IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 23, 2020, accepted September 10, 2020, date of publication September 14, 2020,
date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3023902

Deep Radiomic Analysis to Predict Gleason Score in Prostate Cancer

AHMAD CHADDAD[1,2], MICHAEL J. KUCHARCZYK[2], CHRISTIAN DESROSIERS[3], IDOWU PAUL OKUWOBI[1], YOUSEF KATIB[2], MINGLI ZHANG[4], (Member, IEEE), SAIMA RATHORE[5], PAUL SARGOS[2], AND TAMIM NIAZI[2]

[1]School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin 541004, China
[2]Lady Davis Institute for Medical Research, McGill University, Montreal, QC H3T 1E2, Canada
[3]Ecole de Technologie Supérieure, University of Quebec, Montreal, QC H3C 1K3, Canada
[4]Montreal Neurological Institute, McGill University, Montreal, QC H3A 2B4, Canada
[5]Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA 19104, USA

Corresponding author: Ahmad Chaddad (ahmadchaddad@guet.edu.cn)

ABSTRACT Convolutional neural networks (CNNs) require large amounts of data for training, beyond what can be acquired for current radiomics models. We hypothesize that deep entropy features (DEFs) derived from existing CNNs can be applied to MRI images of prostate cancers (PCa) to reliably predict the Gleason score (GS) of PCa lesions. In this study, we analyzed 112 lesions acquired from 99 PCa patients, either pre-biopsy or pre-treatment, their associated GS, and multi-parametric MRI (mpMRI) sequences. Our approach is based on the extraction of DEF features produced in individual layers of 9 pre-trained CNN models. We first analyze DEFs from separate CNNs using the Wilcoxon test and Spearman correlation to find significant features associated with GS. In a multivariate analysis, we then use the combined DEFs of all CNNs as input to a random forest (RF) classifier for predicting the Gleason grade group of patients. Among the 9 pre-trained CNNs, the NASNet-mobile architecture offered the features most correlated to GS ($\rho = 0.47$; $p<0.05$). From the 7,857 combined features, 11 DEFs could differentiate GS < 8 from GS ≥ 8 (corrected $p < 0.05$). Moreover, the RF classifier discerned GS of 6, 3+4, 4+3, 8 and ≥ 9 with an AUC (%) of 80.08, 85.77, 97.30, 98.20, and 86.51, respectively. Our results suggest that the DEFs can be used to differentiate GS of PCa lesions with the highest accuracy of GS ≥ 8 based on mpMRI. DEFs could improve diagnosis accuracy, reduce the risks of misclassification, help to better assess prognosis, and individualize patient care approaches.

INDEX TERMS Prostate cancer, deep learning, Gleason score, radiomics.

I. INTRODUCTION

Radiomics is a technique to extract large number of features from medical image to build prediction models. However, this technique suffers from overfitting when a large number of features are directly used to train and test predictive models [1]. While, convolutional neural networks (CNNs) have shown an outstanding ability to identify complex associations in high-dimensional data for disease diagnosis and treatment planning [2]. In order to get the benefit of the representational capacity of well-known deep CNN designs (e.g., ResNet, GoogleNet, etc.) and overcome the issue of overfitting and limited datasets, we propose to encode the CNN features which are a key challenge in personalized medicine of grading prostate cancer (PCa). For devising a personalized approach to patients with PCa, the diagnosis and management depend on the assessment of biological aggressiveness of the malignancy, for which the gold standard is prostate biopsy [3], [4]. The biopsy specimen is evaluated in a standardized fashion by specialized physicians, i.e. the pathologists, for assigning a Gleason Score (GS) to the malignancy [5]. However, this procedure can lead to complications [6], incurs a significant cost [7] and may need to be repeated if sampled tissue are inadequate for analysis [8]. Additionally, significant discrepancies can arise between the biopsy-evaluated GS and what is found during surgery (e.g., radical prostatectomy [9]). Important inter-observer variability may also be found in biopsy reports [10]. Hence, there is a

The associate editor coordinating the review of this manuscript and approving it for publication was Sotirios Goudos.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

critical need to develop non-invasive methods that can predict the PCa grades to improve delivery of high precision care for these patients.

Evidence supports the role of multiparametric magnetic resonance imaging (mpMRI) performed before biopsy as a guide for PCa assessment [3], [4]. Recent studies have shown that MRI offers advantages over transrectal ultrasound (TRUS) guided biopsies in ruling out clinically-significant disease, and that MRI followed by targeted biopsies improves the detection rate compared to systematic biopsies [11], [12]. Likewise, the MRI-FIRST study [13] found that combining a MRI-targeted approach with a systematic biopsy provided substantial added value. The standardized method for reporting prostate mpMRI, known as Prostate Imaging Reporting and Data System (PI-RADS), stratifies prostatic lesions by their potential for malignancy [14], [15]. PI-RADS's efficacy ranges from 74-82% for its sensitivity to detect PCa and 65-94% for its negative predictive value [12]. Recently, important efforts have been invested to improve PCa screening, risk stratification and individualized patient management. Radiomics and CNNs offers an effective and non-invasive way to predict oncological outcomes [16]–[20]. For PCa, multiple studies have identified imaging features that correlate with GS [21]–[23]. However, a common limitation of radiomics approaches is the requirement of having enough high-quality data to both train and validate the model.

Our work proposes a novel radiomics method based on deep entropy features (DEFs) to predict the GS of PCa lesions from mpMRI. In contrast to traditional imaging features, DEFs are learned from a convolutional neural network (CNN) and thus have the potential of capturing more informative characteristics of an image. In a previous study, DEFs obtained from a three-dimensional CNN were shown to be capable of describing differences between brain MRI of patients with Alzheimer's and healthy control subjects [24]. Expanding on this study, the current work evaluates the potential of DEFs, extracted from all layers of multiple network architectures, to offer a more reliable prediction of GS in prostate mpMRI. To overcome the challenge of limited training data, the proposed approach exploits a transfer learning strategy where pre-trained CNNs are used to extract generic imaging features, which are then summarized into a small set of DEFs. These radiomic descriptors offer a highly-compact representation of image texture which captures the heterogeneity of imaged tissues. To ensure reproducibility, our study employs a publicly-accessible and verifiable database of prostate cancer images.

The major contributions of our paper are as follows:

- This work is the first comprehensive work in encoding well-known CNNs with quantifier function (Shannon entropy) for predicting the GS of patients with PCa.
- We demonstrate the effectiveness of using the deep entropy features to deal with RF in predicting the GS.
- We propose a small set of DEFs that encode multi-scale (e.g., deep feature maps) PCa-related information.
- We present the classification performance of different prostatic zone and its relationship with GS.

The rest of this article is structured as follows. Section II describes the data used in this study as well as the proposed pipeline. We then present the experimental results in Section III and discuss our main findings in Sections IV. Finally, Section V concludes with a summary of our work's main contributions and results

II. MATERIALS AND METHODS

This section describes the public dataset of PCa and explains the steps data acquisition in preprocessing procedures, the proposed pipeline and the performance metrics.

A. PATIENTS AND DATA ACQUISITION

The Cancer Imaging Archive (TCIA) was accessed to acquire patient data and mpMRI images for this study. TCIA hosts a publicly-accessible repository of labelled imaging data sponsored by the International Society for Optics and Photonics (SPIE), National Cancer Institute/National Institutes of Health (NCI/NIH), and the American Association of Physicists in Medicine (AAPM) [25]. Our study uses the labeled training data of the SPIE-AAPM-NCI Prostate MR Gleason Grade Group Challenge (PROSTATEx-2), comprising a total of 112 PCa lesions from 99 different subjects [26]. The testing data of the PROSTATEx-2 dataset was not considered in our work since it does not contain labels. The GS of each tumor was determined via MRI-localization. Specifically, MR studies were read and reported by an expert radiologist (>20 years of experience in prostate MR) who indicated areas of suspicion with a score per modality using a point marker. A biopsy was then performed for areas considered as cancer. The biopsy process was performed under MR-guidance and confirmation scans of the biopsy needle in situ were done to achieve the highest localization accuracy [25]. At each stage, a physician with relevant expertise in the procedure was involved. Based on the biopsy specimens and the mpMRI report, tumors were partitioned into different Epstein grade groups [27], as per their GS: G1 (GS ≤ 6), G2 (GS $3+4=7$), G3 (GS $4+3=7$), G4 (GS $=8$), or G5 (GS ≥ 9) (Table 1). As all patient data were accessed through an anonymized public resource, no institutional review board or Health Insurance Portability and Accountability Act approval was required.

