with biopsy-confirmed cancer prior to prostatectomy using a 3T GE Discovery MR750 (GE Healthcare, Waukesha, WI, USA) with an endorectal coil (Prostate eCoil, Medrad Inc., Warrendale, PA, USA). In particular, T2W, DW, DCE images were obtained using the following acquisition parameters: T2W 2D fast spin echo: TR: 4–9 s, TE: 158–163 msec and slice thickness: 2.2 mm; DW 2D echo-planar: TR: 4 s, TE: 70–77 msec, slice thickness: 3.3–3.6 mm, b-value: 600–800. DCE spoiled gradient-recalled echo: TR: 5.6–5.9 msec, TE: 2.1–2.2 msec, flip angle: 15°, 90-sec intervals, slice thickness: 2.8 mm. The apparent diffusion coefficient (ADC) was computed according to Ref. [9] using the MR750 console and displayed as a parametric map for subsequent analysis. Images were assessed by 4 observers (1 radiology resident and 3 radiologists with 5, 6, 2.5 and 2.5 years of experiences in prostate MRI assessment) following the Prostate Imaging-Reporting and Data System (PI-RADS) detection guidelines (Version 1) [13]. Each observer delineated lesions that were equivocally, likely and highly likely to be clinically significant (PI-RADS of 3–5 respectively) and assigned a PI-RADS score to each delineated lesion on T2W, DW and DCE MR images separately. Typical prostate cancer diagnosis workflow involves interpretation of the three MRI sequences
and assignment of PI-RADS scores, but does not include contouring of lesions. The time required for this workflow was 5–10 min for each prostate. Delineation of lesions took an additional 5 min per prostate.

Images acquired for each patient were then registered by an expert observer [35]. The images along with the contours delineated by the observers were resampled to a standard size to allow for pixel-by-pixel feature analysis. After resampling, 10 image slices with size 256 × 256 are obtained for each prostate with an in-plane pixel size of 0.5 × 0.5 mm and interslice distance of 3 mm. Since 75% of tumours develop in the peripheral zone, this study analyzed the peripheral zone only, which was represented by approximately 2000 pixels in each image slice as shown in Fig. 4. As histology results were not available for this study, a surrogate gold standard of PI-RADS scores was required to be defined for training and validation purposes. This gold standard should take into account of the different regions delineated on three imaging sequences by each observer and the PI-RADS score associated with each region as illustrated in Fig. 1. In the following discussion, we denote the PI-RADS scores assigned by different radiologists and in different MR sequences at Pixel \( i \) by \( E^a_b(i) \), where \( a \in \{1, 2, 3, 4\} \) and \( b \in \{1, 2, 3\} \) were used to index radiologists and MR sequences respectively. There are a total of 12 (3 sequences × 4 observers) scores assigned for each pixel. The overall pixel-wise likelihood score served as the surrogate ground truth for training and evaluation of the algorithm, and was determined by taking the maximum of the 12 PI-RADS scores, i.e.,

\[
E_i = \max \{E^a_b(i), a = 1, 2, 3, 4, b = 1, 2, 3\}. \quad (1)
\]

A lesion with PI-RADS score \( \geq 3 \) is considered as an identified lesion [36–38]. Eq. (1) is a conservative way of defining the overall likelihood score since the lesion would be deemed significant if any observer in any sequence scores \( \geq 3 \). In clinical practice, lesions deemed significant by any observer in any sequence are flagged and further tests are required to confirm or rule out cancer. The choice of using Eq. (1) to define the overall likelihood score was made to mimic this diagnostic procedure.

2.2. Features extracted from mpMRI

The following groups of features were extracted on a pixel-by-pixel basis. The parentheses show the number of features associated with each pixel:

- Physical coordinates of each voxel (3). This feature is important because the neighbors of a cancerous location are more likely to be cancerous. The physical coordinates were represented by a triple \((x, y, z)\).
- Grayscale values (9). As prostate lesions typically appear as homogeneous low-intensity regions in T2W images and have low ADC values, gray levels in T2W and DCE images were extracted to train and validate the proposed algorithm. As DCE MR images were generated in a temporal sequence, each pixel is equipped with multiple grayscale values representing temporal signal variation after the injection of a contrast agent.
- Sorted grayscale values in the neighbourhood (16). For the T2W and ADC images, grayscale values in the 8 neighbouring pixels were obtained and sorted in the ascending order. The reason that we chose to sort the grayscale values here was to make the feature and the lesion localization result rotation-invariant (i.e., the same region should be detected as cancerous even though the images are rotated).
- Semi-quantitative parameters for DCE MR sequences (6). Lesions are associated with increased vascularity, which leads to early enhancement and rapid washout of the contrast agent as characterized by DCE MR images. We computed the slope of the time-intensity curve by \( \Delta I_t = \frac{I_{t+1} - I_t}{t} \), where \( I_t \) and \( I_{t+1} \) are gray levels in the two adjacent time points, and \( \Delta t \) is the time interval between two consecutive scans.

2.3. A graph-based regression model for PI-RADS score prediction

The central idea of the proposed PI-RADS score prediction framework is to promote label consistency (i.e., pixels with similar features should be assigned a similar PI-RADS score – the label in graph theory), while minimizing the difference between the estimated and surrogate ground truth PI-RADS score derived from observer ratings for the training data as described in Sec. 2.1. Each pixel is equipped with (i) the input features described in Sec. 2.2, which for Pixel \( i \), are collectively denoted by the vector \( \nu_i \) and (ii) \( E_i \), the surrogate ground truth PI-RADS score derived from observer ratings as described in Sec. 2.1. A portion of data was reserved for training, which we referred to as the labelled data set and denoted it by \( X_L \), and the remaining data are unlabelled, denoted by \( X_U \). The whole set of data points is denoted by \( X \), which consists of the labelled and unlabelled data set.

