Journal of
Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI

Prathyush Chirra
Patrick Leo
Michael Yim
B. Nicolas Bloch
Ardeshir R. Rastinehad
Andrei Purysko
Mark Rosen
Anant Madabhushi
Satish E. Viswanath

Prathyush Chirra, Patrick Leo, Michael Yim, B. Nicolas Bloch, Ardeshir R. Rastinehad, Andrei Purysko, Mark Rosen, Anant Madabhushi, Satish E. Viswanath, "Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI," *J. Med. Imag.* **6**(2), 024502 (2019), doi: 10.1117/1.JMI.6.2.024502.

Multisite evaluation of radiomic feature reproducibility and discriminability for identifying peripheral zone prostate tumors on MRI

Prathyush Chirra,[a,*] Patrick Leo,[a] Michael Yim,[b] B. Nicolas Bloch,[c] Ardeshir R. Rastinehad,[d] Andrei Purysko,[e] Mark Rosen,[f] Anant Madabhushi,[a,g] and Satish E. Viswanath[a]
[a]Case Western Reserve University, Department of Biomedical Engineering, Cleveland, Ohio, United States
[b]Northeast Ohio Medical University, College of Medicine, Rootstown, Ohio, United States
[c]Boston University School of Medicine, Department of Radiology, Boston, Massachusetts, United States
[d]Icahn School of Medicine at Mount Sinai, Department of Urology, New York, New York, United States
[e]Cleveland Clinic, Department of Radiology, Cleveland, Ohio, United States
[f]Hospital of the University of Pennsylvania, Department of Radiology, Philadelphia, Pennsylvania, United States
[g]Louis Stokes Cleveland Veterans Administration Medical Center, Cleveland, Ohio, United States

Abstract. Recent advances in the field of radiomics have enabled the development of a number of prognostic and predictive imaging-based tools for a variety of diseases. However, wider clinical adoption of these tools is contingent on their generalizability across multiple sites and scanners. This may be particularly relevant in the context of radiomic features derived from T1- or T2-weighted magnetic resonance images (MRIs), where signal intensity values are known to lack tissue-specific meaning and vary based on differing acquisition protocols between institutions. We present the first empirical study of benchmarking five different radiomic feature families in terms of both reproducibility and discriminability in a multisite setting, specifically, for identifying prostate tumors in the peripheral zone on MRI. Our cohort comprised 147 patient T2-weighted MRI datasets from four different sites, all of which are first preprocessed to correct for acquisition-related artifacts such as bias field, differing voxel resolutions, and intensity drift (nonstandardness). About 406 three-dimensional voxel-wise radiomic features from five different families (gray, Haralick, gradient, Laws, and Gabor) were evaluated in a cross-site setting to determine (a) how reproducible they are within a relatively homogeneous nontumor tissue region and (b) how well they could discriminate tumor regions from nontumor regions. Our results demonstrate that a majority of the popular Haralick features are reproducible in over 99% of all cross-site comparisons, as well as achieve excellent cross-site discriminability (classification accuracy of ≈0.8). By contrast, a majority of Laws features are highly variable across sites (reproducible in <75% of all cross-site comparisons) as well as resulting in low cross-site classifier accuracies (<0.6), likely due to a large number of noisy filter responses that can be extracted. These trends suggest that only a subset of radiomic features and associated parameters may be both reproducible and discriminable enough for use within machine learning classifier schemes. [DOI: 10.1117/1.JMI.6.2.024502]

Keywords: radiomics; discriminability; reproducibility; multisite; magnetic resonance imaging; prostate; feature analysis; stability.

Paper 18275R received Dec. 28, 2018; accepted for publication May 15, 2019; published online Jun. 14, 2019.

1 Introduction

The extraction of quantitative descriptors of image intensity, appearance, shape, gradient, structure, and texture (termed radiomics) from medical imaging[1,2] has enabled the development of machine learning tools for disease detection,[2] characterization,[3] outcome prediction,[4] and prognosis.[5] While there is an increasing interest in using radiomic tools in a clinical setting, this is contingent on benchmarking these features in terms of (a) reproducibility (i.e., numeric consistency and variability of radiomic feature values associated with a specific tissue region) and (b) discriminability (how well radiomic features can distinguish pathologically different tissue regions) across a variety of clinically acquired images. This may be especially relevant when utilizing imaging data [magnetic resonance image (MRI), computed tomography (CT), and positron-emission tomography (PET)] from multiple different institutions, each of which may utilize different scanners as well as potentially different sequences and acquisition parameters.

Radiomic features (quantified as responses to texture and wavelet operators, such as gray, Haralick,[6] gradient,[7] Laws,[8] and Gabor[9]) have primarily been evaluated in terms of how well they differentiate between tumor and nontumor regions on imaging.[10–12] The reproducibility of these feature families has primarily been evaluated through well-controlled single-site studies in the context of CT imaging, where a majority of radiomic features did demonstrate consistency across scanners[13] and acquisition settings.[14] However, these findings have not generalized to similar single-site studies of radiomic features derived from MRI.[12,15,16] This may be because most previous work has primarily focused on feature reproducibility within a defined tumor region. However, radiomic features are known to be sensitive to subtle pathological differences in tumor phenotype (grade and aggressiveness[17]). Therefore, when assessing radiomic feature performance, one may need to consider both normal (or nontumor) regions and diseased regions on MRI, in order to benchmark feature performance in a more controlled fashion across different sites and scanners.

*Address all correspondence to Prathyush Chirra, E-mail: pvc5@case.edu

2329-4302/2019/$25.00 © 2019 SPIE

The performance of radiomic features on MRI is also dependent on common sources of acquisition variance between sites, including voxel resolutions, image reconstruction methods, magnetic field strengths, scanner hardware, and sequence parameters (echo times, repetition times, and slice thicknesses), as well as acquisition artifacts, such as bias field,[18] noise,[19] and intensity drift.[20] When benchmarking radiomic features on MRI, it may thus be critical to first account for these sources of variance between different sites and scanners. To our knowledge, benchmarking of commonly used radiomic features on MRI to determine their reproducibility or discriminability has not been widely attempted, especially in a multisite setting. It is, thus, also unknown which specific classes and parameters of radiomic features offer optimal performance in a multisite setting.

In this paper, we present the first detailed study of both reproducibility and discriminability of five different radiomic feature families in a multisite setting, using clinical prostate T2-weighted (T2w) MRIs. After correcting for known MR acquisition-related artifacts, we evaluated the reproducibility of over 400 radiomic features within a pathologically defined "nontumor" region in the peripheral zone (PZ) of the prostate. Two different measures of feature reproducibility have been utilized: (1) multivariate coefficient of variation[21] (CV) to evaluate the overall dispersion (or relative variability) of radiomic feature distributions across sites, and (2) instability score[22] (IS) to quantify the overlap in radiomic feature distributions between sites. Radiomic feature discriminability was evaluated via the receiver-operator characteristic (ROC) analysis of a quadratic discriminate analysis (QDA) classifier in distinguishing between nontumor and tumor regions within the prostatic PZ, in a multisite experiment. The specific questions we posed are as follows: (a) What are the cross-site reproducibility and discriminability characteristics of different radiomic features and feature families? (b) Which radiomic feature parameters result in the best trade-off between reproducibility and discriminability? The answers to these questions could help us understand the basis for why certain radiomic features or feature families generalize or discriminate better than others in a multisite setting. Further, evaluating the performance of individual radiomic operators may also help provide an intuition for benchmarking differences between radiomic feature families.

The rest of the paper is organized as follows: we next describe the overall experimental design of this work (illustrated in Fig. 1), including the data, preprocessing, and feature extraction, as well as the experimental methodology for evaluating cross-site radiomic feature reproducibility and discriminability. Finally, our experimental results for multisite benchmarking of different radiomic feature families are presented and discussed, followed by the concluding remarks.