Images were acquired by either a Siemens 3T MAGNETOM Trio or Skyra MRI [27]. Pixel spacing, slice thickness, and contrast varied within the included cohort. Image heterogeneity was corrected via resampling all the images to an ordinary voxel resolution of 1 mm^3, for a total size of $320 \times 320 \times 19$ voxels. The unified grey image adjustment to the [0-255] range for normalization where the maximum grayscale value is 255, the minimum value is 0, and the rest have been linearly transformed. We identified an 84×84 pixel 2D ROI from each MRI sequence. The ROI selected included the abnormal area expertly identified by

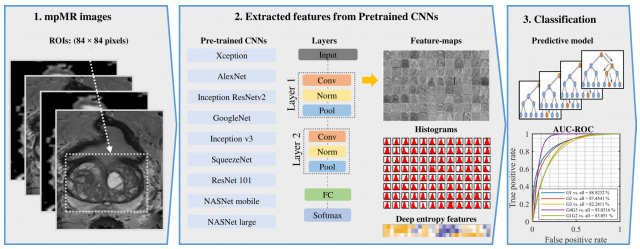


FIGURE 1. Illustration of the proposed methodology to generate and evaluate deep entropy features. 1) T2-WI, ADC and DCE MRI images of 99 patients with 112 lesions; ROIs determined by size of 84 × 84 pixels. 2) Upsampling of the ROI (e.g., 224 × 224 pixels) for processing by the 9 pre-trained CNNs. The texture of the CNN layer-blocks (e.g., convolution layers, max pooling layers, ReLU, normalization and fully-connected layer) is quantified using entropy. 3) Features capacity to predict the GS evaluated via uni- and multivariate analyses.

the PROSTATEx-2 Challenge. Using these ROIs, we derived DEFs from the T2-WI, ADC, and DCE images.

B. DEEP ENTROPY FEATURE EXTRACTION

The deep entropy features (DEFs) employed in the proposed radiomics pipeline (**Figure 1**) measure the spatial heterogeneity (i.e., texture) of feature maps computed by a pre-trained deep CNN. In this study, we considered 9 well-known 2D CNN architectures that were pre-trained on natural images from the ImageNet database: Xception [28], AlexNet [29], Inception ResNet-v2, GoogleNet, Inception-v3 [30], SqueezeNet [31], ResNet101 [32], NASNet-mobile [33] and NASNet-large [33]. Considered the 2D ROIs extracted from mpMRI, each network was applied separately on T2-WI, ADC, and DCE MRI series to obtain imaging descriptors corresponding to the feature maps of convolutional blocks. In most CNNs, a convolutional block is composed of the following sequence of operations: convolution, pooling, normalization, and rectified linear unit (ReLU) activation. For extracting DEFs, we computed the entropy in each feature map of the 9 CNNs. Toward this goal, the values at each position of a feature map are aggregated into a discrete probability distribution (i.e., histogram) by grouping them into 256 equal-sized bins. Let p_i be the ratio of values of a feature map falling into bin i, entropy is computed as

$$H = -\sum_{i=1}^{256} p_i log p_i$$

Feature maps with high entropy correspond to textures having more pronounced heterogeneity. The number of obtained DEFs can vary for each CNN, according to the number of convolution blocks in the network.

C. DEEP ENTROPY FEATURE EVALUATION AND MODELING

Uni- and multivariate analyses were performed to assess the relationship between DEFs and GS. First, we used Spearman correlation to identify the features most correlated to GS [34]. The Wilcoxon rank-sum test [35] was then employed to compare the distribution of features in lesion groups defined based on GS. For this second analysis, we considered five different partitions of lesions in two separate groups: G1 vs all (G2-5); G2 vs all (G1+G3+G4-5); G3 vs all (G1-G2+G4-5); G4G5 vs all (G1-3); G1G2 vs all (G3-5). For each of these binary partitions, we performed a Wilcoxon rank-sum test on individual features to identify those having a significantly different distribution across the two lesion groups. To account for multiple comparisons, the p-values of all Wilcoxon tests and Spearman correlation estimates were adjusted using the Holm-Bonferroni correction. Statistical significance was defined as corrected $p < 0.05$ [36].

In a multivariate analysis, we used the DEFs as input to a random forest (RF) classifier for predicting different combinations of Gleason grade groups. While different classifiers could be used for the same tasks, we have chosen the RF classifier as it is performing well when training data is small and has an optimized selection mechanism that allows interpretability [37]. By integrating decision tree bagging with random subspace search, it decreases errors due to the heterogeneity of training data and offers a strong generalization for new samples [38]. RF classifiers also have a relatively small number of hyper-parameters to tuning compared to more complex models such as neural networks, the major factors being the number of trees, the maximum tree depth, the minimum number samples in a node. In our experiments, these hyper-parameters were selected using grid search on a validation set. In this context, we set 500, 15 and 4, respectively, for the number of trees, the maximum tree depth, and the minimum number samples in a node.

In this analysis, we considered the same partitions as before, i.e. G1 vs all, G2 vs all, G3 vs all, G4G5 vs all, and G1G2 vs all to define five binary classification problems. A 5-fold cross-validation (CV) was performed to obtain performance measures. In this internal validation technique, data samples are randomly divided into five folds. Each of these folds is then used, in turn, to calculate the area under the ROC curve (AUC) of an RF model trained with remaining samples (those in the 4 other folds). To generate a quantifiable performance metric, we then computed the average of AUC values across all five folds. The out-of-bag sample permutation error of the RF classifier was used to measure the relative importance of each feature for predicting the Gleason grade group. Importance values were computed for every RF tree and then averaged over the entire ensemble. To obtain normalized values, we divided them by the standard deviation of the ensemble. Features are considered to be predictive of the grade group if they have a positive importance value [39].

To further validate results, for each classification task (G1 vs all, G2 vs all, etc.), we randomly divided the datasets into a training (70%) and testing (30%) cohort using balanced populations of each grade group in training. The performance of predictive models was measured based on the AUC and the confusion matrix obtained on the test samples. Moreover, we analyzed the localized relationship between DEFs and GS by considering separately the lesions located in three different anatomical zones of the prostate, i.e. peripheral

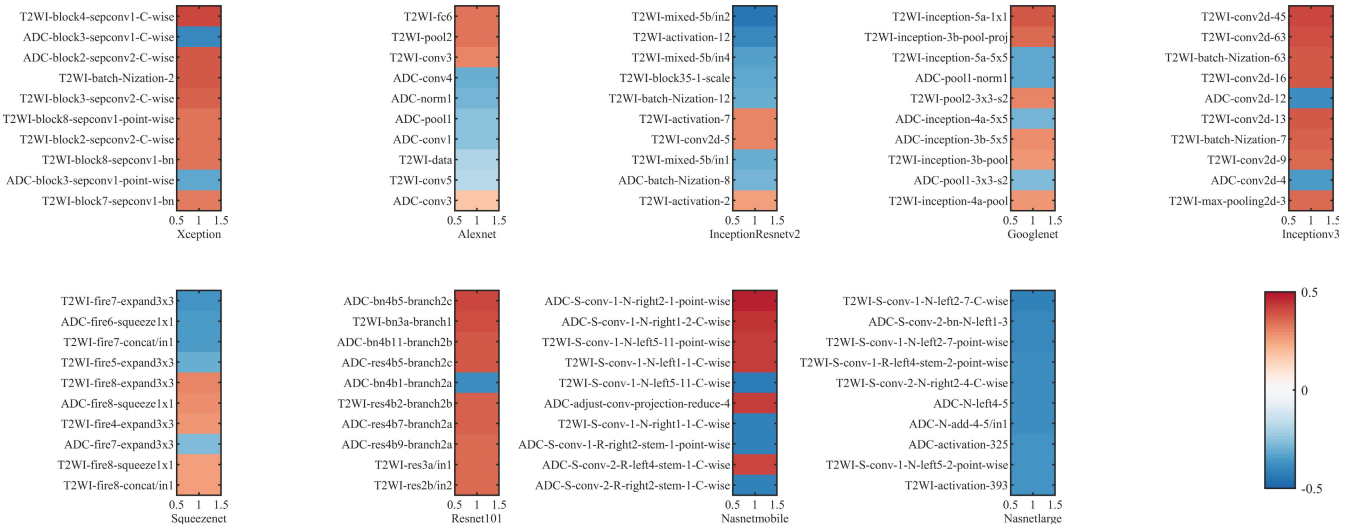


FIGURE 2. Spearman's correlation analysis of the top 10 deep entropy features (DEFs). Color-coded from -0.5 (dark blue) to 0.5 (dark red). *x* represents the CNN architecture and the *y*-axis represents the DEFs best correlated with the GS among each CNN architectures.

TABLE 1. Distribution of prostate cancer patients by Gleason score groups.

Gleason Score Group	*n*	Gleason score
G1: Grade Group 1	36	≤ 6
G2: Grade Group 2	41	7 (3+4)
G3: Grade Group 3	19	7 (4+3)
G4: Grade Group 4	8	8 (4+4; 3+5; 5+3)
G5: Grade Group 5	8	≥ 9

zone (PZ), transitional zone (TZ), and anterior zone (AZ). The zone labels of PCa lesions were provided by TCIA in the dataset. Among the 112 lesions, 50 where located in the PZ, 17 in the TZ, and 45 in the AZ. For each zone, we measured the Spearman correlation between DEFs and GS and used the Kruskall-Wallis test to establish significant differences between the feature distributions of distinct Gleason grade groups. Once again, Holm-Bonferroni correction of p-values was used to account for multiple comparisons. All our processing/analysis steps were performed using the Matlab Statistics and Machine Learning Toolbox.