Linear regression models the dependent variables as a weighted sum of input variables. This assumption of linear dependence between input and output variables may not hold, thereby compromising the prediction accuracy of the model. In addition, with many predictors, a linear model without regularization may not provide a unique solution. One way to solve this issue is using the Tikhonov regularization [39], which is applied in ridge regression [40]. To obviate the requirement for linearity, the Laplacian regularized regression model was applied in this study, which is an extension of Tikhonov regularization on manifold [41]. While conventional regressors such as linear regression, logistic regression and robust regression [42] established a model in the training phase and applied the trained model in the test data without considering unlabeled data points nearby in the feature space, a major advantage of the graph-based regression model proposed in this study is that it takes advantage of the feature vectors of neighbouring data points when predicting the score of a particular data point. The proposed graph-based model establishes the similarity with neighbouring labelled and
unlabeled data points and makes use of this contextual information to enhance prediction accuracy through maintaining label consistency in the neighbourhood.

Suppose the first \( n \) samples are labelled data, and the next \( N - n \) samples are unlabeled. The Laplacian regularized regression model was designed to minimize

\[
\mathcal{F}(\mathbf{E}_i) = \sum_{i=1}^{n} [\mathbf{E}_i(i) - \mathbf{E}_s(i)]^2 + \frac{\gamma}{2} \sum_{i=1}^{N} w_i [\mathbf{E}_i(i) - \mathbf{E}_s(j)]^2.
\]

(2)

where \( \gamma \) is a weighting parameter that will be tuned to optimize accuracy as described in Sec. 3.1. To simplify Eq. (2), we defined the similarity matrix \( \mathbf{W} = (w_{ij}) \),

\[
\mathbf{E}_s = \begin{bmatrix} E_s(1) & E_s(2) & \cdots & E_s(n) \end{bmatrix}^T \quad \text{and} \quad \mathbf{E}_i(i) = \begin{bmatrix} E_i(1) & E_i(2) & \cdots & E_i(n) \end{bmatrix}^T,
\]

and the surrogate ground truth PI-RADS score of Pixel \( i \) as defined in Sec. 2.1, \( \mathbf{E}_i(i) \) is the score for Pixel \( i \) and the diagonal matrix \( \mathbf{U} = \text{Diag}(1, 1, \ldots, 1, 0, 0, \ldots, 0) \). With these definitions, Eq. (2) can be simplified to:

\[
\mathcal{F}(\mathbf{E}_i) = (\mathbf{E}_i - \mathbf{E}_s)^T \mathbf{U} (\mathbf{E}_i - \mathbf{E}_s) + \gamma \mathbf{E}_i^T \mathbf{L} \mathbf{E}_i,
\]

(3)

where \( \mathbf{L} \) is the Laplacian matrix of the graph defined as \( \mathbf{L} = \mathbf{D} - \mathbf{W} \) and \( \mathbf{D} \) is the diagonal degree matrix defined by \( \mathbf{D} = \text{Diag}(d_1, d_2, \ldots, d_N) \), where \( d_i \) is the degree of Pixel \( i \), defined as \( d_i = \sum_j w_{ij} \). Here \( \mathbf{E}_i^T \mathbf{L} \mathbf{E}_i \) is a regularization term used to approximate the Laplace-Beltrami operator on the manifold [41]. Since both \( \mathbf{U} \) and \( \mathbf{L} \) are positive semi-definite matrices, Eq. (3) can be minimized by taking the derivative of \( \mathcal{F} \) and setting it to 0, resulting in the following solution:

\[
\mathbf{E}_i = (\mathbf{U} + \gamma \mathbf{L})^{-1} \mathbf{U} \mathbf{E}_s.
\]

(4)

2.4. Framework for simultaneous feature selection and graph construction

Two important issues have to be addressed before solving Eq. (2) for the pixel-based PI-RADS score. First, existing graph-based approaches are sensitive to the choice of the edge weight, denoted by \( w_{ij} \) in Eq. (2) [33]. The Gaussian function of the pairwise difference between each of the two nodes is commonly used as the similarity function, but this function is highly dependent on the variance of the Gaussian function, which is typically determined empirically in previous studies [33]. Second, among the four groups of features extracted in Sec. 2.2, some would be more useful than others in estimating the PI-RADS score. The contribution by each feature group to the final estimation should be modelled by weight parameters that can be automatically tuned by the regression framework in order to maximize PI-RADS score prediction accuracy. A major contribution of this paper is the development of a framework capable of tuning the edge weights and the feature weights simultaneously.

In this work, we estimated the feature weights and computed the similarity matrix simultaneously based on multi-kernel locally linear embedding. The proposed algorithm was inspired by the Locally Linear Embedding (LLE) scheme [43] and shares with LLE the assumption that the labelling and the feature spaces have the same graph structure, i.e., the graph generated for the labels has the same affinity matrix with the graph of the features as shown in Fig. 5. The reason for using multiple kernels to map features was that features in different groups are highly inhomogeneous in terms of the unit of measurement and the domain range. Similarities in different feature groups were mapped using different kernels to the range between 0 and 1 as detailed in Sec. 2.5. The proposed framework for simultaneous feature selection and graph construction consists of three steps detailed below. Note that \( n \) and \( N \) denote the number of labelled samples and all available samples respectively as defined in Sec. 2.3 and \( \mathcal{F}(i) \) denotes the set containing the \( K \) nearest neighbours of Pixel \( i \).

