Diagn Interv Radiol 2024; DOI: 10.4274/dir.2024.242856

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Machine learning models for discriminating clinically significant from clinically insignificant prostate cancer using bi-parametric magnetic resonance imaging

Hakan Ayyıldız[1]
Okan İnce[2]
Esin Korkut[3]
Merve Gülbiz Dağoğlu Kartal[4]
Atadan Tunacı[4]
Şükrü Mehmet Ertürk[4]

[1]Kars Harakani State Hospital, Clinic of Radiology, Kars, Türkiye

[2]Rush University Medical Center, Department of Radiology, Division of Vascular and Interventional Radiology, Chicago, Illinois

[3]University of Health Sciences Türkiye, Bağcılar Training and Research Hospital, Clinic of Radiology, İstanbul, Türkiye

[4]İstanbul University, İstanbul Faculty of Medicine, Department of Radiology, İstanbul, Türkiye

Corresponding author: Hakan Ayyıldız

E-mail: hakanayyildiz77@gmail.com

Received 25 May 2024; revision requested 28 June 2024; last revision received 14 August 2024; accepted 29 August 2024.

Epub: 01.10.2024

Publication date: xx.xx.2024

DOI: 10.4274/dir.2024.242856

PURPOSE

This study aims to demonstrate the performance of machine learning algorithms to distinguish clinically significant prostate cancer (csPCa) from clinically insignificant prostate cancer (ciPCa) in prostate bi-parametric magnetic resonance imaging (MRI) using radiomics features.

METHODS

MRI images of patients who were diagnosed with cancer with histopathological confirmation following prostate MRI were collected retrospectively. Patients with a Gleason score of 3+3 were considered to have clinically ciPCa, and patients with a Gleason score of 3+4 and above were considered to have csPCa. Radiomics features were extracted from T2-weighted (T2W) images, apparent diffusion coefficient (ADC) images, and their corresponding Laplacian of Gaussian (LoG) filtered versions. Additionally, a third feature subset was created by combining the T2W and ADC images, enhancing the analysis with an integrated approach. Once the features were extracted, Pearson's correlation coefficient and selection were performed using wrapper-based sequential algorithms. The models were then built using support vector machine (SVM) and logistic regression (LR) machine learning algorithms. The models were validated using a five-fold cross-validation technique.

RESULTS

This study included 77 patients, 30 with ciPCA and 47 with csPCA. From each image, four images were extracted with LoG filtering, and 111 features were obtained from each image. After feature selection, 5 features were obtained from T2W images, 5 from ADC images, and 15 from the combined dataset. In the SVM model, area under the curve (AUC) values of 0.64 for T2W, 0.86 for ADC, and 0.86 for the combined dataset were obtained in the test set. In the LR model, AUC values of 0.79 for T2W, 0.86 for ADC, and 0.85 for the combined dataset were obtained.

CONCLUSION

Machine learning models developed with radiomics can provide a decision support system to complement pathology results and help avoid invasive procedures such as re-biopsies or follow-up biopsies that are sometimes necessary today.

CLINICAL SIGNIFICANCE

This study demonstrates that machine learning models using radiomics features derived from bi-parametric MRI can discriminate csPCa from clinically insignificant PCa. These findings suggest that radiomics-based machine learning models have the potential to reduce the need for re-biopsy in cases of indeterminate pathology, assist in diagnosing pathology–radiology discordance, and support treatment decision-making in the management of PCa.

KEYWORDS

Prostate, magnetic resonance imaging, prostate cancer, radiomics, machine learning, bi-parametric magnetic resonance imaging

You may cite this article as: Ayyıldız H, İnce O, Korkut E, Dağoğlu Kartal MG, Tunacı A, Ertürk ŞM. Machine learning models for discriminating clinically significant from clinically insignificant prostate cancer using bi-parametric magnetic resonance imaging. *Diagn Interv Radiol*. 01 October 2024 DOI: 10.4274/dir.2024.242856 [Epub Ahead of Print].

Prostate cancer (PCa) is the second most common cancer in men, with a rising incidence.[1] The prostate-specific antigen (PSA) test remains a commonly used screening method, although recent studies suggest it has a limited impact on survival outcomes.[2,3] In the modern medical landscape, the significance of prostate imaging, particularly with magnetic resonance imaging (MRI), has grown. Imaging plays an important role in the diagnosis of PCa, and multiparametric prostate magnetic resonance imaging (mpMRI) is the most commonly used imaging modality for diagnosis. Different versions of Prostate Imaging–Reporting and Data System (PI-RADS®) have been published to standardize mpMRI radiology reports.[4] In cases where the Gleason score is 6, the Gleason grade group (GGG) is 1, and in cases where the Gleason score is 3+4 or higher, the GGG is 2 or above. In PCa, the prognosis is expected to be better if GGG = 1.[5] However, active surveillance can be applied to patients with GGG <2.[6] Treatment management varies with the GGG. Although dynamic contrast-enhanced imaging is considered a "safety zone," bi-parametric magnetic resonance imaging (bpMRI) is increasingly favored due to its speed and, in some studies, comparable diagnostic performance to mpMRI.[7,8] PI-RADS® version 2.1 indicates that bpMRI may be a viable option for decreasing the use of ga-dolinium contrast media, associated adverse reactions, and examination time.[4] This can lead to greater accessibility to prostate MRI for patients. Nevertheless, the PI-RADS® version 2.1 suggests mpMRI for patients with a high likelihood of cancer based on factors such as PSA levels, family history, or genetic predisposition. It also recommends mpMRI in cases where image quality may be compromised, such as in patients with hip prostheses.