2 Experimental Design

2.1 *Data Description*

This Institutional Review Board-approved retrospective study utilized 147 T2w prostate MRI datasets from four different institutions. This study was limited to using T2w acquisitions as these are routinely used in prostate radiomic analysis[23]

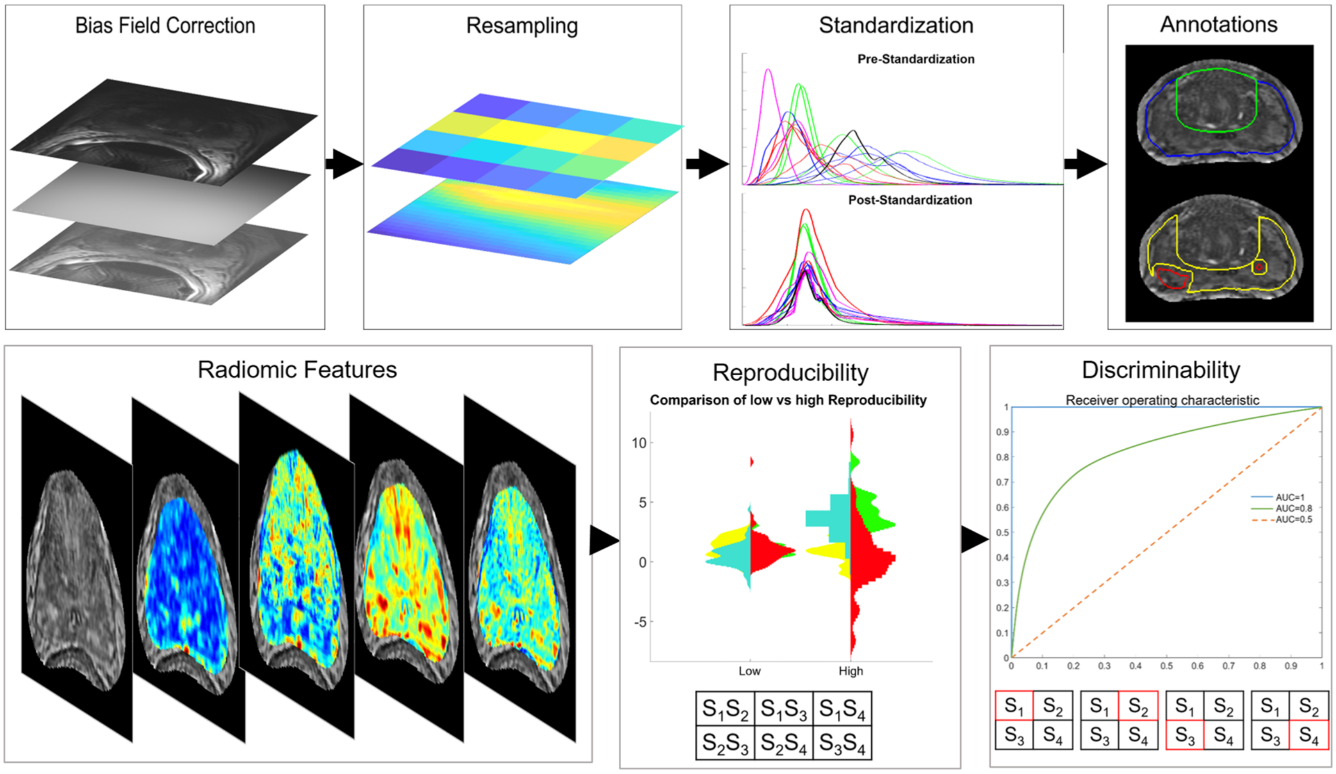

Fig. 1 Overview of the experimental workflow. The upper row shows the preprocessing steps applied to each MRI dataset. The lower row shows experimental benchmarking of radiomic features in terms of multisite reproducibility and discriminability. The table under reproducibility shows all the cross-site pairings used to calculate the two different reproducibility measures. The tables under discriminability show four training and testing set used to calculate the cross-site discriminability measure, with the red box corresponding to the held-out testing site in each run.

Table 1 Summary of multisite prostate 3T T2w MRI data, as originally acquired from each institution.

Site	$[x, y, z]$ voxel dimensions (mm)	Manufacturer	TR/TE(ms)	Number of datasets
S_1	[0.27, 0.27, 2.20]	GE Medical	4216-8266/ 155-165	15
S_2	[0.41, 0.41, 3.00]	Siemens	2840-7500/ 107-135	11
S_3	[0.27, 0.27, 3.00]	Philips	4754/115	56
S_4	[0.36, 0.36, 2.97]	Siemens	4000/120-122	65

(due to offering excellent structural detail and contrast), as well as being most commonly available from all four sites. Each MRI dataset had been acquired preoperatively using an endorectal coil on a three Tesla MRI scanner. The included patients from each institution had confirmed prostate cancer and an evaluation of their prostate-specific antigen levels and biopsy reports resulted in them undergoing a radical prostatectomy. In the absence of additional clinical markers, this inclusion criterion was to ensure that the patient populations at each site had undergone clinically comparable management and thus may be expected to have similar tumor phenotypes.

T2w MRI data from each patient was acquired as a series of DICOM images, which were directly saved from the scanner (acquisition parameters summarized in Table 1). Datasets from each site were then annotated for tumor extent in the PZ (based on available pathology sections and reports), the outer boundaries of the PZ and central gland (CG), as well as the prostate capsule. Each site was annotated by a different radiologist and no inter-reader analysis was available.

2.2 *Correction of T2w MRI to Account for Intensity Artifacts and Resolution Differences*

All MRI datasets were first processed to minimize three known sources of noise and variance in MRIs: bias field,[18] differing voxel resolutions (see Table 1), and intensity nonstandardness.[20,24] These artifacts were accounted for by sequentially applying bias field correction, resampling, and intensity standardization. Correcting bias field prior to intensity standardization was based on findings in previous work.[25] Resampling was performed after bias field correction (and prior to standardization) to ensure that intensity variations were not propagated through the volume. Figure 2 depicts a representative two-dimensional (2-D) patient image from each site, both prior to and after all three correction procedures had been applied. Expert annotations for tumor (red), CG (green), and capsule (blue) are also visualized in Figs. 2(b), 2(e), 2(h), and 2(k). With the exception of annotations all prepossessing steps as well as feature extraction and analysis was performed using MATLAB2016b.

1. Bias field correction: This was performed to compensate for inhomogeneity artifacts across the T2w MRI volume due to the use of an endorectal coil during acquisition. This manifested as a nonuniform intensity appearance across the MRI [which can be seen in Figs. 2(a), 2(g), and 2(j)]. For site S_1, low-pass bias filtering[26] was applied, whereas for sites S_3 and S_4, the N4ITK method[27] was utilized to correct this artifact. The site S_2 was found to have been bias field corrected on the scanner, and thus no additional bias correction was applied.
2. Resampling: Datasets were isotropically resampled in all three dimensions via linear interpolation to ensure consistent voxel sizes and resolutions across all the sites and patients. In addition, this enabled the use of “true” 3-D radiomic feature extraction in subsequent steps. The resulting voxel dimensions of each of the 147 T2w MRI datasets were $0.27 \times 0.27 \times 0.27$ mm. All expert annotations were similarly resampled to ensure that they remained in correspondence with T2w MRI volumes.
3. Intensity standardization: T2w MR signal intensities have been shown to lack tissue-specific meaning between patients, sites, and acquisition protocols. Landmark-based histogram transformation[20] was used to align T2w signal intensity distributions across all patient datasets. Five patients from S_1 were selected at random to generate a template distribution. Distributions for all patient volumes from all four sites were then nonlinearly mapped to the template distribution, using deciles as landmarks on both target and template distributions. As a result, distributions for all patient datasets were brought into alignment, thus ensuring that the signal intensities were in tissue-specific correspondence.

2.3 *Tumor and Nontumor Region of Interest Selection*

Resampled annotations were utilized to select the nontumor region of interest (ROI) for every patient volume as follows: first, all tumor annotations per 2-D section in each dataset were selected and dilated by 1.89 mm (≈7 pixels). This region was then removed from the annotated PZ region and the remaining largest contiguous region was then extracted to be used as the nontumor ROI. Similarly, the largest contiguous annotated tumor region within the PZ on each 2-D section was used as the tumor ROI. These are visualized as red (tumor) and yellow (nontumor) outlines in Figs. 2(c), 2(f), 2(i), and 2(l).

2.4 *Radiomic Feature Extraction*

Previous studies have widely demonstrated that prostate appearance within the PZ can be modeled using image texture features.[11,28] A total of 406 radiomic features from across five different families were extracted on a per-voxel basis from each T2w MRI dataset (see Table 2). Features were extracted in 3-D from the entire T2w MRI volume, following which the mean value of each feature was calculated over all the voxels within each of the tumor and nontumor ROIs. We denote the resulting average radiomic feature value for a given ROI c as $f_i(c), i \in \{1, \dots, 406\}$. For each site $S_k, k \in \{1, \dots, 4\}$, and for all ROIs (both tumor and nontumor) $c \in S_k, k \in \{1, \dots, 4\}$, the radiomic feature vectors are denoted as $\bar{\mathbf{F}}_i(k) = [f_i(c) | \forall c \in S^k]$.

