Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images

Duc Fehr[a,1], Harini Veeraraghavan[a,1,2], Andreas Wibmer[b], Tatsuo Gondo[c], Kazuhiro Matsumoto[c], Herbert Alberto Vargas[b], Evis Sala[b], Hedvig Hricak[b], and Joseph O. Deasy[a]

[a]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY; [b]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY; and [c]Department of Urology, Memorial Sloan Kettering Cancer Center, New York, NY

Edited by Grace Wahba, University of Wisconsin–Madison, Madison, WI, and approved September 22, 2015 (received for review March 26, 2015)

Noninvasive, radiological image-based detection and stratification of Gleason patterns can impact clinical outcomes, treatment selection, and the determination of disease status at diagnosis without subjecting patients to surgical biopsies. We present machine learning-based automatic classification of prostate cancer aggressiveness by combining apparent diffusion coefficient (ADC) and T2-weighted (T2-w) MRI-based texture features. Our approach achieved reasonably accurate classification of Gleason scores (GS) 6(3 + 3) vs. ≥7 and 7(3 + 4) vs. 7(4 + 3) despite the presence of highly unbalanced samples by using two different sample augmentation techniques followed by feature selection-based classification. Our method distinguished between GS 6(3 + 3) and ≥7 cancers with 93% accuracy for cancers occurring in both peripheral (PZ) and transition (TZ) zones and 92% for cancers occurring in the PZ alone. Our approach distinguished the GS 7(3 + 4) from GS 7(4 + 3) with 92% accuracy for cancers occurring in both the PZ and TZ and with 93% for cancers occurring in the PZ alone. In comparison, a classifier using only the ADC mean achieved a top accuracy of 58% for distinguishing GS 6(3 + 3) vs. GS ≥7 for cancers occurring in PZ and TZ and 63% for cancers occurring in PZ alone. The same classifier achieved an accuracy of 59% for distinguishing GS 7(3 + 4) from GS 7(4 + 3) occurring in the PZ and TZ and 60% for cancers occurring in PZ alone. Separate analysis of the cancers occurring in TZ alone was not performed owing to the limited number of samples. Our results suggest that texture features derived from ADC and T2-w MRI together with sample augmentation can help to obtain reasonably accurate classification of Gleason patterns.

Gleason score classification | learning from unbalanced data | multiparametric MRI | PCa Gleason 6 vs. ≥7 | PCa Gleason (3+4) vs. (4+3) cancers

Prostate cancer (PCa) is among the most common cancers and a leading cause of cancer-related death in men in the United States (1). In general, patients diagnosed with PCa with a Gleason score (GS) (≤6) have better 5- and 10-y survival rates, lower biochemical recurrence rates, and lower prostate cancer-specific mortality than do patients with GS ≥7 (2). Similarly, compared with patients with GS 7(4 + 3), those with GS 7(3 + 4) have better outcomes (2). The GS and prostate specific antigen (PSA) level are clinically used to determine PCa aggressiveness (3). GS is a well-validated factor and known to be a powerful predictor of disease progression, mortality, and outcomes (4, 5). However, owing to random sampling, the GS determined through biopsies is known to differ from those determined following radical prostatectomy (6, 7), as well as between immediate repeat biopsies (8). Therefore, the ability to automatically detect the GS with high accuracy from the diagnostic MRIs would have a significant impact on clinical decision making, treatment selection, and prediction of outcomes for patients and spare them from painful biopsies and their accompanying risk of complications. Noninvasive and accurate techniques that determine the aggressiveness of PCa are needed to enhance the quality of patient care.

Previously, MRI has been investigated (9) as a modality for determining PCa aggressiveness. Although MRI has been shown to be a valuable tool for PCa detection (10–13), there is no clear consensus on the specific imaging biomarker that is most effective in distinguishing the aggressiveness of PCa lesions. In addition to MR spectroscopic and T2-weighted (T2-w) MR imaging, the apparent diffusion coefficient (ADC) from diffusion-weighted MRI has been confirmed to be valuable for differentiating PCa aggressiveness (14–17). However, studies differ in the specific ADC value used to distinguish between the cancers. The features used have included ADC mean computed from a single slice region of interest (ROI) (15, 16, 18), ADC mean computed from the entire volume using different sets of diffusion b-values (all vs. fast vs. slow) (19), 10th percentile of the ADC computed from the entire lesion (20), 10th percentile and ADC mean (21), and ADC mean computed over the entire lesion (22). Furthermore, none of the aforementioned studies used more than five imaging features for the analysis.

Texture-based imaging features in conjunction with machine learning-based classification have predominantly been applied for classifying malignant from noncancerous prostate tissues (23–25) with one exception (26). Linear discriminant analysis (LDA)-based classification of various histogram-based ADC measures, namely,

Significance

Gleason scores and ultimately the aggressiveness of prostate cancers determined using transrectal ultrasound (TRUS)-guided biopsy procedures could result in incorrect diagnosis in addition to patient discomfort. The Gleason scores determined from TRUS-guided biopsies often differ from immediate repeat biopsies and the biopsies determined following whole excision of the prostate. Our approach presents a highly accurate and automated method for differentiating between the high ≥7 and low Gleason score 6(3+3), as well as between 7(3+4) and 7(4+3) Gleason score cancers through multiparametric MRI combined with texture features computed on the same images. Noninvasive and accurate techniques such as ours can benefit patient care without subjecting them to unnecessary interventions.

Author contributions: D.F., H.V., E.S., H.H., and J.O.D. designed research; D.F., H.V., and J.O.D. performed research; D.F. and H.V. contributed new reagents/analytic tools; D.F., H.V., A.W., T.G., K.M., H.A.V., and J.O.D. analyzed data; D.F., H.V., and J.O.D. wrote the paper; H.V. advised D.F. in the research and development of the analysis methods; A.W. segmented and labeled the tumors in the MR T2 and apparent diffusion coefficient (ADC) images; T.G. classified and helped delineate tumors in the MR images using the pathology images; K.M. helped delineate and classify tumors in MR by correlating with the pathology images; H.A.V. advised and mentored A.W.; H.H. provided the data for the research and provided financial support for A.W.; and J.O.D. provided financial support to D.F. and guidance during research to D.F. and H.V.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]D.F. and H.V. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: veerarah@mskcc.org.

www.pnas.org/cgi/doi/10.1073/pnas.1505935112

ADC mean, 10th percentile, T2-w histogram-based skewness, and k-trans were used to distinguish between cancer vs. benign and between cancer GSs (21). Our work builds on the aforementioned work by classifying GS $7(3+4)$ vs. $7(4+3)$ in addition to classifying cancer vs. noncancerous prostate and GS $6(3+3)$ vs. GS ≥ 7 with texture-based features derived from ADC and T2-w MR images. Furthermore, our work addresses an important problem of obtaining highly accurate machine learning despite severe class imbalance between the different groups of cancers by using sample augmentation with feature selection.

Our work demonstrates that PCa diagnosis can be improved by combining data-augmented classification together with more of the latent information in standard MRIs (the so-called "radiomics hypothesis") (27, 28) compared with using ADC mean or T2 signal intensities alone, thereby reducing the potential for under- or overdiagnosis. Fig. 1 *A* and *B* show the ADC energy, ADC entropy, T2 energy, and T2 entropy overlaid on a slice of the ADC and corresponding T2-w MR image for two different patients: one with a tumor of GS $6(3+3)$ and the other with a tumor of GS $9(4+5)$. As shown in Fig. 1 *A* and *B*, the energy and entropy values computed from different tumor types appear to be very different, which suggests that textures, in combination with ADC, can help to differentiate between the cancer types.

Materials and Methods

The retrospective study used for the analysis in this work was approved by the Institutional Review Board, which waived written informed consent. The study population used in this study was the same as the one used in our previous work (29).