1. Step 1: Compute the edge weights of the graph of labels by solving the following quadratically constrained quadratic program (QCQP):

Fig. 5. Illustration of the edge weight and feature weight learning framework: the labels and the features share the same graph structure. (a) The graph structure in the label space, and the PI-RADS score \( E_s(i) \) of the central voxel of Pixel \( i \) could be locally approximated by the PI-RADS scores of its \( K \) nearest neighbors, which are indexed by \( \{i_1, i_2, \ldots, i_K\} \) in this figure. (b) Assuming that the affinity matrix \( \mathbf{W} \) was estimated by (a), the feature weights could be estimated by minimizing the reconstruction error of the features.
\[
\begin{align*}
\min_{W} & \quad \ell_1(W) = \sum_{i=1}^{n} \left( E_i(i) - \sum_{j \neq i} w_{ij} E_j(j) \right)^2 \\
\text{subject to} & : \\
& 1) \sum_{j \neq i} w_{ij} = 1, \ i = 1, 2, \ldots, n, \\
& 2) 0 \leq w_{ij} \leq 1, \ i = 1, 2, \ldots, n.
\end{align*}
\]  

In Eq. (5), \( \ell_1(W) \) is the reconstruction error for the pixel-wise PI-RADS scores associated with labelled samples. Constraints 1 and 2 stipulate that no sample makes negative contribution.

Step 2: Estimate the weights of grouped features of labelled samples through the following QCQP:

\[
\begin{align*}
\min_{\Lambda} & \quad \ell_2(\Lambda) = \sum_{i=1}^{G} \left\| \frac{1}{N} \sum_{j \in i} \left[ \lambda_1 \phi_1(v_{ij}) - \lambda_2 \phi_2(v_{ij}) - \cdots - \lambda_k \phi_k(v_{ij}) \right] \right\|^2 \\
\text{subject to} & : \\
& 1) \sum_{k=1}^{G} \lambda_k = 1, \\
& 2) 0 \leq \lambda_k \leq 1, k = 1, 2, \ldots, G.
\end{align*}
\]

where \( G \) is the number of feature groups, \( \Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_k]^T \) are the weights for features indexed from 1 to \( G \), \( \{ \phi_1, \phi_2, \ldots, \phi_k \} \) are the kernel functions associated with \( G \) feature groups and \( v_{ij} \) is the vector of Pixel \( i \) containing only features in the \( g \)th feature group. \( W \) was computed in Eq. (5). Since \( \sum_{j \in i} w_{ij} = 1 \), Eq. (6) can be written as:

\[
\begin{align*}
\min_{\Lambda} & \quad \ell_2(\Lambda) = \sum_{i=1}^{G} \left\| \frac{1}{N} \sum_{j \in i} \left[ \lambda_1 \phi_1(v_{ij}) - \lambda_2 \phi_2(v_{ij}) - \cdots - \lambda_k \phi_k(v_{ij}) \right] \right\|^2 \\
\text{subject to} & : \\
& 1) \sum_{k=1}^{G} \lambda_k = 1, \\
& 2) 0 \leq \lambda_k \leq 1, k = 1, 2, \ldots, G.
\end{align*}
\]

Using the kernel trick [44] and defining \( K_{ij} = \mathcal{K}(v_{ij}, v_{jl}) \), \( \ell_2 \) can be simplified to:

\[
\begin{align*}
\ell_2(\Lambda) &= \sum_{i=1}^{G} \sum_{j=1}^{N} \sum_{l \in i} w_{jl} \sum_{g=1}^{G} \lambda_g \left[ \phi_g(v_{ij}) - \phi_g(v_{jl}) \right]^2 \left[ \phi_g(v_{ij}) - \phi_g(v_{jl}) \right]^T \\
&= \sum_{g=1}^{G} \lambda_g \sum_{i=1}^{G} \sum_{j=1}^{N} \sum_{l \in i} w_{jl} \left[ \phi_g(v_{ij}) - \phi_g(v_{jl}) \right]^T \left[ \phi_g(v_{ij}) - \phi_g(v_{jl}) \right] \\
&= \sum_{g=1}^{G} \lambda_g \sum_{i=1}^{G} \sum_{j=1}^{N} \sum_{l \in i} w_{jl} \left( K_{ij} - K_{il} - K_{lj} + K_{ll} \right). 
\end{align*}
\]

Here \( \mathcal{K}(\cdot, \cdot) \) is the kernel function for the \( g \)th group of features. Weights of the group features \( \Lambda \) were obtained after solving Eq. (6).

Step 3: Edge weights were computed for all the labelled and unlabelled samples by solving the following problem:

\[
\begin{align*}
\min_{W} & \quad \ell_3(W) = \sum_{i=1}^{n} \left\| \frac{1}{N} \sum_{j \neq i} w_{ij} \left[ \lambda_1 \phi_1(v_{ij}) - \lambda_2 \phi_2(v_{ij}) - \cdots - \lambda_k \phi_k(v_{ij}) \right] \right\|^2 \\
\text{subject to} & : \\
& 1) \sum_{j \neq i} w_{ij} = 1, \ i = 1, 2, \ldots, N, \\
& 2) 0 \leq w_{ij} \leq 1, \ i = 1, 2, \ldots, N.
\end{align*}
\]

Again, using the kernel trick, \( \ell_3 \) were simplified to:

\[
\ell_3(W) = \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} \sum_{g=1}^{G} \lambda_g \left( K_{ij} - K_{il} - K_{lj} + K_{ll} \right). 
\]

The proposed algorithm involves solving three consecutive QCQPs, which have been studied thoroughly and can be solved in polynomial time [45,46]. Computational times required to solve these three problems are detailed in Sec. 4.