III. RESULTS

A. CHARACTERISTICS OF THE STUDY POPULATION

Histopathological data was available to confirm the GS of the 112 malignant lesions identified in mpMRI. All mpMRI images had the same three series available, i.e. T2 weighted imaging (T2 WI), apparent diffusion coefficient (ADC), and dynamic contrast enhancement (DCE) series. Among these 112 lesions/findings, there were 36, 41, 19, 8, 8 tumors with GS ≤ 6 (G1), GS $= 7$ (3+4; G2), GS $= 7$ (4+3; G3), GS $= 8$ (4+4, 3+5, or 5+3; G4), and GS ≥ 9 (G5), respectively (**Table 1**). In the cohort of 99 patients (average age 65 years, range 42–78 years), 87 patients had one lesion, 11 patients had two, and a single patient had three [40].

TABLE 2. Deep Entropy Features derived from pre-trained convolutional neural networks.

CNNs	Layers (n)	Features (n)
Xception	71	133
AlexNet	8	13
InceptionResNet-v2	164	52
GoogleNet	22	82
Inception-v3	48	219
SqueezeNet	18	38
ResNet101	101	244
NASNet-mobile	914	703
NASNet-large	1244	1135

B. ANALYSIS OF DEEP ENTROPY FEATURES

Table 2 reports the number of layers in each of the 9 pre-trained CNN architectures and their corresponding number of unique DEFs. The layer names of these architectures are reported in *Supplementary Table S1*. Combining features of all 9 networks yields a total of 7,857 unique DEFs. The Spearman rank correlation (ρ) between GS and all significant DEFs (determined by input modality and layer name) is given in **Table 3** and **Figure 2**. A DEF is significant if it has a correlation p-value < 0.05 after correction. The correlation ρ and corrected p-values of all layers can be found in *Supplementary Table S2*. It can be seen that the *NASNet-mobile* architecture yields the most correlated DEFs and the feature with the highest absolute correlation of $\rho = 0.47$. After Holm-Benferroni correction, a total of 5, 4, 3, 2, 6, 3, 5, 16 and 2 DEFs extracted from *Xception, AlexNet, InceptionResNet-v2, GoogleNet, Inception-v3, SqueezeNet, ResNet101, NASNet-mobile* and *NASNet-large* architectures, respectively, were statistically correlated with Gleason grade groups. Statistically-correlated DEFs are found for both T2-WI and ADC modalities in all 9 pre-trained CNNs.

Results of the Wilcoxon rank sum test comparing the distribution of DEF values across Gleason grade groups are shown

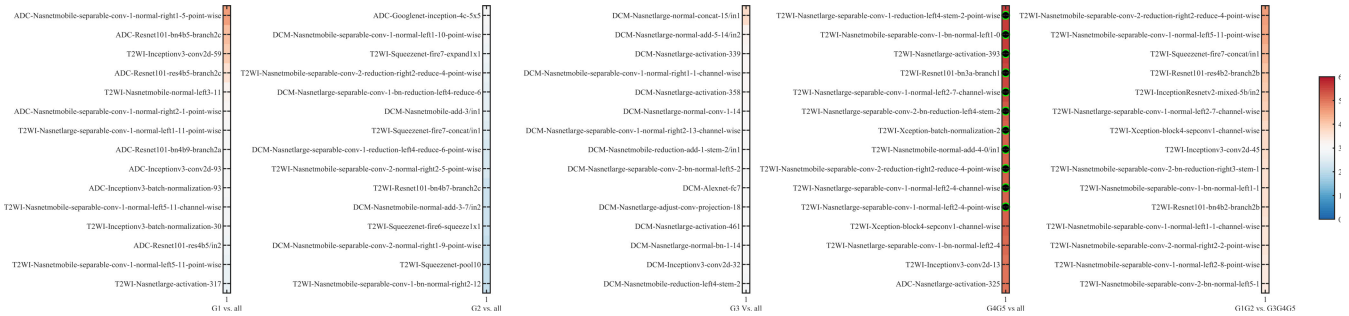


FIGURE 3. Statistical significance heatmap of DEFs in discriminating Gleason grade groups. Each column represents the 15 most significant DEFs per Gleason grade group comparison, p-values as per the Wilcoxon test. Significant DEFs (corrected $p < 0.05$) are identified by a green-black circle. Color-coded from 0 (dark blue: least significance) to 6 (dark red: greatest significance). *x* represents the compared Gleason grade groups and the *y*-axis corresponds to the most significant DEFs per Gleason grade group comparison, for each CNN architecture.

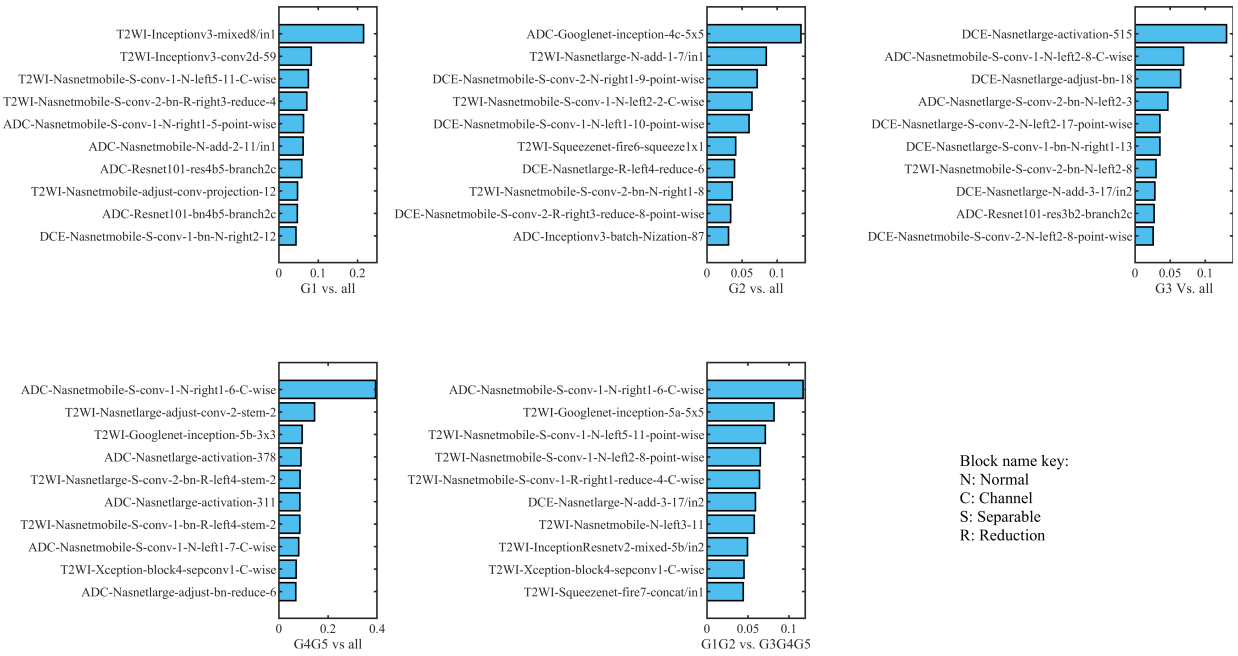


FIGURE 4. Importance values of deep entropy features which predicted the Gleason score groups. The 10 most important deep entropy features for each Gleason grade group prediction task, using the RF model's out-of-bag sample permutation error as importance measure.

in **Figure 3**. We find 11 DEFs with statistically-significant differences for the GS ≥ 8 (i.e., G4G5) vs GS < 8 lesion partition, with p-value < 0.05 following correction. All these significant DEFs were derived from T2-WI images. No significant DEFs were found when comparing between other lesion partitions due to the p-value correction on a large number of comparisons. The full set of p-values is provided in *Supplementary Table S3*.

C. PREDICTIVE MODELS FOR GLEASON GRADE GROUP

Table 4 summarizes the results of the 5-fold CV analysis evaluating the RF classifier's ability to predict the Gleason grade group of the 112 lesions. When considering DEFs of each network architecture separately, the *NASNet-mobile* yields the best prediction in all but one case (i.e., G3 vs all), a result which is consistent with the previous correlation analysis. The highest accuracy is obtained when discriminating between G4G5 and other Gleason grade groups (G4G5 vs all), with an AUC of 92.68%. Furthermore, combining DEFs derived from all 9 CNNs (7,857 features) into the same RF model boosts performance in all but one classification tasks compared to *NASNet-mobile*, with relative AUC improvements of 0.38, 4.80, 3.90, 0.35 and -1.67, for G1 vs all, G2 vs all, G3 vs all, G4G5 vs all, and G1G2 vs all, respectively.