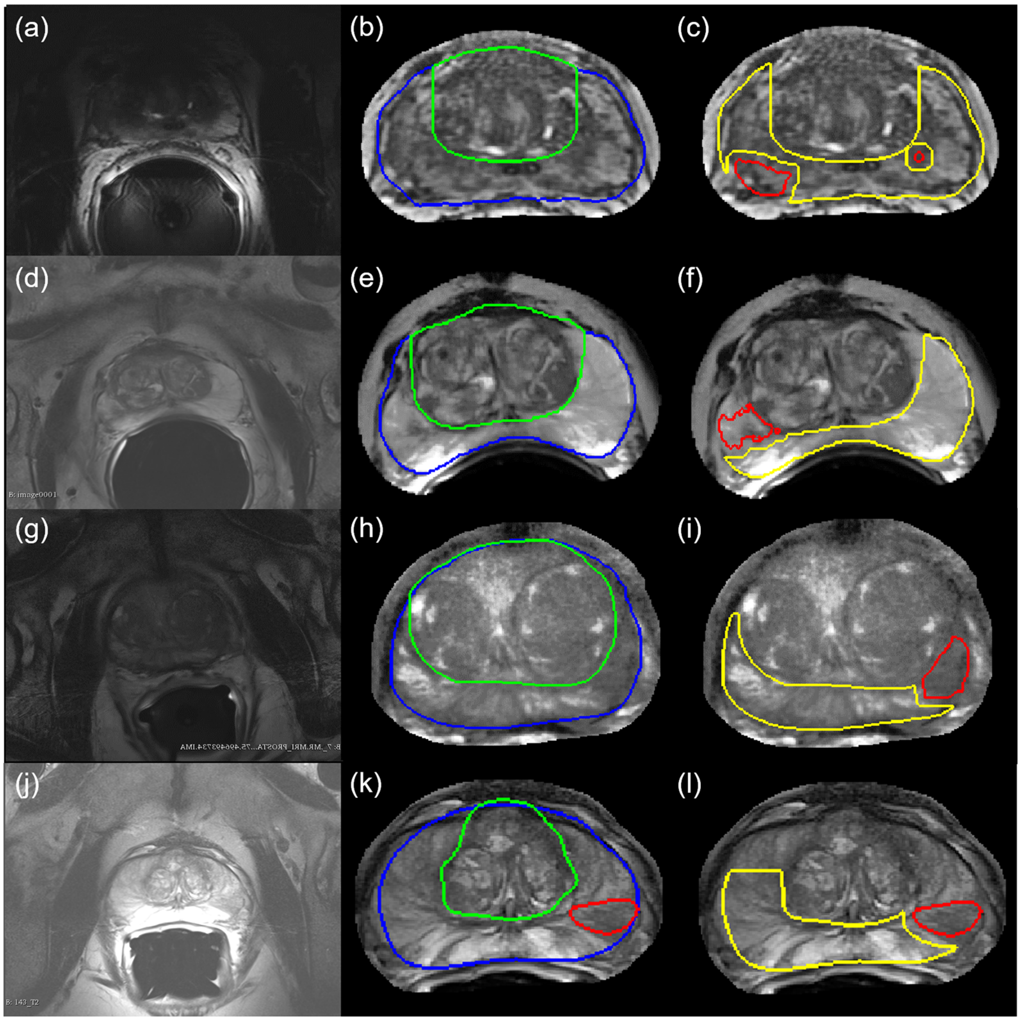

Fig. 2 Results of annotating and correcting representative T2w MRI datasets, each row corresponding to a different site. (a), (d), (g), and (j) Original field-of-view prostate T2w MRIs. (b), (e), (h), and (k) Same 2-D MRIs after bias field correction, resampling, and intensity standardization, but cropped to around the prostate capsule alone. Note the relatively uniform appearance of the image. Also shown are the expert annotations for the prostate capsule (in blue) and the central gland (in green). (c), (f), (i), and (l) Expert-annotated tumor region in red and the "nontumor" region in yellow are shown for each of these images.

Table 2 Summary of 3-D radiomic features and associated parameters extracted from each T2w MRI dataset.

Feature type	Parameters	Window sizes (WS)	Total number
Gray	Mean, median, variance, range	3, 5, 7, 9, 11	20
Haralick	Co-occurrence features for entropy, homogeneity, contrast, etc.	3, 5, 7, 9, 11	60
Gradient	Sobel, Kirsch, gradient operators in X, Y, Z directions	—	13
Laws	Edge (E), ripple (R), spot (S), level (L), wave (W) operators	3, 5	152
Gabor	XY and XZ orientations $(\theta_{XY}, \theta_{XZ})$	3, 5, 7, 9, 11	160

2.5 *Feature Normalization*

Feature normalization was applied to ensure that radiomic features extracted from different sites lie within a comparable range of values. For each site S_k, each feature vector $\bar{\mathbf{F}}_i(k)$ is normalized as

$$\mathbf{F}_i(k) = \frac{\bar{\mathbf{F}}_i(k) - \mu_i(k)}{\sigma_i(k)}, \tag{1}$$

where $\mu_i(k)$ is the mean and $\sigma_i(k)$ is the mean absolute deviation (MAD) of $\bar{\mathbf{F}}_i(k)$ (over all the samples $c \in S_k, k \in \{1, \dots, 4\}$). This process was repeated for each site $S_{1,\dots,4}$ individually so that all features within a site have a mean of 0 and a MAD of 1. Then, for each site $S_k, k \in \{1, \dots, 4\}$, we denote the normalized radiomic feature vectors for tumor ROIs $c^t \in S_k$ as $\mathbf{F}_i^t(k)$, and for nontumor ROIs $c^b \in S_k$ as $\mathbf{F}_i^b(k), i \in \{1, \dots, 406\}$.

2.6 *Generation of Bootstrapped Subsets*

To ensure more robust estimation of the evaluation measures used in this study, bootstrapping was utilized as follows. For each site S_k, $k \in \{1, \dots, 4\}$, $N = 100$ bootstrapped subsets were generated. Each bootstrapped subset, $s_{k,n}, n \in \{1, \dots, N\}$, comprised 75% of the samples from each site S_k. The associated radiomic feature vectors for each subset $s_{k,n}$ are denoted as $F_i^b(k,n)$ (for nontumor ROIs) and $F_i^t(k,n)$ (for tumor ROIs), $i \in \{1, \dots, 406\}$.

2.7 *Quantifying Cross-Site Reproducibility*

The two different reproducibility measures utilized were multivariate coefficient of variation (denoted CV) and instability (denoted IS). These measures were chosen as quantifying different aspects of feature reproducibility. The CV considered features with a large standard deviation across sites as being poorly reproducible, whereas IS considered features with dissimilar distributions across sites as exhibiting poor reproducibility. Both measures were only evaluated within nontumor regions to remove any potential confounding factors introduced by tumor region heterogeneity.

Multivariate CV: The ratio of the standard deviation to the population mean is known as the CV.[29] A multivariate extension[21] of this measure (denoted CV) is utilized to assess the variability of radiomic feature between sites, where a lower value indicates a better cross-site reproducibility. For instance, given sites S_1 and S_2 with radiomic feature vectors $\mathbf{F}_i^b(1)$ and $\mathbf{F}_i^b(2)$ (corresponding to nontumor ROIs), the mean vector is defined as $\bar{\mu}_{1,2} = [\mu_1, \mu_2]$ [where μ_1 and μ_2 correspond to the means of $\mathbf{F}_i^b(1)$ and $\mathbf{F}_i^b(2)$], and the covariance matrix is denoted as $\Sigma_{1,2}$. Then, for every $i \in \{1, \dots, 406\}$

$$CV_i(1,2) = \left[\frac{(\bar{\mu}_{1,2}^T * \Sigma_{1,2} * \bar{\mu}_{1,2})}{(\bar{\mu}_{1,2}^T * \bar{\mu}_{1,2})^2}\right]^{1/2}. \tag{2}$$

Instability: We have utilized preparation-induced instability (denoted IS), as previously presented by Leo et al.,[22] for multisite comparison of histomorphometric features from pathology data. For example, for a pair of sites S_1 and S_2 with corresponding feature vectors $\mathbf{F}_i^b(1)$ and $\mathbf{F}_i^b(2)$, IS is computed as the percentage of bootstrapped pairwise comparisons in which $\mathbf{F}_i^b(1)$ is different from $\mathbf{F}_i^b(2)$ $\forall i \in \{1, \dots, 406\}$. Further algorithmic details for the implementation of IS are provided in the original paper.[22] The IS has the benefit of being more directly interpretable than CV. For example, if $IS_i = 0.1$, this indicates $\mathbf{F}_i^b(1)$ and $\mathbf{F}_i^b(2)$ are significantly different in 10% of cross-site comparisons (i.e., they are reproducible 90% of the time). Features with an IS closer to 1 are considered to be more unstable (and hence less reproducible).

Bootstrapped computation of *CV* and *IS*: For a given pair of sites S_p, S_q, consider the bootstrapped subsets $s_{p,n}$ and $s_{q,n}$, $n \in \{1, \dots, N\}$. The corresponding radiomic feature vectors are denoted as $F_i^b(p,n)$ (associated with all nontumor ROIs $c^b \in s_{p,n}$) and $F_i^b(q,n)$ (associated with $c^b \in s_{q,n}$), $i \in \{1, \dots, 406\}$. The $CV_{i,n}$ is computed for each pairwise comparison of $F_i^b(p,n)$ and $F_i^b(q,n)$, $\forall n \in \{1, \dots, N\}$. As there are six unique pairs of sites and $N = 100$ subsets, there are 600 comparisons for each feature. The cumulative CV for each feature is calculated as the average of all 600 comparisons.