2.5. Kernel functions

Due to the inhomogeneous properties of the grouped features, we mapped the similarity between samples into the same domain range in the proposed algorithm (Eqs. (8) and (10)). Kernel trick [44] is a widely used technique used for adjusting the inhomogeneity of the features of the data, since the implicit inner product defined by the same category of kernel functions can map features into the same range [44]. The Gaussian heat kernel is typically used if there is no prior information on the distribution of a data set and was used in the proposed algorithm:

\[
K^g(v_{ij}, v_{jl}) = \exp \left[ -\frac{\|v_{ij} - v_{jl}\|^2}{2\sigma_g^2} \right],
\]

where \( K^g \) is the kernel function for the \( g \)th group of features and \( \sigma_g \) is the scaling parameter for \( K^g \). In this study, \( \sigma_g \) was defined as \( \sigma_g = \sigma \times \text{median}(D_{ij}) \) where \( \sigma \) is a scaling coefficient, \( D_{ij} \) is a distance matrix with \( D_{ij} \) representing the Euclidean distances between \( v_{ij} \) and \( v_{jl} \) and median(\( D_{ij} \)) is the median of all elements in the matrix operand. In this setting, \( \sigma \) is the only user specified parameter, which will be tuned according to the procedure described in Sec. 3.1. The Gaussian heat kernel functions map the similarity between groups of features into the same range [0,1].

3. Experimental methods

Since the peripheral zone of each prostate was represented by approximately 20,000 pixels, typically spanning 10 transverse slices, generating results simultaneously for all slices was computationally impractical for a single prostate, let alone for all prostates in the study population. The proposed algorithm involves the similarity matrix, and for a regression problem involving 20,000 pixels, the similarity matrix is of size 20,000 by 20,000, containing 400 million elements, requiring an amount of memory not supportable by most PCs. In addition, the similarity matrix of this size would be sparse, and eigenvalue decomposition of the matrix would be prohibitively slow for clinical uses. The experimental workflow consists of two phases: (a) Parameter tuning and (b) validation. Both phases involve training and testing on a prostate-by-prostate basis. In both phases, feature and edge weights involved in the proposed algorithm were required to be trained as specified in Sec. 2.4.
3.1. Parameter tuning

The proposed regression framework involves three user-defined parameters: (i) $k$, the number of nearest neighbors considered in the graph construction; (ii) $\gamma$, the regularized coefficient in GFHF (Eq. (2)) and (iii) $\sigma$, the coefficient for defining the standard deviations of $k^2$ in Eq. (11). For this reason, there was a need to tune these parameters to optimize lesion localization accuracy. The optimized parameters were then used in the validation phase. One prostate from 13 prostates involved in this study was randomly selected for parameter tuning and this prostate was excluded in the subsequent evaluation. To evaluate the sensitivity of the three tuned parameters to the prostate used for tuning, tuning was repeated in a second prostate. Two transverse slices of the chosen prostate were selected as the training set to train the feature and edge weights in the proposed algorithm. Results were generated for another two image slices in a total of 216 settings with $k \in \{5, 10, 20, 30, 50, 80\}$, $\gamma \in \{10^{-6}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and $\sigma \in \{0.1, 0.2, 0.3, 0.5, 0.8, 1.3\}$. The parameters associated with the minimum sum of the false negative and false positive rates (FNR and FPR respectively) were used in the testing phase. As later demonstrated in the Results section, the optimum values of these three parameters were not sensitive to the prostate chosen for tuning. For this reason, the parameters obtained from the first tuning experiment were used in the validation phase. It should be emphasized that the prostate used in tuning was not involved in the validation phase, thereby leaving 12 prostates for validation.

3.2. Validation

For each prostate, one slice was randomly selected as the training set and the PI-RADS score was evaluated for the remaining slices on a slice-by-slice basis. A constraint for selecting the training slice was that it must have at least an identified lesion (i.e., a lesion with an overall PI-RADS $\geq 3$). To take into account of the variability of the results arising from the use of different training sets, we repeated the experiments with five different training sets denoted by TS1, TS2, TS3, TS4 and TS5. Since a training slice must include cancerous pixels, we randomly selected a slice with at least one identified lesion for each prostate to form a training set. Each training set consists of 12 slices, one from each of the 12 prostates being evaluated. We also ensured that a slice included in one training set is not included in the other.

Classification performance was evaluated by sensitivity, specificity, accuracy and receiver-operator characteristic (ROC) analysis. These evaluations require the PI-RADS score distribution generated either manually or by the algorithm to be binarized into cancerous and non-cancerous regions and this operation depends on the threshold PI-RADS score used. Previous investigation shows that histologically validated tumour lesions have a mean PI-RADS scores between 3 and 4; and if the score associated with lesions was reported by an accumulative score obtained by adding the PI-RADS scores assigned to T2W, ADC and if the score associated with lesions was reported by an accumulative score obtained by adding the PI-RADS scores assigned to T2W, ADC and ictological ground truth established by thresholding the overall likelihood map (Eq. (11)) by either 3 and 4, resulting in 10 ROC curves (5 training sets $\times$ 2 thresholds). The area under each ROC curve (AUC) served as another metric to quantify classification performance. It is of practical importance in focal therapies to know what margin is necessary to achieve full coverage of the tumour. For this reason, we expanded the algorithm identified lesion with a PI-RADS score $\geq 3$ by in a 1-pixel interval and plotted how detection sensitivity increased with margin width for the five experimental settings trained by different datasets.