D. FEATURE IMPORTANCE FOR GLEASON SCORE PREDICTION

Figure 4 compares the DEFs, with the greatest importance values from the pretrained CNN's application to the mpMRI's T2-WI, ADC, and DCE series. Specifically, **Figure 4** shows the 10 DEFs with the highest importance value (i.e., permutation error on out-of-bag samples) for each classification task. Overall, the image modality (i.e., T2-WI, ADC or DCE) and network architecture leading to the most predictive features

TABLE 3. Correlation of deep entropy features with the Gleason grade group.

Pre-trained CNNs	DEFs (modality-layer)	ρ	p-value
Xception	T2WI-block4-sepconv1-C-wise	0.42	0.001
	ADC-block3-sepconv1-C-wise	-0.39	0.007
	ADC-block2-sepconv2-C-wise	0.38	0.011
	T2WI-batch-Nization-2	0.38	0.013
	T2WI-block3-sepconv2-C-wise	0.36	0.034
AlexNet	T2WI-fc6	0.33	0.014
	T2WI-pool2	0.32	0.015
	T2WI-conv3	0.30	0.040
	ADC-conv4	-0.30	0.040
InceptionResNet-v2	T2WI-mixed-5b/in2	-0.44	0.0001
	T2WI-activation-12	-0.40	0.001
	T2WI-mixed-5b/in4	-0.34	0.035
GoogleNet	T2WI-inception-5a-1x1	0.37	0.009
	T2WI-inception-3b-pool-proj	0.34	0.048
Inception-v3	T2WI-conv2d-63	0.39	0.01
	T2WI-batch-Nization-63	0.38	0.017
	T2WI-conv2d-16	0.38	0.018
	ADC-conv2d-12	-0.37	0.023
	T2WI-conv2d-13	0.37	0.024
	T2WI-batch-Nization-7	0.36	0.041
SqueezeNet	T2WI-fire7-expand3x3	-0.36	0.009
	ADC-fire6-squeeze1x1	-0.35	0.016
	T2WI-fire7-concat/in1	-0.34	0.017
ResNet101	ADC-bn4b5-branch2c	0.41	0.002
	T2WI-bn3a-branch1	0.39	0.009
	ADC-bn4b11-branch2b	0.38	0.018
	ADC-res4b5-branch2c	0.38	0.022
	ADC-bn4b1-branch2a	-0.37	0.027
NASNet-mobile	ADC-S-conv-1-N-right2-1-point-wise	**0.47**	0.0003
	ADC-S-conv-1-N-right1-2-C-wise	0.43	0.002
	T2WI-S-conv-1-N-left5-11-point-wise	0.43	0.003
	T2WI-S-conv-1-N-left1-1-C-wise	0.43	0.004
	T2WI-S-conv-1-N-left5-11-C-wise	-0.42	0.006
	ADC-adjust-conv-projection-reduce-4	0.42	0.006
	T2WI-S-conv-1-N-right1-1-C-wise	-0.42	0.008
	ADC-S-conv-1-R-right2-stem-1-point-wise	-0.41	0.01
	ADC-S-conv-2-R-left4-stem-1-C-wise	0.41	0.013
	ADC-S-conv-2-R-right2-stem-1-C-wise	-0.41	0.014
	ADC-N-left4-6	0.4	0.02
	ADC-S-conv-1-bn-R-right3-stem-2	-0.4	0.021
	T2WI-S-conv-1-N-left1-1-point-wise	0.4	0.022
	ADC-S-conv-1-N-right1-5-point-wise	0.4	0.024
	T2WI-S-conv-2-bn-R-right3-stem-1	0.39	0.043
	T2WI-R-add4-stem-2/in1	0.38	0.047
NASNet-large	T2WI-S-conv-1-N-left2-7-C-wise	-0.41	0.022
	ADC-S-conv-2-bn-N-left1-3	-0.40	0.032

*Bold font and underscore represent the highest correlated feature with GS.

varies from one classification task to another. For instance, DCE-based features obtain a high importance value for G3 vs all, however DEFs derived from this modality are not so predictive for other GS partitions. Moreover, this analysis demonstrates the RF model's ability to select relevant features from a large set, with most DEFs obtaining a null importance value. The importance values of DEFs derived from all MRI modalities and CNN layers are reported in *Supplementary Table S4*.

E. PREDICTION ACCURACY ON CLASS-BALANCED SAMPLES

A possible confound in the previous analysis is the class imbalance when splitting lesions based on Gleason grade groups. In the next analysis, we evaluate the RF model's performance on binary partitions having the same number of samples. Specifically, we use a balanced sample partition of 36/36 for G1 vs all, 41/41 for G2 vs all, 19/19 for G3 vs all, 16/16 for G4G5 vs all, and 35/35 for G1G2 vs all. For each classification task, we then use 70% of samples in each partition for training and 30% for testing. **Figure 5** gives the confusion matrix and ROC curves of the RF model for the five tasks. The model achieves an accuracy of 80.95%, 83.33%, 100.00%, 100.00% and 47.62%, respectively, for G1 vs all, G2 vs all, G3 vs all, G4G5 vs all and G1G2 vs all. Correspondingly, the highest AUC is obtained for discriminating between GS $<$ 8 and GS $\geq$ 8 (G4G5 vs all), with an AUC of 98.20%.In addition, **Table 5** illustrates the performance

TABLE 4. Area under the ROC curve (%) of the RF classifier for discriminating between different Gleason grade group partitions, when using DEFs from 9 pre-trained CNNs.

Pretrained CNNs	G1 vs all	G2 vs all	G3 vs all	G4G5 vs all	G1G2 vs all
Xception	77.20	72.51	75.30	87.35	77.54
AlexNet	68.49	68.92	68.78	86.49	71.57
InceptionResNet-v2	68.67	63.57	76.82	81.46	70.89
GoogleNet	81.41	81.12	79.92	85.47	79.35
Inception-v3	83.72	79.54	81.76	82.90	72.09
SqueezeNet	76.70	71.50	74.42	83.43	77.38
ResNet101	83.46	73.73	73.91	82.87	72.29
NASNet-mobile	88.44	82.65	78.38	92.68	84.72
NASNet-large	68.75	79.68	81.48	90.69	64.06
Combined features	88.82	87.45	82.28	93.03	83.05
*Combined features represent the concatenation of 7857 features that derived from 9 pretrained CNNs.					

TABLE 5. Performance comparison of AUC with existing techniques.

	Feature methods/techniques	G1 vs all	G2 vs all	G3 vs all	G4G5 vs all	G1G2 vs all
Our work	Deep entropy features (5-fold CV)	88.82	87.45	82.28	93.03	84.72
	Training 70%, testing30%	80.08	85.77	97.30	98.20	85.11
Chaddad et al. **[41]**	Histogram+GLCM+NGTDM+GLSZM	83.40	72.71	77.35	-	-
Chaddad et al. **[42]**	Joint intensity matrices (JIM)+GLCM	78.40	82.35	64.76	-	-
Toivonen et al. **[43]**	GLCM+ LBP+HOG+Gabor+Haar+filters	88.00	-	-	-	-
Jesen et al. **[40]**	Histogram+GLCM+ GLSZM	85.00	89.00	94.00	86.00	83.00
Cao et al. **[44]**	FocalNet	-	81.00	79.00	-	-

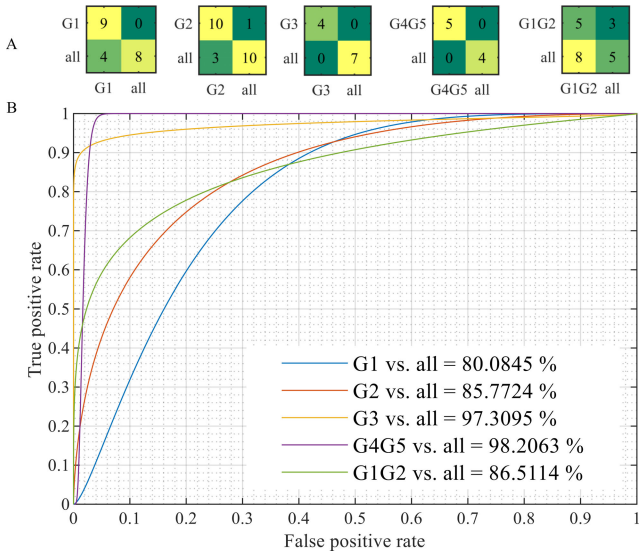


FIGURE 5. Random forest models implementing deep entropy features. Models predicting the Gleason score group of each PCa lesion, with their associated (A) confusion matrix (x and y coordinates represents the true class and predicted class, respectively) and (B) receiver operator characteristic curve.

comparison of AUC with the previous studies in predicting the GS of PCa. Except the G2 vs. all, our approaches showed a better AUC value in predicting the G1 vs. all, G3 vs. all, G4G5 vs. all and G1G2 vs. all comparing to the existing techniques.

To assess the impact of splitting samples, we repeat 70-30% split random samples 20 times. For G1 vs. all, G2 vs. all, G3 vs all, G4G5 vs all and G1G2 vs all, respectively, the average AUC value of 86.40, 86.35, 95.40, 94.70 and 85.05% is achieved.