In PI-RADS® version 2.1, for clinically significant prostate cancer (csPCa), at least one of the following must be present: GGG >2, volume >0.5 cc, or extra-prostatic extension. As a result, frequent distinction between clinically significant and clinically insignificant prostate cancer (ciPCa) is achieved by histopathological verification as a result of prostate biopsy, which is an invasive method. Gleason score may need to be re-evaluated by pathology in some cases.[9] In our study, we aimed to show the role of machine learning algorithms created from radiomics features obtained from T2-weighted (T2W) and apparent diffusion coefficient (ADC) sequences in MRI. A review of the literature reveals numerous machine learning-based studies on PCa detection, particularly csPCa detection.[10,11] Our study aims to make a significant contribution to the existing literature by providing an easily applicable, reproducible, and more accurate model that facilitates the distinction between csPCa and ciPCa. This model is particularly focused on improving the management of patient populations who may require multiple biopsies over time. In daily practice, this could impact a considerable number of patients.

Main points

- This study employed machine learning algorithms [support vector machine (SVM) and logistic regression (LR)] to differentiate between clinically significant prostate cancer (csPCa) and clinically insignificant prostate cancer (ciPCa) using radiomics features from bi-parametric magnetic resonance imaging images.
- Feature selection yielded 5 key features from T2-weighted (T2W) images, 5 from apparent diffusion coefficient (ADC) images, and 15 from the combined dataset, which was critical for model accuracy.
- A total of 77 patients were analyzed, with the SVM model achieving area under the curve (AUC) values of 0.64 for T2W, 0.86 for ADC, and 0.86 for combined images, whereas the LR model achieved AUC values of 0.79 for T2W, 0.86 for ADC, and 0.85 for combined images.
- The findings suggest that machine learning models using radiomics can significantly aid in distinguishing csPCa from ciPCa, potentially reducing the need for invasive biopsy procedures.

Methods

Ethics and data source

This study was approved by the İstanbul University, İstanbul Faculty of Medicine Ethics Committee (decision number: 2021/676, date: 28/05/2021). Since it was a retrospective study, informed consent was waived. The dataset was obtained by retrospectively scanning the images of patients >18 years of age, whose lesions detected after mpMRI were confirmed histopathologically by systematic core and targeted biopsy at the department of radiology of the institution between 2016 and 2022. Fusion biopsy, which combines MR and ultrasound imaging, was used as the biopsy technique in all patients. The data were obtained on a lesion-by-lesion basis to avoid possible bias during the data collection. The exclusion criteria encompassed the following conditions: 1) elimination of cases with imaging artifacts that hindered the accurate segmentation of cancer lesions and 2) exclusion of instances with incomplete MRI data, including scenar-ios where essential images were missing.

Magnetic resonance imaging

The primary MRI sequences chosen for radiomic input in prostate imaging included axial T2W images and ADC images. In this study, two distinct MR technologies were employed: the Magnetom Aera 1.5 Tesla (Siemens Healthcare, Germany) and the Achieva 3.0 Tesla (Philips Medical Systems, the Netherlands). ADC images were acquired from diffusion-weighted imaging (DWI) with a b-value of 0 and 1400 sec/mm^2 on both devices. For the axial T2W images and DWI protocols, the repetition time/echo time parameters were set at 7500/100 and 5000/70 ms for the 1.5 Tesla system and 4200/100 and 3600/70 ms for the 3.0 Tesla system. The field of view (FOV) differed between the devices, with an 18 mm x 18 mm FOV for the 3.0 Tesla system and a 20 mm x 20 mm FOV for the 1.5 Tesla system. A slice thickness of 3.0 mm was maintained consistently on both devices, with no slice gap, to ensure homogeneity in imaging parameters throughout the study. Table 1 provides detailed parameters related to MRI.

Table 1. Parameters of the magnetic resonance imaging

	T2W (1.5 Tesla system)	DWI (1.5 Tesla system)	T2W (3.0 Tesla system)	DWI (3.0 Tesla system)
TR	7500	5000	4200	3600
TE	100	70	100	70
FOV area	20 mm x 20 mm	20 mm x 20 mm	18 mm x 18 mm	18 mm x 18 mm
Matrix	320 x 320	256 x 256	230 x 180	64 x 64
Voxel size (x, y, z; mm)	0.6 x 0.6 x 3 (mm)	0.8 x 0.8 x 3 (mm)	0.8 x 1 x 3 (mm)	3 x 3 x 3 (mm)
Slice thickness	3 mm	3 mm	3 mm	3 mm
Slice gap	-	-	-	-
Sequence	Turbo spin echo	Echo planar imaging	Turbo spin echo	Echo planar imaging

T2W, T2-weighted; DWI, diffusion-weighted imaging; TR, repetition time; TE, echo time; FOV, field of view.

Image preprocessing and feature extraction

The acquired images underwent normalization through the proprietary algorithm embedded in Olea Sphere® software (Olea Medical, La Ciotat, France). Despite ADC being a computationally derived sequence, it underwent normalization in a manner consistent with the axial T2W series, aligning with the recommendations in radiomics studies.[12] Subsequently, outlier pixels were systematically eliminated using the ±3 sigma technique.[13] Following normalization and the removal of outlier pixels, pixel sizes were standardized to a 1 × 1 mm^2 scale using cubic B-spline interpolation. The gray levels were then discretized uniformly for both series with a bin width of 0.05.[14] Utilizing PyRadiomics, Laplacian of Gaussian (LoG) filter images were extracted from the original images with logarithmic values of 2, 4, and 6. Consequently, four images were derived from a single original image, where one of them represented the original unaltered image.