The PI-RADS score variability obtained manually and by the algorithm were quantified by the pixel-by-pixel variance estimated by ANOVA [47], denoted by $s$, and subsequently compared. $S$ is defined in the following equation:

$$S = \frac{1}{N_{\text{pixels}}} \sum_{i=1}^{N_{\text{pixels}}} \left( \sum_{j=1}^{N_{\text{obs}}} \frac{1}{N_{\text{obs}}} \sum_{k=1}^{N_{\text{pixels}}} \left( P_{jk} - \bar{P}_{jk} \right)^2 \right),$$

where $P_{jk}$ is the PI-RADS score assigned by Observer $i$ for Pixel $j$ in the $k$th prostate, $\bar{P}_{jk}$ is the average PI-RADS at Pixel $j$ of the $k$th prostate, $N_{\text{obs}}$ is the number of observers, $N_{\text{pixels}}$ is total number of pixels in the $k$th prostate and $N_{\text{prostate}}$ is total number of prostates. There are four observers for manual delineation and five observers for algorithm (i.e., TS1-TSS).

4. Results

In the first tuning experiment, the minimum sum of FNR and FPR was 0.39, which corresponds to the FNR and FPR of 0.27 and 0.12 and was achieved with $k_{\text{opt}} = 50, \gamma_{\text{opt}} = 1, \sigma_{\text{opt}} = 0.8$. In the tuning experiment involving the second prostate, the minimum sum of FNR and FPR was 0.59 (FNR: 0.44, FPR: 0.15), which was achieved with $k_{\text{opt}} = 50, \gamma_{\text{opt}} = 10^{-4}, \sigma_{\text{opt}} = 0.8$. Although the optimization was performed exhaustively involving 216 settings, it is useful to understand the effect of the three parameters on classification performance separately. Fig. 6(a) and (b) show the plots of FNR + FPR against $k$ in the two tuning experiments with $\gamma$ and $\sigma$ fixed at $\gamma_{\text{opt}}$ and $\sigma_{\text{opt}}$ respectively. Fig. 6(c) and (d) show the plots of FNR + FPR against $\sigma$ in the two tuning experiments with $k$ and $\gamma$ fixed at $k_{\text{opt}}$ and $\gamma_{\text{opt}}$ respectively.

The proposed algorithm was applied to estimate PI-RADS score using five different training sets (i.e., TS1-TSS) as described in Sec. 3.2. Table 1 shows the means and standard deviations of the evaluation metrics computed based on the surrogate gold standards generated by thresholding the overall PI-RADS score (Eq. (11)) by the thresholds of 3 and 4. The low standard deviations for sensitivity, specificity, accuracy and DSC indicate that the classification performance was not sensitive to the training sets used. Tukey’s multiple comparison tests were performed to compare the sensitivity, specificity, accuracy and DSC generated by TS1 to TS5 and no significant difference was detected in all metrics at a significant level of 5%.

Fig. 7 shows the delineation results for three sample slices. The first, second and third columns correspond to the example slices in which the proposed algorithm produced good, average and lower than average

where $A_1$ and $A_2$ denote the regions enclosed by the algorithm and manually identified boundaries respectively. $|\cdot|$ denotes the area of the region specified by the operand.
The sum of false negative and false positive rates (FNR and FPR respectively) in the tuning experiments involving two prostates versus the three internal parameters of the framework $k$, $\sigma$ and $\gamma$. (a, b) fixing $\sigma$ and $\gamma$, varying $k$, in the two experiments respectively; (c, d) fixing $k$ and $\gamma$, varying $\sigma$, in the two experiments respectively; (e, f) fixing $k$ and $\sigma$, varying $\gamma$, in the two experiments respectively.

Fig. 6. The sum of false negative and false positive rates (FNR and FPR respectively) in the tuning experiments involving two prostates versus the three internal parameters of the framework $k$, $\sigma$ and $\gamma$. (a, b) fixing $\sigma$ and $\gamma$, varying $k$, in the two experiments respectively; (c, d) fixing $k$ and $\gamma$, varying $\sigma$, in the two experiments respectively; (e, f) fixing $k$ and $\sigma$, varying $\gamma$, in the two experiments respectively.
Table 1

The means and standard deviations of sensitivity, specificity, accuracy and DSC obtained in 5 experiments involving different training sets (i.e., TS1 to TS5). The performance metrics were obtained by comparison with the overall PI-RADS score (Eq. (1)) binarized at the thresholds of 3 and 4.

<table>
<thead>
<tr>
<th>PI-RADS threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>DSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.77 ± 0.04</td>
<td>0.86 ± 0.01</td>
<td>0.85 ± 0.01</td>
<td>0.65 ± 0.01</td>
</tr>
<tr>
<td>4</td>
<td>0.65 ± 0.05</td>
<td>0.93 ± 0.01</td>
<td>0.88 ± 0.01</td>
<td>0.64 ± 0.03</td>
</tr>
</tbody>
</table>

The first row[i.e., Fig. 7 (a, e, i)] shows the three slices in axial T2W images. The second row[i.e., Fig. 7 (b, f, j)] shows the T2W images with the surrogate ground truth PI-RADS scores superimposed. The third row[i.e., Fig. 7 (c, g, k)] shows the T2W images with predicted PI-RADS scores superimposed. The fourth row[i.e., Fig. 7 (d, h, l)] shows a comparison of the lesions identified by the proposed algorithm with the surrogate ground truth. Red indicates true positive, green indicates false negative and blue indicates false positive.