F. ZONE-SPECIFIC RELATIONSHIP BETWEEN DEFs AND GLEASON SCORE

It is unknown if peripheral zones (PZ) are biologically different than transitional Zones (TZ). With radomic analysis applied on mpMRI in understanding the mechanism of PCa, the impact of tumor location on biological behavior may have significant implications for optimum treatment modalities [40]. **Figure 6A** shows the DEFs most correlated with GS, for lesions located in the three anatomical zones of the prostate (i.e., PZ, TZ, and anterior – AZ). We see that features with moderate ($0.3 \leq |\rho| \leq 0.7$) or high correlation ($|\rho| > 0.7$) are found in all three zones. However, after p-value correction, statistical significance can only be established for DEFs in peripheral lesions, with absolute correlation in the 0.6-0.63 range. Although more pronounced correlation is found for transitional lesions, statistical significance could not be confirmed due to the smaller number of lesions in this zone (i.e., 17 compared to 50 for PZ). Similarly, **Figure 6B** displays the results of the Kruskal-Wallis test comparing the DEFs among each Gleason grade groups by anatomic zone. Following Holm-Bonferroni correction, 39 DEFs derived from PZ were statistically significant, the highest significance obtained for ADC-NASNet-mobile features. The complete set of corrected p-values can be found in *Supplementary Table S6*.All correlation coefficients and corrected p-values are reported in *Supplementary Table S5*.

IV. DISCUSSIONS

We implemented a novel approach using deep entropy features (DEFs) derived from all layers of 9 different pre-trained CNNs to analyze mpMRI images of PCa lesions.

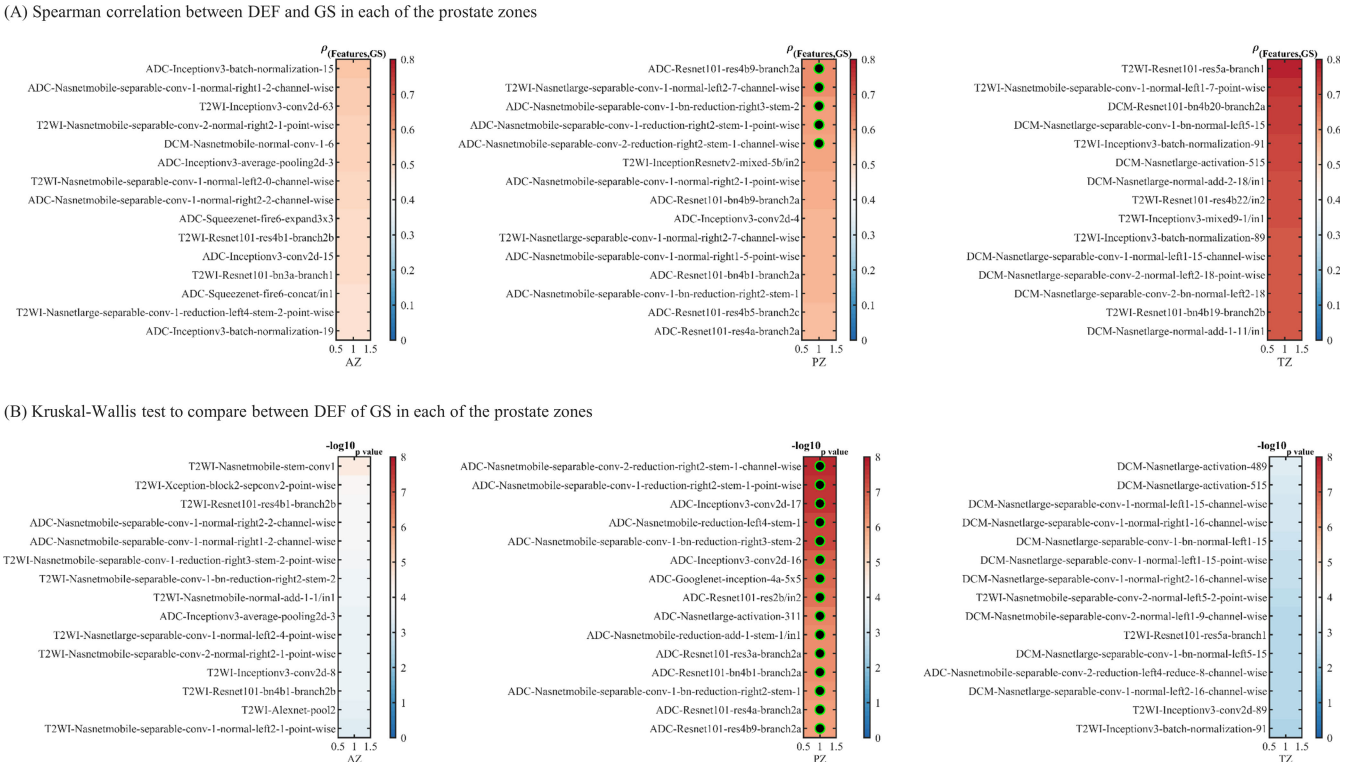


FIGURE 6. **Sensitivity analysis of deep entropy features applicability to distinct regions of prostate anatomy. (A) Spearman's correlation analysis of the 15 DEFs most correlated to GS in the AZ, PZ and TZ regions, respectively. (B) The 15 most significant DEFs for discriminating Gleason grade groups in the AZ, PZ and TZ zones, respectively. Black-green circles in A and B represent a significant p-value following Holm-Bonferroni correction.**

This contrasts with our previous study on brain MRI, where features extracted only from the most superficial layers of a single CNN were used to predict Alzheimer's or mild cognitive impairment [45]. Experiments in the current study identified 46 DEFs derived from the pre-trained CNNs that were significantly correlated to GS. Furthermore, when given as input to a RF classifier, these combined features led to a highly accurate prediction of the Gleason grade group. There currently is a clinical need for radiomic tools that can predict the aggressiveness of PCa lesions with high reliability. Even with modern advances in targeting, the existing diagnostic gold standard of TRUS biopsy is not as reliable as definitive resection [46], while also carrying risks. Reported morbidity includes pain, bleeding, lower urinary tract symptoms, erectile dysfunction, and infection which can be life threatening in some cases [6], [47]. Cost is another limiting factor of biopsy. Thus, a single biopsy requires a medical specialist to acquire a sample then a separate specialized pathologist to evaluate the resultant specimen with reasonable accuracy [48]. In many centers, TRUS biopsy will already be preceded by an MRI for guidance or to determine the necessity of additional biopsy [49]. Therefore, the implementation of an MRI-derived radiomics approach would not add significant cost, and could replace two expensive steps for the diagnosis and/or re-evaluation of prostate cancer. The proposed method based on DEFs compares favorably with previous approaches for predicting the Gleason grade group of PCa lesions. In the PROSTATEx Challenge [26], the best-performing method among 32 submissions achieved an AUC of 87% for discriminating between clinically significant and non-significant lesions. Since almost all clinically-significant lesions (i.e., 71 out of 73) have a Gleason grade group > 1, we can compare this value with the AUC of 88.8% obtained by our method for the G1 vs all task.

Our experiments also confirm results of previous works showing the efficacy of radiomics for analyzing PCa images [40], [50]–[55]. In [56], entropy-based texture features extracted from gray-level co-occurrence matrix (GLCMs) were found to be related to GS, more specifically, that a higher GS is associated with a higher ADC entropy and low ADC energy. Likewise, the average ADC image/maps are thought to be a biomarker for GS, combinations of the ADC volume and average held an AUC value of 74.9% to discriminate a biologically low risk PCa (GS=6) from higher risk malignancies (GS≥7) PCa [21]. Other methods have similarly discriminated between a low and high GS, such as combined T2-WI and spectroscopy images [57]. Strategies which compensated for unbalanced samples, via imputation, have found that texture features could reliably discriminate intermediate-risk prostate cancers (GS 6, 3+4, and 4+3) from each other [58]. When implemented as combined radiomic feature models (joint intensity matrices and GLCM), AUC values were 78.40% (GS=6), 82.35% (GS=3+4), and 64.76% (GS≥4+3) [42].

Along the same vein, a model which combined 45 radiomic features achieved AUC values of 83.40% (GS=6), 72.71% (GS=3+4), and 77.35% (GS≥4+3) [41]. Moreover, volume derived from mpMRI has shown a moderate correlation with GS [59], while the combined volume and mean ADC features achieved an AUC of 70.4% for classifying GS 6 from GS $\geq$ 7 tumors [21]. Compared to features based on texture, shape features are more sensitive to the manual segmentation of ROIs. The current study overcomes this problem by applying a fixed ROI size of 84 $\times$ 84 pixels.

Our experiments showed that certain CNN architectures provide more discriminative features for analyzing PCa lesions. In particular, the *NASNet-mobile* network yielded the features most correlated to GS, with the highest correlation of 0.47 (p=0.003) obtained for ADC image in layer *Separable-Conv-1-Normal-right2-1-point-wise*. This architecture, as well as the *NASNet-large* network, also showed prominent differences when comparing lesions grouped based on Gleason score. Notably, a total of 9 DEFs computed by these two networks from T2-WI images gave statistically-significant differences when comparing lesions with GS $<$ 8 vs GS $\geq$ 8, with corrected p $<$ 0.05. Furthermore, among the 9 pre-trained CNNs, the *NASNet-mobile* network gave the most predictive DEFs when used as input to a RF classifier (**Figure 4** and **Figure 5**). Hence, the features of this model achieved the highest AUC for identifying G1, G2, G4G5, and G1G2 lesions.

