Viswanath *et al. BMC Medical Imaging* (2019) 19:22
https://doi.org/10.1186/s12880-019-0308-6

BMC Medical Imaging

RESEARCH ARTICLE Open Access

Comparing radiomic classifiers and classifier ensembles for detection of peripheral zone prostate tumors on T2-weighted MRI: a multi-site study

Satish E. Viswanath[1*], Prathyush V. Chirra[1], Michael C. Yim[2], Neil M. Rofsky[3], Andrei S. Purysko[4], Mark A. Rosen[5], B Nicolas Bloch[6] and Anant Madabhushi[1]

Abstract

Background: For most computer-aided diagnosis (CAD) problems involving prostate cancer detection via medical imaging data, the choice of classifier has been largely ad hoc, or been motivated by classifier comparison studies that have involved large synthetic datasets. More significantly, it is currently unknown how classifier choices and trends generalize across multiple institutions, due to heterogeneous acquisition and intensity characteristics (especially when considering MR imaging data). In this work, we empirically evaluate and compare a number of different classifiers and classifier ensembles in a multi-site setting, for voxel-wise detection of prostate cancer (PCa) using radiomic texture features derived from high-resolution in vivo T2-weighted (T2w) MRI.

Methods: Twelve different supervised classifier schemes: Quadratic Discriminant Analysis (QDA), Support Vector Machines (SVMs), naïve Bayes, Decision Trees (DTs), and their ensemble variants (bagging, boosting), were compared in terms of classification accuracy as well as execution time. Our study utilized 85 prostate cancer T2w MRI datasets acquired from across 3 different institutions (1 for discovery, 2 for independent validation), from patients who later underwent radical prostatectomy. Surrogate ground truth for disease extent on MRI was established by expert annotation of pre-operative MRI through spatial correlation with corresponding *ex vivo* whole-mount histology sections. Classifier accuracy in detecting PCa extent on MRI on a per-voxel basis was evaluated via area under the ROC curve.

Results: The boosted DT classifier yielded the highest cross-validated AUC (= 0.744) for detecting PCa in the discovery cohort. However, in independent validation, the boosted QDA classifier was identified as the most accurate and robust for voxel-wise detection of PCa extent (AUCs of 0.735, 0.683, 0.768 across the 3 sites). The next most accurate and robust classifier was the single QDA classifier, which also enjoyed the advantage of significantly lower computation times compared to any of the other methods.

Conclusions: Our results therefore suggest that simpler classifiers (such as QDA and its ensemble variants) may be more robust, accurate, and efficient for prostate cancer CAD problems, especially in the context of multi-site validation.

Keywords: Classifiers, Radiomics, Prostate cancer, MRI, Comparison

*Correspondence: sev21@case.edu
[1]Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH, USA
Full list of author information is available at the end of the article

Background

Pattern recognition approaches for distinguishing between object classes (diseased versus normal or cancerous versus benign) on bioinformatics [1, 2] or medical imaging [3, 4] data typically involve first extracting informative features, which are then used to train a machine learning classifier. In the case of medical imaging data, depending on the specific classes to be discriminated, a variety of computerized image-derived (i.e. *radiomic*) features have been proposed and evaluated [5–7]. A problem that has perhaps not received as much attention is the choice of classifier scheme for a particular computer aided detection (CAD) problem. The advent of ensemble schemes (bagging [8], boosting [9]) to overcome known shortcomings of classifier algorithms with respect to bias and variance [10] have further expanded the choices available when choosing an optimal classifier.

While several classifier comparison studies [11–16] have been reported using large standardized datasets, there has been some lack of concordance with regard to their recommendations for choice of optimal classifier scheme or how classifier trends generalize as in the presence of noise. Most recently, one of the largest comparison studies evaluated 179 classifiers on the popular UCI machine learning repository [17] and determined that random forests may be the most effective choice for most problems [18]. Thus far, medical imaging CAD studies [3, 4, 19] have also arrived at similar conclusions when identifying the optimal classifier scheme or when reporting trends between classifiers for a specific problem. In comprehensive comparisons of 12 different classifier methods on large databases of lung cancers as well as head & neck cancers when using radiomic features (including independent training and validation cohorts), it was reported that the random forest classifier yielded the best predictive performance in both problems [20, 21]. Notably, both these studies also acknowledged that the choice of classifier method had the most dominant effect on predictive performance (i.e. it was a larger source of performance variability as compared to feature selection method or size of cohort).

In this work we aim to compare classifier performance in the specific context of voxel-wise detection of tumors in the peripheral zone (PZ) of the prostate via "radiomic" texture-based features derived from T2w MRI. While a variety of approaches have been proposed for prostate cancer CAD [22] recently, they have typically been evaluated via cross-validation using data from a single site or scanner. While acquisition-related noise artifacts and acquisition protocols are relatively homogeneous and well understood in the single-center setting, these issues are less studied when considering imaging data from multiple institutions. Multi-site data suffers from differences in scanners, acquisition protocols, and resolution differences - all of which can affect classifier performance, and hence choice of classifier, significantly. To address this question, we will perform an empirical examination of how classifier trends and performance generalize for prostate cancer (PCa) detection on MRI data that has been acquired from across multiple different institutions and scanners. The classifiers we will consider include discriminant analysis (DA) [23], support vector machines (SVMs) [24], naïve Bayes [25], and decision trees (DTs) [26], as well as their bagging and boosting variants; thus covering most popular families of classifier methods. We will attempt to identify the optimal classifier for prostate cancer CAD in terms of cross-site detection accuracy, while considering computational complexity as well. The overall workflow and experimental design of our paper is depicted in Fig. 1.

Methods

Data description

A total of 85 patient datasets were considered in this study. These were retrospectively obtained after de-identification from 3 different institutions, under previous IRB-approved research protocols. For the current study, informed consent was waived by the IRB because the data was obtained retrospectively and did not include any protected health information, by the University Hospitals IRB. All patients included had first been confirmed to have prostate cancer via positive core needle biopsies and later underwent a radical prostatectomy. Prior to surgery, all patients had been imaged via MRI using a combined torso-phased array and endorectal coil, at their respective institutions. Further details of the imaging acquisition at each institution are summarized in Table 1. Note that D_1 was utilized for discovery and optimization of the classifiers alone. D_2 and D_3 were then used for independent evaluation and validation of classifier performance and trends.

Expert annotation of central gland, PZ, and PCa extent on T2w MRI

For all 85 datasets considered, the central gland and the PZ were annotated on the the axial endorectal T2w MRI image by a radiologist (a different expert annotated data from each institution). Regions of cancer extent within the PZ were annotated as follows:

Cohorts D_1, D_2: As the radical prostatectomy specimens had been processed as whole-mount histology sections at these sites, it was possible to identify corresponding WMHS and MRI sections. Each pair of corresponding WMHS and MR images were first affinely aligned to enable correction of large translations, rotations, and differences in image scale. A non-linear alignment of WMHS and MR images was then performed via fully automated, non-linear hierarchical (multi-scale) B-spline registration driven by a higher-order variant of mutual information