Fig. 7. Comparison of the algorithm PI-RADS scores with the surrogate ground truth PI-RADS scores derived from observers’ ratings (Eq. (1)) for example axial locations for three patients. Each column shows the images obtained in the same axial location for a patient. (a), (e), (i) are the axial T2W images. (b), (f), (j) are the images with the surrogate ground truth PI-RADS scores superimposed. (c), (g), (k) are the images with the algorithm PI-RADS scores superimposed. (d), (h), (l) shows a comparison of the lesions identified by the algorithm with the surrogate ground truth. Red indicates true positive, green indicates false negative and blue indicates false positive. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)
the contrast agent is completely cleared out through urine within 48 h, there is a rate of 7/5,000,000 that serious allergic reaction would occur. For this reason, patients were required to be monitored by a doctor for this reaction throughout the MRI scanning session. If DCE imaging did not contribute much to the classification performance, it will be advisable to remove DCE from the protocol to reduce the scanning cost. Although we have established that the temporal DCE change contributed the least among the four groups of features, DCE intensity was incorporated in the intensity feature group and further evaluation will be required to assess the contribution of DCE intensity features before making a conclusion on the value of DCE imaging in the proposed lesion delineation framework.

Fig. 10 shows the sensitivity versus the expanding margins for TS1 to TS5. A coverage ranging from 88 to 95% was achieved when the margin size reaches 5 pixels (2.5 mm) and the coverage changed very little with the margin increasing. Fig. 11 shows one example with margins of size from 1 to 6 superimposed. Although only the centre of the lesion on the left was detected by the proposed algorithm as shown in Fig. 11(c), 100% coverage for all lesions in this slice were achieved after adding a 2.5 mm (5 pixels) margin. The coverage never reached 100% in the five experiments involving different training sets because 8 lesions were missed by the algorithm in all experiments. The volumes of these lesions were estimated by thresholding the overall observer rating (Eq. (1)) by 3. The areas of each lesion as appeared in contiguous slices were summed together to obtain a total area, which was then multiplied by the slice thickness (i.e., 3 mm) to arrive at a volume estimation. According to this estimation, the 8 lesions had volumes ranging from 0.03 to 0.16 cm^3, which are smaller than 0.2 cm^3 and are not considered clinically significant according to Epstein’s criteria [48].

The PI-RADS score variabilities (\(S\) in Eq. (13)) for manual and algorithm segmentations were 1.26 and 0.67 respectively. It should be noted that \(S\) was computed within cancerous regions detected either manually or by the algorithm (i.e., the regions where either the surrogate ground truth or the algorithm scores were greater than or equal to 3). Fig. 12 shows \(S\) obtained manually and by the algorithm in two example prostate images. Fig. 12(a)-(g) are images of one example, where (a) to (d) show the T2W images with the lesion boundaries outlined by four radiologists superimposed, (e) is the image with the surrogate ground truth PI-RADS scores superimposed and (f) and (g) show the images with \(S\) associated with manual and algorithm segmentations superimposed. Fig. 12(h)-(n) are corresponding images generated for another example. Both examples demonstrated that the segmentation variability for the algorithm is lower as compared to manual segmentation. To emphasize the fact that \(S\) was generated on a pixel-by-pixel basis for the entire peripheral zone instead of only at the lesions, the colour bars in the \(S\) maps encode a zero inter-observer variability by dark red instead of black, as black could not be visualized on the superimposed maps.

An average time of 8s was required to train the model for each prostate, and 84s was required to evaluate the PI-RADS score on all slices of a prostate, which averaged to about 10s to evaluate the PI-RADS score on an axial slice, as compared to 15 min for an human observer to assess and delineate each prostate. The experiments were performed on Intel(R)

---

**Table 2**

Area Under the ROC curves (AUC) obtained in 5 experiments involving different training sets.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>TS1</th>
<th>TS2</th>
<th>TS3</th>
<th>TS4</th>
<th>TS5</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.92</td>
<td>0.90</td>
<td>0.89</td>
<td>0.90</td>
<td>0.91</td>
</tr>
<tr>
<td>4</td>
<td>0.93</td>
<td>0.91</td>
<td>0.92</td>
<td>0.91</td>
<td>0.92</td>
</tr>
</tbody>
</table>

---
Core (TM) i7 CPU @ 3.0 GHz platform with 8 GB memory. Features were extracted from mpMR images using a custom-built C++ program with frequent uses of the Visualization Toolkit (VTK) [49] and Insight Segmentation and Registration Toolkit (ITK) libraries [50,51]. The feature weight and edge weight learning framework and the Laplacian regularized regression model were implemented in MATLAB (Natick, MA).

5. Discussion and conclusion

Although mpMRI has shown promising results in detection and localization of prostate cancer [12], there was a lack of standardization in acquiring mpMRI as well as a structured reporting scheme to communicate observations from the images [52]. In order to improve the diagnostic value of mpMRI and improve communications between clinicians and radiologists, the ESUR and ACR established consensus guidelines for the acquisition protocol and a structured interpretation and reporting scheme named PI-RADS [13,32]. However, inter-observer agreement of the PI-RADS scoring system is only moderate [36]. Inter-observer variability in lesion delineation is another factor that may impact assessment reproducibility as the lesion volume is an important criterion in determining whether the lesion is clinically significant [48]. The ultimate goal of the proposed algorithm is to reduce the impact of observer variability in lesion delineation as well as PI-RADS score assignment. Since results can be generated efficiently using the algorithm, the human effort involved in the analysis are also reduced.

