SN Computer Science (2026) 7:430
https://doi.org/10.1007/s42979-026-04983-w

ORIGINAL RESEARCH

Automated PI-RADS 3–5 Classification Using Multiparametric MRI: A Comparative Study of Radiomics and Deep Learning Approaches

Saman Fouladi[1,2] · Isa Bossi Zanetti[2] · Fatemeh Darvizeh[2] · Rosario Di Meo[2] · Luca Di Palma[2] · Eros Cambie[2] · Antonino Licata[2] · Alessandro Maiocchi[3] · Ernesto Damiani[1] · Corrado Mio[1] · Gabriele Gianini[4] · Marco Ali[2] · Deborah Fazzini[2]

Received: 22 November 2025 / Accepted: 10 April 2026

Abstract

Accurate classification of clinically significant prostate cancer remains a major challenge. While multiparametric MRI (mpMRI) has improved lesion detection, effective categorization in accordance to the Prostate Imaging Reporting and Data System (PI-RADS) remains complex. In this study, we propose and evaluate three complementary approaches for automated PI-RADS classification differing in the way in which the features are extracted from the mpMRI imaging sequences. The first approach extracts hand-crafted radiomic features from manually segmented lesions using the PyRadiomics library. The second approach extends this by integrating fully automated lesion and zonal segmentation to simulate a real-world, manual-free pipeline. The third approach utilizes a custom convolutional neural network (CNN) to learn high-level features images and lesion masks directly. The images come from Apparent diffusion coefficient (ADC), diffusion-weighted imaging (DWI), and T2-weighted (T2W) imaging. The features issued by the three methods were used to train a set of machine learning models for multi-class PI-RADS classification, specifically targeting the clinically relevant categories 3, 4, and 5. Results show that ADC-derived features consistently yield superior performance, with one of the ensemble models reaching an AUC of 0.83. Combining features across all sequences further improved robustness (AUC = 0.84). PI-RADS 5 classification was most reliable (AUC ≥ 0.94), whereas PI-RADS 3 remained the most difficult to distinguish. Our findings highlight the effectiveness of ADC features and the advantage of combining automated and deep learning-based strategies for robust prostate cancer risk stratification.

Keywords Prostate cancer · PI-RADS classification · Risk stratification · Multiparametric MRI · Radiomics features · Pyradiomics · Automated lesion segmentation

Introduction

One of the most pressing challenges in prostate cancer care today is not just detecting cancer, but identifying which cancers matter. Many prostate tumors grow so slowly that they would never cause harm during a man's lifetime, while others are aggressive and potentially life-threatening. Yet the approaches traditionally used for diagnosis, mainly based on prostate-specific antigen (PSA) testing, and on trans-rectal ultrasound (TRUS)-guided biopsy, often fail to make this distinction, leading to missing diagnosis or overtreatment of indolent cancers [1].

Over the last decade, multiparametric magnetic resonance imaging (mpMRI) has emerged as a powerful solution to this diagnostic gap. It provides detailed anatomical and functional information using a combination of imaging

Saman Fouladi, Fatemeh Darvizeh and Rosario Di Meo contributed equally to this work.

✉ Gabriele Gianini
gabriele.gianini@unimib.it

1 Department of Computer Science, University of Milano, via Celoria 18, Milan 20133, Italy

2 Present address: CDI Centro Diagnostico Italiano S.p.A, via Saint Bon 20, Milan 20147, Italy

3 Bracco S.p.A, Via Egidio Folli, 50, Milan 20134, Italy

4 Department of Informatics, Systems and Communication, University of Milano-Bicocca, viale Sarca 336, Milan 20126, Italy

Published online: 19 May 2026

sequences. T2W imaging highlights the prostate's internal architecture, diffusion-weighted imaging (DWI) assesses tissue cellularity, and dynamic contrast-enhanced (DCE) imaging evaluates vascular properties [2, 3]. Understanding the prostate's internal structure is key to interpreting MRI correctly. The gland is structured into a peripheral zone (PZ) and a central gland (CG), which includes the transition zone (TZ) and central zone (CZ), and can be more challenging to evaluate because of its complex and variable appearance on MRI [4]. The majority of prostate cancers (approximately 70% to 75%) arise in the PZ. The TZ is the second most frequently affected region, accounting for about 20% to 25% of cases. In contrast, the CZ is rarely involved.