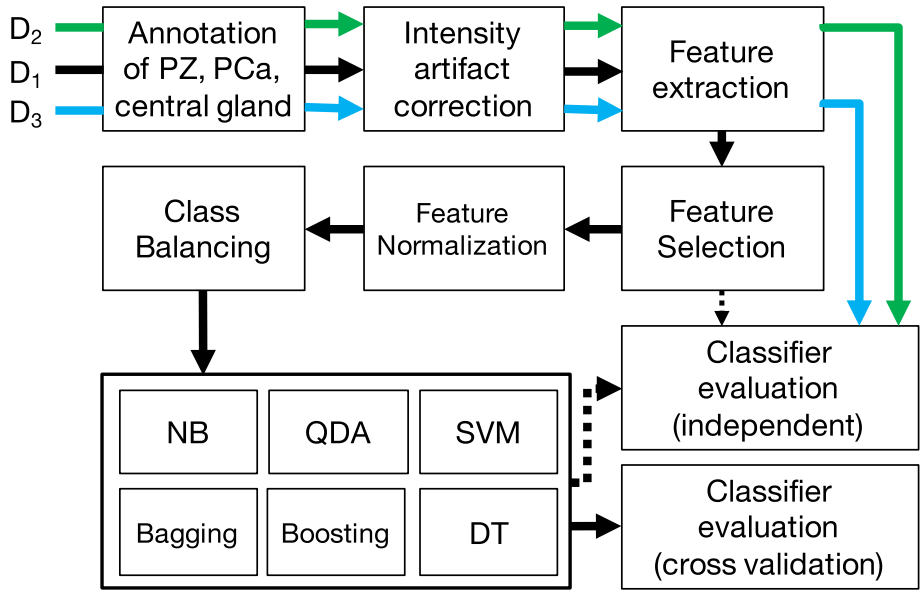

Fig. 1 Overview of experimental workflow. Arrow colors show how training (D_1, black) and independent validation cohorts (D_2 and D_3 in green and blue) go through different stages

[27]. This allowed for spatial mapping of pathologic annotations of PCa extent onto T2w MRI, which were further examined and manually corrected for registration artifacts (as required) by the radiologist for that site.

Cohort D_3: As only low-resolution photographs of digitized whole-mount sections were available from this site, these had to be visually correlated with corresponding MRIs by a radiologist. Based on this information, they annotated PCa extent on the T2w MR images.

For the purposes of this study, only PZ regions within the midgland region of the prostate were considered in all 85 datasets. This was to ensure that there was maximum confidence in the PCa annotations as well as consistency in non-tumor and tumor characteristics across sites. Any regions within the PZ that were not annotated as cancer were thus considered to be "non-tumor" regions. In total, we had 116 prostate tumors annotated across 85 patient datasets. Representative 2D T2w MRI midgland sections from each site together with annotations for central gland, PZ, and PCa are depicted in Fig. 2a–c.

Table 1 Summary of multi-site prostate T2w MRI data used in this study

Site	[x,y,z] voxel dimensions (mm)	Magnetic field strength (T)	TR/TE (ms)	Number of datasets
D_1	[0.27,0.27,2.20]	3	4216 - 8266/ 155 - 165	16
D_2	[0.41,0.41,3.00]	3	2840 - 7500/ 107 - 135	13
D_3	[0.27,0.27,2.96]	3	4754/115	56

D_1 was used as the discovery cohort, while D_2 and D_3 were used as independent validation cohorts

Post-processing of T2w MRIs to account for intensity-based artifacts

The prostate ROI was corrected for known acquisition-based intensity artifacts; bias field inhomogeneity [28] and intensity non-standardness [29]. The effects of bias field occur due to the usage of an endorectal probe [28], and manifest as a smooth variation of signal intensity across the T2w MR image. Bias field has been shown to significantly affect the automated classification of tissue regions [30], and was corrected for via the popular N3 algorithm [31]. Intensity non-standardness [29] refers to the issue of MR signal "intensity drift" across different imaging acquisitions, resulting in MR signal intensities lacking tissue-specific numeric meaning between images obtained on different scanners, despite using the same MRI protocol for the same body region. This can be seen the widely varying intensity distributions depicted in Fig. 2d (different colors correspond to different sites). This was corrected using an automated implementation of the method presented by Nyul et al. [29], whereby the signal intensity histograms across different patient MRI studies were non-linearly aligned to a common template. As a result of standardization, the distributions in Fig. 2e appear to be more consistent and aligned to the common template (shown in black).

Each of the 85 datasets were post-processed for these artifacts independent of each other. We denote $\mathcal{C} = (C, f)$ as representing the corrected, standardized prostate ROI

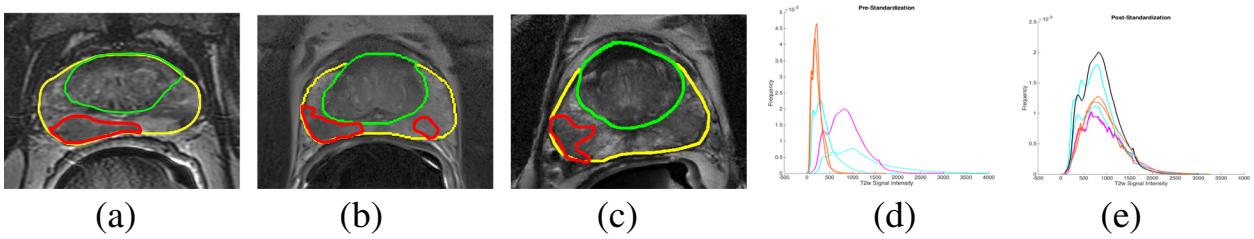

Fig. 2 Annotated prostate T2w MRI midgland sections from (**a**) D_1, (**b**) D_2, (**c**) D_3; overlaid with delineations of peripheral zone (yellow), PCa (red), and central gland (green). Also shown are representative T2w intensity distributions from different sites (**d**) before standardization, and (**e**) after correcting intensity drift, where colors correspond to: D_1 (orange), D_2 (cyan), D_3 (pink), template (black)

comprising voxels (samples) $c \in C$, where $f(c)$ represents the MR image intensity value at voxel c. The set of voxels in PCa regions (as annotated by experts) are denoted as $G(C) = \{c|l(c) = 1\}$ (also called the *target class*, denoted ω_{+1}). Table 2 summarizes commonly used notation and symbols appearing in this paper.

Extracting PZ tumor specific radiomic texture features from T2w MRI

It has previously been demonstrated that PZ tumor appearance on T2w MRI may be specifically modeled by image texture (radiomic) features [32, 33]; many of which have been widely used in the prostate cancer CAD literature [22]. A total of 116 image features corresponding to 4 different types of texture were extracted, including Gabor [34] and Haar [35] wavelet features, as well as first and second order texture [36, 37] features. After feature extraction, every voxel $c \in C$ is associated with a 116-dimensional feature vector denoted $\widehat{\mathbf{F}}(c) = \{f_1(c), f_2(c), \ldots, f_{116}(c)\}$, for every $c \in C$.

In order to determine radiomic features specific to PZ PCa regions, feature selection was performed using the minimum Redundancy Maximum Relevance (mRMR) method [38], using voxel-wise features and corresponding voxel-wise labels for tumor and non-tumor regions from site D_1 alone. The mRMR scheme attempts to simultaneously optimize two distinct criteria: (a) selecting features that have the maximal mutual information (MI) with respect to the corresponding set of labels, and (b) that selected features are those have the minimum MI with respect to each other. Use of mRMR ensured that bias towards a particular classifier was prevented as the feature selection step utilized an independent objective function (MI). This is in direct contrast to forward or backward feature selection [39] where the classifier is more integral to the selection process. The result of mRMR feature selection was a subset of 25 voxel-wise radiomic features characterizing PZ PCa appearance (denoted $\mathbf{F}(c)$, complete listing in Appendix A).