Study Population. The study population consisted of T2-w and ADC MR images acquired from 217 men subjected to MR imaging with the following inclusion criteria: (*i*) patients with biopsy-proved PCa, (*ii*) radical prostatectomy performed in our institution between January and December 2011, (*iii*) endorectal 3T prostate MRI performed within 6 mo of prostatectomy, and (*iv*) with whole-mount step-section pathological tumor maps. Patients with prior treatment for prostate cancer ($n = 7$), those with cancers <0.5 mL on histopathology ($n = 51$), those with imaging artifacts making segmentation of cancer difficult ($n = 8$), and those whose cancer location precluded segmentation of normal structures ($n = 7$) were excluded from study. The final number of male patients in the study population was 147. More details about patient selection are provided in ref. 29.

MR Image Acquisition and Histopathological Image Analysis. All MR images were acquired on a 3.0-T MR imaging system (Signa HDX; GE Medical Systems), with a pelvic phased-array coil in combination with an endorectal coil (Medrad) for improved signal reception. Transverse T1-w images were acquired by using the following parameters: repetition time (milliseconds)/echo time (milliseconds), 467–1,349/6.6–10.2; section thickness, 5 mm; intersection gap, 1 mm; field of view, 22–40 cm; and matrix, 256 × 192–448 × 224. Transverse, coronal, and sagittal T2-w fast spin-echo images were acquired with the following parameters: 2,500–7,700/83.3–143.5; section thickness, 3–4 mm; intersection gap, 0–1 mm; field of view, 14–24 cm; and matrix, 288 × 288–448 × 224. Diffusion-weighted sequences were performed in the transverse plane by using a single-shot spin-echo echo-planar imaging sequence with two b values (0 and 1,000 s/mm^2) (3,500–5,675/70.3–105.6; section thickness, 3–4 mm; no intersection gap; field of view, 14–24 cm; matrix, 96 × 96–128 × 128) and with the same orientation and location used to acquire transverse T2-w images. The ADC maps were computed from Advanced Workstation (GE Medical Systems). The excised prostates, following the amputation of seminal vesicles, were serially sectioned from apex to base at 3- to 5-mm intervals and submitted as whole-mount sections for histopathologic examination. The Gleason grade patterns in each lesion were determined, and the corresponding lesion borders were outlined on each slide. More details of MR image acquisition and the histopathological analysis are provided in ref. 29.

Image Segmentation. Tumors and normal structures were identified and volumetrically segmented on both the T2-w and ADC MR images by three readers in consensus: one genitourinary imaging research fellow (A.W.), one clinical urology research fellow (T.G.), and one pathology research fellow (K.M.), using 3DSlicer (30) as described in ref. 29. PCa foci ≥0.5 mL were first identified from the pathology whole-mount step-section tumor images. Given the similar slice thickness of the step-section (3–5 mm) and the MR images (5 mm), visual coregistration was used to find the corresponding slices on the T2-w and ADC MR images. Furthermore, anatomical landmarks including urethra, ejaculatory ducts, prostatic capsule, and well-delineated hyperplastic nodules were used to pinpoint the appropriate tumor. The draw tool available in the Editor module of the 3DSlicer was used to delineate the tumors in multiple slices. In addition to tumors, a noncancerous prostate region was delineated in both the peripheral zone (PZ) and the transition zone (TZ) of each patient and marked. To avoid any errors from automatic registration, the tumors and normal structures were drawn on both T2-w and ADC images.

Texture Features. First- and second-order texture features were computed from the T2-w and ADC MR images following preprocessing and intensity rescaling (0–255). The first-order features consisted of the moments of the intensity volume histogram (mean, SD, skewness, and kurtosis) computed from the structure ROI. The second-order features, namely the Haralick features (31), were computed using the gray level co-occurrence matrix (GLCM) with 128 bins and consisted of energy, entropy, correlation, homogeneity, and contrast. The first-order features were computed from an in-house software implemented in Matlab (32) and the Haralick features from an in-house software implemented in C++ using the Insight Toolkit (ITK) (33).

Sample Augmentation Through Oversampling. Class imbalance can adversely impact the performance of a classifier wherein all of the samples are classified as the majority class, thereby obtaining fairly good classification accuracy, albeit with low specificity or sensitivity. Oversampling (34) and sample weighting (35) are two solutions to address this problem. Our work builds on ref. 34 and used two different sampling approaches: (*i*) sample generation from joint weighting of multiparametric features using synthetic minority

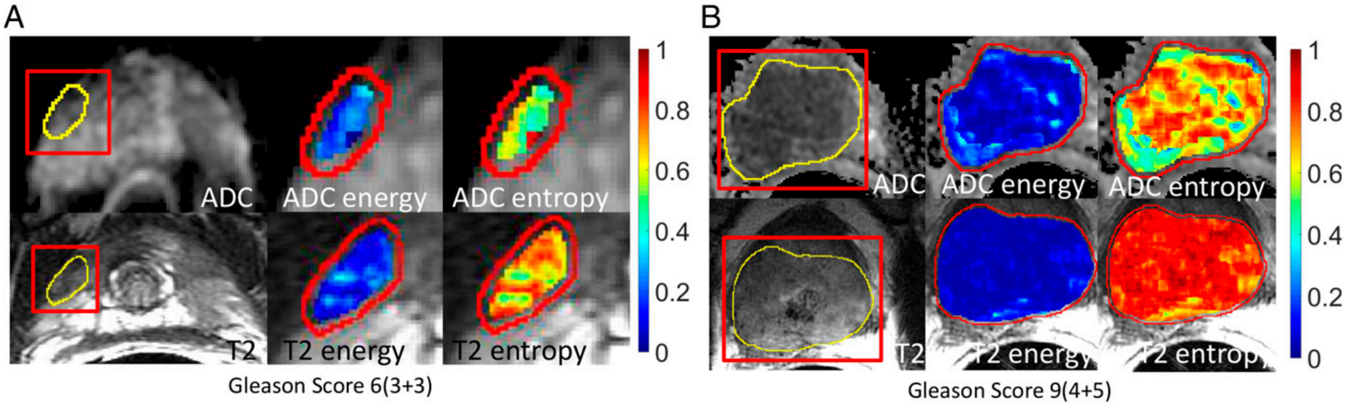

Fig. 1. Example of (*A*) a GS $6(3+3)$ tumor and (*B*) a GS $9(4+5)$ tumor. The top row shows the ADC image with the computed energy and entropy values overlaid on the tumor. The bottom row shows the T2-w MR image with the computed energy and entropy values overlaid on the same tumor on the corresponding slice. The texture features were computed per voxel by using a 5×5×5 patch centered at each voxel.

oversampling technique (SMOTE) (34) and (*ii*) sample generation through conditionally independent features using Gibbs sampling (36). The former technique requires all features to be scaled to the same level, whereas Gibbs sampling avoids this requirement. To prevent influence of outliers from shrinking the decision boundary between two classes, we used outlier rejection using $K = 5$ nearest neighbors. Gibbs sampling has not previously been explored in the context of sample augmentation for improving machine learning robustness.