To bring structure and consistency to MRI interpretation, the Prostate Imaging Reporting and Data System (PI-RADS) was introduced. Now in its second major revision (PI-RADS v2.1), the system assigns a score from 1 to 5 to describe the likelihood that a lesion is a clinically significant cancer, based on findings across T2W, DWI, and DCE sequences [5]. This classification helps radiologists communicate findings clearly and supports urologists in deciding whether to perform a biopsy or proceed with surveillance [6].

While PI-RADS has greatly improved communication and diagnostic consistency among radiologists, its interpretation still depends heavily on expert visual analysis, which may be subjective and variable. As a result, there has been growing interest in developing computer-aided diagnostic (CAD) tools and AI-based models that can automate or assist in PI-RADS classification, using quantitative features extracted from mpMRI [7].

In this study, we explore three automated approaches for feature extraction aimed at PI-RADS classification, leveraging the individual and combined contributions of the following key sequences:

- T2W images, which provide high-resolution anatomical detail, are especially useful in assessing structural distortion in the transition zone.
- DWI, which captures the diffusion of water molecules in tissue and is highly sensitive to cancer-related changes in cell density.
- ADC maps, derived from DWI, offering a quantitative measure of diffusion restriction, a hallmark of aggressive prostate cancer in the peripheral zone [8].

By analyzing these sequences both independently and in combination, our goal is to develop and evaluate models capable of accurately predicting PI-RADS scores. To this end, we apply a range of Machine Learning (ML)/Deep Learning (DL) techniques, which have shown growing promise in automating radiological assessment and reducing inter-reader variability [9]. We aim to contribute toward more objective, reproducible, and potentially real-time PI-RADS classification tools that support clinical decision-making and improve the consistency of prostate cancer diagnosis.

Related Work

Several studies have investigated the application of radiomic and other computational features for assessing prostate cancer and PI-RADS classification. Many works employed Pyradiomic features extracted from multiparametric MRI to characterize lesions, while others incorporated advanced ML or DL approaches. The classification tasks varied among studies: some focused on differentiating between benign and malignant lesions, others aimed to discriminate between specific PI-RADS categories, such as PI-RADS 4 versus PI-RADS 5, and a few attempted to classify all PI-RADS classes. These studies highlight the potential of radiomic analysis in supporting clinical decision-making, although differences in datasets, feature sets, and classification strategies make direct comparisons challenging. Collectively, they demonstrate the growing interest in leveraging quantitative imaging features for prostate lesion characterization and risk stratification.

The study [10] developed a DL model to classify prostate lesions into PI-RADS categories 2 to 5 using multiparametric MRI sequences (T2W, DWI, ADC) from 687 scans. On a slice basis, the AI agreed with radiologist scores with 58% accuracy, and lesion-based agreement was also 58%, showing moderate agreement (kappa = 0.40). Agreement was highest for PI-RADS 5 (80%) and lowest for PI-RADS 2 (6%). The model tended to upgrade lesions, especially in PI-RADS 3 cases. When allowing a ± 1 category margin, accuracy improved to 86%, indicating the model's potential to assist but not fully replace radiologist assessment.

The study by Winkel et al. [11] proposed a DL algorithm (DLA) for automated PI-RADS classification (1–5) of focal prostate lesions using bi-parametric MRI (T2W and DWI with ADC maps). A total of 121 patients who underwent mpMRI and biopsy were included. The DLA was compared with five radiologists of varying experience. For detecting clinically significant prostate cancer (Gleason ≥ 7), the DLA achieved an AUROC of 0.83, outperforming less experienced radiologists but underperforming compared to an expert (AUROC 0.91). At PI-RADS ≥ 4, the DLA showed 76.7% sensitivity and 85.9% specificity. While the DLA matched average clinical performance, it did not improve inter-reader variability.