The proposed algorithm enables objective prediction of the PI-RADS score distribution throughout the peripheral zone for the first time. This distribution shows the likelihood of each lesion being clinically significant and provided more information to assist treatment decision making than previously introduced binary classifiers [27–30]. Although the predicted PI-RADS score distribution offered information regarding how likely each lesion was clinically significant, which was not provided by previously introduced binary classifiers [27–30], there is a requirement to evaluate the tumour classification capability of the proposed method and compare this classification performance with existing binary classifiers. To make such evaluation possible, there was a need to binarize the surrogate gold standard PI-RADS distribution maps (Eq.(1)) and those by the algorithm. No cutoff value was provided in the consensus guideline [13,32] and the determination of which has been the topic of a number of recent investigations [36–38,53]. Schimmoller et al. [36] reported that histologically verified malignant lesions have mean PI-RADS scores of 4.2, 4.5 and 3.5 in T2W, DW and DCE images respectively. A number of investigations [36–38] concluded that a sum score of 10 from the three contrasts was optimal for classification of malignant and benign tumours.

Fig. 10. Change of sensitivity with the margin widths in five experiments involving different training sets (i.e., TS1 to TS5).

Fig. 11. One example with margins of size from 1 to 6 to cover the missed regions. (a) The axial T2W MR image; (b) the axial T2W MR image with the surrogate ground truth PI-RADS scores derived from observers’ ratings (Eq. (1)) superimposed and (c) the image with predicted PI-RADS scores superimposed, which are colour-coded according to the colour bar; (d-i) the images show the coverage of the surrogate for ground truth lesions by the algorithm identification with margins of size from 0 to 5 pixels superimposed. Red indicates true positive, green indicates false negative. The lesions were fully covered by the algorithm identification when using margins of a size larger than 4 pixels. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)
while some suggest to use a threshold sum score of 9 to increase the sensitivity. In the current study, we evaluated the sensitivity, specificity, accuracy and DSC of the proposed algorithm in two settings using thresholds of 3 and 4 to binarize the manual and algorithm distribution maps. The delineation sensitivity, specificity and accuracy achieved by the proposed algorithm were 65–77%, 86–93% and 85–88% respectively, which are comparable to those produced by previously proposed binary classifiers evaluated with quantitative T2 maps available, which were associated with sensitivities of 64–78% and specificities of 78–90%

Although graph-based regression frameworks have been previously introduced, the proposed graph-based framework was the first algorithm capable of determining both the edge weights in the graph and the feature weights simultaneously based on the structure of the data being analyzed. A generic strategy was introduced in Ref. to estimate edge weights, but the feature weights were not tuned and uniform feature weights were used in the proposed algorithm. Tuning of feature weights is important in our applications of PI-RADS score prediction from mpMRI as different groups of features contribute differently to the prediction accuracy as shown in Fig. 9. Ref. introduced a multi-kernel graph embedding algorithm for dimensionality reduction and random forest classifiers for classification. Although the proposed algorithm could learn the kernel weights, the graph similarity matrix was defined in a fixed parametric form, which is not as flexible as the similarity matrix defined non-parametrically in our proposed framework. The use of the hierarchical brute force algorithm for kernel weight optimization was also time-consuming. The kernel weights was optimized iteratively with a step size of 0.001, which implies that, on average, 500 multi-kernel eigenvalue decomposition problems were required to be solved throughout the entire training process. In contrast, our framework involves 3 quadratically constrained quadratic problems that could be efficiently solved in polynomial time. Our algorithm also guaranteed global optimization of the convex constrained problems, whereas the iterative optimization algorithm described in Ref. did not.

This study is associated with a number of limitations. The first limitation is related to the small sample size analyzed in this study in which we pilot the proposed self-tuned graph-based approach. As this study requires four expert observers to delineate around 10 image slices for each prostate for three image contrasts, manual delineation and assignment of PI-RADS score have already taken a considerable amount of time. While we acknowledge that a larger population will be required to take into account a larger spectrum of biological variability across patients, it is appropriate to evaluate the proposed graph-based approach in a pilot study before expanding the validation study to a larger population. In fact, the number of subjects involved in this study is similar to previous studies that focused on algorithm development for prostate delineation and we evaluated much more transverse images per prostate than these prior studies. Another limitation is related to the use of the lesion boundaries delineated by radiologists and the PI-RADS scores assigned by them to train and evaluate the proposed algorithm, instead of histologically obtained lesion boundaries, which are typically considered as the gold standard for lesion detection. Since the algorithm was trained...

---

**Fig. 12.** Two examples showing a comparison of the inter-observer variabilities associated with manual and algorithm segmentations. (a)–(g) show the T2W image of one example, with the lesion contours outlined by four radiologists superimposed on (a)–(d), the surrogate ground truth PI-RADS scores superimposed on (e) and the pixel-by-pixel inter-observer variabilities (S in Eq. (13)) associated with manual and algorithm delineations superimposed on (f) and (g) respectively. (h)–(n) show the corresponding images of another example. To emphasize the fact that S was generated on a pixel-by-pixel basis for the entire peripheral zone instead of only at the lesions, the colour bars in the S maps encode a zero inter-observer variability by dark red instead of black, as black could not be visualized on the superimposed maps. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)
using the surrogate gold standard derived from radiologists’ ratings in this pilot study, the best the algorithm can do was to reproduce the lesion boundaries delineated by the radiologists and their PI-RADS rating with reduced analysis time and inter-observer variability. However, if provided the histologically obtained lesion boundaries and more training samples, the proposed algorithm has the potential to avoid false positive detection of prostate lesions caused by the lesion-mimicking appearances of normal anatomic structures, such as the central zone and neurovascular bundles, and non-cancerous abnormalities, such as post-biopsy hemorrhage, acute and chronic prostatitis and post-inflammatory scars (all listed among the ten pitfalls of prostate mpMRI in Ref. [56]), thereby improving the specificity of prostate mpMRI.