Segmentations were performed manually by two radiologists using the freehand method, prior to the steps described in the previous paragraph. Each radiologist had 5 and 4 years of experience, respectively, and performed the segmentations independently using axial T2W and ADC images (Figure 1). When necessary for improved tumor orientation, DWI with b-values of 0 and 1400, as well as sagittal T2W images, were incorporated. However, for objective bi-parametric modeling, contrast-enhanced series were intentionally omitted and not reviewed during the segmentation process. During segmentation, the lesion area with high suspicion of tumor was included, whereas areas of uncertainty were excluded. The suspicious lesion underwent volumetric 3D segmentation using Olea Sphere® software. Subsequently, feature extraction was performed from the original image, as well as from three LoG-filtered series within each set, following the steps described in the previous paragraph. The radiomics workflow is summarized in Figure 2.

Data preprocessing and feature selection

To ensure consistency and dependability of machine learning models, meticulous data pre-processing steps were performed.[15] After standardization and discretization were applied uniformly to all data using a consistent bin width, the data were divided into 20 bins. The dataset was randomly split into training and test sets with a 70/30 ratio. To prevent contamination of the test dataset with the training dataset, data splitting was conducted before any data augmentation. This approach ensured the integrity and independence of the training and test datasets, and T2W and ADC series were combined to construct a unified dataset.

Pearson's correlation coefficient was employed to identify and remove redundant features. Pairs of features with a correlation coefficient exceeding a threshold of 0.80 were identified and subsequently removed.[16] The remaining features, which met these criteria, served as input for the next stage.

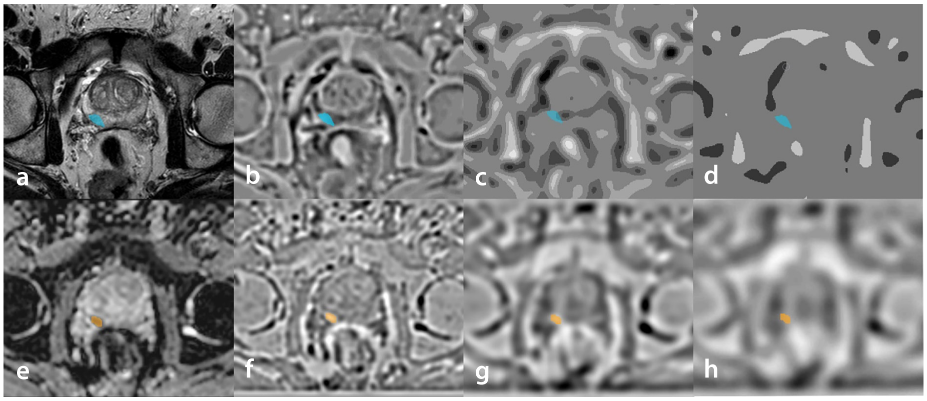

Figure 1. The figure shows examples of the segmentation of the original **(a)** T2-weighted images with Laplacian of Gaussian (LoG) filters using sigma values of 2 **(b)**, 4 **(c)**, and 6 **(d)**, as well as the original **(e)** apparent diffusion coefficient images with LoG filters using sigma values of 2 **(f)**, 4 **(g)**, and 6 **(h)**.

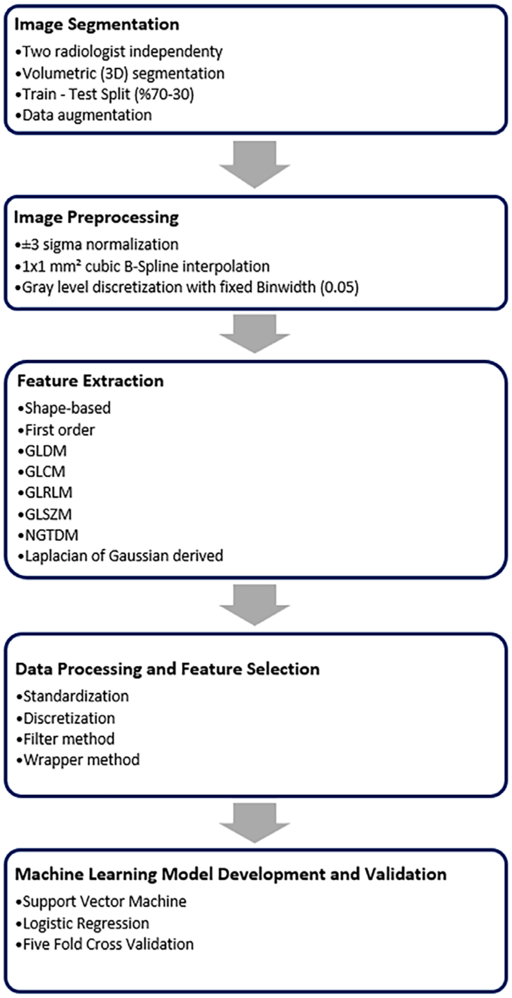
Image Segmentation
•Two radiologist independenty
•Volumetric (3D) segmentation
•Train - Test Split (%70-30)
•Data augmentation
Image Preprocessing
•±3 sigma normalization
•1x1 mm² cubic B-Spline interpolation
•Gray level discretization with fixed Binwidth (0.05)
Feature Extraction
•Shape-based
•First order
•GLDM
•GLCM
•GLRLM
•GLSZM
•NGTDM
•Laplacian of Gaussian derived
Data Processing and Feature Selection
•Standardization
•Discretization
•Filter method
•Wrapper method
Machine Learning Model Development and Validation
•Support Vector Machine
•Logistic Regression
•Five Fold Cross Validation

Figure 2. The figure illustrates the radiomics workflow. GLDM, gray-level dependence matrix; GLCM, gray-level co-occurrence matrix; GLRLM, gray-level run length matrix; GLSZM, gray-level size zone matrix; NGTDM, neighboring gray-tone difference matrix.