Instability was calculated as a single value for each pair of sites, where the bootstrapped subsets $s_{k,n}, n \in \{1, \dots, N\}$ were utilized for pairwise comparison between sites. The cumulative IS for each feature was computed as the average of six values (one for each unique pair of sites).

2.8 *Quantifying Cross-Site Discriminability*

A QDA classifier[30] was trained to distinguish between tumor and nontumor ROIs, using each radiomic feature individually. The choice of classifier was based on the fact that single features were being evaluated in all experiments and because simpler classifiers may provide a more direct evaluation of a feature's discriminatory performance. The area under the ROC curve (denoted AUC) was used to quantify classifier performance, where an AUC $= 1$ implied perfect classification, whereas an AUC $= 0.5$ implied random guessing.[10]

Bootstrapped computation of AUC: AUC was calculated in a hold-one-site-out fashion, in order to determine the effect of utilizing different sites for training the classifier. Thus, when S_1 is the testing cohort, the corresponding training cohort comprised data from $\{S_2, S_3, S_4\}$. This was repeated four times so that each of the four sites $S_1, \dots, S_4$ in turn is considered to be the testing cohort once.

Bootstrapping was integrated into this process as follows. Consider when $\{S_2, S_3, S_4\}$ formed the training cohort and S_1 was the testing cohort. For each bootstrap iteration $n, n \in \{1, \dots, N\}$, tumor and nontumor ROIs from the training subsets $s_{q,n}, q \in \{2,3,4\}$ were used to train a QDA classifier $h_i(1,n)$, for each radiomic feature $i \in \{1, \dots, 406\}$. The $h_i(1,n)$ was then evaluated on samples in $s_{1,n}$ (from the testing cohort S_1) to compute $\text{AUC}_{i,n}$. This was repeated for each of the four possible testing cohorts and for each of $N = 100$ subsets so that the cumulative AUC for each feature was calculated as the average of 400 values.

2.9 *Experimental Evaluation*

As most existing studies have focused on intrasite reproducibility or discriminability,[15] the following experiments were constructed to assess cross-site trends in radiomic feature performance. To gain further insight into the optimal choice of radiomic features and associated parameters in a cross-site setting, the trade-off between complementary measures of reproducibility and discriminability was also examined. All processing, feature extraction, and analyses were performed using MATLAB2017a (The MathWorks, Inc., Natick, Massachusetts).

2.9.1 *Experiment 1: cross-site reproducibility and discriminability of radiomic feature families*

Radiomic features were ranked and evaluated in terms of each CV, IS, and AUC. Cumulative bar plots for each measure were visualized as follows. First, the range of each measure was divided into four bins. The proportion of features from each feature family that fell into each bin was then cumulatively plotted. This yielded a single bar per bin, comprising different colors for different feature families. A complementary set of bar plots were also visualized to show the percentage of total features found within each of the four bins. In addition, box-and-whisker plots were generated for each measure on a per-family basis to facilitate a direct comparison of trends across radiomic feature families.

2.9.2 *Experiment 2: relationship between cross-site reproducibility and discriminability and radiomic feature parameters*

To determine which features were both discriminable and reproducible, scatter plots of AUC versus IS and AUC versus CV were generated. Each point in the scatter plot was an individual radiomic feature with a unique color assigned for each feature family. Based on the trade-off between different evaluation measures, relevant "feature clusters" were identified, such as groups of features that were most reproducible (low CV or IS) as well as most discriminative (high AUC) across sites. Such clusters were further evaluated to determine any common parameters or window sizes that can be linked to their benchmarked performance.

3 Results

3.1 *Experiment 1: Radiomic Feature Family Trends in Cross-Site Reproducibility and Discriminability*

Figure 3 illustrates the performance of different radiomic feature families via cumulative bar plots in terms of each IS, CV, and AUC, and Fig. 4 similarly depicts this performance via box-and-whisker plots for each evaluation measure.

The gray features primarily had cross-site IS scores of under 0.2 with a median $IS = 0.1$ [Figs. 3(a) and 4(a)]. This suggests that the gray feature family is relatively stable, demonstrating cross-site reproducibility of at least 80% across all comparisons. As revealed by Figs. 3(b)–3(c) and Figs. 4(b)–4(c), the median performance of the gray feature family is among the best in terms of CV as well as AUC, albeit with a fair amount of variance.

The Haralick feature family comprises the majority of features with an $IS = 0$ [Figs. 3(a)–3(d)] with an upper bound of $IS < 0.01$ [Fig. 4(a)]. In other words, most Haralick features are reproducible in >99% of all cross-site comparisons. Based on the box-and-whisker plots in Figs. 4(a)–4(c), the Haralick feature family had the lowest median IS, as well as the second-lowest median CV and AUC values (only marginally worse than gray features).

The gradient features perform comparably to the Haralick feature family in terms of IS [Fig. 3(a)], indicating that they are relatively reproducible across all sites. However, the gradient feature family exhibits the worst median performance for CV across all feature families, appearing primarily within the bottom 50% of features [Fig. 3(b)] as well as the second worst performance in terms of AUC [Fig. 4(b)].

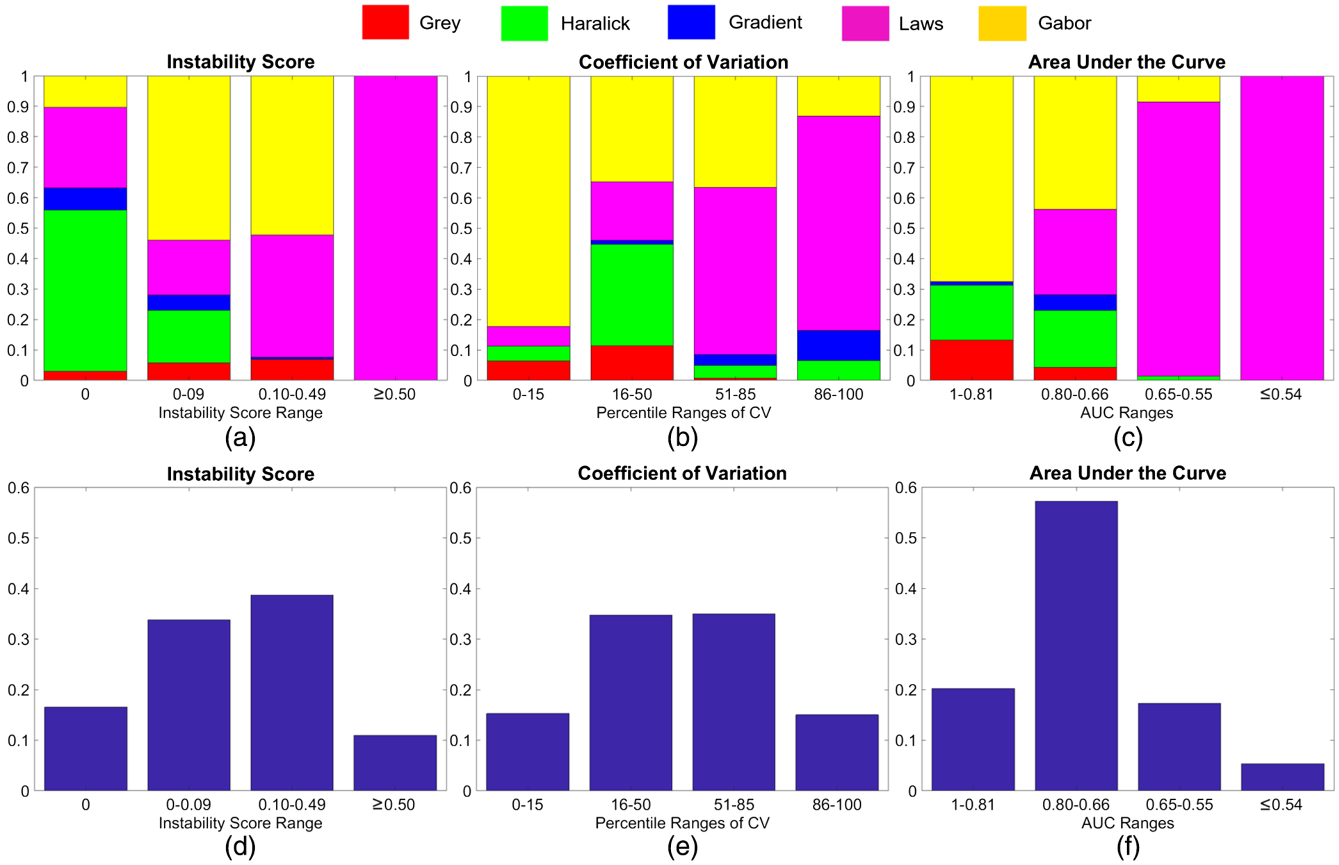

Fig. 3 Cumulative bar plots depicting the proportion of radiomic features that lie within specified bins for (a) and (d) IS, (b) and (e) CV, and (c) and (f) AUC. Bins in each plot are chosen to roughly correspond to 0 to 15th percentile, 16 to 50th percentiles, 51 to 85th percentiles, and 86 to 100th percentiles of performance for each measure. Note that the X-axis of CV (b) comprises percentiles as this measure is not bounded, unlike IS (a) and AUC (c), which are bounded and thus comprise the original values. The top row of bar plots (a)–(c) shows the proportion of each radiomic feature family (in different colors) that lies within each bin. The bottom row of bar plots (d)–(f) shows the percentage of the total number of features found within each bin.