Table 2 Summary of commonly used notation and symbols in this paper

c	Samples in set C
n	Number of samples in C
$\mathbf{F}(c)$	N-dimensional (texture) feature vector
$\mathcal{F}$	Set of all feature vectors
$l(c)$	Class label of sample c
ω_{+1}, ω_{-1}	Classes associated with $l(c) = 1, l(c) = 0$
$\mathbf{h}^{\beta}$	Classifier, $\beta \in \{QDA, SVM, Bay, DT\}$
h_t^{β}	Component classifier within $\mathbf{h}^{Bag,\beta}$, $\mathbf{h}^{Boost,\beta}$
$\mathbf{h}^{Bag,\beta}$	Bagged classifier
$\mathbf{h}^{Boost,\beta}$	Boosted classifier

Classifier construction

The feature set $\mathbf{F}(c)$ was input to the different classification algorithms and their ensemble variants (summarized in Table 3). The specific steps for classifier construction are described below. All classifiers were constructed using datasets from discovery site D_1 alone.

Feature normalization

Normalization of radiomic features ensures that different feature values lie in a comparable range of values when input to a classifier. Given a feature vector $\mathbf{F}(c)$, this can be done for each $f_i(c) \in \mathbf{F}(c)$ as follows,

$$f_i(c) = \frac{f_i(c) - \mu_{\mathbf{i}}}{\sigma_i}, \tag{1}$$

where μ_i is the mean and σ_i is the mean absolute deviation (MAD) corresponding to feature $i, i \in \{1, \ldots, N\}$. As a result of normalization, $\forall c \in C$, each feature in $\mathbf{F}(c)$ was transformed to have a mean of 0 and a MAD of 1. Note that radiomic features from D_2 and D_3 were normalized with respect to the mean and MAD of corresponding radiomic features from D_1.

Class balancing

A significant issue when training a supervised classifier is the *minority class problem* [40], wherein the target class

Table 3 Machine learning classification algorithms evaluated in this work, together with their associated parameters and notation

QDA [23]	$\mathbf{h}^{QDA}$	-	MATLAB
	$\mathbf{h}^{Bag,QDA}$, $\mathbf{h}^{Boost,QDA}$	$T = 50$	MATLAB
SVM [24, 49]	$\mathbf{h}^{SVM}$	Ω, λ	LIBSVM
	$\mathbf{h}^{Bag,SVM}$, $\mathbf{h}^{Boost,SVM}$	$\Omega, \lambda, T = 50$	LIBSVM [49], MATLAB
Naïve Bayes [50]	$\mathbf{h}^{Bay}$	-	MATLAB
	$\mathbf{h}^{Bag,Bay}$, $\mathbf{h}^{Boost,Bay}$	$T = 50$	MATLAB
Decision Trees [26]	$\mathbf{h}^{DT}$	-	C4.5
	$\mathbf{h}^{Bag,DT}$, $\mathbf{h}^{Boost,DT}$	$T = 50$	MATLAB TreeBagger, PBTs [51]

SVM parameters include Ω (trade-off between training error and model complexity) and λ (normalization factor for inputs), which are determined via a grid search strategy. For ensemble approaches, T refers to the number of component classifiers

(in this study ω_{+1}) has significantly fewer samples compared to the other class (ω_{-1}), i.e. $|\omega_{+1}| \ll |\omega_{-1}|$. Weiss et al. [40] and Doyle et al. [41] previously showed that using an imbalanced training set will likely result in a lower classifier accuracy compared to balanced training sets ($|\omega_{+1}| = |\omega_{-1}|$). The class balance problem was addressed for each of the base classifiers, as well as their ensemble variants. Note that class balancing and data sub-sampling was only applied to the training data in each case.

(a) *QDA, DTs*: For classifiers corresponding to these two families $\left(\mathbf{h}^{QDA}, \mathbf{h}^{Bag,QDA}, \mathbf{h}^{Boost,QDA}, \mathbf{h}^{DT}, \mathbf{h}^{Bag,DT}, \mathbf{h}^{Boost,DT}\right)$, class imbalance was accounted for by randomized under-sampling of the majority class (ω_{-1}) such that $|\omega_{+1}| = |\omega_{-1}|$, i.e. an equal class balance was maintained when training the classifier.
(b) *SVMs*: Due to the complex nature of this algorithm, not only did a class balance have to be ensured in the training data, but the number of samples (voxels) used to train the classifier had to be reduced to ensure convergence within a reasonable amount of time. When training an SVM classifier, an equal number of voxels (not less than $0.7 \times |\omega_{+1}|$) were randomly sub-sampled from both ω_{+1} and ω_{-1} classes to form the training dataset. The number of samples was empirically decided based on a trade-off between execution time, classifier accuracy, and memory constraints specific to the SVM classifier. This procedure was adopted for all classifiers in the SVM family $\left(\mathbf{h}^{SVM}, \mathbf{h}^{Bag,SVM}, \mathbf{h}^{Boost,SVM}\right)$.
(c) *Naïve Bayes*: Training of the naïve Bayes classifier was implemented by directly estimating distributions for each of the classes, ω_{+1} and ω_{-1}, based on all the samples present. Such an estimate is most accurate when the maximal number of samples is utilized in calculating the distribution. Thus, no sub-sampling of the data was performed when constructing these classifiers $\left(\mathbf{h}^{Bay}, \mathbf{h}^{Bag,Bay}, \mathbf{h}^{Boost,Bay}\right)$.

Classifier training

All classifiers were trained and evaluated via 2 approaches:

(a) *Three Fold Cross Validation (3FCV)*: In a single cross-validation run for 3FCV, all the datasets from cohort D_1 were divided into 3 random subsets (comprising 6, 5, and 5 studies). 2 subsets were considered for training while the third was held out for independent testing. This process was repeated until all 3 subsets were classified at least once within each cycle. Each cycle was repeated 25 times. Within each cross-validation cycle, the classification results were cumulatively evaluated over all testing results to obtain a single AUC value, in addition to estimating lower and upper bounds on the AUC.
(b) *Multi-Site Validation (MSV)*: The entire discovery cohort D_1 was utilized to train a classifier model. This trained model was then evaluated for detecting PCa on a voxel-wise basis within the PZ for each dataset from validation cohorts D_2 and D_3. Classification results were evaluated to obtain a per-dataset AUC value, which were then averaged to obtain a per-cohort AUC value (and standard deviation).

Note that feature selection and classifier construction were done separately for each set of training data so constructed, with corresponding testing data only used for evaluation of classifier performance. All classifications were performed and evaluated on a per-voxel basis.

Evaluation of voxel-wise PCa classifiers

Classifier accuracy

In the case of $\mathbf{h}^{QDA}(c)$, $\mathbf{h}^{SVM}(c)$, $\mathbf{h}^{Bay}(c)$, $\mathbf{h}^{Bag,\beta}(c)$, $\mathbf{h}^{Boost,\beta}(c)$, $\beta \in \{QDA, Bay, SVM, DT\}$, which yield a probabilistic result (SVM hard decisions were also converted to a probabilistic result [42]), a binary prediction result at every $c \in C$ can be obtained by thresholding the associated probability value $\mathbf{h}(c) \in [0, 1]$. These classifier can be evaluated via Receiver Operating Characteristic

(ROC) curves [25], representing the trade-off between classification sensitivity and specificity of voxel-wise PCa detection. In the case of $\mathbf{h}^{DT}(c)$, the output is a single hard partitioning of the sample $c \in C$ into one of the two classes under consideration. In this case, a single detection result is calculated at a single threshold, based on which a single value for specificity and sensitivity can be calculated. It is assumed that the remaining points on the ROC curve for $\mathbf{h}^{DT}(c)$ are at $[0,0]$ and $[1,1]$, hence allowing the construction of a pseudo-ROC curve.