Feature Selection and Classification. We evaluated the efficacy of three different methods for classifying the PCa GSs. The methods incorporated different levels of interaction between feature selection and classifier training. The methods were as follows:

- *t* Test Support Vector Machine (*t* test SVM), wherein the set of features were selected before classifier training through a two-sided, unpaired *t* test (37). Only those features that were significantly different at 95% confidence level ($P \leq 0.05$) were selected.
- Adaptive Boosting (AdaBoost), wherein the feature selection was integrated with the training, although the features were selected linearly (one feature at a time). AdaBoost (38) extracts a "strong" classifier from linear combinations of "weak" classifiers such that the strong classifier has a higher accuracy than any of the individual classifiers.
- Recursive Feature Selection Support Vector Machine (RFE-SVM), wherein the feature selection was integrated with the classifier training and the interaction of features was modeled. By allowing feature interactions, RFE-SVM (39, 40) explicitly models the correlations between features, thereby leading to robust classifier performance. Feature selection involved backward elimination where in each iteration, the feature that had the least impact on improving the performance of the classifier was removed. The algorithm continued until the desired number of features *r* was reached; *r* can be chosen empirically or through a separate cross-validation model selection as used in this work. The RFE-SVM is essentially a method for ranking the relative importance of a set of features such that the top-few features, i.e., those that remain through the longest number of iterations are chosen for training the SVM. A particular feature is deemed irrelevant, when the margin of two SVM classifiers, one trained with and the other trained without the same feature is the smallest compared with the margin differences of all of the features remaining in the given iteration.

Cross-Validation. Each method was evaluated independently through stratified ($K = 10$)-fold cross-validation. The goal of cross-validation was not to select among the different methods. The hyperparameters for each classifier were selected separately through *K*-fold cross-validation–based model selection. The SVM classifiers used radial basis function (RBF) kernels and the hyperparameters consisted of the RBF kernel's width parameter $\gamma = \{0.01, 0.02, \ldots, 0.3\}$, and the misclassification penalty $C = \{0.1, 1,5,10\}$. The RFE-SVM included a third parameter, namely, the number of features *r*. The model parameters were selected to be those that resulted in the best overall accuracy over all of the *K* folds. Following model selection, the individual classifiers were evaluated separately using repeated stratified ($K = 10$) cross-validation with 100 trials. Stratified cross-validation ensures that each fold of the classifier has equal proportion of data from each class. Repeated cross-validation helps to estimate any error due to particular partitioning of the data. To avoid overoptimistic cross-validation results owing to hyperparameter selection bias, the final cross-validation accuracies were reduced by a bias computed from the cross-validation model selection as used in the Tibshirani and Tibshirani method (41). The Tibshirani bias is the mean of the difference between the error on all of the folds using the parameters that minimize overall error and the parameters that result in minimum error in each fold. AdaBoost did not use any hyperparameter model selection, and hence, bias correction was unnecessary.

Results

Experimental Description. We analyzed the efficacy of classifying GS $6(3+3)$ ($n = 34$) vs. GS ≥ 7 ($n = 159$) cancers and GS $7(3+4)$ ($n = 114$) vs. $7(4+3)$ ($n = 26$) cancers using (*i*) *t* test SVM, (*ii*) RFE-SVM, and (*iii*) AdaBoost trained without and with sample augmentation using (*i*) Gibbs oversampling and (*ii*) synthetic minority oversampling technique (SMOTE) oversampling. To compare our results with those of previous works (23–25), we also applied the same methods for distinguishing between noncancerous structures and prostate cancers. The number of samples for noncancerous structures ($n = 158$) and cancer ($n = 198$, the GS for five tumors were not provided) were balanced, and hence, did not require sample augmentation. Finally, we compared the performance of texture feature-based classifiers with SVM classifiers trained with (*i*) ADC mean and (*ii*) ADC mean and T2 mean computed from inside the segmented tumor volumes.

All classifiers were trained with identical samples. In the sample augmentation experiments, the samples in the minority and majority class were oversampled to 200 samples in each class. We also experimented with different ratios of samples (minority class sampled to exact number of majority samples, 400 samples in the minority and majority class). We present only the results for 200 samples as the classifier performance increased with the increasing number of samples.

Cancer vs. Noncancerous Tissue Classification. We analyzed the classification performance of the three methods for classifying noncancerous prostate from cancers that occurred in both the PZ and TZ. SVM trained using ADC mean achieved an accuracy of 0.84 (or 84%) with a sensitivity of 0.87 and specificity of 0.84, resulting in a Youden index of 0.71. T2 mean added to SVM with ADC mean achieved a similar accuracy of 0.81 with a Youden index of 0.64. Youden index, also referred to as Youden (J)-statistic (42), summarizes the ROC curve and is an indicator for the performance of a classifier. It is expressed by combining the specificity (*sp*) and sensitivity (*se*) as $sp + se - 1$.

In comparison, the classifiers trained using 18 different texture features with feature selection achieved the following accuracies: *t* test SVM, 0.95 accuracy and Youden index of 0.91; RFE-SVM, 0.96 accuracy and Youden index of 0.91; AdaBoost, 0.95 accuracy and Youden index of 0.89. ADC mean was selected as a significant feature by the *t* test in addition to ADC entropy,

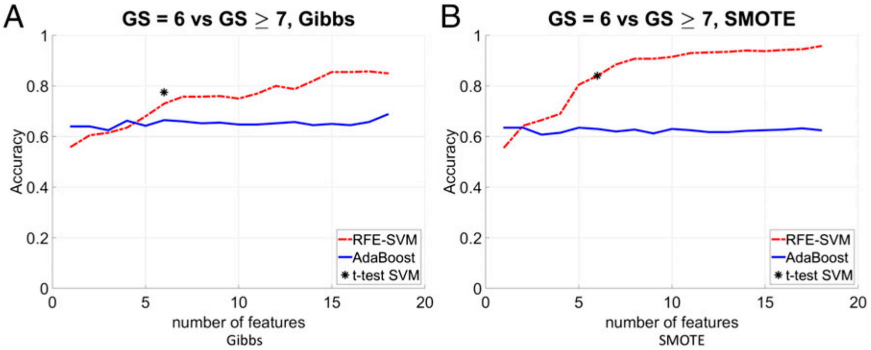

Fig. 2. Classification accuracies for the *t* test SVM, RFE-SVM, and AdaBoost for separating lesions by their Gleason score, GS $6(3+3)$ vs. GS ≥ 7 in the PZ and TZ using Gibbs oversampling with 200 samples in each class (*A*) and SMOTE oversampling with 200 samples in each class (*B*).

Table 1. Accuracy results for GS 6(3 + 3) vs. GS ≥7 classification for tumors with and without oversampling

	PZ and TZ sites			PZ site only		
	34/159 samples	200/200 samples	200/200 samples	23/120 samples	200/200 samples	200/200 samples
Method	not augmented	Gibbs	SMOTE	not augmented	Gibbs	SMOTE
t test SVM	0.83 (0.06)	0.73 (0.53)	0.82 (0.65)	0.86 (0.24)	0.70 (0.46)	0.79 (0.62)
RFE-SVM	0.83 (0.03)	**0.83** (0.71)	**0.93** (0.91)	0.84 (0.00)	**0.89** (0.83)	**0.92** (0.89)
AdaBoost	0.73 (0.11)	0.69 (0.38)	0.64 (0.28)	0.79 (0.34)	0.72 (0.44)	0.72 (0.44)
ADC-SVM	0.82 (0.00)	0.57 (0.22)	0.58 (0.22)	0.84 (0.00)	0.63 (0.27)	0.59 (0.23)
ADC&T2-SVM	0.82 (0.00)	0.57 (0.22)	0.64 (0.38)	0.84 (0.00)	0.63 (0.29)	0.62 (0.29)

The numbers in parentheses correspond to the Youden Index. The bold fonts indicate the best classification accuracies for the augmented cases.

ADC homogeneity, ADC SD, and ADC energy. ADC mean was also selected among the relevant features chosen by the RFE-SVM, although it was ranked as a low impact feature. The AdaBoost method selected ADC entropy, ADC mean, ADC energy, T2 kurtosis, and ADC contrast as the top five features. Although ADC mean was selected as a relevant feature, the addition of texture features improved the classification performance of all of the classifiers compared with the SVM using ADC mean alone.