A wrapper-based feature selection algorithm was then developed using backward elimination and five-fold cross-validation. Logistic regression (LR) was chosen as the estimator for feature selection. This wrapper method evaluates different models by iteratively including or excluding features to determine the optimal feature combination. Each model was analyzed by iteratively removing one feature at a time. Through multiple evaluations, the most relevant features were identified. Crucial features were selected exclusively from the training folds using cross-validation, thereby avoiding the "double-dipping" phenomenon.[17] As previously indicated, the test set remained untouched throughout the feature selection process due to the prior data division into training and test sets.

Machine learning algorithms

T2W, ADC, and the combined dataset were incorporated into the machine learning modeling. The finalized set of features was used for implementing machine learning algorithms, which were executed using Python (version 3). The first model employed was a support vector machine (SVM) with hyperparameters set to C: 1.0 and kernel: linear. Another model, LR, was used with hyperparameters configured as C: 25, solver: liblinear, and regularization penalty: L2 (Ridge). The performance of the models was evaluated using five-fold cross-validation. Metrics including accuracy, sensitivity, specificity, precision, recall, F1 score, and the area under the curve (AUC) were calculated.

Results

Patients

The study involved a total of 108 patients. However, 14 patients were excluded due to incomplete pathology results, 3 patients had incomplete images, and 14 patients had artifacts in their images (Figure 3). Of the remaining patients, 61% (47 patients) were diagnosed with csPCa, whereas 39% (30 patients) were classified as having ciPCa. Table 2 provides a summary of the patients' characteristics.

Feature extraction and selection

A total of 444 features were extracted from each sequence. These features were categorized as follows: 17 (15.32%) shape, 19 (17.12%) first-order, 24 (21.62%) gray-level co-occurrence matrix, 16 (14.41%) gray-level run-length matrix, 16 (14.41%) gray-level size-zone matrix, 14 (12.61%) gray-level dependence matrix, and 5 (4.50%) neighboring gray-tone difference matrix features. Subsequently, a combined dataset was generated by concatenating features from both T2W and ADC sequences.

Pearson's correlation coefficient identified 28, 31, and 50 features as non-overlapping in T2W, ADC, and the combined group, respectively. Following the wrapper-based sequential feature selection step, the final feature subsets consisted of 5 features in T2W, 5 in ADC, and 15 in the combined group, details of which are shown in Table 3 and Figure 4.

Models performance

The SVM models demonstrated accuracy scores of 75%, 85%, and 91% in the training group and 64%, 76%, and 72% in the test group for the T2W, ADC, and combined groups, respectively. The corresponding AUC values with 95% confidence intervals (CI) were 0.75 (0.74–0.76), 0.89 (0.88–0.89), and 0.95 (0.95–0.96) in the training group, and 0.64 (0.62–0.65), 0.86 (0.85–0.88), and 0.86 (0.85–0.88) in the test group for the T2W, ADC, and combined groups, respectively.

The LR models in the T2W, ADC, and combined groups had accuracy scores of 74%, 84%, and 86% in the training group, and 70%, 79%, and 77% in the test group, respectively. The AUC values with 95% CI were as follows: for the T2W, ADC, and combined groups in the training group, 0.83 (0.82–0.83), 0.89 (0.88–0.89), and 0.95 (0.94–0.95); and in the test group, 0.79 (0.78–0.80), 0.86 (0.84–0.88), and 0.85 (0.83–0.87), respectively. Detailed performance analyses for the training group and the test group are presented in Table 4, and Figure 5 shows the receiver operating characteristic curves for all models.

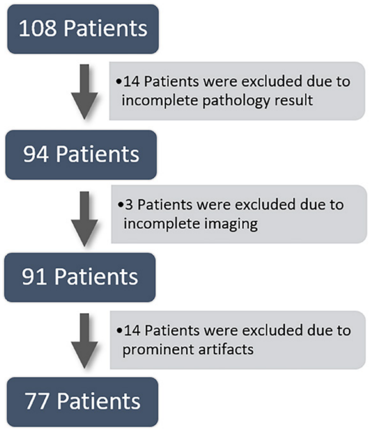
108 Patients
•14 Patients were excluded due to incomplete pathology result
94 Patients
•3 Patients were excluded due to incomplete imaging
91 Patients
•14 Patients were excluded due to prominent artifacts
77 Patients

Figure 3. The figure demonstrates the patient selection algorithm. (F, feature; as listed in Table 2).