ROC curves were visualized for the training cohort D_1 by fitting a smooth polynomial through each set of sensitivity and specificity values calculated for each of the 3FCV runs, and averaging over all the curves generated for each classifier considered. The area under the ROC curve (AUC) was used as a measure of classification performance for both 3FCV and MSV, as is commonly reported in the literature [20, 22, 43].

While analyzing ROC results, the 12 classifiers were segregated into 3 groups, (1) single classification strategies (comprising $\mathbf{h}^{QDA}$, $\mathbf{h}^{Bay}$, $\mathbf{h}^{SVM}$, $\mathbf{h}^{DT}$), (2) bagging strategies (comprising $\mathbf{h}^{Bag,QDA}$, $\mathbf{h}^{Bag,Bay}$, $\mathbf{h}^{Bag,SVM}$, $\mathbf{h}^{Bag,DT}$), and (3) boosting strategies (comprising $\mathbf{h}^{Boost,QDA}$, $\mathbf{h}^{Boost,Bay}$, $\mathbf{h}^{Boost,SVM}$, $\mathbf{h}^{Boost,DT}$). Classifier comparisons were first made within each group (e.g. which of the single classification strategies $\mathbf{h}^{QDA}$, $\mathbf{h}^{Bay}$, $\mathbf{h}^{SVM}$, and $\mathbf{h}^{DT}$ performed best), following which classifier performance across groups was examined. These trends were first examined for the 3FCV results, followed by examining them separately for MSV results. When evaluating MSV results, comparisons were also made to determine how well classifier performance and trends generalized across the 3 sites.

Statistical testing

For the 3FCV procedure, each classifier yielded a set of 25 AUC values (corresponding to each cycle of the procedure). For the MSV procedure, each classifier yielded 13 and 56 AUC values (corresponding to the number of datasets in cohorts D_2 and D_3).

Multiple comparison testing to determine statistically significant differences in performance within groups (e.g. between all of $\mathbf{h}^{QDA}$, $\mathbf{h}^{Bay}$, $\mathbf{h}^{SVM}$, $\mathbf{h}^{DT}$) was performed using the Kruskal—Wallis (K-W) one-way analysis of variance (ANOVA) [44]. The K-W ANOVA is a non-parametric alternative to the standard ANOVA test which does not assume normality of the distributions when testing. The null hypothesis for the K-W ANOVA was that the populations from which the AUC values originate have the same median. Based off the results of a K-W ANOVA, multiple comparison testing was performed to determine which groups (single classification strategies, bagging strategies, boosting strategies) show significant differences in performance. Similar multiple comparison testing was also performed to determine significant differences in classifier performance between sites.

Pairwise comparisons were performed for classifiers across groups (e.g. between $\mathbf{h}^{QDA}$ and $\mathbf{h}^{Bag,QDA}$) to identify statistically significant differences in performance. This was done using the non-parametric Wilcoxon rank-sum test [44]. The null hypothesis in such a case was that there were no statistically significant differences in AUC values between the 2 classifiers being compared.

The Bonferroni correction [44] was applied to correct the p-value within all statistical comparisons considered (whether pairwise or other).

Computation time

For each of the classifiers compared, $\mathbf{h}^{\beta}, \mathbf{h}^{Bag,\beta}, \mathbf{h}^{Boost,\beta}$, $\beta \in \{QDA, Bay, SVM, DT\}$, the total amount of time required during each 3FCV cycle of D_1 for (i) classifier construction, and (ii) for executing the constructed classifier on testing data, was recorded in seconds. The execution time for each classifier was averaged over all cross-validation runs. All algorithms were implemented and evaluated using built-in or publicly available implementations for MATLAB®9.10 (The Mathworks, MA).

Results

Classification accuracy

Figure 3 shows average ROC curves (over all 3FCV runs) for voxel-wise PCa classification performance in the PZ, using the training cohort D_1. Figure 4 depicts bar plots of AUC values (with standard deviations as error bars) comparing voxel-wise PCa classification performance in the training cohort D_1 (purple bars) against the 2 independent validation cohorts D_2 (yellow bars) and D_3 (orange bars). Note that both single and ensemble SVM classifiers only reached convergence in D_1 but not in D_2 and D_3, and thus had to be omitted. This was likely due to the large number of training samples (voxels), which in conjunction with the grid search optimization of SVMs, caused these methods to error out before completion.

Comparing single classifier strategies

Figure 3a shows that all three of $\mathbf{h}^{QDA}$ (red), $\mathbf{h}^{Bay}$ (blue), and $\mathbf{h}^{SVM}$ (green) performed comparably for PCa classification in the PZ in D_1, with no statistically significant differences in their AUC values. $\mathbf{h}^{DT}$ (black) demonstrated significantly poorer performance than $\mathbf{h}^{QDA}$, $\mathbf{h}^{Bay}$, and $\mathbf{h}^{SVM}$ in a multiple comparison test of 3FCV AUC values (based off K-W ANOVA) in D_1.

As shown in Fig. 4a, this trend continued to hold in cohorts D_2 and D_3, where $\mathbf{h}^{DT}$ again performed signifi-

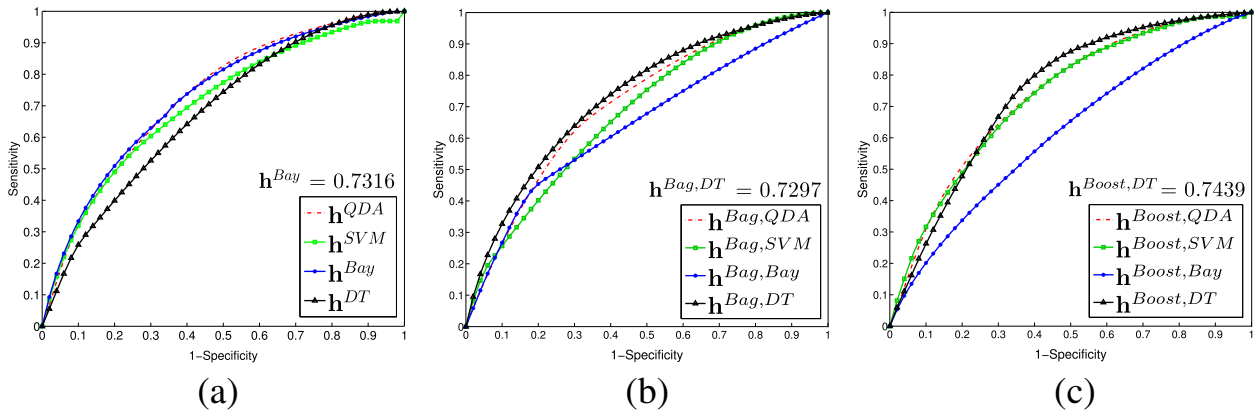

Fig. 3 (**a**)-(**c**) ROC curves obtained by averaging over 25 runs of 3FCV for PCa classification in the PZ for discovery cohort D_1. In each graph different colors correspond to different classifier strategies: (**a**) $\mathbf{h}^{QDA}$ (red), $\mathbf{h}^{SVM}$ (green), $\mathbf{h}^{Bay}$ (blue), $\mathbf{h}^{DT}$ (black); (**b**) $\mathbf{h}^{Bag,QDA}$ (red), $\mathbf{h}^{Bag,SVM}$ (green), $\mathbf{h}^{Bag,Bay}$ (blue), $\mathbf{h}^{Bag,DT}$ (black), and (**c**) $\mathbf{h}^{Boost,QDA}$ (red), $\mathbf{h}^{Boost,SVM}$ (green), $\mathbf{h}^{Boost,Bay}$ (blue), $\mathbf{h}^{Boost,DT}$ (black)

cantly worse than $\mathbf{h}^{QDA}$ and $\mathbf{h}^{Bay}$. In fact, $\mathbf{h}^{DT}$ performed significantly worse in D_2 and D_3 compared to D_1, with an $\approx$ 8% drop in AUC in independent validation. As DTs often performed on par with guessing (AUC range 0.54-0.58), they may be extremely suboptimal as a single classifier, and may be best used within ensembles (as typically formulated [10]). When comparing classifier performance between sites, all the classifiers performed worse in D_2 than in D_3 (though this was not always statistically significant).