GS 6(3 + 3) vs. GS ≥7 Classification. We analyzed the performance of the classifiers for (*i*) cancers occurring in both the PZ and TZ and (*ii*) cancers occurring in the PZ alone. The number of cancers in the TZ was too small for analysis. In both the aforesaid scenarios, the number of cancers that occurred in the GS 6(3 + 3) class was much smaller than the number of cancers in the GS ≥7 category. We analyzed the performance of the various classifiers trained without any sample augmentation and with samples augmented using Gibbs and SMOTE methods. In general, the classification accuracy, sensitivity, and specificity of the classifiers regardless of the cancer location were much lower without sample augmentation in comparison with classifiers trained using samples generated by either the Gibbs or SMOTE methods.

Table 1 shows the accuracies of the classifiers applied to GS classification using textures (*t* test SVM, RFE-SVM, AdaBoost), SVM trained using ADC mean, and SVM trained using ADC mean and T2 mean. Results are also shown when the classifiers were trained without and with augmented samples obtained from Gibbs and SMOTE methods. The Youden index is shown next to the accuracy. As shown in Table 1, when the classifiers were trained without any sample augmentation, there was no real difference in the accuracies between the classifiers using textures vs. the SVM classifiers using ADC mean or ADC mean and T2 mean. Furthermore, even though the accuracies were seemingly high, ranging from 0.73 for AdaBoost to 0.83 for *t* test SVM and RFE-SVM (for cancers occurring in the PZ and TZ), the Youden index was close to 0 for all of the classifiers. A similar result was seen for cancers occurring in the PZ alone, although the achieved accuracies were higher ranging from 0.79 for the AdaBoost to 0.86 for the *t* test SVM and with all of the classifiers having a low Youden index, the highest being for the AdaBoost, namely, 0.34. The high accuracy but low Youden index suggests that the classifiers learned a biased model, wherein the majority of the data were classified as the majority class.

On the other hand, when the classifiers were trained using augmented samples obtained either from the Gibbs or SMOTE methods, the performance of the classifiers, particularly, the RFE-SVM, improved drastically. As shown in Table 1, the classification accuracy of the RFE-SVM for cancers occurring in the PZ and TZ was 0.83 when trained using samples generated using Gibbs sampling and 0.93 when trained using samples generated by using SMOTE sampling. The Youden index of the RFE-SVM improved from 0.03 without sample augmentation to 0.71 with Gibbs sampling and 0.91 with SMOTE sampling. A similar result was seen for cancers occurring in the PZ only using the RFE-SVM classifier. Additionally, the difference in performance when using the texture features in comparison with ADC mean or ADC mean and T2 mean alone was apparent when the classifiers were trained with augmented samples. As seen, the performance of the classifiers trained without textures (ADC mean and ADC mean and T2 mean) was much worse than when the classifiers were trained using the texture features. The texture-based AdaBoost method resulted in a low accuracy of 0.64 for the PZ and TZ and 0.72 for cancers in the PZ using SMOTE sampling. The corresponding accuracy of the SVM trained with ADC mean and the same sampling technique was 0.58 for the PZ and TZ cancers and 0.59 for PZ only cancers. Whereas no difference in the accuracy was observed with the addition of T2 mean for Gibbs' sampling, a slight improvement in

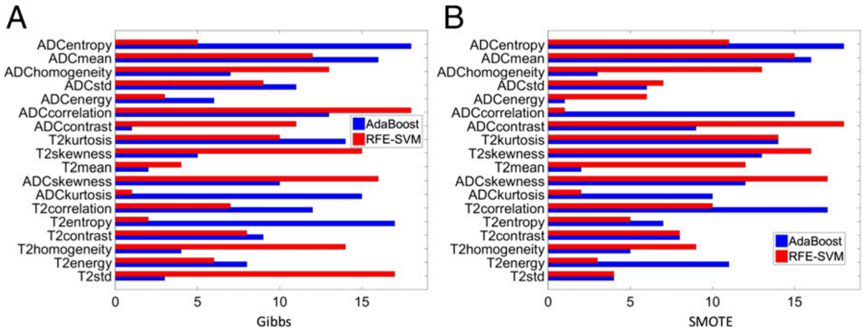

Fig. 3. Features chosen by the AdaBoost and the RFE approach for the GS 6(3 + 3) vs. GS ≥ 7 classification in the PZ and TZ. The blue bars represent the relative feature importance computed by AdaBoost. The red bars represent the iteration at which the features were rejected by RFE. The results using Gibbs and SMOTE oversampling are given in *A* and *B* respectively.

Table 2. Accuracy results for GS 7(3 + 4) vs. GS 7(4 + 3) classification for tumors with and without oversampling

	PZ and TZ sites			PZ site only		
	114/26 samples	200/200 samples	200/200 samples	80/25 samples	200/200 samples	200/200 samples
Method	not augmented	Gibbs	SMOTE	not augmented	Gibbs	SMOTE
t test SVM	0.81 (0.00)	0.66 (0.39)	0.76 (0.55)	0.76 (0.00)	0.65 (0.32)	0.63 (0.32)
RFE-SVM	0.83 (0.11)	**0.86** (0.76)	**0.92** (0.88)	0.81 (0.23)	**0.85** (0.76)	**0.93** (0.91)
AdaBoost	0.79 (0.41)	0.76 (0.52)	0.73 (0.46)	0.76 (0.36)	0.74 (0.48)	0.75 (0.50)
ADC-SVM	0.81 (0.00)	0.59 (0.24)	0.56 (0.19)	0.76 (0.00)	0.60 (0.24)	0.60 (0.22)
ADC&T2-SVM	0.81 (0.00)	0.64 (0.32)	0.57 (0.24)	0.76 (0.00)	0.61 (0.26)	0.54 (0.17)

The numbers in parentheses correspond to the Youden Index. The bold fonts indicate the best classification accuracies for the augmented cases.

accuracy was noted when using SMOTE oversampling for cancers in the PZ and TZ.

Fig. 2 shows a comparison in the performance of the texture-based classifiers RFE-SVM, *t* test SVM, and AdaBoost trained using samples generated from Gibbs sampling in Fig. 2*A* and SMOTE sampling in Fig. 2*B*. The accuracies are shown with increasing number of features starting from the most important to the least important feature added to the feature set. *t* Test SVM used six features that were found to be significantly different between the two classes. As shown in Fig. 2, the accuracy of the RFE-SVM and *t* test SVM is better when trained using the SMOTE sampling method than when using the Gibbs sampling method. The accuracy of the AdaBoost was the same regardless of the sampling method and the number of features. On the other hand, the accuracy of the RFE-SVM ranged from 0.58 with 1 feature to 0.83 with 15 features, when trained using samples generated from Gibbs and from 0.56 with 1 to 0.93 with 18 features using the SMOTE method.

Fig. 3 shows the relative ranking of the various features as selected by the AdaBoost and RFE-SVM methods trained with samples generated from the Gibbs and SMOTE methods for separating GS 6(3 + 3) vs. GS ≥7 cancers occurring in the PZ and TZ. As shown in Fig. 3*A*, the ADC correlation was the highest ranked feature selected by the RFE-SVM method with Gibbs oversampling. The next four important features were T2 SD, ADC skewness, T2 skewness, and T2 homogeneity. ADC kurtosis and T2 entropy were among the least important features. Similarly, when using the SMOTE sampling method, RFE-SVM ranked ADC contrast, ADC skewness, T2 skewness, ADC mean, and T2 kurtosis among the most important. AdaBoost and RFE-SVM result in different ordering of the relative importance of features. For instance, with both the sampling methods, ADC entropy was the most relevant feature for the AdaBoost method, whereas it was not as important when using the RFE-SVM method. Incidentally, ADC entropy was among the features that were found to be significantly different between the GS 6(3 + 3) vs. ≥7 cancers using the *t* test. The relative ranking of features was obtained by applying the RFE-SVM with the selected hyperparameters on the entire dataset. Clearly, when applying the cross-validation–based model assessment, the relative ranking of features may vary from fold to fold.