Table 2. Demographic and patient characteristics for both groups

	csPCa	ciPCa	*P* value
Age (mean ± SD) (95% CI)	65.22 ± 8.85 (62.59–67.84)	61.61 ± 6.8 (58.97–64.24)	0.086
PSA level (median) (min–max)	7.46 (1.22–38.67)	5.95 (2.0–45.0)	0.044*
Localization (n)			
Peripheric zone	40	24	0.560
Transitional zone	7	6	
MRI technology			
1.5 T scanner	31	19	0.009†
3.0 T scanner	16	11	
Gleason score (n)			
Gleason 3+3	NA	30	
Gleason 3+4	24	NA	
Gleason 4+3	9	NA	
Gleason 4+4	9	NA	
Gleason 4+5	3	NA	
Gleason 5+4	1	NA	
Gleason 5+5	1	NA	

*A significant difference was found between both groups by Mann–Whitney U test revealing a higher value in the CsPCa group. †1.5 Tesla scanners have a higher number of patients and a significant difference was found in the Pearson's chi-square test. csPCa, clinically significant prostate cancer; ciPCa, clinically insignificant prostate cancer; SD, standard deviation; CI, confidence interval; PSA, prostate-specific antigen; min–max, minimum–maximum; MRI, magnetic resonance imaging; NA, not available.

Discussion

In our investigation, the efficacy of machine learning models employing prostate bpMRI radiomics analysis for predicting csPCa was explored, revealing promising predictive capabilities. As the two different algorithms work on different principles, an attempt was made to minimize the possibility of overfitting by using them in the algorithms created and to evaluate the usability of the different models. The reasonable and comparable accuracy rates of these algorithms in this study demonstrate the feasibility of using machine learning algorithms to identify csPCa.

In the literature, radiomics studies conducted using ultrasonography and computed tomography in prostate imaging are available.[18,19] Nevertheless, the popularity of radiomics studies in prostate MRI is notably increasing. The field of radiomics studies conducted in MR is expansive, encompassing endeavors to differentiate extraprostatic extension, discern normal tissue from cancer, identify recurrence post-radical prostatectomy, and distinguish recurrence after treatment.[20] Notably, the treatment approaches for csPCa and ciPCa can vary significantly.[21-23] Hence, there is a growing significance in conducting studies aimed at differentiating csPCa and ciPCa. Some of these studies have been performed with mpMRI and some with bpMRI. Our study was conducted with bpMRI, which is more accessible, has a shorter duration, and is considered suitable for acquisition with certain criteria in PI-RADS® version 2.1, and studies are being conducted to disseminate it.[4]

Zhang et al.[24] used GGG 1 and GGG >1 groups in their nomogram study of 159 patients with radiomics, similar to our study. Similar to our study, only bpMRI images were used, and segmentation was performed on DWI, ADC, and T2W. Although the use of internal validation was the advantage of the study, this study was performed with a single 3.0T MR technology. In addition, this study was performed with a radiomic nomogram, and machine learning modeling was not applied. In a retrospective radiomics study of 489 patients, Gong et al.[25] derived models from bpMRI data (T2W and DWI). They incorporated clinical modeling by including PSA data in the study. Performed on a single 3.0T MRI machine, they reported an AUC of 0.811 in the training group and 0.788 in the test group for the combined model, which was created without integrating clinical

Table 3. Selected features and their classifications for T2W, ADC, and combined datasets

Selected features					
T2W		ADC		Combined dataset	
Image type	Feature name (feature class)	Image type	Feature name (feature class)	Image type	**Feature name (feature class)**
Original	**Original shape surface area to volume ratio** (shape)	Original	Original shape mesh volume (shape)	T2W - original	**Original shape surface area to volume ratio** (shape)
Original	**Original shape sphericity** (shape)	Original	Original shape surface area to volume ratio (shape)	T2W - original	**Original shape sphericity** (shape)
Original	Original first order root mean squared (first order)	Original	*Original first order entropy* (shape)	T2W - original	Original shape elongation (shape)
Original	**Original GLCM correlation** (GLCM)	Original	Original first order skewness (first order)	T2W - original	Original shape flatness (shape)
LoG (Sigma: 4)	Original GLCM informal measure of correlation 1 (GLCM)	LoG (Sigma: 4)	Original first order kurtosis (first order)	T2W - original	**Original GLCM correlation** (GLCM)
				T2W - laplacian of gaussian (Sigma: 2)	Original first order 90th percentile (first order)
				ADC - original	*Original first order entropy* (first order)
				ADC - original	Original first order minimum (first order)
				ADC - original	Original GLCM inverse variance (GLCM)
				ADC - original	Original GLSZM small area emphasis (GLSZM)
				ADC - original	Original GLDM large dependence low gray level emphasis (GLDM)
				ADC - LoG (Sigma: 4)	Original GLCM inverse variance (GLCM)
				ADC - LoG (Sigma: 6)	Original GLCM informal measure of correlation 1 (GLCM)
				ADC - LoG (Sigma: 6)	Original GLCM maximal correlation coefficient (GLCM)
				ADC - LoG (Sigma: 6)	Original GLSZM variance (GLSZM)

The same features included in the combined group and T2W are shown in bold; the combined group and ADC are shown in italic. T2W, T2-weighted; ADC, apparent diffusion coefficient; GLCM, gray-level co-occurrence matrix; LoG, Laplacian of Gaussian; GLSZM, gray-level size-zone matrix; GLDM, gray-level dependence matrix.