Comparing bagged classifier strategies

Figure 3b demonstrates that $\mathbf{h}^{Bag,DT}$ yielded a significantly improved PCa classification performance compared to all of $\mathbf{h}^{Bag,QDA}$, $\mathbf{h}^{Bag,SVM}$, and $\mathbf{h}^{Bag,Bay}$, in D_1. Additionally $\mathbf{h}^{Bag,SVM}$ was the worst performing bagged classifier, with significantly lower AUC values compared to all of $\mathbf{h}^{Bag,QDA}$, $\mathbf{h}^{Bag,Bay}$, and $\mathbf{h}^{Bag,DT}$.

Figure 4b depicts the change in these trends for independent evaluation of the bagged classifiers. In D_2, $\mathbf{h}^{Bag,QDA}$, $\mathbf{h}^{Bag,Bay}$, and $\mathbf{h}^{Bag,DT}$ demonstrated no significant differences in performance. By contrast, $\mathbf{h}^{Bag,Bay}$ performed significantly better in D_3 compared to either of $\mathbf{h}^{Bag,QDA}$ and $\mathbf{h}^{Bag,DT}$. Notably, the performance of $\mathbf{h}^{Bag,DT}$ was significantly better in the training cohort D_1 than in the validation cohorts D_2 and D_3 (corresponding to $6-10$% higher AUC in 3FCV evaluation). Conversely, while performance of $\mathbf{h}^{Bag,QDA}$ and $\mathbf{h}^{Bag,Bay}$ on D_1 were reflective of their performance on D_2 (i.e. no significant differences), both these classifiers showed a significant improvement in D_3.

A significant improvement in performance for $\mathbf{h}^{Bag,SVM}$ and $\mathbf{h}^{Bag,DT}$ was seen on D_1, compared to using $\mathbf{h}^{SVM}$ or $\mathbf{h}^{DT}$ (Wilcoxon test $p < 0.01$). This observation held when examining corresponding results on D_2 and D_3, where $\mathbf{h}^{Bag,DT}$ performed significantly better than

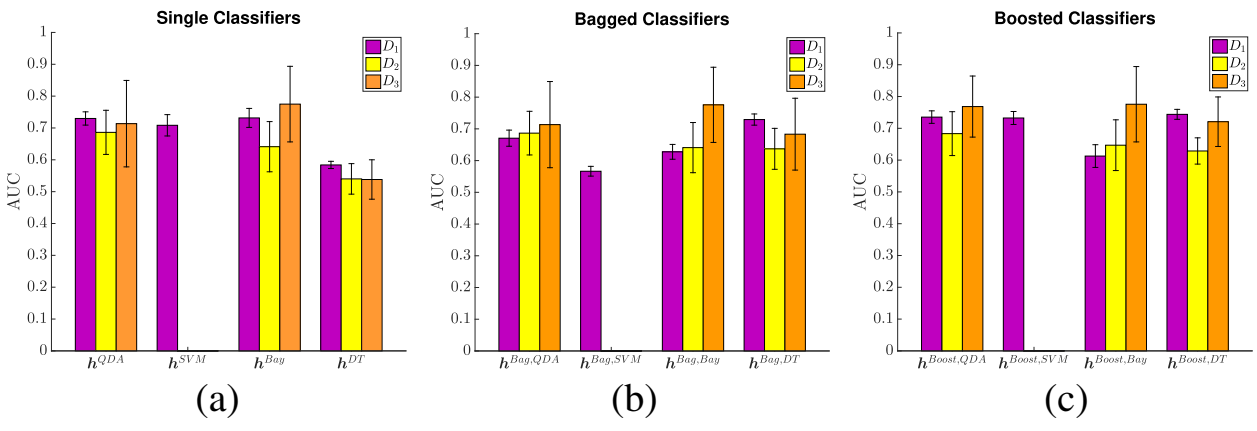

Fig. 4 Bar plots comparing voxel-wise PCa classifier performance between D_1 (purple), D_2 (yellow), and D_3 (orange) for (**a**) single classifiers, (**b**) bagged classifiers, and (**c**) boosted classifiers. As all variants of the SVM classifier only converged in the training cohort D_1; corresponding bars for D_2 and D_3 have been omitted. Error bars correspond to the standard deviation in AUC across each cohort

$\mathbf{h}^{DT}$ (Wilcoxon test $p << 0.01$). This result can be explained by the fact that SVMs and DTs are known to have high variance [45], which would imply they are most likely perform better in conjunction with bagging.

Comparing boosted classifier strategies

All 3 of $\mathbf{h}^{Boost,SVM}$, $\mathbf{h}^{Boost,DT}$, and $\mathbf{h}^{Boost,QDA}$ showed no significant differences in performance in 3FCV evaluation in D_1 (Fig. 3c), though $\mathbf{h}^{Boost,Bay}$ performed significantly worse by comparison. As seen in Fig. 4c, these trends did not hold in D_2 and D_3, where $\mathbf{h}^{Boost,Bay}$ performed on par with the other boosted classifiers. When comparing classifier performance between sites, $\mathbf{h}^{Boost,QDA}$ and $\mathbf{h}^{Boost,DT}$ performed worse in D_2 than in D_3 (though this was not always statistically significant).

$\mathbf{h}^{Boost,QDA}$, $\mathbf{h}^{Boost,DT}$, and $\mathbf{h}^{Boost,SVM}$ yielded a marginal but significantly improved performance compared to $\mathbf{h}^{Bag,QDA}$, $\mathbf{h}^{Bag,DT}$, and $\mathbf{h}^{Bag,SVM}$ on D_1. However this trend did not hold in D_2 and D_3, where the bagged and boosted variants of the different classifiers did not perform significantly differently from each other.

$\mathbf{h}^{Boost,DT}$ did yield a significantly improved performance compared to $\mathbf{h}^{DT}$ in all 3 cohorts (as did $\mathbf{h}^{Bag,DT}$). While optimal performance of bagging is highly dependent on the component classifiers exhibiting high variance [8], boosting is dependent on the component classifiers having low bias [45]. Thus classifiers which enjoy both these advantages (such as DTs) thus show significant performance improvements when combined within an ensemble.

Classifier execution time

Figure 5 depicts the computation times for training and evaluating the different classifiers considered in this study for all 3 folds of a single 3FCV run (averaged over all 25 3FCV runs). All computations were performed on 32 or 72 GB RAM quad core 64-bit Intel cluster computers. Note that the X-axis of Fig. 5 has been log-scaled for ease of display.