In [12], the authors evaluated 252 PI-RADS 1–5 lesions from 188 patients using bi-parametric MRI (T2W and ADC maps). The goal was to improve PI-RADS-based classification by incorporating mean ADC (mADC) values and radiomic features. Lesions initially rated PI-RADS 3–5 were selectively downgraded, and PI-RADS 1–2 lesions were upgraded based on quantitative thresholds. The "Down (ADC)" method improved specificity (from 39.2% to 56.7%) with a slight decrease in sensitivity. Radiomic-based classification achieved similar performance. These methods helped refine PI-RADS assessment without needing full ML models.

Radiomic features extracted using PyRadiomics from T2W and ADC MRI sequences, combined with clinical data, were employed to improve the detection of clinically significant prostate cancer (Gleason score ≥ 7) in PI-RADS 4 and 5 lesions. ML models trained on 111 lesions from 99 patients, with 71% confirmed as significant, achieved an overall accuracy of 79%, with sensitivities and specificities of approximately 79% and 80%, respectively. When focusing on peripheral zone lesions, the accuracy increased to 84%, and sensitivity reached 86%, indicating that integrating radiomic and clinical information enhances classification performance in high PI-RADS categories [13].

The study [14] investigated a ML-based analysis of MRI radiomics to improve the diagnostic performance of PI-RADS v2 in identifying clinically significant prostate cancer. The dataset included MRI scans from several patients (exact number depending on the specific study) using multiparametric MRI sequences such as T2W, DWI, and ADC maps. The model classified prostate lesions between clinically significant prostate cancer (PI-RADS categories 4 and 5) and non-clinically significant or benign cases (PI-RADS categories 1 to 3). The combined model achieved an AUC of 0.85, sensitivity of 83%, and specificity of 78%, outperforming PI-RADS v2 alone and improving diagnostic accuracy for clinically relevant prostate cancer.

In a retrospective study of 203 patients (141 training, 62 validation), radiomic features were extracted from T2W, DWI, ADC, and DCE MRI using PyRadiomics [15]. A ML–based Rad-score was developed and combined with PI-RADS v2.1. The combined model outperformed PI-RADS alone in detecting prostate cancer: AUC 0.99 vs. 0.90 (training) and 0.931 vs. 0.845 (validation). It also improved sensitivity (from 92.3% to 79.4%) and specificity (from 98.4% to 96.4%), particularly for PI-RADS 3 and peripheral zone lesions. Adding PSA did not improve performance. Radiomics helped address the limitations of PI-RADS in terms of specificity and lesion stratification.

In a multicenter study, Zhu et al. [9] analyzed 1,186 lesions from 927 patients to predict clinically significant prostate cancer using radiomics. Lesions were classified between PI-RADS 3, 4, and 5 using a logistic regression model. The radiomics model achieved AUCs of 0.85 (training), 0.87 (internal test), and 0.83–0.85 (external validation cohorts). Accuracy ranged from 75.5% to 81.4%, with sensitivities up to 89.2%. Feature selection was performed using Least Absolute Shrinkage and Selection Operator (LASSO), and the model combined radiomics features with radiologist-assigned PI-RADS scores. Results showed improved detection of clinically significant cancer, especially for PI-RADS 3 lesions, supporting its potential for real-world clinical use.

In the study [16], 615 patients were classified into clinically significant (grade group ≥ 2) and non-significant/benign lesions using a deep radiomics pipeline that combined nnU-Net segmentation and an XGBoost classifier. The model achieved a patient-level AUROC of 0.91, with 90% sensitivity and 73% specificity, comparable to PI-RADS ≥ 3 (AUROC 0.94). However, lesion-level sensitivity was lower (68% vs. 84% for PI-RADS). Results suggest deep radiomics may complement PI-RADS but not replace it in clinical practice.

A multiparametric MRI radiomic approach was evaluated on 102 patients with PI-RADS 3 and upgraded PI-RADS 4 lesions. Radiomic features from T2W and ADC images were analyzed using a random forest classifier to distinguish clinically significant prostate cancer (Gleason score ≥ 7) from non-significant cases. The model achieved an AUC of 0.82, demonstrating improved stratification compared to PI-RADS alone. Feature selection and image preprocessing were applied to enhance performance. This method focused on refining classification within the PI-RADS 3 and 4 categories [17].