Fig. 4 shows the ROC curves for RFE-SVM, *t* test SVM, and AdaBoost trained from samples generated using SMOTE for distinguishing GS 6(3 + 3) vs. GS ≥7 PZ and TZ cancers in Fig. 4*A* and PZ-only cancers in Fig. 4*B*. As shown in Fig. 4*A*, the RFE-SVM produced the best classification performance, and the AdaBoost resulted in the worst performance. The RFE-SVM and the *t* test SVM achieved an area under the curve (AUC) of 0.99 and 0.90, respectively, with SMOTE and 0.91 and 0.83 with Gibbs sampling for PZ and TZ cancers. AdaBoost achieved an AUC of 0.60 with SMOTE and 0.74 using Gibbs sampling. The AUC of the same classifiers for PZ only cancers were as follows: RFE-SVM, 0.99 (SMOTE), 0.97 (Gibbs); *t* test-SVM, 0.89 (SMOTE), 0.80 (Gibbs); and AdaBoost, 0.79 for SMOTE and Gibbs sampling. The bias for all classifiers ranged from 0.01 to 0.05, with a bias of 0.03 and 0.02 for RFE-SVM using Gibbs and SMOTE oversampling, respectively.

GS 7(3 + 4) vs. GS 7(4 + 3) Cancer Classification. Table 2 shows the classification accuracies together with the Youden index for the various classification methods for distinguishing GS 7(3 + 4) from GS 7(4 + 3) for cancers that occurred in (*i*) PZ and TZ and (*ii*) in PZ only when trained without and with samples augmented using the Gibbs and SMOTE methods. Similar to the GS 6(3 + 3) vs. GS ≥7 cancers, all classifiers achieved comparably poor performance without sample augmentation. However, all of the classification methods' performance improved following either sampling techniques. The RFE-SVM method achieved the best performance, regardless of the sampling technique and location of cancers (PZ and TZ or PZ only). The classifiers (*t* test SVM, RFE-SVM, AdaBoost) using texture features achieved better classification performance compared with SVM trained with ADC mean or SVM trained with ADC mean and T2 mean.

Fig. 5 shows a comparison in the performance of the texture-based classifiers RFE-SVM, *t* test SVM, and AdaBoost trained using samples generated from Gibbs sampling (Fig. 5*A*) and SMOTE sampling (Fig. 5*B*), for classifying cancers occurring in the PZ and TZ into GS 7(3 + 4) vs. GS 7(4 + 3). The accuracies are shown with an increasing number of features starting from the most important to the least important. Three features were selected as being significantly different by the *t* test. As shown in Fig. 5, the accuracy of the RFE-SVM and *t* test SVM were better when trained using the SMOTE sampling method than when using the Gibbs sampling method. The AdaBoost method was relatively less impacted by the features added to the classifier. On the other hand, the accuracy of the RFE-SVM varied from 0.60 with 1 feature to a maximum of 0.86 and 0.92 with 15 features when using Gibbs (Fig. 5*A*) and SMOTE oversampling (Fig. 5*B*).

Fig. 6 shows the relative importance of the various features when trained with the RFE-SVM and AdaBoost methods using the two different sample augmentation techniques Gibbs (Fig. 6*A*) and SMOTE (Fig. 6*B*). ADC contrast, T2 entropy, ADC entropy, T2 skewness, and ADC mean were among the top five ranked features when using Gibbs sampling with RFE-SVM, whereas SMOTE-based RFE-SVM selected T2 homogeneity, ADC contrast, T2 kurtosis, ADC energy, and T2 mean as the top five features. AdaBoost, on the other hand, selected ADC entropy as the third most relevant feature after ADC kurtosis and T2 SD for Gibbs oversampling and as the second most relevant feature after ADC homogeneity for SMOTE.

Fig. 7 shows the ROC curves for RFE-SVM, *t* test SVM, and AdaBoost trained from samples generated using SMOTE for classifying GS 7(3 + 4) vs. GS 7(4 + 3) cancers in the PZ and TZ (Fig. 7*A*) or the PZ alone (Fig. 7*B*). The AUC when trained

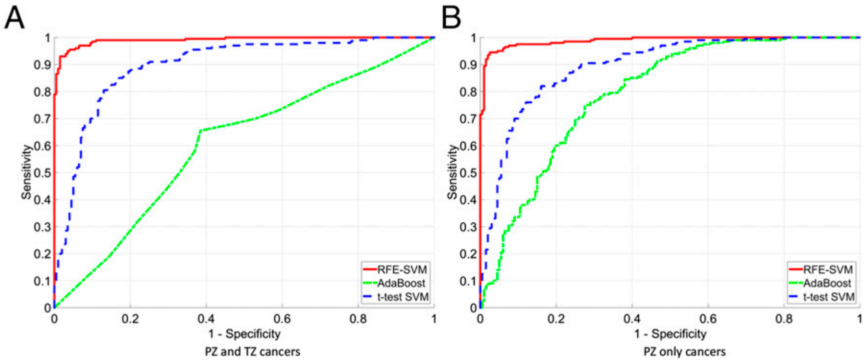

Fig. 4. ROC curves for RFE-SVM, *t* test, SVM, and AdaBoost performances when using SMOTE augmented samples (200 samples for each class) for GS 6(3 + 3) vs. GS ≥ 7 occurring in (*A*) PZ and TZ and (*B*) PZ only.

using SMOTE and Gibbs sampling were as follows: RFE-SVM, 0.99, 0.94 (PZ and TZ), 0.98, 0.95 (PZ only); *t* test SVM, 0.80, 0.75 (PZ and TZ), 0.72, 0.66 (PZ only); AdaBoost, 0.77, 0.81 (PZ and TZ), 0.82, 0.80 (PZ only), respectively. The bias for all classifiers ranged from 0.01 to 0.05 with a bias of 0.02 for RFE-SVM in all cases but Gibbs oversampling in the PZ only (0.03).

Discussion and Future Work

Whereas expert users can detect malignant cancers with very high accuracy, visually determining the cancers' aggressiveness from MR images is a challenging problem. Automatic classification techniques can simultaneously analyze a large number of imaging features that are beyond the scope of visual analysis by a clinician. Furthermore, algorithms that can achieve robust and consistent classification can be a very valuable tool that can aid clinicians in identifying appropriate treatment options for patients without subjecting them to unnecessary interventions.

Our work used three different methods: (*i*) *t* test SVM, (*ii*) AdaBoost, and (*iii*) RFE-SVM. We compared the performance of the aforementioned methods with SVMs trained using (*i*) ADC mean and (*ii*) ADC mean and T2 mean. The RFE-SVM achieved the best classification performance with the highest specificity and sensitivity for all of (*i*) cancer vs. noncancerous, (*ii*) GS 6(3 + 3) vs. GS ≥7, and (*iii*) GS 7(3 + 4) vs. GS 7(4 + 3) for both sample augmentation methods. RFE-SVM outperformed or achieved about the same accuracy and performance as the other methods even when not using any sample augmentation. The reason for the better performance of the RFE-SVM method in comparison with the other methods is that the RFE-SVM method incorporates the interaction of features when selecting features as opposed to *t* test, which treats the features as being independent and AdaBoost, which uses conditional independence of the features during feature selection. The poor performance of the AdaBoost method can be explained by overfitting (43), particularly in the presence of insufficient and unbalanced training data, as was the case in our dataset. Furthermore, as the tumors were segmented in MR images by matching their approximate location in histology, any errors resulting from correlating MR images obtained through an endorectal coil (which distorts the shape of the prostate) with the whole-mount step-section specimens would adversely affect AdaBoost performance. The poor performance of the Adaboost method results from the fact that the method focuses more on the problematic examples for classification in each iteration. The RFE-SVM and *t* test SVM were less impacted by such errors compared with the AdaBoost as all examples were treated equally.