$\mathbf{h}^{QDA}$ required the least amount of computation time, followed by $\mathbf{h}^{DT}$. $\mathbf{h}^{SVM}$ required the most time for training and evaluating the classifier amongst all algorithms; training and testing times were even longer for $\mathbf{h}^{Bag,SVM}$ and $\mathbf{h}^{Boost,SVM}$. Thia is likely because of the additional grid search required to estimate the SVM parameters Ω and λ, which significantly increased the amount of time required. Note that SVM classifiers also required more careful memory management for voxel-wise classification, compared to the other methods.

Bagging was seen to typically increase computation time by a factor of 5, while boosting increased the computation time by a factor of 20. This may be because with bagging, the component classifiers are trained on smaller bootstrapped sample subsets, whereas with boosting, the component classifiers are trained on the entire set of training samples.

Discussion

Our primary findings from this work were the following,

- We identified the most consistently performing method across all 3 sites as the boosted QDA classifier. It had a relatively high AUC (= 0.735) in

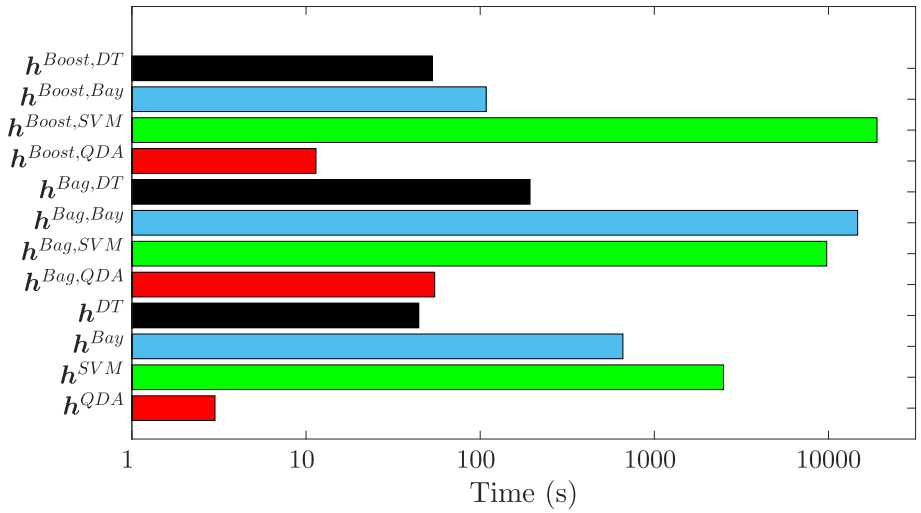

Fig. 5 Runtimes for different classifier strategies, averaged over all 25 3FCV runs on training cohort D_1. Colors correspond to different classifier families: QDA (red), SVM (green), naïve Bayes (blue), DTs (black). Note X-axis is log-scaled for display purposes

the training cohort as well as in both validation cohorts (average AUCs of 0.683 and 0.768, respectively). Coupled with its relatively quick execution time (2nd lowest among all methods), we believe this would make it the best classifier overall. Our second choice would be the single QDA classifier, which did not perform significantly worse (average AUCs of 0.730, 0.686, 0.713 for each of the sites) than the boosted QDA classifier.

- The performance of all variants of the decision tree classifier (single, bagged, boosted) were overestimated by $\approx$ 10%, when compared between the training and validation cohorts. In fact, the top-performing classifier identified in the training cohort was the boosted decision tree classifier (AUC=0.744), but this classifier performed more variably when evaluated on multi-site data. This clearly indicates the need for independent validation when building CAD models, as otherwise these perhaps less generalizable models would have been identified as the top performer.
- The popular SVM classifier achieved reasonable classification performance in the training cohort alone (similar to previous SVM-based PCa detection schemes for prostate T2w MRI [22]). However, they took the longest to train and test, and did not achieve convergence in multi-site validation. This may be a consideration to take into account as prostate CAD schemes start to undergo larger scale multi-site validation.
- We could not reach a clear conclusion regarding which of boosting and bagging yielded better performance across the classifier strategies. There were no significant differences in their performance in multi-site validation.
- Satisfying the conditions of bias and variance were extremely crucial when constructing classifier ensembles. While SVMs and DTs show significant improvements within both bagging and boosting frameworks, Bayesian and QDA classifiers provided a more mixed performance as they suffered from low variance and/or high bias. However, not all of these trends generalized in multi-site validation.
- For all the classifiers considered, performance in the 2 validation cohorts D_2 and D_3 fell within the confidence bounds of their performance in the discovery cohort D_1. Thus, despite heterogeneous acquisition and imaging characteristics across the 3 sites, our post-processing steps (correcting for bias field and non-standardness) appear to have enabled some degree of harmonization in terms of radiomic features and associated classifier models. Appropriate post-processing of multi-site imaging data may therefore be critical when evaluating radiomic classifiers in this fashion.
- In terms of the site-specific performance trends, it is interesting to note that all classifiers performed worse in D_2 than in D_3. While all 3 sites used a 3 T magnet, D_2 had a lower voxel resolution than D_1 and D_3 (which were similar to each other). This seems to indicate that voxel resolution may have a marked effect on classifier performance. This result has also been observed in previous phantom studies of texture analysis in medical imaging [46].

In the context of the specific problem of voxel-wise PZ PCa detection via T2w MRI, we achieved classification accuracies comparable to the literature: optimal AUCs between 0.683-0.768 across 3 different sites. To our knowledge, Rampun et al. [43] have performed the only other such classifier comparison study, where they evaluated 11 different classifiers for voxel-wise prostate cancer detection in the PZ while using 45 patient studies from a single site. While they reported a Bayesian Network classifier as their top performer (AUC $= 90.0 \pm 7.6$), in our experiments both single and ensemble Bayesian classifiers performed well in either the discovery or the testing cohorts (but not both). Dinh et al. [47] utilized 106 patients imaged on scanners from 2 different manufacturers to identify a robust set of statistics from multi-parametric MRI for prostate cancer diagnosis. As opposed to the voxel-wise detection problem examined by us, they developed a linear mixed model to identify which expert-delineated lesions within the PZ had a Gleason score of at least 7 (i.e. a region-wise classification) which achieved per-site AUCs of 0.85 and 0.90. In a more limited study of 18 patients imaged on 2 different scanners [48], a cross-scanner MR intensity normalization technique was presented for detecting which pixels within expert-annotated regions in the PZ corresponded to cancerous or benign. In the current study, we corrected for cross-site intensity drift via histogram standardization [29]; likely due to which all the classifiers performed relatively consistently across the 3 sites. Most recently, a multi-institutional study of radiomic features across 3 different sites (80 patients) was able to identify a PZ tumor-specific set of features for voxel-wise detection of prostate cancer regions [32]. While multi-parametric MRI data was utilized, radiomic features from T2w MRI were most often identified as discriminatory; and they reported comparable AUCs to our own (between 0.61-0.71).