modeling. However, in this study, PCa was separated into low-grade and high-grade, and patients with a Gleason score <8 were considered low-grade. Li et al.[26] used T2W and ADC series in their retrospective study of 381 patients to differentiate csPCa, but 199 patients were selected from the benign patient group. Clinical modeling was also included in the study, and they reported the AUC value obtained without clinical modeling as 0.99 in the training group and 0.98 in the test group. However, this study used a single MR scanner, and lesion segmentation was supported by pathological data and dynamic contrast-enhanced images.[26]

In current clinical practice, almost all patients with suspected PCa require a biopsy. The advantage of conducting our study only with patients who have a Gleason score of 6 or higher is to avoid the possibility that these patients, diagnosed with cancer, might require re-biopsies during follow-up under current conditions or even immediately after the initial biopsy. Thus, the aim is to create a decision support system to aid the pathology result or to identify patients who need re-biopsy. The use of two MR scanners, 3.0T and 1.5T, and the modeling of both peripheral and transitional zone lesions are advantageous in our study. In addition to its rapid applicability, another advantage of our model for bpMRI over other studies is that it relies solely on ADC series and does not use contrast-enhanced sequences. Furthermore, the significance of our study lies in the high reproducibility of the technique, along with its favorable accuracy rates and AUC values, which are relatively high compared to other studies.[27] Other studies in the literature used more images than T2 and ADC and achieved similar accuracy rates to those in our study.[11] Additionally, some studies with high accuracy rates focused only on the peripheral or transitional zones. For instance, Fehr et al.[28] reported high accuracy rates but performed segmentation in conjunction with pathological results. Chen et al.[29] also reported high accuracy rates, but their study did not perform an interobserver analysis.

Our study has several limitations. First, as a retrospective study sourcing data from past registries, it may introduce selection bias. Second, although segmentation was performed independently by two radiologists, the manual nature of this process can introduce subjectivity. Third, the patient population was relatively small, raising concerns about a potential imbalance between groups. Class imbalance can challenge many machine learning algorithms, which typically assume that all classes are equally distributed.[15] In cases of imbalanced classes, models tend to favor predictions for the majority class. To address class imbalance and reduce the risk of overfitting, especially with a limited number of samples, data augmentation is a validated technique. The use of different synthetic over-sampling methods can provide a more efficient and effective approach.[30,31] However, this would result in a substantial portion of the data being synthetic. Furthermore, despite employing a systematic and

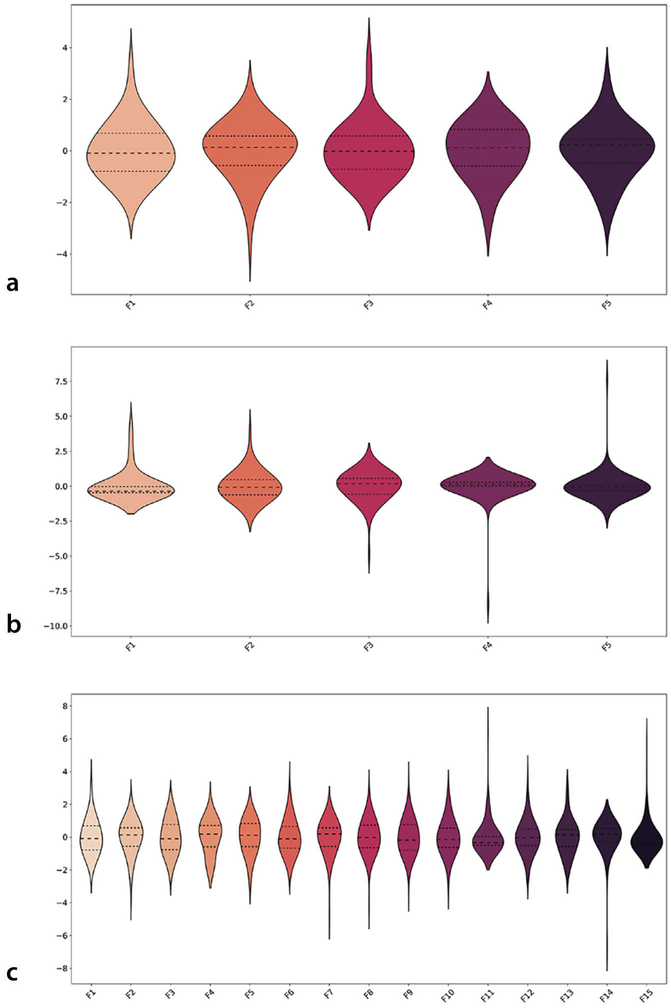

Figure 4. The selected features for T2-weighted images **(a)**, apparent diffusion coefficient images **(b)**, and the combined dataset **(c)** are shown.

Table 4. Detailed performance statistics for machine learning algorithms on T2W, ADC, and combined dataset

	Group	Accuracy	Sensitivity	Specificity	Recall	F1	AUC (95% CI)
LR-T2W	Train Test	74% 70%	77% 76%	57% 56%	84% 79%	0.80 0.76	0.83 (0.82–0.83) 0.79 (0.78–0.80)
SVM-T2W	Train Test	75% 64%	77% 69%	52% 46%	85% 75%	0.81 0.71	0.75 (0.74–0.76) 0.64 (0.62–0.65)
LR-ADC	Train Test	84% 79%	85% 82%	69% 67%	90% 87%	0.87 0.84	0.89 (0.88–0.89) 0.86 (0.84–0.88)
SVM-ADC	Train Test	85% 76%	85% 80%	65% 63%	90% 85%	0.88 0.82	0.89 (0.88–0.89) 0.86 (0.85–0.88)
LR-combined	Train Test	86% 77%	89% 85%	73% 70%	95% 81%	0.92 0.80	0.95 (0.94–0.95) 0.85 (0.83–0.87)
SVM-combined	Train Test	91% 72%	90% 75%	68% 66%	95% 75%	0.93 0.75	0.95 (0.95–0.96) 0.86 (0.85–0.88)

T2W, T2-weighted image; ADC, apparent diffusion coefficient; F1 score, the harmonic mean of precision and recall; AUC, area under the curve; CI, confidence interval; LR, logistic regression; SVM, support vector machine.