Compared to applications involving natural images, deep learning models like CNNs have had a more limited success for classifying medical images. This is largely due to the much smaller amount of training data in clinical applications, but also to the particularity of medical images which often have poor contrast and low resolution. A recent survey on deep learning for Alzheimer's prediction [60] found that most studies reporting high accuracy suffered from some form of data leakage (e.g., using images from the same subject in both training and testing).

In the current work, we alleviate the problem of overfitting when training with a small dataset via a transfer learning strategy that computes a compact set of informative features from pre-trained CNNs. The proposed DEFs are based on entropy, a well-known concept of information theory to measure uncertainty of random variables. In our radiomics model, entropy is used to assess the heterogeneity of CNN feature maps considered as image textures. Information theory has been explored for various applications in computational biology [61], for example, in a maximal information transduction estimation approach to reduce uncertainty in transcriptome analyses [62]. To our knowledge, this is the first work proposing DEFs from different pre-trained CNNs for PCa analysis.

This work has notable limitations, the foremost being the limited number of retrospectively evaluated PCa lesions (n=112) and patients (n = 99). Validation steps on a larger scale and prospective design are required before broader clinical application. Furthermore, images derived from multiple medical centers would be an essential step in demonstrating the generalizability of DEFs to predict PCa lesion aggressiveness. A larger patient cohort would also enable a better quantification of the variability between CNN features and their relationship to GS. Another key limitation is our reliance on biopsy data, which has an understandable potential for sampling error despite modern targeting approaches [46]. This limitation is also present in other radiomics works for PCa [53] and is likely due to the logistical difficulties of acquiring pre-operative mpMRI alongside an anatomically-correlated intact pathological specimen. To bridge this gap in the literature, a study with complete prostatectomy specimens would likely have a limited sample size, but the application of our pre-established CNN methodology could minimize this limitation while validating the approach in an external data set. Future work will be focused on testing GS using the DEF before therapy and the post-diagnostic prognostic test.

V. CONCLUSION

In this study, we generated and evaluated novel radiomic features based on the entropy of features maps in 9 pre-trained CNNs fed with mpMRI data of 112 PCa lesions and GS $\geq$ 9 with an AUC of 80.08, 85.77, 97.30, 98.20, and 86.51 %, respectively. Our results surpass, via an indirect assessment, the published performance of the clinically-implemented PIRADS as well as recent radiomics models in the literature. We conclude that the use of pre-trained CNNs to generate DEFs is an efficient method to empower radiomics analysis for PCa. The potential clinical yield of this work is a tool that can not only limit misclassification but could be refined to optimize non-invasive evaluations of a PCa's malignant potential. Next steps will include combining DEFs with other novel imaging features or in prospective assessments that can quantify its clinical applicability.

REFERENCES

[1] J. E. van Timmeren, "Test-retest data for radiomics feature stability analysis: Generalizable or study specific?" *Tomography*, vol. 2, no. 4, pp. 361–365, 2016.

[2] B. Abdollahi, A. El-Baz, and H. B. Frieboes, "Overview of deep learning algorithms applied to medical images," in *Big Data in Multimodal Medical Imaging*. London, U.K.: Chapman & Hall/CRC, 2019, pp. 225–237.

[3] S.-O. Professionals. *EAU Guidelines: Prostate Cancer*. Accessed: Nov. 29, 2019. [Online]. Available: https://uroweb.org/guideline/prostate-cancer/

[4] *Overview | Prostate Cancer: Diagnosis and Management | Guidance | NICE*. Accessed: Nov. 29, 2019. [Online]. Available: https://www.nice.org.uk/guidance/ng131

[5] J. I. Epstein, "Prostate cancer grading: A decade after the 2005 modified system," *Mod. Pathol.*, vol. 31, no. 1, pp. 47–63, Jan. 2018, doi: 10.1038/modpathol.2017.133.

[6] K. Braun, Y. Ahallal, D. D. Sjoberg, T. Ghoneim, M. D. Esteban, J. Mulhall, A. Vickers, J. Eastham, P. T. Scardino, and K. A. Touijer, "Effect of repeated prostate biopsies on erectile function in men on active surveillance for prostate cancer," *J. Urol.*, vol. 191, no. 3, pp. 744–749, Mar. 2014, doi: 10.1016/j.juro.2013.08.054.

[7] C. Lao, R. Edlin, P. Rouse, C. Brown, M. Holmes, P. Gilling, and R. Lawrenson, "The cost-effectiveness of active surveillance compared to watchful waiting and radical prostatectomy for low risk localised prostate cancer," *BMC Cancer*, vol. 17, no. 1, p. 529, Aug. 2017, doi: 10.1186/s12885-017-3522-z.

[8] L. Klotz, D. Vesprini, P. Sethukavalan, V. Jethava, L. Zhang, S. Jain, T. Yamamoto, A. Mamedov, and A. Loblaw, "Long-term follow-up of a large active surveillance cohort of patients with prostate cancer," *J. Clin. Oncol.*, vol. 33, no. 3, pp. 272–277, Jan. 2015, doi: 10.1200/JCO.2014.55.1192.

[9] I. Heidegger, V. Skradski, E. Steiner, H. Klocker, R. Pichler, A. Pircher, W. Horninger, and J. Bektic, "High risk of under-grading and -staging in prostate cancer patients eligible for active surveillance," *PLoS ONE*, vol. 10, no. 2, Feb. 2015, Art. no. e0115537, doi: 10.1371/journal.pone.0115537.

[10] A. Harbias, E. Salmo, and A. Crump, "Implications of observer variation in Gleason scoring of prostate cancer on clinical management: A collaborative audit," *Gulf J. Oncol.*, vol. 1, no. 25, pp. 41–45, Sep. 2017.

[11] V. Kasivisvanathan, A. S. Rannikko, M. Borghi, V. Panebianco, L. A. Mynderse, M. H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, R. G. Hindley, and M. J. Roobol, "MRI-targeted or standard biopsy for prostate-cancer diagnosis," *New England J. Med.*, vol. 378, no. 19, pp. 1767–1777, May 2018, doi: 10.1056/NEJMoa1801993.

[12] H. U. Ahmed, A. E. S. Bosaily, L. C. Brown, R. Gabe, R. Kaplan, M. K. Parmar, Y. Collaco-Moraes, K. Ward, R. G. Hindley, A. Freeman, and A. P. Kirkham, "Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): A paired validating confirmatory study," *Lancet*, vol. 389, no. 10071, pp. 815–822, Feb. 2017, doi: 10.1016/S0140-6736(16)32401-1.

[13] O. Rouvière, P. Puech, R. Renard-Penna, M. Claudon, C. Roy, F. Mège-Lechevallier, M. Decaussin-Petrucci, M. Dubreuil-Chambardel, L. Magaud, L. Remontet, and A. Ruffion, "Use of prostate systematic and targeted biopsy on the basis of multiparametric MRI in biopsy-Naive patients (MRI-FIRST): A prospective, multicentre, paired diagnostic study," *Lancet Oncol.*, vol. 20, no. 1, pp. 100–109, Jan. 2019, doi: 10.1016/S1470-2045(18)30569-2.

[14] S. Mehralivand, J. H. Shih, S. Rais-Bahrami, A. Oto, S. Bednarova, J. W. Nix, J. V. Thomas, J. B. Gordetsky, S. Gaur, S. A. Harmon, and M. M. Siddiqui, "A magnetic resonance imaging–based prediction model for prostate biopsy risk stratification," *JAMA Oncol.*, vol. 4, no. 5, pp. 678–685, May 2018, doi: 10.1001/jamaoncol.2017.5667.

[15] E. Hassanzadeh, D. I. Glazer, R. M. Dunne, F. M. Fennessy, M. G. Harisinghani, and C. M. Tempany, "Prostate imaging reporting and data system version 2 (PI-RADS v2): A pictorial review," *Abdominal Radiol.*, vol. 42, no. 1, pp. 278–289, Jan. 2017, doi: 10.1007/s00261-016-0871-z.

[16] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn, R. H. Mak, R. M. Tamimi, C. M. Tempany, C. Swanton, U. Hoffmann, L. H. Schwartz, R. J. Gillies, R. Y. Huang, and H. J. W. L. Aerts, "Artificial intelligence in cancer imaging: Clinical challenges and applications," *CA, Cancer J. Clinicians*, vol. 69, no. 2, pp. 127–157, Feb. 2019, doi: 10.3322/caac.21552.

[17] Q. Li, H. Lu, J. Choi, K. Gage, S. Feuerlein, J. M. Pow-Sang, R. Gillies, and Y. Balagurunathan, "Radiological semantics discriminate clinically significant grade prostate cancer," *Cancer Imag.*, vol. 19, no. 1, p. 81, Dec. 2019, doi: 10.1186/s40644-019-0272-y.