A radiomics study on 90 subjects performed multiclass classification of prostate tumors using PI-RADS 2, 3, 4, and 5 lesions on multiparametric MRI. From each lesion, 609 texture features were extracted, and eight classifiers were evaluated; the Naïve Bayes model using 120 selected features achieved the highest mean AUC of 0.744 ± 0.088, while a Random Forest classifier with 52 features achieved an AUC of 0.739 ± 0.096. Although detailed accuracy, sensitivity, and specificity were not reported, the performance indicates moderate discrimination across all four PI-RADS classes [18].

The study [19] developed a semi-automated ML system for PI-RADS v2.1 scoring using multiparametric MRI data from 59 patients (PI-RADS 2:18, 3:10, 4:16, 5:15). Two classification approaches were evaluated: a multiclass classification across four PI-RADS categories (2, 3, 4, 5) and a binary classification grouping low-risk (PI-RADS 2 + 3) versus high-risk (PI-RADS 4 + 5) lesions. The system combined prostate segmentation, 3D co-registration, and lesion ROI extraction with classifiers including LDA, linear SVM,

and Gaussian SVM. For multiclass classification, the best model achieved 88.0% ± 0.98% accuracy and an AUC of 0.94. The binary classifier performed even better, reaching 93.2% ± 2.1% accuracy with an AUC of 0.99.

The study [20] used data from 91 patients across three centers to evaluate a commercially available DL–based AI algorithm for automated PI-RADS v2.1 score assignment and lesion detection on multiparametric MRI. The AI achieved lesion-level sensitivity of 81% and specificity of 78% for clinically significant prostate cancer (PI-RADS ≥ 4), comparable to radiologists (sensitivity 90%, specificity 70%). The method employed convolutional neural networks trained on large datasets for lesion segmentation and classification.

Multiparametric MRI data from 346 patients were used to develop a DL ensemble model combining ResNet and DenseNet architectures for classifying clinically significant prostate cancer (csPCa, PI-RADS ≥ 4) versus non-significant cases (PI-RADS ≤ 3). ResNet al.one achieved 87% accuracy, DenseNet 85%, while the ensemble improved performance to 89% accuracy, with 91% sensitivity and 87% specificity [21].

The study [22] Involving 453 patients with PSA levels between 4 and 10 ng/mL, this study evaluated the diagnostic performance of combining PI-RADS v2.1 scores and prostate-specific antigen density (PSAD) for detecting clinically significant prostate cancer (csPCa). The analysis revealed that both PI-RADS v2.1 and PSAD were independent predictors of csPCa. A logistic regression model incorporating both factors achieved an area under the ROC AUC of 0.935, indicating excellent diagnostic accuracy. The model demonstrated improved sensitivity and specificity compared to using either PI-RADS or PSAD alone. Notably, patients with a PI-RADS score ≤ 2 or a PI-RADS score = 3 combined with a PSAD value ≤ 0.33 ng/mL/mL had low detection rates for csPCa, suggesting that biopsies in these cases may be unnecessary.

Using 287 PI-RADS 3 lesions, logistic regression, SVM, XGBoost, and random forest classifiers were evaluated to distinguish clinically significant prostate cancer (csPCa) from non-significant or benign lesions. Random forest showed superior performance with an AUC of 0.832 (internal) and 0.688 (external), sensitivity of 87%, and specificity of 50%. Clinical and radiological features were combined to reduce unnecessary biopsies [23].

A radiomics-based ML model using 107 features from T2W MRI was developed to classify clinically significant prostate cancer (csPCa) versus non-significant lesions within equivocal PI-RADS 3 cases. A random forest classifier achieved an AUC of 0.76, outperforming clinical parameters like PSA density (AUC 0.61) and prostate volume (AUC 0.62). This approach improves diagnostic accuracy, specifically for stratifying csPCa in PI-RADS 3 lesions [24].