Previously, in ref. 26, they used a hierarchical machine learning-based automatic voxel-wise classification of GS [6(3 + 3) and 7(3 + 4)] vs. GS [7(4 + 3) and ≥8] cancers by combining T2-w MR images with MR spectroscopy images. The number of samples used, namely, 29 patients in ref. 26, was much smaller than the 147 patients used in our work. Furthermore, our approach combined the ADC textures with the T2-w MR texture features for classifying both GS 6(3 + 3) from GS ≥7 and differentiating between the GS 7 cancers by their primary Gleason subtype, namely, (3 + 4) vs. (4 + 3). Our work also incorporated sample augmentation methods to tackle the problem of highly unbalanced data and achieved a much higher accuracy of 0.93 for cancers in the PZ and TZ and 0.92 for cancers occurring only in the PZ following reduction by the hyperparameter selection bias. The majority of the prior works have been focused on classifying cancerous regions from benign structures (24, 25, 43, 44). Our results for classifying noncancerous prostate from malignant cancers

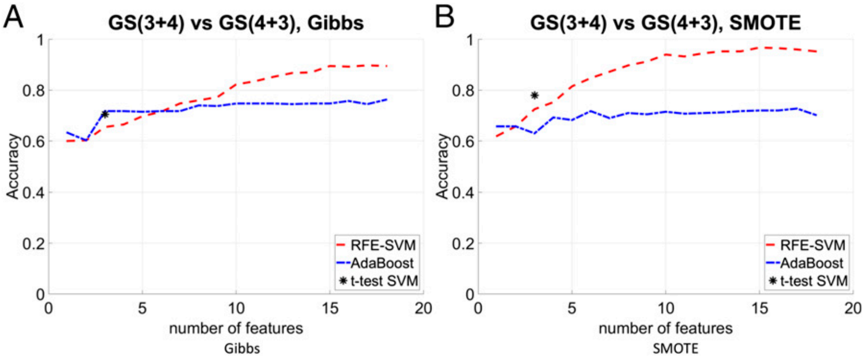

Fig. 5. Classification accuracies for the *t* test SVM, RFE-SVM, and AdaBoost for separating lesions by their Gleason score, GS 7(3 + 4) vs. GS 7(4 + 3) in the PZ and TZ using Gibbs oversampling with 200 samples in each class (*A*) and SMOTE oversampling with 200 samples in each class (*B*).

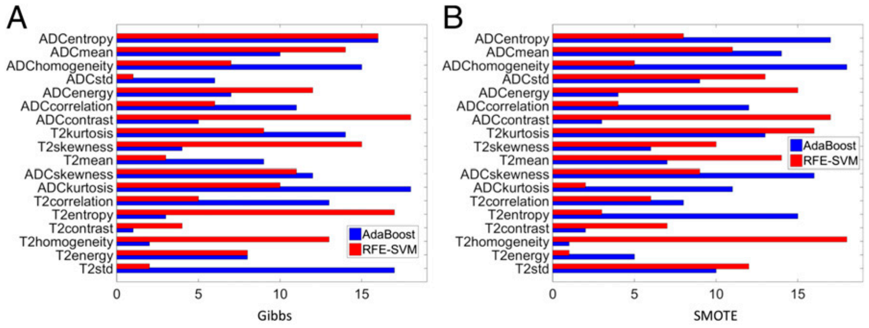

Fig. 6. Features chosen by the AdaBoost and the RFE approach for the GS 7(3 + 4) vs. GS 7(4 + 3) classification in the PZ and TZ. The blue bars represent the relative feature importance computed by AdaBoost. The red bars represent the iteration at which the features were rejected by RFE. The results using Gibbs and SMOTE oversampling are given in *A* and *B* respectively.

were similar or better. The accuracies of our methods were as follows: *t* test, SVM (0.95) compared with 0.89 using the same method in ref. 44 (with 42 cancer region of interests); the AdaBoost (0.95); and RFE-SVM (0.96) for cancers that occurred in both the PZ and TZ. In ref. 24, the classification accuracy of cancerous vs. noncancerous structures for PZ tumors was 0.73 with 22 patients.

ADC mean has been shown to be a viable biomarker for differentiating cancers by their aggressiveness with reasonable accuracy of 0.75 (22). Our work investigated the efficacy of ADC mean as a feature when used with automatic machine learning-based classification and compared its performance to three different classification methods that used texture features derived from ADC and T2-w MRI. Our results suggest that texture features drastically improved the classification performance of the automatic methods in comparison with using ADC mean alone for both GS 6(3 + 3) vs. GS ≥7 and GS 7(3 + 4) vs. GS 7(4 + 3) cancers. Classifiers trained using more features than just the ADC mean or T2 mean resulted in higher accuracies particularly with augmented samples. ADC mean was ranked among the top 10 features by the RFE-SVM method for classifying GS 6(3 + 3) vs. GS ≥7 cancers regardless of the sampling method. ADC mean was found as a significant feature by the *t* test. In the analysis involving the GS 7(3 + 4) vs. GS 7(4 + 3) cancers, the ADC mean was the 8th and 5th most important feature selected by the RFE-SVM method when trained using SMOTE and Gibbs sampling (for PZ and TZ cancers).

Previously, the authors of ref. 29 investigated the efficacy of Haralick texture features for differentiating between GS 6(3 + 3) and GS ≥7 cancers and found that the higher GS cancers were associated with relatively high ADC entropy and low ADC energy in comparison with low GS cancers that occurred in the PZ on 147 patients with ≥0.5 mL histology volume. Our work analyzed the dataset with the same patients. The main difference in the samples used in this work from ref. 29 is that more TZ cancer examples were available for analysis. ADC entropy and ADC energy were found to be significantly different between the GS 6(3 + 3) and GS ≥7 cancers when using a two-sided, unpaired *t* test and found to be the top feature by AdaBoost, which was consistent with the results from ref. 29 for separating low from high GS cancers. ADC entropy and ADC energy were not ranked among the top five features by RFE-SVM.

Classifier accuracy in previous efforts has been restrained due to the common machine learning obstacle of class imbalance: less aggressive samples are typically fewer in number than highly aggressive cancer samples, resulting in a bias in performance (43). Class imbalance is very common in medical applications of machine learning, compared with, say, financial risk modeling where available datasets are typically much larger. Our work addressed class imbalance through an innovative and general sample augmentation/feature selection method that increases classifier accuracy. Ignoring issues of class imbalance leads to poor classifiers with low generalization capability. Our results indicate that sample augmentation combined with machine learning and feature selection help to improve the classification performance of GS from MRI. All source code and anonymized texture features data used for analysis will be made publicly available on acceptance for publication. Owing to privacy concerns, the patient MR and histopathology images cannot be made publicly available.