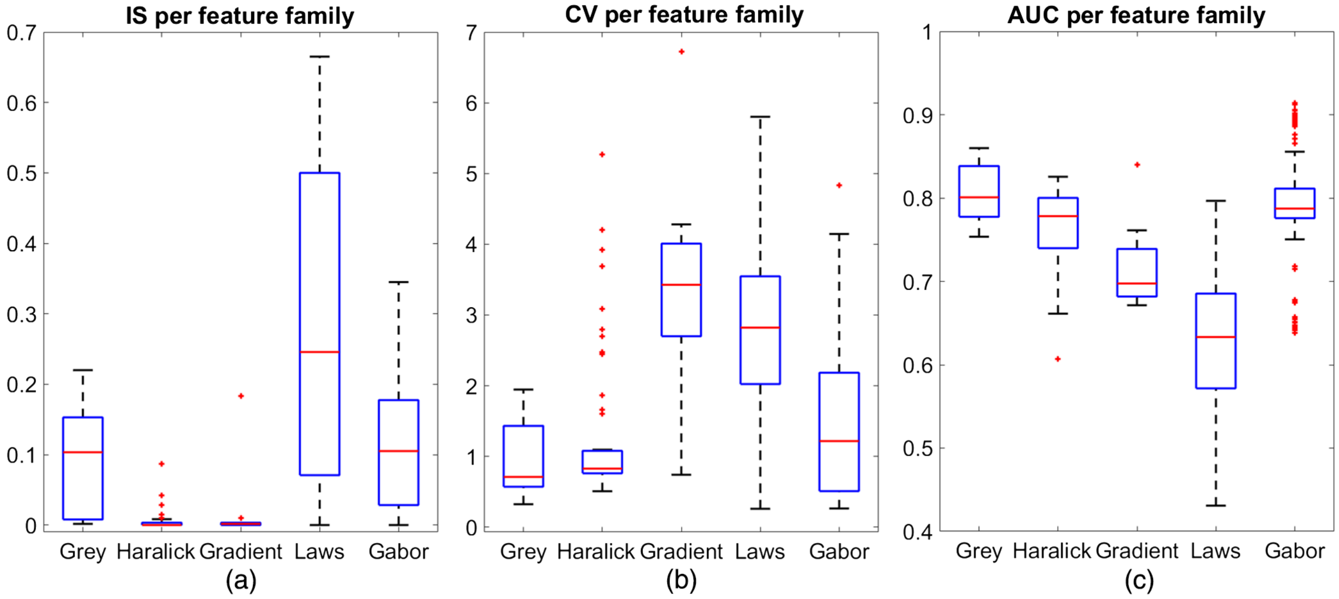

Fig. 4 Box-and-whisker plots illustrating overall trends for each feature family (along the X-axis), in terms of (a) IS, (b) CV, and (c) AUC. Note that the red line in the middle of each box reflects the median value, and the box is bounded by 25th and 75th percentiles. The whisker plot extends to the minimum and maximum values (obtained across all features) outside the box and outliers are denoted via the red plus symbol.

The Laws features are among the worst performing in term of IS, CV, and AUC [Figs. 3(a)–3(c)]. It is the only feature family that consistently appears in the lowest ranked bin across all three measures, as well as exhibits a large range in performance [seen in Figs. 4(a)–4(c)]. As shown in Figs. 4(d)–4(f), this means that Laws features comprise the majority of the bottom 10% to 15% of all features in each measure.

The Gabor feature family demonstrates performance similar to that of Haralick features. It is second-ranked in terms of IS [together with gray features, Figs. 3(a) and 4(b)] and comprises the majority of the top 15% of features in terms of CV [Figs. 3(b) and 3(e)] and AUC [Figs. 3(c) and 3(f)].

3.2 *Experiment 2: Relationship between Cross-Site Reproducibility and Discriminability*

Figure 5(a) depicts the scatter plot for AUC versus IS, based on which five distinctive “feature clusters” can be identified (highlighted via boxes, denoted as A–E). Similarly, Fig. 5(b) depicts a scatter plot for AUC versus $\ln(CV)$ (natural log used to ensure scaled visualization), within which three feature clusters have been denoted as F–H. A comprehensive list of all features in each of these clusters is provided in the Appendix Table 3.

Cluster A [yellow, upper left, Fig. 5(a)] comprises the best-performing Gabor features, all of which have a θ_{XZ} of 0 and were primarily extracted at window sizes of 7, 9, and 11. This set of features also appears in cluster F [yellow, upper left, Fig. 5(b)]. Cluster B [$\text{AUC} = 0.75-0.82$, $IS = 0.15-0.2$, Fig. 5(a)] also comprises Gabor features but with a θ_{XZ} greater than 0. While the Gabor features in clusters A and F are in-plane 2-D responses, cluster B corresponds to cross-plane 3-D responses. Cluster C [green, $IS = 0-0.01$, Fig. 5(a)] comprises Haralick features with $\text{AUC} = 0.79-0.83$, all of which are either inertia- or entropy-based features extracted at different window sizes. In Fig. 5(b), this group of Haralick features appears to have been split into clusters G (green, $CV = 1.65-1.85$) and H (green, $CV = 1.86-2.60$) based on differences in cross-site reproducibility. The marginally more reproducible features in cluster G correspond to entropy features extracted at larger window sizes (9 and 11). Clusters D (pink, $IS = 0$) and E (pink, $IS = 0.5$) comprise Laws features that have opposing trends in cross-site reproducibility but perform equally poorly in terms of cross-site discriminability ($\text{AUC} = 0.45-0.65$).

4 Discussion

Our overall experimental objective in this work was to provide insight into how radiomic features generalized across different sites in terms of specific benchmarking measures. In addition, we tried to determine how similar or different these benchmarking trends were within each radiomic feature family (i.e., features that share a common formulation) as well as across parameter choices.

We hypothesize that the excellent performance of the gray feature family observed in experiment 1 may be due to the extensive preprocessing applied to the MRIs. As this radiomic feature family is the most directly dependent on the underlying MR signal intensity values, it is also likely to be most affected by correcting the major sources of noise and variance on T2w MRI. Incidentally, similar findings on gray features have also been reported in the context of lung PET imaging.[31,32]

Our findings that Haralick features were associated with the excellent performance in IS and CV mirror the findings by Zhao et al.[14] on radiomic features derived from CT, who reported that Haralick features were resilient to variance in slice thickness and image reconstruction. However, when analyzing radiomic feature reproducibility on ADC MRI, Brynolfsson et al.[33] reported that Haralick features varied as a function of image noise, resolution, and intensity range. Our contrasting results in the current study may be due to our extensive preprocessing of T2w MRI to specifically correct the major sources of noise.