Materials and methods

In this study, we utilized a prostate MRI dataset comprising three key sequences: ADC, T2W, and DWI. To ensure consistency and enhance the quality of the data, we applied a comprehensive preprocessing pipeline, which included normalization, registration, resizing, and other standard steps. For PI-RADS classification, we implemented three distinct approaches, each leveraging different aspects of the imaging data to differentiate between PI-RADS categories.

Subsequently, we employed a set of ML models to perform the classification tasks, allowing us to evaluate the effectiveness of the proposed approaches and compare their performance across different models.

In the following sections, we provide a detailed description of the dataset and preprocessing steps, followed by an explanation of the three classification approaches, and finally outline the ML models used along with the evaluation strategy applied to assess their performance.

Dataset Description

In this study, a dedicated prostate MRI dataset was developed by systematically collecting and processing raw multiparametric scans. The scans were originally acquired at the Department of Diagnostic Imaging and Stereotactic Radiosurgery, Centro Diagnostico Italiano (CDI), located in Milan, Italy.

MRI data were acquired using a 3T Philips Achieva dStream scanner, with a slice thickness of 3.3 mm and an in-plane resolution of 0.2557×0.2557 mm².

To comply with privacy standards, all DICOM files were anonymized. Using customized Python scripts, we extracted the three key MRI sequences: T2W, ADC, and DWI. During this step, DWI sequences with a b-value of 1600 were specifically selected, as higher b-values are known to enhance lesion visibility and improve contrast between malignant and benign tissue [25, 26]. The corresponding ADC maps were also extracted for these sequences. All images were converted from DICOM to NIfTI format using the dicom-2nifti Python package.

Some cases included multiple lesions (2, 3, or 4 per patient). For PI-RADS classification and model training, each lesion was handled separately, and we trained the models only on images containing a single lesion, allowing for more focused and controlled learning.

Table 1 Summary of MRI Dataset by PI-RADS Score and Lesion Location

Number of	PI-RADS 3	PI-RADS 4	PI-RADS 5	Located in PZ	Located in CG	Located in Both
Training Set	91	196	76	283	60	20
Test Set	12	22	15	35	9	5