However, all of the classifier performance results reported in this work are cross-validation results, which is less ideal compared with using a separate validation dataset. Over-optimistic results owing to selection bias, as has been previously reported in refs. 45 and 46, can be avoided through nested cross-validation and by reporting

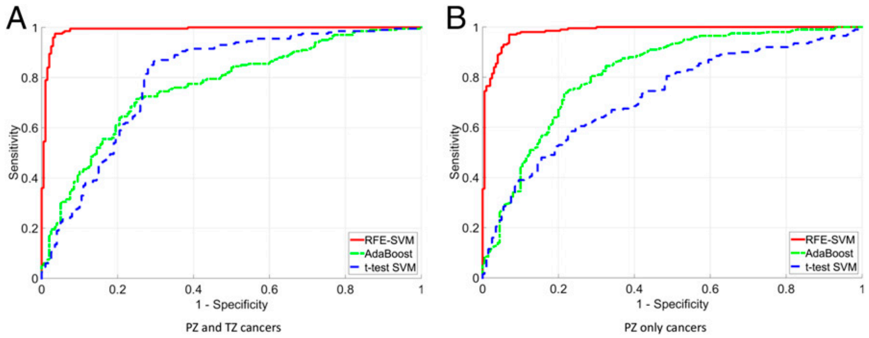

Fig. 7. ROC curves for RFE-SVM, *t* test SVM, and AdaBoost performances when using SMOTE augmented samples (200 samples for each class) for GS 7(3 + 4) vs. GS 7(4 + 3) occurring in (*A*) PZ and TZ and (*B*) PZ only.

model assessment accuracy (47). As has been shown in ref. 48, there is little difference between nested cross-validation and the Tibshirani (41) bias correction method. Given the already limited number of samples in this work, using an additional internal cross-validation would overly limit the data size, leading to a large variance in the classifier performance. Therefore, we chose the bias correction approach to report the repeated cross-validation results following cross-validation–based hyperparameter selection. Furthermore, we used $K = 10$-fold cross-validation as has been shown to be robust in ref. 49. We used multiple classifiers for assessing their performance on the dataset. However, obtaining statistically meaningful comparisons as in ref. 50 was difficult because there was only a single dataset. Besides, the goal of this work was not to compare different classifiers but instead to assess whether reasonable classification of cancer aggressiveness could be obtained from multiparametric MR images alone when having highly unbalanced training data.

Our work has several limitations. First, all of the analysis was performed on retrospective data, and no true validation dataset was available for evaluating the various methods. Second, our classifications were performed per tumor instead of a voxel-wise classification as in refs. 24, 26, and 43. The reason for this was that texture parameters, in particular the Haralick textures computed over the whole volume of the tumor, are more descriptive of the underlying heterogeneity of the tumors, given the larger number of voxels available to compute the texture values. In the future, we plan to explore multiscale texture extraction methods (using different neighborhoods around voxels) so that voxel-wise classification can be explored with the goal of generating GS maps of the tumor. Automated classification generated GS maps can potentially aid in guiding surgical biopsy procedures. Third, in this work, we explored only 18 different image-based features including first- and second-order textures. Prior works (26, 43, 44) used first- and second-order texture features and additionally used gradient (Sobel)-based and Gabor edge-based features at multiple scales resulting in more than 100 different features. We considered only the first- and second-order texture features primarily because they seemed to be the most relevant texture features for characterizing tumor heterogeneity, as shown in ref. 43. Furthermore, when using about the same number of features as the examples, it is not clear how well the learning method generalizes, besides making the feature selection more difficult. Our work also did not explore morphological features including volume, shape characteristics of tumors, including solidity, convexity, and eccentricity of tumors, and their relevance to aggressiveness of cancers. We excluded such features to eliminate any bias resulting from the manual segmentation. Exploration of additional imaging features and shape-based features with prospective data are work for the future. Finally, the ordering of the features was specific to the available data, and confirmation of the ranking on more data should also be done in the future.

In summary, highly accurate classification of PCa GS from T2-w and ADC MR images is feasible despite highly imbalanced data. Addition of texture-based features drastically improves the classification accuracy of GS in comparison with using ADC mean or T2 mean alone.

Conclusions

In this work, we presented multiple machine learning- and feature selection-based methods for classifying (*i*) cancer vs. noncancerous prostate, (*ii*) low GS $6(3+3)$ vs. high GS ≥ 7, and (*iii*) GS $7(3+4)$ vs. GS $7(4+3)$ cancers. Our results suggest that with sample augmentation, reasonably accurate classification with high sensitivity and specificity can be obtained, even for highly imbalanced data such as in the classification of cancer GS. Our work also showed that texture features computed from both ADC and T2-w images drastically improve the classification performance than just the ADC mean and T2 mean. Finally, incorporating the interaction of image features for feature selection followed by a classification method (RFE-SVM) achieved the highest classification accuracy.

Appendix: *t* Test Significance Values. The significance values for the cancer vs. noncancerous class separation were as follows: ADCentropy (<0.001), ADCmean (<0.001), ADChomogeneity (<0.001), ADCSD (<0.001), ADCenergy (<0.001), ADCcorrelation (<0.001), ADCcontrast (<0.001), T2kurtosis (<0.001), T2skewness (<0.001), T2mean (<0.001), ADCskewness (<0.001), ADCkurtosis (<0.001), T2correlation (0.008), T2entropy (0.015), T2contrast (0.018), T2homogeneity (0.075), T2energy (0.350), and T2SD (0.739).

The significance values for the low and high aggressive [GS $6(3+3)$ vs. GS ≥ 7] cancers in the PZ and TZ were as follows: ADCmean (0.003), ADCenergy (0.003), ADCentropy (0.008), ADCskewness (0.016), T2entropy (0.031), T2SD (0.038), ADCkurtosis (0.089), T2energy (0.129), T2mean (0.133), T2contrast (0.147), ADCcorrelation (0.208), ADChomogeneity (0.250), T2correlation (0.321), T2homogeneity (0.479), ADCSD (0.490), T2kurtosis (0.520), ADCcontrast (0.838), and T2skewness (0.925).

The significance values for the low and high aggressive [GS $6(3+3)$ vs. GS ≥ 7] cancers in the PZ were as follows: ADCenergy (0.002), ADCmean (0.002), ADCentropy (0.005), ADCskewness (0.018), T2SD (0.096), T2entropy (0.136), ADCcorrelation (0.138), ADCkurtosis (0.166), T2kurtosis (0.167), ADChomogeneity (0.191), T2contrast (0.202), T2mean (0.314), ADCcontrast (0.508), T2skewness (0.613), ADCSD (0.728), T2homogeneity (0.851), T2energy (0.908), and T2correlation (0.994).

The significance values for the GS 7 [GS (3+4) vs. GS (4+3)] cancers in the PZ and TZ were as follows: ADCmean (0.006), ADCskewness (0.008), ADChomogeneity (0.019), T2SD (0.066), T2entropy (0.093), T2energy (0.109), ADCentropy (0.119), T2mean (0.156), ADCcorrelation (0.261), ADCcontrast (0.349), ADCkurtosis (0.387), T2correlation (0.454), T2kurtosis (0.504), T2contrast (0.602), ADCenergy (0.707), T2homogeneity (0.712), T2skewness (0.723), and ADCSD (0.823).

The significance values for the GS 7 [GS (3+4) vs. GS (4+3)] cancers in the PZ were as follows: ADCmean (0.004), ADCskewness (0.015), ADChomogeneity (0.054), ADCentropy (0.106), T2SD (0.327), ADCcontrast (0.366), T2homogeneity (0.390), ADCkurtosis (0.462), ADCcorrelation (0.500), T2entropy (0.561), T2energy (0.618), ADCenergy (0.628), T2correlation (0.681), T2mean (0.718), ADCSD (0.746), T2skewness (0.771), T2contrast (0.912), and T2kurtosis (0.935).

ACKNOWLEDGMENTS. This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748.