[18] Y.-F. Lu, Q. Zhang, W.-G. Yao, H.-Y. Chen, J.-Y. Chen, C.-C. Xu, and R.-S. Yu, "Optimizing prostate cancer accumulating model: Combined PI-RADS v2 with prostate specific antigen and its derivative data," *Cancer Imag.*, vol. 19, no. 1, p. 26, May 2019, doi: 10.1186/s40644-019-0208-6.

[19] C. De Vente, P. Vos, M. Hosseinzadeh, J. Pluim, and M. Veta, "Deep learning regression for prostate cancer detection and grading in bi-parametric MRI," *IEEE Trans. Biomed. Eng.*, early access, May 8, 2020, doi: 10.1109/TBME.2020.2993528.

[20] Y. Shao, J. Wang, B. Wodlinger, and S. E. Salcudean, "Improving prostate cancer (PCa) classification performance by using three-player minimax game to reduce data source heterogeneity," *IEEE Trans. Med. Imag.*, early access, Apr. 15, 2020, doi: 10.1109/TMI.2020.2988198.

[21] O. F. Donati, A. Afaq, H. A. Vargas, Y. Mazaheri, J. Zheng, C. S. Moskowitz, H. Hricak, and O. Akin, "Prostate MRI: Evaluating tumor volume and apparent diffusion coefficient as surrogate biomarkers for predicting tumor Gleason score," *Clin. Cancer Res.*, vol. 20, no. 14, pp. 3705–3711, Jul. 2014, doi: 10.1158/1078-0432.CCR-14-0044.

[22] Q. Wang, H. Li, X. Yan, C.-J. Wu, X.-S. Liu, H.-B. Shi, and Y.-D. Zhang, "Histogram analysis of diffusion kurtosis magnetic resonance imaging in differentiation of pathologic Gleason grade of prostate cancer," *Urol. Oncol., Seminars Original Investigations*, vol. 33, no. 8, pp. 337.e15–337.e24, Aug. 2015, doi: 10.1016/j.urolonc.2015.05.005.

[23] S. O. S. Osman, R. T. H. Leijenaar, A. J. Cole, C. A. Lyons, A. R. Hounsell, K. M. Prise, J. M. O'Sullivan, P. Lambin, C. K. McGarry, and S. Jain, "Computed tomography-based radiomics for risk stratification in prostate cancer," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 105, no. 2, pp. 448–456, Oct. 2019, doi: 10.1016/j.ijrobp.2019.06.2504.

[24] A. Chaddad, M. Toews, C. Desrosiers, and T. Niazi, "Deep radiomic analysis based on modeling information flow in convolutional neural networks," *IEEE Access*, vol. 7, pp. 97242–97252, 2019, doi: 10.1109/ACCESS.2019.2930238.

[25] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, "Computer-aided detection of prostate cancer in MRI," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1083–1092, May 2014, doi: 10.1109/TMI.2014.2303821.

[26] S. G. Armato, H. Huisman, K. Drukker, L. Hadjiiski, J. S. Kirby, N. Petrick, G. Redmond, M. L. Giger, K. Cha, A. Mamonov, J. Kalpathy-Cramer, and K. Farahani, "PROSTATEx challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images," *Proc. SPIE*, vol. 5, no. 4, Oct. 2018, Art. no. 044501, doi: 10.1117/1.JMI.5.4.044501.

[27] J. I. Epstein, L. Egevad, M. Amin, B. Delahunt, J. Srigley, J. R. Humphrey, and A. Peter, "The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system," *Amer. J. Surg. Pathol.*, vol. 40, no. 2, pp. 244–252, Feb. 2016, doi: 10.1097/PAS.0000000000000530.

[28] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[31] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: http://arxiv.org/abs/1602.07360

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[33] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.

[34] J. H. Zar, "Significance testing of the Spearman rank correlation coefficient," *J. Amer. Stat. Assoc.*, vol. 67, no. 339, pp. 578–580, 1972, doi: 10.2307/2284441.

[35] J. W. Pratt, "Remarks on zeros and ties in the Wilcoxon signed rank procedures," *J. Amer. Stat. Assoc.*, vol. 54, no. 287, pp. 655–667, 1959, doi: 10.2307/2282543.

[36] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandin. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.

[37] D. Dittman, T. M. Khoshgoftaar, R. Wald, and A. Napolitano, "Random forest: A reliable tool for patient response prediction," in *Proc. IEEE Int. Conf. Bioinf. Biomed. Workshops (BIBMW)*, Nov. 2011, pp. 289–296.

[38] R. Pal, *Predictive Modeling of Drug Sensitivity*. New York, NY, USA: Academic, 2016.

[39] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Comput. Statist. Data Anal.*, vol. 52, no. 4, pp. 2249–2260, Jan. 2008, doi: 10.1016/j.csda.2007.08.015.

[40] C. Jensen, J. Carl, L. Boesen, N. C. Langkilde, and L. R. Østergaard, "Assessment of prostate cancer prognostic Gleason grade group using zonal-specific features extracted from biparametric MRI using a KNN classifier," *J. Appl. Clin. Med. Phys.*, vol. 20, no. 2, pp. 146–153, Feb. 2019, doi: 10.1002/acm2.12542.

[41] A. Chaddad, T. Niazi, S. Probst, F. Bladou, M. Anidjar, and B. Bahoric, "Predicting Gleason score of prostate cancer patients using radiomic analysis," *Frontiers Oncol.*, vol. 8, p. 630, Dec. 2018.

[42] A. Chaddad, M. Kucharczyk, and T. Niazi, "Multimodal radiomic features for the predicting Gleason score of prostate cancer," *Cancers*, vol. 10, no. 8, p. 249, Jul. 2018, doi: 10.3390/cancers10080249.

[43] J. Toivonen, I. Montoya Perez, P. Movahedi, H. Merisaari, M. Pesola, P. Taimen, P. J. Boström, J. Pohjankukka, A. Kiviniemi, T. Pahikkala, H. J. Aronen, and I. Jambor, "Radiomics and machine learning of multisequence multiparametric prostate MRI: Towards improved non-invasive prostate cancer characterization," *PLoS ONE*, vol. 14, no. 7, Jul. 2019, Art. no. e0217702, doi: 10.1371/journal.pone.0217702.

[44] R. Cao, A. Mohammadian Bajgiran, S. Afshari Mirak, S. Shakeri, X. Zhong, D. Enzmann, S. Raman, and K. Sung, "Joint prostate cancer detection and Gleason score prediction in mp-MRI via FocalNet," *IEEE Trans. Med. Imag.*, vol. 38, no. 11, pp. 2496–2506, Nov. 2019, doi: 10.1109/TMI.2019.2901928.

[45] A. Chaddad, C. Desrosiers, and T. Niazi, "Deep radiomic analysis of MRI related to Alzheimer's disease," *IEEE Access*, vol. 6, pp. 58213–58221, 2018, doi: 10.1109/ACCESS.2018.2871977.

[46] S. Goel, J. E. Shoag, M. D. Gross, B. A. H. A. Awamlh, B. Robinson, F. Khani, B. B. Nelson, D. J. Margolis, and J. C. Hu, "Concordance between biopsy and radical prostatectomy pathology in the era of targeted biopsy: A systematic review and meta-analysis," *Eur. Urology Oncol.*, vol. 3, no. 1, pp. 10–20, Feb. 2020.

[47] B. Ehdaie, E. Vertosick, M. Spaliviero, A. Giallo-Uvino, Y. Taur, M. O'Sullivan, J. Livingston, P. Sogani, J. Eastham, P. Scardino, and K. Touijer, "The impact of repeat biopsies on infectious complications in men with prostate cancer on active surveillance," *J. Urol.*, vol. 191, no. 3, pp. 660–664, Mar. 2014.

[48] L. J. Raff, G. Engel, K. R. Beck, A. S. O'Brien, and M. E. Bauer, "The effectiveness of inking needle core prostate biopsies for preventing patient specimen identification errors: A technique to address joint commission patient safety goals in specialty laboratories," *Arch. Pathol. Lab. Med.*, vol. 133, no. 2, pp. 295–297, Feb. 2009. [Online]. Available: https://meridian.allenpress.com/aplm/article/133/2/295/64084/The-Effectiveness-of-Inking-Needle-Core-Prostate

[49] J. L. Mohler, E. S. Antonarakis, A. J. Armstrong, A. V. D'Amico, B. J. Davis, T. Dorff, J. A. Eastham, C. A. Enke, T. A. Farrington, C. S. Higano, and E. M. Horwitz, "Prostate cancer, version 2.2019, NCCN clinical practice guidelines in oncology," *J. Nat. Comprehensive Cancer Netw.*, vol. 17, no. 5, pp. 479–505, 2019.