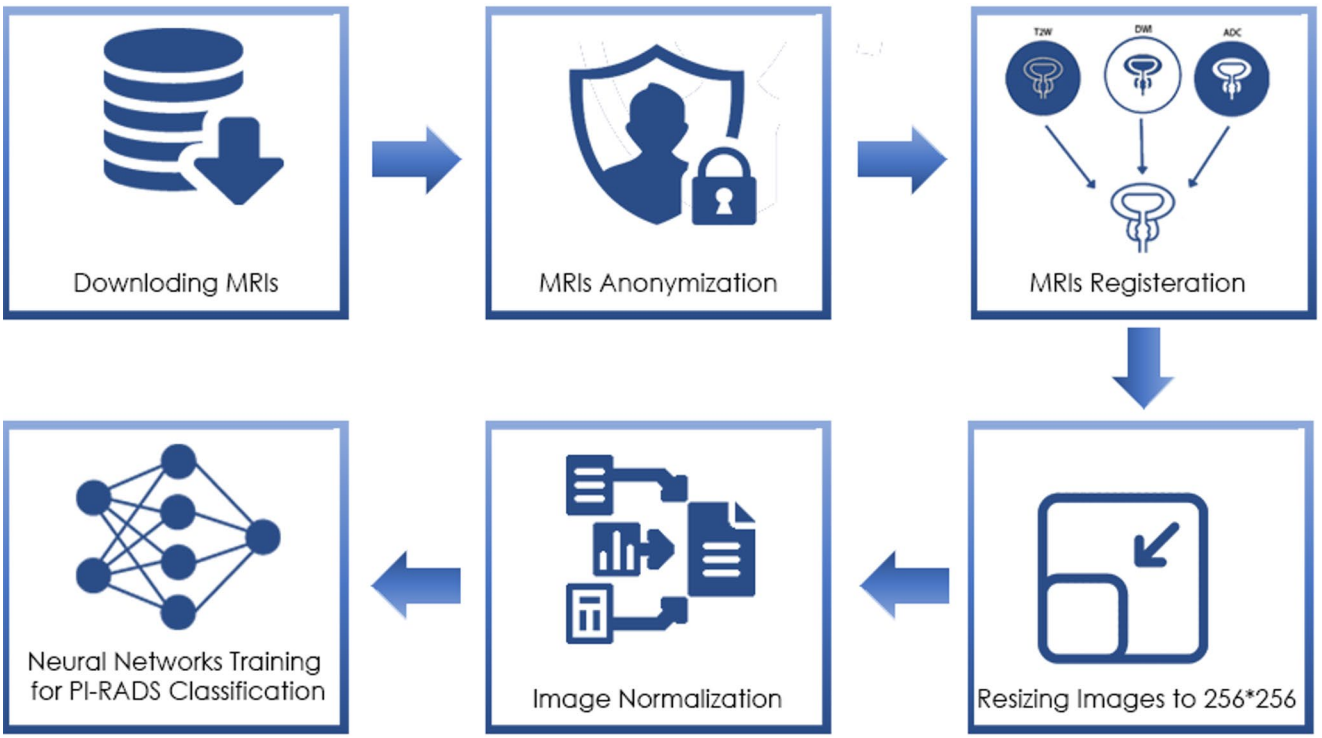


Fig. 1 Overview of the MRI dataset preprocessing pipeline

Lesion segmentation was initially performed using 3D Slicer [27] by a trained operator, under the supervision of a certified radiologist. Lesions were manually annotated on all three sequences (T2W, ADC, and DWI) to ensure complete lesion characterization. Two experienced radiologists then independently reviewed and refined these annotations. The final segmentations served as the ground truth for evaluating lesion segmentation models, ensuring high consistency and expert-level accuracy.

Our dataset consists of 412 MRI images, each containing one lesion, with 363 used for training and 49 for testing. In the training set, there are 91 MRIs with PI-RADS 3, 196 with PI-RADS 4, and 76 with PI-RADS 5. Regarding lesion location, 283 lesions are located in the PZ, 60 in the CG, and 20 in either the intermediate or both regions. In the test set, there are 12 MRIs with PI-RADS 3, 22 with PI-RADS 4, and 15 with PI-RADS 5. Concerning lesion location, 35 lesions are in the PZ, 9 in the CG, and 5 in either the intermediate or both regions.

A summary of these details is provided in Table 1.

Dataset Preprocessing Pipeline

Figure 1 demonstrates the process of the dataset preprocessing pipeline. Initially, MRI images were downloaded and underwent anonymization, during which all patient-identifiable information (e.g., names, IDs, and metadata) was removed to ensure privacy and compliance with data protection standards. Then, given the differences in resolution, orientation, and field of view between MRI sequences and potential artifacts due to patient motion, careful preprocessing and alignment were required before model training [28].

We used the SimpleITK [29] Python library for rigid and affine registration of the ADC and DWI images to the T2W sequence, which was selected as the fixed reference. Mutual information was used as the similarity metric during the optimization process. A multi-resolution strategy was applied to improve robustness by refining alignment across multiple image scales [30]. After registration, all sequences were resampled to match the T2W image dimensions and resolution, ensuring spatial consistency across modalities.

Before registration, the images were normalized and standardized [31]. All were converted to a common format, and their intensity values were adjusted for consistency. After registration and alignment, all images were resized to 256 × 256 pixels to meet the input size requirements of the neural networks.

To address differences in intensity distributions across modalities and scanners, modality-specific normalization techniques were applied [32]:

- T2W and ADC images were normalized using Min-Max scaling, rescaling intensities to the [0, 1] range.
- DWI images were processed using Z-score normalization, where each voxel intensity was standardized by subtracting the mean and dividing by the standard deviation.

These preprocessing steps helped improve both training stability and overall model performance in multi-modal medical image analysis.

PI-RADS Classification Approaches

For PI-RADS classification, we implemented three complementary approaches, each leveraging different strategies to extract informative features from multiparametric MRI scans. In the first approach, we extracted hand-crafted radiomic features from manually segmented lesions using the PyRadiomics library, capturing intensity, shape, and texture patterns from multiple MRI sequences. The second approach extended this pipeline by incorporating fully automated lesion and prostate zone segmentations, enabling a more clinically realistic workflow without manual input. In the third approach, we employed a custom CNN to automatically learn and extract high-level features from ADC images and their corresponding lesion masks.