1. Siegel RL, Miller KD, Jemal A (2015) Cancer statistics, 2015. *CA Cancer J Clin* 65(1):5–29.
2. Wright JL, et al. (2009) Prostate cancer specific mortality and Gleason 7 disease differences in prostate cancer outcomes between cases with Gleason 4 + 3 and Gleason 3 + 4 tumors in a population based cohort. *J Urol* 182(6):2702–2707.
3. Ahmed HU, et al.; Transatlantic Consensus Group on Active Surveillance and Focal Therapy for Prostate Cancer (appendix) (2012) Transatlantic Consensus Group on active surveillance and focal therapy for prostate cancer. *BJU Int* 109(11): 1636–1647.
4. Eggener SE, et al. (2011) Predicting 15-year prostate cancer specific mortality after radical prostatectomy. *J Urol* 185(3):869–875.
5. Lavery HJ, Droller MJ (2012) Do Gleason patterns 3 and 4 prostate cancer represent separate disease states? *J Urol* 188(5):1667–1675.
6. King CR, Long JP (2000) Prostate biopsy grading errors: A sampling problem? *Int J Cancer* 90(6):326–330.
7. Epstein JI, Feng Z, Trock BJ, Pierorazio PM (2012) Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: Incidence and predictive factors using the modified Gleason grading system and factoring in tertiary grades. *Eur Urol* 61(5):1019–1024.
8. Berglund RK, et al. (2008) Pathological upgrading and up staging with immediate repeat biopsy in patients eligible for active surveillance. *J Urol* 180(5):1964–1967, discussion 1967–1968.

9. Moore CM, Petrides N, Emberton M (2014) Can MRI replace serial biopsies in men on active surveillance for prostate cancer? *Curr Opin Urol* 24(3):280–287.
10. Sato C, et al. (2005) Differentiation of noncancerous tissue and cancer lesions by apparent diffusion coefficient values in transition and peripheral zones of the prostate. *J Magn Reson Imaging* 21(3):258–262.
11. Fütterer JJ, et al. (2006) Prostate cancer localization with dynamic contrast-enhanced MR imaging and proton MR spectroscopic imaging. *Radiology* 241(2):449–458.
12. Tanimoto A, Nakashima J, Kohno H, Shinmoto H, Kuribayashi S (2007) Prostate cancer screening: The clinical value of diffusion-weighted imaging and dynamic MR imaging in combination with T2-weighted imaging. *J Magn Reson Imaging* 25(1):146–152.
13. Kitajima K, et al. (2010) Prostate cancer detection with 3 T MRI: Comparison of diffusion-weighted imaging and dynamic contrast-enhanced MRI in combination with T2-weighted imaging. *J Magn Reson Imaging* 31(3):625–631.
14. Mazaheri Y, et al. (2008) Prostate cancer: Identification with combined diffusion-weighted MR imaging and 3D 1H MR spectroscopic imaging–correlation with pathologic findings. *Radiology* 246(2):480–488.
15. Langer DL, et al. (2010) Prostate tissue composition and MR measurements: Investigating the relationships between ADC, T2, K(trans), v(e), and corresponding histologic features. *Radiology* 255(2):485–494.
16. Oto A, et al. (2011) Diffusion-weighted and dynamic contrast-enhanced MRI of prostate cancer: Correlation of quantitative MR parameters with Gleason score and tumor angiogenesis. *AJR Am J Roentgenol* 197(6):1382–1390.
17. Nagarajan MB, et al. (2013) Classification of small lesions in breast MRI: Evaluating the role of dynamically extracted texture features through feature selection. *J Med Biol Eng* 33(1):33.
18. Vargas HA, et al. (2011) Diffusion-weighted endorectal MR imaging at 3 T for prostate cancer: Tumor detection and assessment of aggressiveness. *Radiology* 259(3): 775–784.
19. deSouza NM, et al. (2008) Diffusion-weighted magnetic resonance imaging: A potential non-invasive marker of tumour aggressiveness in localized prostate cancer. *Clin Radiol* 63(7):774–782.
20. Donati OF, et al. (2014) Prostate cancer aggressiveness: Assessment with whole-lesion histogram analysis of the apparent diffusion coefficient. *Radiology* 271(1):143–152.
21. Peng Y, et al. (2013) Quantitative analysis of multiparametric prostate MR images: Differentiation between prostate cancer and normal tissue and correlation with Gleason score–a computer-aided diagnosis development study. *Radiology* 267(3): 787–796.
22. Donati OF, et al. (2014) Prostate MRI: Evaluating tumor volume and apparent diffusion coefficient as surrogate biomarkers for predicting tumor Gleason score. *Clin Cancer Res* 20(14):3705–3711.
23. Tiwari P, Viswanath S, Kurhanewicz J, Sridhar A, Madabhushi A (2012) Multimodal wavelet embedding representation for data combination (MaWERiC): integrating magnetic resonance imaging and spectroscopy for prostate cancer detection. *NMR Biomed* 25(4):607–619.
24. Viswanath SE, et al. (2012) 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.
25. Moradi M, et al. (2012) Multiparametric MRI maps for detection and grading of dominant prostate tumors. *J Magn Reson Imaging* 35(6):1403–1413.
26. Tiwari P, Kurhanewicz J, Madabhushi A (2013) Multi-kernel graph embedding for detection, Gleason grading of prostate cancer via MRI/MRS. *Med Image Anal* 17(2): 219–235.
27. Lambin P, et al. (2012) Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 48(4):441–446.
28. Aerts HJ, et al. (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 5:4006.
29. Wibmer A, et al. (2015) Haralick texture analysis of prostate MRI: Differentiating normal prostate from clinically significant prostate cancer and associations with cancer aggressiveness. *Eur Radiol* 25(10):2840–2850.
30. Fedorov A, et al. (2012) 3D Slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging* 30(9):1323–1341.
31. Haralick RM, Shanmugam K, Dinstein IH (1973) Textural features for image classification. *IEEE Trans Syst Man Cybern* 3(6):610–321.
32. MATLAB (2014) *Version 8.3.0 (R2014a)* (MathWorks, Natick, MA).
33. Yoo TS, et al. (2002) Engineering and algorithm design for an image processing Api: A technical report on ITK–the Insight Toolkit. *Stud Health Technol Inform* 85:586–592.
34. Chawla NV, Japkowicz N, Kotcz A (2004) Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsl - Special Issue on Learning from Imbalanced Datasets* 6:1–6.
35. Weiss GM (2004) Mining with rarity: A unifying framework. *ACM SIGKDD Explorations Newsl - Special Issue on Learning from Imbalanced Datasets* 6:7–19.
36. Resnik P, Hardisty E. Gibbs sampling for the uninitiated. College Park (MD): Institute of Advanced Computer Studies, University of Maryland; 2010. Report No.: UMIACS-TR-2010-04.
37. Gosset WS (1908) The probable error of a mean. *Biometrika* 6(1):1–25.
38. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. *Proceedings of the 13th International Conference on Machine Learning* (*ICML1996*) (Morgan Kaufmann, San Francisco), pp 148–156.
39. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Mach Learn* 46(1):389–422.
40. Rakotomamonjy A (2003) Variable selection using SVM based criteria. *J Mach Learn Res* 3(1):1357–1370.
41. Tibshirani R, Tibshirani R (2009) A bias correction for the minimum error rate in cross validation. *Ann Appl Stat* 3(2):822–829.
42. Youden WJ (1950) Index for rating diagnostic tests. *Cancer* 3(1):32–35.
43. Madabhushi A, Feldman MD, Metaxas DN, Tomaszeweski J, Chute D (2005) Automated detection of prostatic adenocarcinoma from high-resolution ex vivo MRI. *IEEE Trans Med Imaging* 24(12):1611–1625.
44. Niaf E, Rouvière O, Mège-Lechevallier F, Bratan F, Lartizien C (2012) Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI. *Phys Med Biol* 57(12):3833–3851.
45. Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA* 99(10):6562–6566.
46. Cawley G, Talbot N (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 11(1):2079–2107.
47. Krstajic D, Buturovic L, Leahy D, Thomas S (2014) Cross-validation pitfalls when selecting and assessing regression and classification methods. *J Cheminformatics* 6(1):1–15.
48. Tsamardinos I, Rakshani A, Lagani V (2014) Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization. *Artif Intell: Methods Appls* 8445(1):1–14.
49. Kohavi R (1995) A study of cross validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artif Intell* (*IJCAI*) (Morgan Kaufmann, San Francisco), Vol 2, pp 1137–1143.
50. Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7(1):1–30.