The evaluation of the algorithm was limited to the peripheral zone. It was a priority to evaluate the proposed algorithm in the peripheral zone as 75% of cancerous tumours develop in this zone [57,58] and these tumours are most susceptible to adverse clinical outcomes, such as extracapsular extension, seminal vesicle and lymphovascular invasion [59]. Although we hypothesize that our framework would be capable of detecting transition zone tumours after it has been trained as described in this paper, lesion delineation in the peripheral and transition zones must be performed separately as tumours in the two zones exhibit markedly different characteristics [13,60,61]. The PI-RADS guidelines score lesions in the peripheral and central zones appeared in the T2W images using separate criteria [13,61]. The localization of peripheral zone tumours involves discrimination of the hypo-intense homogeneous regions from the hyper-intense homogeneous normal regions, whereas central gland tumour localization is achieved by separation of hypo-intense homogeneous lenticular shaped regions from the normal heterogeneous nodular structures of the central gland. The capability of the proposed algorithm in delineating transition zone tumours will be required to be evaluated in a future study.

This algorithm was trained using a slice from a series of consecutive transverse images acquired for each prostate. For this reason, this algorithm supports a semi-automatic delineation workflow in which an expert observer is required to segment one slice to initialize the algorithm. Because of semi-automatic nature of the algorithm and the correlation of the tumours’ appearance in different image slices of the same prostate, although only 12 prostates were evaluated in this study, we do not expect the lesion localization performance to be drastically different if more prostates are evaluated, although further evaluation should be performed to verify this hypothesis. T2W, DW and DCE images involved in this study are associated with subject-specific contrast and intensity even after normalization. The application of the endorectal coil in this study is a major contributing factor for this subject-specific variability. The bias field inhomogeneity introduced at locations close to the coil leads to a spatial variation in the MR intensity with regions close to the coil showing higher intensity and this effect varies across patients depending on the positioning of the coil. In addition, different patients have different amount of scarring, inflammation and prostatic intraepithelial neoplasia, which would affect the appearance of normal tissues. A fully automated method would require a model to be developed to characterize these variations, and training such model would require more data than available in this study. Nonetheless, the proposed algorithm has substantially reduced the time required for assessing lesions in mpMRI. Further reduction of user interaction can be achieved; for example, instead of requiring a user to delineate lesions in an image slice, a workflow can be designed requesting users to provide a few seed points inside and outside each lesion and specify the PI-RADS score for each seed point inside a lesion. In addition, the semi-automatic tool introduced herein has made a contribution in reducing the inter- and intra-observer assessment of mpMRI. Manually assigned PI-RADS score has been previously reported to have a moderate inter-observer agreement with $\kappa$ for interpreting T2W, ADC and DCE ranging from 0.55 to 0.65 [36]. Observer lesion segmentation variability would also have an impact on lesion assessment. The proposed framework took these two sources of variabilities into account in training the algorithm and came up with PI-RADS distribution maps that are highly reproducible as shown in the five experiments in which the algorithm was trained by different data-sets. The results generated by this framework can be used as a reference for training less experienced observers.

The application of this graph-based regression model is not limited to the prediction of the PI-RADS score in mpMRI images. In addition of its use in prostate cancer detection and localization, mpMRI has also shown to be able to identify rupture-prone plaque components such as the lipid-rich necrotic core, intraplaque hemorrhage, calcification and thin/rupture fibrous cap [62]; the hazard ratios for cerebral ischemic outcomes associated with these high-risk features range from 3 to 6 [63]. Although these vulnerable plaque components can be delineated from mpMRI manually, an automated framework for segmenting plaque components such as the proposed method could reduce the segmentation time, thereby making the analysis time more affordable in clinical practices. Observer variability would also be reduced by an automated segmentation method. Although we believe the current approach can be directly adopted for plaque component segmentation by replacing the PI-RADS scores with labels indexing different components, extensive validation has to be performed in a future study.

In conclusion, we developed and validated the first framework that is capable of predicting the PI-RADS score from mpMRI acquired according to the consensus protocol. In addition, this algorithm is the first graph-based method that has the ability to optimize the edge weights of the graph and the weights of the feature groups used in PI-RADS score prediction simultaneously, thereby obviating the need for manually tuning these parameters as required in previous studies. Fast computation could be achieved because the algorithm only involves solving a linear system of equations (Eq. (2)) and a few QCQP problems (Eqs. (6)–(10)). The efficiency afforded by this framework suggests that it can be developed into an MR prostate analysis system suitable for clinical use.

Acknowledgements

Dr. Chiu is grateful for the funding support from the Basic Research Free Exploration Program of the Science Technology and Innovation Committee of Shenzhen Municipality, China (Project No. JCYJ20160428155118212) and the City University of Hong Kong Strategic Research Grants (Nos. 7004425 and 7004617). Dr. Chen and Prof. Feng are partially supported by the NSFC under grants No. 61703443.

References