[50] T. Chen, M. Li, Y. Gu, Y. Zhang, S. Yang, C. Wei, J. Wu, X. Li, W. Zhao, and J. Shen, "Prostate cancer differentiation and aggressiveness: Assessment with a radiomic-based model vs. PI-RADS v2," *J. Magn. Reson. Imag.*, vol. 49, no. 3, pp. 875–884, Mar. 2019, doi: 10.1002/jmri.26243.

[51] X. Min, M. Li, D. Dong, Z. Feng, P. Zhang, Z. Ke, H. You, F. Han, H. Ma, J. Tian, and L. Wang, "Multi-parametric MRI-based radiomics signature for discriminating between clinically significant and insignificant prostate cancer: Cross-validation of a machine learning method," *Eur. J. Radiol.*, vol. 115, pp. 16–21, Jun. 2019, doi: 10.1016/j.ejrad.2019.03.010.

[52] W. Li, J. Li, K. V. Sarma, K. C. Ho, S. Shen, B. S. Knudsen, A. Gertych, and C. W. Arnold, "Path R-CNN for prostate cancer diagnosis and Gleason grading of histological images," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 945–954, Apr. 2019, doi: 10.1109/TMI.2018.2875868.

[53] J. Chen, D. Remulla, J. H. Nguyen, D. Aastha, Y. Liu, P. Dasgupta, and A. J. Hung, "Current status of artificial intelligence applications in urology and their potential to influence clinical practice," *BJU Int.*, vol. 124, no. 4, pp. 567–577, Oct. 2019.

[54] M. Moradi, S. E. Salcudean, S. D. Chang, E. C. Jones, N. Buchan, R. G. Casey, S. L. Goldenberg, and P. Kozlowski, "Multiparametric MRI maps for detection and grading of dominant prostate tumors," *J. Magn. Reson. Imag.*, vol. 35, no. 6, pp. 1403–1413, Jun. 2012, doi: 10.1002/jmri.23540.

[55] G. Almeida and J. M. R. S. Tavares, "Deep learning in radiation oncology treatment planning for prostate cancer: A systematic review," *J. Med. Syst.*, vol. 44, no. 10, pp. 1–15, Oct. 2020, doi: 10.1007/s10916-020-01641-3.

[56] A. Wibmer, H. Hricak, T. Gondo, K. Matsumoto, H. Veeraraghavan, D. Fehr, J. Zheng, D. Goldman, C. Moskowitz, S. W. Fine, V. E. Reuter, J. Eastham, E. Sala, and H. A. Vargas, "Haralick texture analysis of prostate MRI: Utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores," *Eur. Radiol.*, vol. 25, no. 10, pp. 2840–2850, Oct. 2015, doi: 10.1007/s00330-015-3701-8.

[57] P. Tiwari, J. Kurhanewicz, and A. Madabhushi, "Multi-kernel graph embedding for detection, Gleason grading of prostate cancer via MRI/MRS," *Med. Image Anal.*, vol. 17, no. 2, pp. 219–235, Feb. 2013, doi: 10.1016/j.media.2012.10.004.

[58] D. Fehr, H. Veeraraghavan, A. Wibmer, T. Gondo, K. Matsumoto, H. A. Vargas, E. Sala, H. Hricak, and J. O. Deasy, "Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 46, pp. E6265–E6273, Nov. 2015, doi: 10.1073/pnas.1505935112.

[59] A. Chatterjee, R. M. Bourne, S. Wang, A. Devaraj, A. J. Gallan, T. Antic, G. S. Karczmar, and A. Oto, "Diagnosis of prostate cancer with noninvasive estimation of prostate tissue composition by using hybrid multidimensional MR imaging: A feasibility study," *Radiology*, vol. 287, Feb. 2018, Art. no. 171130, doi: 10.1148/radiol.2018171130.

[60] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-Gonzalez, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, and O. Colliot, "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," 2019, *arXiv:1904.07773*. [Online]. Available: http://arxiv.org/abs/1904.07773

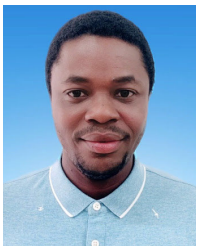
[61] Y. Deng, F. Bao, X. Deng, R. Wang, Y. Kong, and Q. Dai, "Deep and structured robust information theoretic learning for image analysis," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4209–4221, Sep. 2016, doi: 10.1109/TIP.2016.2588330.

[62] Y. Deng, F. Bao, Y. Yang, X. Ji, M. Du, Z. Zhang, M. Wang, and Q. Dai, "Information transduction capacity reduces the uncertainties in annotation-free isoform discovery and quantification," *Nucleic Acids Res.*, vol. 45, no. 15, p. e143, Sep. 2017, doi: 10.1093/nar/gkx585.

AHMAD CHADDAD received the Ph.D. degree in engineering systems from the University of Lorraine, Metz, France, in 2012. He has worked for seven years with McGill University, the Ecole de Technologie Superieure, The University of Texas MD Anderson Cancer Center, and Villanova University. In 2020, he joined the School of Artificial Intelligence, Guilin University of Electronic University, as a Professor. He is currently the Project Director of the Lady Davis Institute for Medical Research, McGill University. He has authored over 60 research articles. His current research interests include AI and radiomics analysis in order to improve personalized medicine strategies, by allowing clinicians to monitor disease in real time as patients move through treatment. He is a member of several international technical and organizational committees.

MICHAEL J. KUCHARCZYK is currently a Radiation Oncologist with Halifax Nova Scotia, affiliated with Dalhousie University as an Assistant Professor. With an academic mandate, he seeks to expand the capacity of the department to pursue grassroots clinical research and pursue international collaboration. He maintains an active clinical practice in genitourinary, gastrointestinal, and breast malignancies. His research interests include optimizing care in advanced prostate cancer, radiomics, oligometastatic disease, and clinical trials.

CHRISTIAN DESROSIERS received the Ph.D. degree in computer engineering from Polytechnique Montreal, in 2008. He was a Postdoctoral Researcher with the University of Minnesota on the topic of machine learning. In 2009, he joined the Department of Software and IT Engineering, ÉTS, University of Quebec, as a Professor. He is also the Co-Director of the Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA). His main research interests include machine learning, image processing, computer vision, and medical imaging. He is a member of the REPARTI research network.



IDOWU PAUL OKUWOBI received the Ph.D. degree in computer science and technology, with a strong focus on pattern recognition and artificial intelligence. From 2012 to 2015, he researches at the College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, where he works with top experts to solve the current problems faced in the mechanical field. From 2015 to 2019, he researches at the School of Computer Science and Engineering, Nanjing University of Science and Technology, where he works with world-class computer experts and clinicians to solve several problems faced by ophthalmologists in their daily clinical routine. He is currently with SAI, GUET, and oversees the VIP Laboratory. His current research interest includes develop new intelligent algorithms for medical image processing.

YOUSEF KATIB has worked on the prostate cancer with McGill University, Montreal, Canada. He is currently a Radiation-Oncologist in Saudi Arabia. His clinical activity is mainly oriented in the management of prostate tumors. He coordinates several prospective studies in prostate cancers.

MINGLI ZHANG (Member, IEEE) received the Ph.D. degree in image processing from the Ecole de Technologie Superieure, University of Quebec, Montreal, in 2017. She is currently the Scientist of the Shandong Co-Innovation Center of Future Intelligent Computing, Yantai. She is also a Postdoctoral Research Fellow with the McGill Centre for Integrative Neuroscience/Ludmer Centre for NeuroInformatics and Mental Health, Montreal Neurological Institute, McGill University, working on many biomedical imaging projects. Her research interests include designing and application of high-performance machine learning models to solve problems in the fields of computer vision, biomedical imaging, and natural images.

SAIMA RATHORE received the Ph.D. degree in computer science from the Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan, in 2015. She is currently a Research Fellow with the Center for Biomedical Image Computing and Analytics (CBICA), Radiology Department, University of Pennsylvania. She has industry software design and development experience of 11 years. At CBICA, she is the Lead Scientific Developer of Cancer Imaging Phenomics Toolkit. Her research interests include medical image analysis, segmentation, classification and evolutionary algorithms. She has published her work in leading scientific journals and presented it at various conferences and universities around the world.

PAUL SARGOS is currently a Radiation-Oncologist working with the Institut Bergonié, Comprehensive Cancer Care Center, Bordeaux, France. His clinical activity is mainly oriented in the management of sarcomas and genito-urinary tumors, where he develops technical innovation. He coordinates several prospective studies in soft tissue sarcomas, prostate, bladder, and testis cancers. He is the author or a coauthor of more than 50 referenced publications.

TAMIM NIAZI received the Doctor of Medicine and Master of Surgery (M.D.C.M.) degrees from McGill University, in 2001. He continued his internship and residency with the McGill University Health Centre. From 2006 to 2007, he was the Clinical Trials Fellow with the National Cancer Institute of Canada Clinical Trials Group (NCIC-CTG). He is currently an Adjoint Professor with the Department of Oncology, McGill University. He is also the Fellowship Co-Director of the Division of Radiation Oncology, Department of Oncology, McGill University. He has presented and published over 50 peer reviewed academic works.