We do acknowledge a few limitations of our study. Despite being one of the first multi-site studies examining voxel-wise PCa detection, our cohort size is still somewhat limited (85 studies across 3 sites). However, this is among the larger cohort sizes when compared to a majority of PCa detection studies in the literature [22] (median cohort size $\approx$ 30 studies). The central gland, PZ, and tumor regions were manually annotated on MRI by expert

radiologists. While these annotations were done based on corresponding excised pathology images and reports, there may be some error in terms of how precise the delineations are. Notably, such expert annotations have been popularly used in the PCa detection literature [22], perhaps in acknowledgement of how difficult it is to get precise "ground truth" for this problem. However, for two of our cohorts, the expert annotations were only used for independent evaluation of the classifier, potentially lessening the impact of annotation error. With the availability of more comprehensive "ground truth" pathologic information, an avenue of future work could be to identify which radiomic classifiers best generalize for characterizing Gleason grade or benign confounders (e.g. prostatitis) on multi-site MRI data. We also limited ourselves to the use of T2w MRI as opposed to a multi-parametric (MP) MRI exam. The reason for the choice of T2w MRI alone was dictated by non-availability of all MP-MRI protocols across 3 different sites. Additionally, we opted to utilize a specific set of radiomic descriptors of T2w MRI [33] for the construction of PZ-specific PCa classifiers. Additional features may also be employed in this regard, and could be an avenue for future work. We also limited ourselves to empirically comparing 12 classifier strategies. However, based on the 4 distinct types of classifiers considered in this study, we believe our results may be generalized to other classifier families (e.g. relevance vector machines are similar to SVMs).

Conclusions

In this work, we empirically compared and evaluated 12 different radiomic classifier ensembles derived from 4 classifier families (QDA, Bayesian learners, Decision Trees, and Support Vector Machines), in terms of accuracy and computation time, for voxel-wise detection of prostate cancer in 85 high resolution T2w MRI patient datasets curated from across 3 different sites. A secondary motivation of this study was to investigate whether classifier trends on data curated from a single site generalize to data acquired from different sites and scanners. Our results suggest that simpler classifiers (such as QDA and its ensemble variants) may be more robust, accurate, and efficient for prostate cancer CAD problems, especially in the context of multi-site validation. A more detailed understanding of radiomic feature and classifier trends in large multi-site settings may be crucial for clinical usage of radiomics-based PCa detection on MRI.

Supporting information: Appendix A

List of selected radiomic features derived from T2w MRI- Below, we enumerate the result of feature selection, using data from the training cohort D_1. Parameters incude: (a) for 1st and 2nd order statistics, window size ($WS \in \{3,5,7\}$); (b) for Gabor wavelets, orientation $\left(\theta \in \left\{0, \frac{\pi}{8}, \frac{2\pi}{8}, \frac{3\pi}{8}, \frac{4\pi}{8}, \frac{5\pi}{8}, \frac{6\pi}{8}, \frac{7\pi}{8}, \pi\right\}\right)$ and wavelength ($\lambda \in \{2.83, 5.66, 8.20, 11.31, 22.63, 45.25\}$). Haar wavelets did not have an associated parameter, and were also not selected in the top 25 radiomic features.

1 2nd Order Statistic Difference Entropy ($WS = 7$)
2 Gabor $\theta = 1.57, \lambda = 45.25$
3 1st Order Statistic Graylevel Median ($WS = 3$)
4 Gabor $\theta = 0, \lambda = 22.63$
5 Gabor $\theta = 2.75, \lambda = 11.31$
6 Gabor $\theta = 1.18, \lambda = 45.25$
7 2nd Order Statistical Energy (WS=7)
8 Gabor $\theta = 2.75, \lambda = 22.63$
9 Gabor $\theta = 2.36, \lambda = 2.83$
10 Gabor $\theta = 0, \lambda = 2.83$
11 2nd Order Statistic Sum Average ($WS = 3$)
12 Gabor $\theta = 0.3927, \lambda = 11.31$
13 Gabor $\theta = 1.96, \lambda = 45.25$
14 2nd Order Statistic Inverse Difference Moment (WS=7)
15 Gabor $\theta = 2.36, \lambda = 5.66$
16 1st Order Statistic Mean ($WS = 3$)
17 1st Order Statistic Range ($WS = 3$)
18 Gabor $\theta = 1.18, \lambda = 22.63$
19 2nd Order Statistic Sum Entropy ($WS = 7$)
20 Gabor $\theta = 2.75, \lambda = 5.66$
21 Gabor $\theta = 1.96, \lambda = 8.20$
22 Gabor $\theta = 0, \lambda = 5.66$
23 2nd Order Statistic Difference Entropy ($WS = 3$)
24 Gabor $\theta = 2.35, \lambda = 8.20$
25 Gabor $\theta = 2.75, \lambda = 2.83$

Abbreviations

3FCV: Three fold cross validation; ANOVA: One way analysis of variance; AUC: Area under the ROC curve; CAD: Computer aided detection; DA: Discriminant analysis; DT: Decision tree; IRB: Institutional review board; K-W: Kruskal-wallis; MAD: Mean absolute deviation; MP: Multi-parametric; MRI: Magnetic resonance imaging; MSV: Multi-site validation; PCa: Prostate cancer; PZ: Peripheral zone; ROC: Receiver-operating characteristic (curve); SVM: Support vector machine; T2w: T2-weighted

Funding

Research reported in this publication was supported by the NIH/NCI under award numbers 1U24CA199374-01, R01CA202752-01A1, R01CA208236-01A1, R01 CA216579-01A1, R01 CA220581-01A1, the National Center for Research Resources under award number 1 C06 RR12463-01,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 Cleveland Digestive Diseases Research Core Center, the NIH/NIDDK 1P30DK097948 DDRCC Pilot/Feasibility Award Program, the NIH/NIBIB CWRU Interdisciplinary Biomedical Imaging Training Program under award number 5T32EB00750912, 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.
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was also supported by the Office of the Assistant Secretary of Defense for Health Affairs, through different Peer Reviewed Research Programs. The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office for these

Programs. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.

Availability of data and materials
All multi-site radiomic features used in the experiments conducted in this manuscript have been made publicly available at https://doi.org/10.5061/dryad.026cj63.

Authors' contributions
SEV and AM conceived the paper and experimental design. SEV, PVC, MCY performed experimental analyses. NMR, ASP, MAR, BNB provided MR imaging data, expert annotations, and clinical input. All authors contributed to writing and editing, and have approved the final manuscript.

Ethics approval and consent to participate
The patient MRI datasets used in this study were retrospectively acquired from independent patient cohorts curated from different sites, where the data was initially acquired under written informed consent at each collecting institution. All 3 datasets comprised de-identified MRI data together with expert annotations of tumor extent, which were provided to the authors through the IRB protocol # 02-13-42C approved by the University Hospitals of Cleveland Institutional Review Board. Data analysis was waived review and consent by the IRB board, as all data was being analyzed retrospectively, after de-identification.

Consent for publication
Not applicable (see Ethics Statement).

Competing interests
AM is an equity holder in Elucid Bioimaging and Inspirata Inc. He is also a scientific advisory consultant for Inspirata Inc. In addition he currently serves as a scientific advisory board member for Inspirata Inc, Astrazeneca, and Merck. He also has sponsored research agreements with Philips and Inspirata Inc. His technology has been licensed to Elucid Bioimaging and Inspirata Inc. He is also involved in a NIH U24 grant with PathCore Inc and 3 different R01 grants with Inspirata Inc. SV is a scientific advisory board member and equity holder in Virbio, Inc and a member of the Editorial Board of BMC Medical Imaging.

Publisher's Note

Author details
[1]Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH, USA. [2]College of Medicine, Northeast Ohio Medical University, Rootstown, OH, USA. [3]Department of Radiology, UT Southwestern Medical Center, Dallas, TX, USA. [4]Department of Radiology, Cleveland Clinic, Cleveland, OH, USA. [5]Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, PA, USA. [6]Department of Radiology, Boston University School of Medicine, Boston, MA, USA.