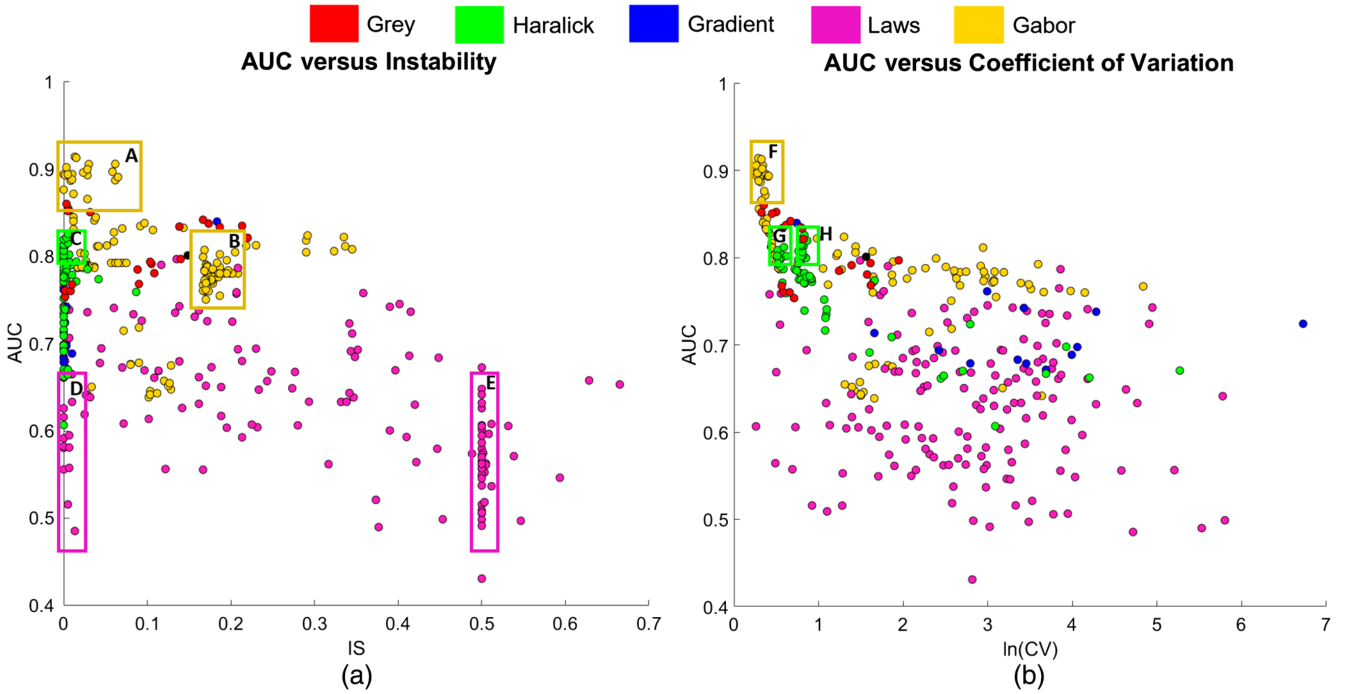

Fig. 5 The 2-D scatter plots illustrating the relationship between measures: (a) IS versus AUC and (b) $\ln(CV)$ versus AUC. Each point in the scatter plot represents an individual feature, and each color represents the feature family. The boxes $A-H$ identify specific feature clusters of interest, which are further discussed in Sec. 4.2.

A majority of the Haralick features in cluster C [Fig. 5(a)] are entropy- or inertia-based and parallels the work of Molina et al.[16] who reported that entropy-based features were relatively resilient to changes in dynamic range and resolution in brain MRIs. Inertia (sum-var, diff-var, sum-av, and diff-av) quantifies the probability of two pixels within a neighborhood either summing or subtracting to specific values, thus, computing the contrast in terms of MR intensity co-occurrences.[30] Subtle variance in image contrast is a well-known hallmark of prostate cancer appearance on MRI,[34] and these radiomic features have been widely used in previous approaches for prostate lesion classification via MRI.[11,28] Entropy-based features are known to capture the dissimilarity in intensity co-occurrences, which has been related to the heterogeneity of prostate cancer lesions on MRI.[17]

Interestingly, the gradient feature family uniquely demonstrates excellent IS performance but relatively poor performance in terms of CV [Figs. 4(a)–4(b)]. Further interrogation revealed that gradient features exhibit a large spread in feature values across sites with a reasonably low mean. In other words, gradient feature distributions markedly overlap between sites (resulting in a low IS), but CV reveals their subpar reproducibility. This could also explain why gradient features perform relatively poorly in cross-site discriminability experiments.

The overall poor reproducibility of Laws features in terms of both CV and IS indicates that there are at least one or more sites that exhibit very little overlap with the others, despite extensive correction of noise and variance on MRI. In a previous work utilizing CT images, Laws features demonstrated similarly poor reproducibility across different slice thicknesses[14] and repeat imaging tests.[35] Figure 3(c) shows that the Laws feature family is the only family with an AUC below 0.6, which is achieved by nearly 50% of all Laws features. Notably, Laws features also comprised ≈38% of our feature set due to the large number of permutations of Laws kernels. A very small subset of Laws features did achieve reasonable CV, IS, and AUC but with no discernible trend in kernels or window sizes among them. We hypothesize that Laws features may thus require careful kernel and parameter selection to ensure that poor features are not utilized. When contrasting Laws features comprising clusters D and E [Fig. 5(a)], we did not observe any clear differences in window sizes or parameters. This further reinforces our observation from experiment 1 that careful parameter selection may be required when using Laws features in a cross-site setting. Notably, Laws features also appear in a much more dispersed fashion in Fig. 5(b) compared to Fig. 5(a), thus highlighting the wide variations in individual feature performance.

While the Gabor feature family exhibit a reasonable trade-off between the three evaluation measures, high-performing Gabor features across all three evaluation measures were found to primarily comprise features extracted at large window sizes (7, 9, and 11) as well as being in-plane 2-D responses (i.e., $\theta_{XZ} = 0$). These findings are consistent with the previous work, where macroscale 2-D Gabor features were similarly found to be discriminatory for prostate cancer detection[28,36] and they exhibit reproducible performance across sites.[11] In addition, while Gabor features comprised as large a proportion of our feature set as Laws features, a majority of Gabor responses are both reproducible and discriminable across sites. The difference in performance between Gabor features in clusters A and F (in-plane 2-D responses) and cluster B (cross-plane 3-D responses) indicates that the 3-D component in Gabor features may be noisier, reflected by their higher instability. This is likely due to isotropic resampling in the Z-direction (from 3 to 0.27 mm) to correct for size differences as well as to use "true" 3-D radiomic features in our experiments. The relatively high discriminability of the 3-D component alone (reflected by high AUC values) does appear to indicate that they may nevertheless provide useful complementary information to the 2-D in-plane response (clusters A and F).

There has been limited recent work on the trade-offs observed between reproducibility and discriminability of radiomic features. In a study of lung CT-based radiomic features to discriminate granulomas from adenocarcinomas,[37] it was found that the top discriminating features were not necessarily the most reproducible across multiple sites (based on an instability measure[22]). Several studies have also looked at multisite discriminatory performance of MRI- and CT-based radiomic features and choice of classifiers.[38–40] Recent work by Ginsburg et al.[11] demonstrated significant variance in the performance of prostate radiomic features in holdout testing (i.e., when evaluated on MRI data from a site not considered for feature selection and classifier training). This swing in performance indicates that cross-site discriminability of radiomic features is likely to be closely tied to their reproducibility across sites. These findings appear to be borne out by our own experiments as well. Our results also paralleled the initial findings by the image biomarker standardization initiative[41] on radiomic feature variability in carefully controlled settings (although primarily for CT imaging thus far).

To quantify the reproducibility of radiomic features, previous studies have primarily focused on measures such as concordance correlation coefficient,[14] CV,[15] or intraclass correlation.[42] However, all three of these measures are based on the availability of repeat (test/retest) acquisitions, i.e., they quantified how reproducible a radiomic feature was between repeated imaging examinations. As our experiments were conducted using clinical multisite data, repeated acquisitions was not retrospectively available. We, therefore, utilized preparation-induced instability,[22] which defines feature reproducibility based on bootstrapped resampling and comparison of feature distributions between different sites. To ensure our findings could be placed in the context of previous work, we also utilized a multivariate extension of the popular CV measure. Between these two measures, we believe we have evaluated radiomic feature reproducibility as best possible using clinically acquired multisite data. An avenue of future work could also be to assess benchmark feature performance between scanners and sites using test-retest MRI scans from patients[16,43] or phantoms,[44] as has been done in PET[31] and CT[13,14] previously.

Previous work has shown that radiomic features may be able to distinguish between different Gleason grades of prostate cancer on MRI.[17] However, our study did not specifically evaluate how feature reproducibility was affected by tumor grade as this information was not consistently available across the different sites on a lesion basis. Instead we examined the reproducibility of radiomic features within relatively well-defined, homogeneous "nontumor" regions within the PZ of the prostate (from which we first excluded annotated tumor regions). While it is difficult to guarantee that there were no benign confounders in these nontumor regions, our approach ensured that tumor tissue heterogeneity[45] did not bias our analysis. Similarly, feature discriminability was considered in the context of separating tumor from "nontumor" independent of the tumor grade. A potential expansion of this study would be to examine radiomic features within similar Gleason grade tumors to determine how the trends from the current study generalize in that setting.

Despite our best efforts in conducting this large-scale multisite study of radiomic features, we do acknowledge several limitations. Though we curated 147 patients across four sites, the retrospective nature of our data collection caused an unequal number of patients per site, which could have affected our cross-site analysis. We did attempt to utilize bootstrapping when computing the evaluation measures to overcome this issue. As PIRADS reports and Gleason score information were not consistently available from all four institutions, we opted to examine feature trends in differentiating nontumor from tumor regions. As all patients had been clinically indicated for radical prostatectomy, our analysis was predicated on the assumption that tumor and nontumor regions would exhibit clear and consistent differences across these patient populations. This also ensured a reasonable number of patient samples for our analysis from across four sites. In addition, as a different expert performed the annotation at each site, inter-reader variability could be a confounding factor in our results. We attempted to account for this to some degree by limiting our analysis to the midgland alone, where experts tend to be most confident in their assessment.[46] In addition, reproducibility was evaluated within a well-defined nontumor region to further minimize the impact of variability in tumor annotations between experts.