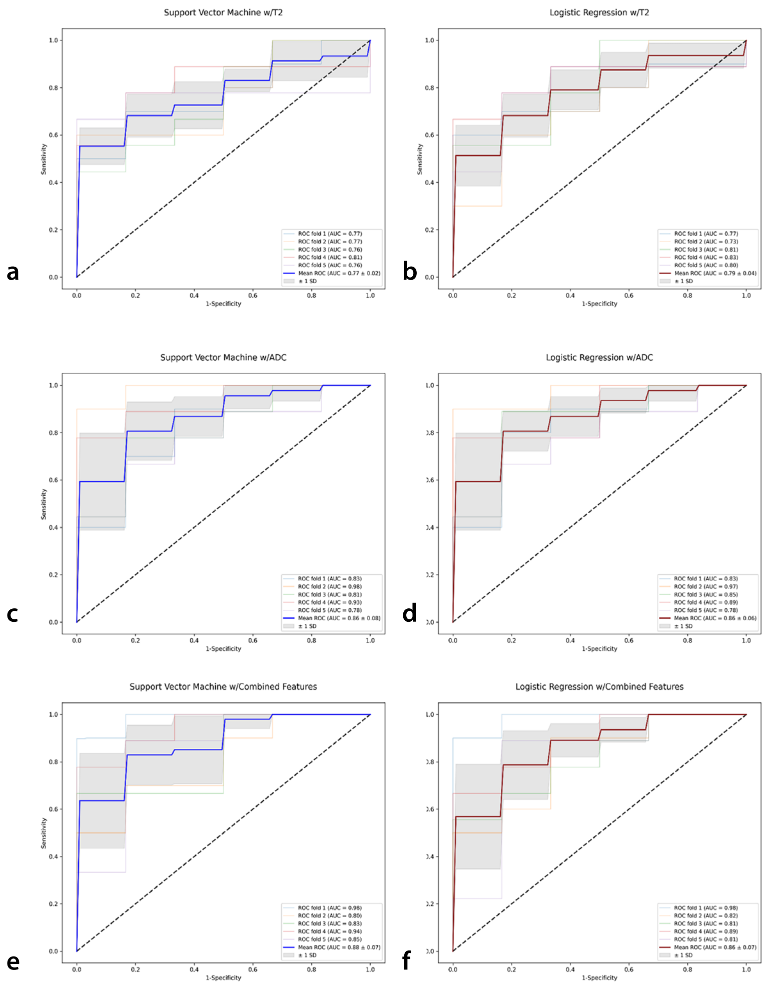

Figure 5. The figure presents the ROC curves, illustrating the ability of the models to differentiate between clinically significant and clinically insignificant prostate cancer across T2W, ADC, and combined datasets. ROC, receiver operating characteristic; T2W, T2-weighted; ADC, apparent diffusion coefficient; AUC, area under the curve; SD, standard deviation.

targeted biopsy approach, the accuracy of the PCa score may be underestimated due to potential limitations in puncture pathology, which might not accurately reflect the true pathological status. Additionally, in PI-RADS® version 2.1, the criteria for csPCa include extraprostatic extension and volume criteria, in addition to the Gleason score. Although patients were selected retrospectively, care was taken to exclude those meeting this criterion from the ciPCa group.

In conclusion, machine learning models utilizing radiomics extracted from prostate bpMRI show promising results in distinguishing between csPCa and ciPCa. However, additional studies with larger datasets are needed to validate these models across external centers before considering their clinical implementation. Incorporating clinical data, such as PSA levels, into these models could lead to the development of more robust tools for clinical practice. The integration of radiomics with artificial intelligence methodologies, including machine learning, holds significant potential for future advancements in prostate imaging.

Conflict of interest disclosure

Şükrü Mehmet Ertürk, MD, is Publication Coordinator in Diagnostic and Interventional Radiology. He had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Other authors have nothing to disclose.

References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *Int J Cancer*. 2015;136(5):359-386. [Crossref]

2. Ilic D, Djulbegovic M, Jung JH, et al. Prostate cancer screening with prostate-specific antigen (Psa) test: a systematic review and meta-analysis. *BMJ*. 2018;5:362:k3519. [Crossref]

3. Eldred-Evans D, Burak P, Connor MJ, et al. Population-based prostate cancer screening with magnetic resonance imaging or ultrasonography: the Ip1-prostagram study. *JAMA Oncol*. 2021;7(3):395-402. [Crossref]

4. Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *Eur Urol*. 2019;76(3):340-351. [Crossref]

5. Epstein JI, Egevad L, Amin MB, et al. The 2014 International Society of Urological Pathology (Isup) consensus conference on gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. *Am J Surg Pathol*. 2016;40(2):244-252. [Crossref]