Received: 11 July 2018 Accepted: 10 January 2019
Published online: 28 February 2019

References
1. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinform. 2003;19(13):1636–43.
2. Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano J, Armananzas R, Santafe G, Perez A, Robles V. Machine learning in bioinformatics. Brief Bioinform. 2006;7(1):86–112.
3. Juntu J, Sijbers J, De Backer S, Rajan J, Van Dyck D. Machine learning study of several classifiers trained with texture analysis features to differentiate benign from malignant soft-tissue tumors in T1-MRI images. J Magn Reson Imaging. 2010;31(3):680–9.
4. Wei L, Yang Y, Nishikawa RM, Jiang Y. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. IEEE Trans Med Imaging. 2005;24(3):371–80.
5. Herlidou-Meme S, Constans JM, Carsin B, Olivie D, Eliat PA, Nadal-Desbarats L, Gondry C, Le Rumeur E, Idy-Peretti I, de Certaines JD. MRI texture analysis on texture test objects, normal brain and intracranial tumors. Magn Reson Imaging. 2003;21(9):989–93.
6. Monaco JP, Tomaszewski JE, Feldman MD, Hagemann I, Moradi M, Mousavi P, Boag A, Davidson C, Abolmaesumi P, Madabhushi A. High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models. Med Image Anal. 2010;14(4):617–29.
7. Prasanna P, Tiwari P, Madabhushi A. Co-occurrence of local anisotropic gradient orientations (collage): a new radiomics descriptor. Scientific Reports. 2016;6:37241.
8. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123–40.
9. Freund Y, Schapire R. Experiments with a New Boosting Algorithm. In: Proc Int'l Conf Mach Learn. San Francisco: Morgan Kaufmann Publishers Inc.; 1996. p. 148–56.
10. Dietterich T. Ensemble Methods in Machine Learning. In: Proc 1st Intl Workshop Mult Class Systems. Berlin: Springer Berlin Heidelberg; 2000. p. 1–15.
11. Lim T-S, Loh W-Y, Shih Y-S. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. Mach Learn. 2000;40(3):203–28.
12. Bauer E, Kohavi R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Mach Learn. 1999;36(1):105–39.
13. Tran QL, Toh KA, Srinivasan D, Wong KL, Shaun Qiu-Cen L. An empirical comparison of nine pattern classifiers. IEEE Trans Sys Man Cybernet. 2005;35(5):1079–91.
14. Dietterich TG. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Mach Learn. 2000;40(2):139–57.
15. Hamza M, Larocque D. An empirical comparison of ensemble methods based on classification trees. JSCS. 2005;75(8):629–43.
16. Opitz D, Maclin R. Popular ensemble methods: An empirical study. JAIR. 1999;11(1):169–98.
17. Frank A, Asuncion A. UCI Machine Learning Repository. 2010. http://archive.ics.uci.edu/ml.
18. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15:3133–81.
19. Schmah T, Yourganov G, Zemel RS, Hinton GE, Small SL, Strother SC. Comparing classification methods for longitudinal fMRI studies. Neural Comput. 2010;22(11):2729–62.
20. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. Sci Rep. 2015;5:13087. https://doi.org/10.1038/srep13087.
21. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJWL. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. Front Oncol. 2015;5:272.
22. Lemaître G, Martí R, Freixenet J, Vilanova JC, Walker PM, Meriaudeau F. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: A review. Comput Biol Med. 2015;60:8–31.
23. McLachlan GJ. Discriminant Analysis and Statistical Pattern Recognition. Hoboken, N.J.: Wiley-Interscience; 2004.
24. Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw. 1999;10(5):988–99.
25. Duda RO, Hart PE, Stork DG. Pattern Classification, 2nd edn. New York: Wiley; 2001.
26. Quinlan J. C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc.; 1993.
27. Chappelow J, Bloch BN, Rofsky N, Genega E, Lenkinski R, DeWolf W, Madabhushi A. Elastic registration of multimodal prostate mri and histology via multiattribute combined mutual information. Med Phys. 2011;38(4):2005–18.
28. Schiebler ML, Schnall MD, Pollack HM, Lenkinski RE, Tomaszewski JE, Wein AJ, Whittington R, Rauschning W, Kressel HY. Current role of MR imaging in the staging of adenocarcinoma of the prostate. Radiology. 1993;189(2):339–52.
29. Nyúl LG, Udupa JK, Zhang X. New variants of a method of mri scale standardization. IEEE Trans Med Imaging. 2000;19(2):143–50.
30. Madabhushi A, Feldman MD, Metaxas DN, Tomaszeweski J, Chute D. Automated detection of prostatic adenocarcinoma from high-resolution ex vivo MRI. IEEE Trans Med Imaging. 2005;24(12):1611–25.

31. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging. 1998;17(1):87–97.
32. Ginsburg SB, Algohary A, Pahwa S, Gulani V, Ponsky L, Aronen HJ, Boström PJ, Böhm M, Haynes A.-M., Brenner P, 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. 2017;46(1):184–93.
33. Viswanath SE, Bloch NB, Chappelow JC, Toth R, Rofsky NM, Genega EM, Lenkinski RE, Madabhushi A. 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. 2012;36(1):213–24.
34. Bovik AC, Clark M, Geisler WS. Multichannel texture analysis using localized spatial filters. IEEE Trans Pattern Anal Mach Intell. 1990;12(1): 55–73.
35. Busch C. Wavelet based texture segmentation of multi-modal tomographic images. Comput Graph. 1997;21(3):347–58.
36. Russ JC. The Image Processing Handbook, 5th edn. Boca Raton: CRC Press Inc.; 2007.
37. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Sys Man Cybernet. 1973;3(6):610–21.
38. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38.
39. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
40. Weiss G, Provost F. The Effect of Class Distribution on Classifier Learning: An Empirical Study. 2001. Technical report, Technical Report Technical Report ML-TR-44, Department of Computer Science, Rutgers University. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.9570.
41. Doyle S, Monaco J, Tomaszewski J, Feldman M, Madabhushi A. An Active Learning Based Classification Strategy for the Minority Class Problem: Application to Histopathology Annotation. BMC Bioinform. 2011;12(1):424. https//doi/10.1186/1471-2105-12-424.
42. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classifiers. 1999;10(3):61–74.
43. Rampun A, Zheng L, Malcolm P, Tiddeman B, Zwiggelaar R. Computer-aided detection of prostate cancer in t2-weighted mri within the peripheral zone. Phys Med Biol. 2016;61(13):4796–825.
44. Demsar J. Statistical comparisons of classifiers over multiple data sets. JMLR. 2006;7:1–30.
45. Breiman L. Arcing classifiers. Ann Stat. 1998;26(3):801–24.
46. Waugh SA, Lerski RA, Bidaut L, Thompson AM. The influence of field strength and different clinical breast mri protocols on the outcome of texture analysis using foam phantoms. Med Phys. 2011;38(9):5058–66.
47. Hoang Dinh A, Melodelima C, Souchon R, Lehaire J, Bratan F, Mège-Lechevallier F, Ruffion A, Crouzet S, Colombel M, Rouvière O. Quantitative analysis of prostate multiparametric mr images for detection of aggressive prostate cancer in the peripheral zone: a multiple imager study. Radiology. 2016;280(1):117–27.
48. Artan Y, Oto A, Yetik IS. Cross-device automated prostate cancer localization with multiparametric mri. IEEE Trans Image Process. 2013;22(12):5385–94.
49. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2011;2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
50. Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, Showe MK. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. Bioinform. 2006;22(11):1325–34.
51. Tu Z. Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. In: Proc Tenth IEEE ICCV. Washington, DC: IEEE Computer Society; 2005. p. 1589–96.