The goal of our analysis was to determine how radiomic features vary as a function of different MRI acquisition parameters resulting in different voxel sizes, slice thicknesses, image appearance, and MR signal intensity ranges. As T2w MRI was both consistently available across all the sites while also being nonquantitative, we opted to limit our study to this one sequence. While our feature trends may hold for other MR sequences, these will have to be studied separately. We attempted to correct for the largest sources of variations in T2w MR appearance via preprocessing (including bias correction, voxel resampling, and intensity standardization), which may have impacted the reproducibility and discriminability of the radiomic features we considered. Past work on the effect of bias field correction on the performance of CAD in prostate cancer detection[47] reported that bias field correction improved classifier AUC as well as intrasite reproducibility. Similarly, when evaluating the effect of sequential preprocessing steps, bias field correction followed by standardization was found to result in optimal image quality and better algorithmic performance.[25] However, the interplay between these different operations in terms of the resulting radiomic features has not been extensively explored. This would require a comprehensive evaluation of different permutations of preprocessing operations, which was out of the scope of the current study and will be examined in future work.

5 Concluding Remarks

In this work, we presented the first empirical, cross-site evaluation of MRI-based radiomic feature reproducibility and discriminability. We evaluated 406 3-D radiomic features from across five distinct feature families in terms of three benchmarking measures, utilizing 147 T2w MRI patient datasets from four different sites. We then attempted to compare different feature families and subgroups to determine how specific filter or parameter choices were linked to their performance. The lessons learned with respect to specific radiomic operators also helped us understand performance differences between feature families, which can help in choosing specific parameters or formulations of radiomic features. The following are our key takeaways.

- The popular Haralick feature family demonstrated the best trade-off in terms of both cross-site reproducibility and discriminability. Further interrogation revealed that specific Haralick features that quantify image contrast and entropy were the most reproducible and discriminable across sites, likely benefiting from using intensity co-occurrences (more resilient to absolute intensity values) as well as the extensive preprocessing applied to the MRIs prior to feature extraction. As these features have been reported as associated with prostate cancer presence[11,28] and aggression,[34] our results appear to confirm their utility for disease characterization on prostate MRI.
- Gabor features extracted as 2-D responses and at larger window sizes also demonstrated both good reproducibility and discriminability.[11] Like the Haralick feature family, these features have been successfully used to detect and segment cancer within the cancerous lesions on prostate MRI.[28]
- Notably, the 3-D component of these Gabor responses was marginally poorer in terms of both reproducibility and discriminability. Despite the addition of complementary information, noise in the 3-D component was likely introduced by isotropic voxel resampling. While 3-D features based on isotropic-resampled volumes have been commonly used in different applications,[48,49] our results indicate that there may be a performance trade-off that needs to be considered when using 2-D versus 3-D features.
- Laws features comprised the majority of the worst-ranked radiomic features in terms of both cross-site reproducibility and discriminability. A possible explanation for the highly variable performance of Laws features may be the large number of parameter combinations (and thus, features), making it hard to identify which of them were optimal for disease characterization. As several Laws features have been associated with specific pathological patterns of disease,[50] our findings indicate that careful parameter selection may be required when considering this feature family.
- Careful preprocessing of the T2w MRI prior to radiomic feature extraction appeared to enable both Haralick and gray radiomic features to benchmark as more reproducible and discriminable in multisite setting than has been reported previously.[16] A more extensive study of the interplay and effectiveness of different processing operations may thus be critical when conducting such large multisite studies.

While our study has examined radiomic features alone, it may also be interesting to consider how our evaluation measures might be utilized in the context of benchmarking deep learning-based approaches. This could provide a better understanding of how convolutional filter responses (commonly used in deep learning) vary as a result of fundamental MRI parameter changes. Other future directions could include examining the effect of cross-site feature reproducibility and discriminability in applications such as patient prognosis and outcome prediction. This could pave the way for more robust and reliable radiomics-based decision support systems to be utilized in clinical practice.

6 Appendix

Table 3 provides a list of features found within each cluster referenced in Fig. 5 and in the accompanying description to provide more detailed information with regards to specific features and parameters.

Table 3 Summary of radiomic features in clusters A–G highlighted in Fig. 5.

Clusters A and F: AUC$(0.86-0.93)$, $IS(0-0.1)$, $CV(1.29-1.53)$
Gabor $\theta_{XY}=0$, $\theta_{XZ}=0$, WS = (9,11)
Gabor $\theta_{XY}=\frac{2\pi}{8}$, $\theta_{XZ}=0$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{2\pi}{8}$, $\theta_{XZ}=0$, WS = (5,7,9,11)
Gabor $\theta_{XY}=\frac{3\pi}{8}$, $\theta_{XZ}=0$, WS = (7)
Gabor $\theta_{XY}=\frac{5\pi}{8}$, $\theta_{XZ}=0$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{6\pi}{8}$, $\theta_{XZ}=0$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{7\pi}{8}$, $\theta_{XZ}=0$, WS = (7,9,11)
Cluster B: $AUC(0.75-0.82)$, $IS(0.15-0.20)$
Gabor $\theta_{XY}=0$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=0$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=0$, $\theta_{XZ}=\frac{3\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{2\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{2\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{2\pi}{8}$, $\theta_{XZ}=\frac{3\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{2\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}\frac{2\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{2\pi}{8}$, $\theta_{XZ}=\frac{3\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{3\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{3\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{3\pi}{8}$, $\theta_{XZ}=\frac{3\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{4\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{4\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{4\pi}{8}$, $\theta_{XZ}=\frac{3\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{5\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{5\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{5\pi}{8}$, $\theta_{XZ}=\frac{3\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{6\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{6\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{6\pi}{8}$, $\theta_{XZ}=\frac{3\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{7\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)
Gabor $\theta_{XY}=\frac{7\pi}{8}$, $\theta_{XZ}=\frac{2\pi}{8}$, WS = (7,9,11)

Table 3 (*Continued*).

Gabor $\theta_{XY} = \frac{7\pi}{8}$, $\theta_{XZ} = \frac{3\pi}{8}$, WS = (7,9,11)
Cluster C: $\text{AUC}(0.79 - 0.83), IS(0 - 0.01)$
Haralick inertia WS = (3,5,7,9,11)
Haralick sum-var WS = (3,5,7,9,11)
Haralick diff-var was = (3,5,7,9)
Haralick diff-av WS = (3,7,9,11)
Haralick entropy WS = (9,11)
Haralick idm WS = (9,11)
Haralick sum-ent WS = (9,11)
Haralick diff-ent WS = (9,11)
Cluster D: $AUC(0.45 - 0.65), IS(0)$
Laws E3S3L3
Laws E5S5L5
Laws E5R5L5
Laws W5R5L5
Laws L5R5E5
Laws S5R5E5
Laws W5R5E5
Laws E5R5S5
Laws S5W5S5
Laws W5R5S5
Cluster E: $\text{AUC}(0.45 - 0.65), IS(0.5)$
Laws E3S3S3
Laws L3E3S3
Laws S3L3S3
Laws S3S3S3
Laws W5E5E5
Laws L5W5S5
Laws E5E5S5
Laws E5W5S5
Laws S5E5S5
Laws S5R5S5
Laws R5L5S5
Laws R5E5S5
Laws R5S5S5
Laws R5R5S5
Laws W5W5S5
Laws L5E5R5
Laws L5S5R5
Laws L5R5R5
Laws L5W5R5
Laws E5S5R5
Laws E5R5R5
Laws S5S5R5
Laws S5R5R5
Laws R5L5R5
Laws R5E5R5
Laws R5S5R5
Laws R5W5R5
Laws W5E5R5
Laws W5S5R5
Laws L5R5W5
Laws L5W5W5
Laws E5L5W5
Laws S5R5W5
Laws R5S5W5
Laws R5R5W5
Laws W5E5W5
Laws W5S5W5
Laws W5R5W5
Cluster G: $\text{AUC}(0.79 - 0.83), CV(1.65 - 1.85)$
Haralick entropy ws = 9,11
Haralick sum-ent ws = 9,11
Haralick diff-ent ws = 9,11
Haralick diff-av ws = 9,11
Haralick idm ws = 9,11
Cluster H: $\text{AUC}(0.79 - 0.83), CV(1.86 - 2.6)$
Haralick inertia ws = 3,5,7,9,11
Haralick sum-var ws = 3,5,7,9,11
Haralick diff-av ws = 3,5,7
Haralick diff-var ws = 3,5,7,9

Disclosures

No conflicts of interest, financial or otherwise, are declared by any authors involved in this work.