6. Evans AJ. Treatment effects in prostate cancer. *Mod Pathol*. 2018;31(Suppl 1):110-121. [Crossref]

7. Woo S, Suh CH, Kim SY, Cho JY, Kim SH, Moon MH. Head-to-head comparison between biparametric and multiparametric MRI for the diagnosis of prostate cancer: a systematic review and meta-analysis. *AJR Am J Roentgenol*. 2018;211(5):226-241. [Crossref]

8. Niu XK, Chen XH, Chen ZF, Chen L, Li J, Peng T. Diagnostic performance of biparametric MRI for detection of prostate cancer: a systematic review and meta-analysis. *AJR Am J Roentgenol*. 2018;211(2):369-378. [Crossref]

9. van Santvoort BWH, van Leenders GJLH, Kiemeney LA, et al. Histopathological re-evaluations of biopsies in prostate cancer: a nationwide observational study. *Scand J Urol*. 2020;54(6):463-469. [Crossref]

10. T JMC, Arif M, Niessen WJ, Schoots IG, Veenland JF. Automated classification of significant prostate cancer on MRI: a systematic review on the performance of machine learning applications. *Cancers (Basel)*. 2020;12(6).1606. [Crossref]

11. Cuocolo R, Cipullo MB, Stanzione A, et al. Machine learning for the identification of clinically significant prostate cancer on MRI: a meta-analysis. *Eur Radiol*. 2020;30(12):6877-6887. [Crossref]

12. Becker AS, Wagner MW, Wurnig MC, Boss A. Diffusion-weighted imaging of the abdomen: impact of b-values on texture analysis features. *NMR Biomed*. 2017;30(1). [Crossref]

13. Collewet G, Strzelecki M, Mariette F. Influence of Mri acquisition protocols and image intensity normalization methods on texture classification. *Magn Reson Imaging*. 2004;22(1):81-91. [Crossref]

14. Shafiq-Ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. *Sci Rep*. 2018;8(1):10545. [Crossref]

15. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: a practical guide for beginners. *Diagn Interv Radiol*. 2019;25(6):485-495. [Crossref]

16. Dormann CF, Elith J, Bacher S, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. 2013;36(1):27-46. [Crossref]

17. Ball TM, Squeglia LM, Tapert SF, Paulus MP. Double dipping in machine learning: problems and solutions. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2020;5(3):261-263. [Crossref]

18. Wildeboer RR, Mannaerts CK, van Sloun RJG, et al. Automated multiparametric localization of prostate cancer based on B-mode, shear-wave elastography, and contrast-enhanced ultrasound radiomics. *Eur Radiol*. 2020;30(2):806-815. [Crossref]

19. Osman SOS, Leijenaar RTH, Cole AJ, et al. Computed tomography-based radiomics for risk stratification in prostate cancer. *Int J Radiat Oncol Biol Phys*. 2019;105(2):448-456. [Crossref]

20. Ferro M, De Cobelli O, Musi G, et al. Radiomics in prostate cancer: an up-to-date review. *Ther Adv Urol*. 2022;14:175628722211090. [Crossref]

21. Miyake H, Sakai I, Inoue TA, Hara I, Fujisawa M. The limited significance of a longer duration of neoadjuvant hormonal therapy prior to radical prostatectomy for high-risk prostate cancer in Japanese men. *Urol Int*. 2006;77(2):122-126. [Crossref]

22. Carroll PH, Mohler JL. NCCN Guidelines Updates: prostate cancer and prostate cancer early detection. *J Natl Compr Canc Netw*. 2018;16(5s):620-623. [Crossref]

23. Hassan O, Han M, Zhou A, et al. IIncidence of extraprostatic extension at radical prostatectomy with pure gleason score 3 + 3 = 6 (grade group 1) cancer: implications for whether gleason score 6 prostate cancer should be renamed “not cancer” and for selection criteria for active surveillance. *J Urol*. 2018;199(6):1482-1487. [Crossref]

24. Zhang Y, Chen W, Yue X, et al. Development of a novel, multi-parametric, MRI-based radiomic nomogram for differentiating between clinically significant and insignificant prostate cancer. *Front Oncol*. 2020;10:888. [Crossref]

25. Gong L, Xu M, Fang M, et al. Noninvasive prediction of high-grade prostate cancer via biparametric MRI radiomics. *J Magn Reson Imaging*. 2020;52(4):1102-1109. [Crossref]

26. Li M, Chen T, Zhao W, et al. Radiomics prediction model for the improved diagnosis of clinically significant prostate cancer on biparametric MRI. *Quant Imaging Med Surg*. 2020;10(2):368-379. [Crossref]

27. Li H, Lee CH, Chia D, Lin Z, Huang W, Tan CH. Machine Learning in Prostate MRI for prostate cancer: current status and future opportunities. *Diagnostics (Basel)*. 2022;12(2):289. [Crossref]

28. Fehr D, Veeraraghavan H, Wibmer A, et al. Automatic classification of prostate cancer gleason scores from multiparametric magnetic resonance images. *Proc Natl Acad Sci U S A*. 2015;112(46):6265-6273. [Crossref]

29. Chen T, Li M, Gu Y, et al. Prostate cancer differentiation and aggressiveness: assessment with a radiomic-based model vs. PI-RADS V2. *J Magn Reson Imaging*. 2019;49(3):875-884. [Crossref]

30. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: Synthetic minority over-sampling technique. *J Artif Intell*. 2002;16:321-357. [Crossref]

31. He H, Bai Y, Garcia EA, Li S. Adasyn: adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence); 2008: Ieee. [Crossref]