Acknowledgments

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Nos. 1U24CA199374-01, R01CA202752-01A1, R01CA208236-01A1, R01CA216579-01A1, R01CA220581-01A1, 1U01CA239055-01; National Center for Research Resources under Award No. 1 C06 RR12463-01; VA Merit Review Award IBX004121A from the United States Department of Veterans Affairs Biomedical Laboratory Research and Development Service; the DOD Prostate Cancer Idea Development Award (W81XWH-15-1-0558); the DOD Lung Cancer Investigator-Initiated Translational Research Award (W81XWH-18-1-0440); the DOD Peer Reviewed Cancer Research Program (W81XWH-16-1-0329); the Ohio Third Frontier Technology Validation Fund; the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering, and the Clinical and Translational Science Award Program (CTSA) at Case Western Reserve University. Initial finding of this research were published in the 2018 SPIE Proceedings along with an oral presentation, though significant changes and additions have been made since that time.[51] The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs, the Department of Defense, the United States Government, or SPIE.

References

1. V. Kumar et al., "Radiomics: the process and the challenges," *Magn. Reson. Imaging* **30**(9), 1234–1248 (2012).
2. H. J. W. L. Aerts, "The potential of radiomic-based phenotyping in precision medicine," *JAMA Oncol.* **2**, 1636 (2016).
3. R. T. H. M. Larue et al., "Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures," *Br. J. Radiol.* **90**(1070), 20160665 (2017).
4. T. P. Coroller et al., "CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma," *Radiother. Oncol.* **114**(3), 345–350 (2015).
5. J. Lao et al., "A deep learning-based radiomics model for prediction of survival in Glioblastoma multiforme," *Sci. Rep.* **7**, 10353 (2017).
6. R. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE* **67**(5), 786–804 (1979).
7. C. Ma et al., "An improved sobel algorithm based on median filter," in *2010 2nd Int. Conf. Mechan. Electron. Eng.* (2010).
8. K. I. Laws, *Textured Image Segmentation* No. USCIPI-940, University of Southern California Los Angeles Image Processing INST, 76–118 (1980).
9. A. Bovik, M. Clark, and W. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(1), 55–73 (1990).
10. A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recogn.* **30**(7), 1145–1159 (1997).
11. S. B. Ginsburg et al., "Radiomic features for prostate cancer detection on MRI differ between the transition and peripheral zones: preliminary findings from a multi-institutional study," *J. Magn. Reson. Imaging* **46**(1), 184–193 (2016).
12. A. H. Dinh et al., "Quantitative analysis of prostate multiparametric MR images for detection of aggressive prostate cancer in the peripheral zone: a multiple imager study," *Radiology* **280**(1), 117–127 (2016).
13. D. Mackin et al., "Measuring computed tomography scanner variability of radiomics features," *Investig. Radiol.* **50**(11), 757–765 (2015).
14. B. Zhao et al., "Reproducibility of radiomics for deciphering tumor phenotype with imaging," *Sci. Rep.* **6**(1), 23428 (2016).
15. S. Gourtsoyianni et al., "Primary rectal cancer: repeatability of global and local-regional MR imaging texture features," *Radiology* **284**(2), 552–561 (2017).
16. D. Molina et al., "Lack of robustness of textural measures obtained from 3D brain tumor MRIs impose a need for standardization," *PLoS One* **12**, e0178843 (2017).
17. D. Fehr et al., "Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images," *Proc. Nat. Acad. Sci.* **112**, E6265 (2015).
18. S. Kahali, S. K. Adhikari, and J. K. Sing, "On estimation of bias field in MRI images: polynomial vs Gaussian surface fitting method," *J. Chemomet.* **30**(10), 602–620 (2016).
19. M. E. Soto, J. E. Pezoa, and S. N. Torres, "Thermal noise estimation and removal in MRI: a noise cancellation approach," *Lect. Not. Comput. Sci.* **7042**, 47–54 (2011).
20. L. Nyul, J. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imaging* **19**(2), 143–150 (2000).
21. S. Aerts, G. Haesbroeck, and C. Ruwet, "Multivariate coefficients of variation: comparison and influence functions," *J. Multivar. Anal.* **142**, 183–198 (2015).
22. P. Leo et al., "Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images," *J. Med. Imaging* **3**(4), 047502 (2016).
23. G. Lemaître et al., "Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review," *Comput. Biol. Med.* **60**, 8–31 (2015).
24. D. Palumbo et al., "Interplay between bias field correction, intensity standardization, and noise filtering for T2-weighted MRI," in *2011 Annual Int. Conf. IEEE Eng. Med. Biol. Soc.* (2011).
25. A. Madabhushi and J. K. Udupa, "Interplay between intensity standardization and inhomogeneity correction in MR image processing," *IEEE Trans. Med. Imaging* **24**(5), 561–576 (2005).
26. M. S. Cohen, R. M. Dubois, and M. M. Zeineh, "Rapid and effective correction of RF inhomogeneity for high field magnetic resonance imaging," *Hum. Brain Mapp.* **10**(4), 204–211 (2000).
27. N. J. Tustison et al., "N4itk: Improved n3 bias correction with robust b-spline approximation," in *2010 IEEE Int. Symp. Biomed. Imaging: From Nano to Macro* (2010).
28. S. E. Viswanath et al., "Central gland and peripheral zone prostate tumors have significantly different quantitative imaging signatures on 3 tesla endorectal, in vivo T2-weighted MR imagery," *J. Magn. Reson. Imaging* **36**(1), 213–224 (2012).
29. A. Albert and L. Zhang, "A novel definition of the multivariate coefficient of variation," *Biomet. J.* **52**(5), 667–675 (2010).
30. R. M. Rangayyan, *Biomedical Image Analysis*, CRC Press, Boca Raton (2005).
31. R. T. H. Leijenaar et al., "Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability," *Acta Oncol.* **52**, 1391–1397 (2013).
32. B. A. Altazi et al., "Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms," *J. Appl. Clin. Med. Phys.* **18**, 32–48 (2017).
33. P. Brynolfsson et al., "Haralick texture features from apparent diffusion coefficient (ADC) MRI images depend on imaging and pre-processing parameters," *Sci. Rep.* **7**(1), 4041 (2017).
34. A. Wibmer et al., "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.* **25**(10), 2840–2850 (2015).
35. Y. Balagurunathan et al., "Test-retest reproducibility analysis of lung CT image features," *J. Dig. Imaging* **27**, 805–823 (2014).
36. A. Algohary et al., "Radiomic features on MRI enable risk categorization of prostate cancer patients on active surveillance: preliminary findings," *J. Mag. Reson. Imaging* **48**(3), 818–828 (2018).
37. M. Orooji et al., "Combination of computer extracted shape and texture features enables discrimination of granulomas from adenocarcinoma on chest computed tomography," *J. Med. Imaging* **5**(02), 1 (2018).

38. Z.-C. Li et al., “Multiregional radiomics features from multiparametric MRI for prediction of MGMT methylation status in glioblastoma multiforme: a multicentre study,” *Eur. Radiol.* **28**(9), 3640–3650 (2018).
39. A. E. Fetit et al., “Radiomics in paediatric neuro-oncology: a multicentre study on MRI texture analysis,” *NMR Biomed.* **31**(1) (2017).
40. C. Parmar et al., “Machine learning methods for quantitative radiomic biomarkers,” *Sci. Rep.* **5**(1), 13087 (2015).
41. A. Zwanenburg et al., “Image biomarker standardisation initiative,” arXiv e-prints (2016).
42. M. Bologna et al., “Stability assessment of first order statistics features computed on ADC maps in soft-tissue sarcoma,” in *39th Ann. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC 2017)* (2017).
43. A. Fedorov et al., “An annotated test-retest collection of prostate multiparametric MRI,” *Sci. Data* **5**(180281) (2018).
44. B. Baebler, K. Weiss, and D. Pinto dos Santos, “Robustness and reproducibility of radiomics in magnetic resonance imaging: a phantom study,” *Invest. Radiol.* **54**(4), 221–228 (2018).
45. R. Stoyanova et al., “Prostate cancer radiomics and the promise of radiogenomics,” *Transl. Cancer Res.* **5**(4), 432–447 (2016).
46. W. L. Smith et al., “Prostate volume contouring: a 3D analysis of segmentation using 3dtrus, CT, and MR,” *Int. J. Radiat. Oncol. Biol. Phys.* **67**(4), 1238–1247 (2007).
47. S. Viswanath et al., “Empirical evaluation of bias field correction algorithms for computer-aided detection of prostate cancer on T2w MRI,” in *Medical Imaging 2011: Computer-Aided Diagnosis* (2011).
48. A. Chaddad, M. Kucharczyk, and T. Niazi, “Multimodal radiomic features for the predicting Gleason score of prostate cancer,” *Cancers* **10**(8), 249 (2018).
49. R. Ortiz-Ramon et al., “A radiomics evaluation of 2D and 3D MRI texture features to classify brain metastases from lung cancer and melanoma,” in *39th Ann. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC 2017)* (2017).
50. N. M. Braman et al., “Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI,” *Breast Cancer Res.* **19** (2017).
51. P. Chirra et al., “Empirical evaluation of cross-site reproducibility in radiomic features for characterizing prostate MRI,” in *Medical Imaging 2018: Computer-Aided Diagnosis* (2018).

Biographies of the authors are not available.