Features obtained from each approach were then used to train ML models for multi-class classification of PI-RADS scores, allowing us to compare the effectiveness of manual, automated, and deep-learning-based feature extraction strategies.

Approach 1: Extracting Features using PyRadiomics

In this study, radiomic features were extracted from multiparametric MRI scans using the PyRadiomics library, an open-source Python package designed for standardized and reproducible radiomics analysis [33]. Radiomics is a computational method that transforms medical images into high-dimensional, mineable data by quantifying patterns, textures, shapes, and intensities within defined regions of interest (ROIs) [34]. This allows for the extraction of hundreds to thousands of features that can capture subtle tissue heterogeneity, which may not be visible to the human eye [3].

Using the PyRadiomics package [33], we extracted features from clinically relevant MRI sequences, T2W, ADC, and DWI, for each lesion.

In the first examination, we extracted features from the following filtered image types: Original, LoG (Laplacian of Gaussian, with sigma values of 2.0, 3.0, 4.0, and 5.0), Wavelet, Square, SquareRoot, and Logarithm. From each image type, we computed a comprehensive set of features, including first-order statistics, shape-based features, and texture features such as those from the Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Gray Level Dependence Matrix (GLDM), and Neighboring Gray Tone Difference Matrix (NGTDM). The extraction settings were: binWidth = 25, resampledPixelSpacing= [1], interpolator='sitkBSpline', normalize=True, removeOutliers = 3, label = 1, correctMask=True, force2D=False, and preCrop=True, we refer to these as *Feature Set 1*. Feature Set 1 initially consisted of 49 features, which were reduced to 38 features after applying LASSO feature selection, improving the relevance and effectiveness of the selected features.

In the second examination, we expanded the feature set by including more image filters: Exponential, Gradient, LBP2D, and LBP3D, in addition to the previously used filters (Original, LoG with an extended range from 1.0 to 5.0, Wavelet, Square, SquareRoot, and Logarithm). This enhanced configuration allowed the extraction of a broader spectrum of intensity, shape, and texture patterns. The same core feature classes were calculated, first-order, shape, GLCM, GLRLM, GLSZM, GLDM, and NGTDM, but on a more diverse set of filtered images. Additional preprocessing settings included normalizeScale = 100, providing a uniform scale across different modalities and scanners, and we refer to these as *Feature Set 2*. Feature Set 2 originally included 110 features, which were refined to 63 after applying LASSO feature selection, resulting in a more compact and informative feature set.

These radiomic features were then used to classify lesions into PI-RADS categories 3, 4, and 5, aiming to support the radiological scoring system with quantitative data. In addition to the image-derived radiomic features, we also included the anatomical location of each lesion as an additional input feature. Specifically, we considered whether the lesion was located in the PZ, CG, or between both zones

simultaneously. To incorporate this information into the model, we assigned numerical labels: 1 for PZ, 2 for CG, and 3 for lesions located across both zones. This spatial context helped enrich the feature set with anatomical relevance, potentially improving the classification of PI-RADS scores. Zone-specific characteristics can significantly influence lesion appearance and are relevant to PI-RADS scoring [35].

By comparing the results from both extractions, we aimed to assess whether the inclusion of more advanced filters and descriptors improves the classification of PI-RADS scores. Notably, the second configuration allowed us to capture more complex patterns, especially from texture features like LBP (Local Binary Patterns) and Gradient, which are known to enhance performance in certain classification tasks [36].

Finally, all extracted features were used to train various ML models for multi-class classification of PI-RADS scores (3, 4, and 5). The models were evaluated on an independent test set, consisting of data that had not been seen during training. Hyperparameter tuning was conducted exclusively on the training data using cross-validation. The test set was not involved in any stage of model training or tuning and was reserved only for the final performance assessment. To address class imbalance, class weights were applied during model training, assigning higher importance to underrepresented classes.

For each test sample, both the lesion masks and the anatomical images in which the lesion was located were available, allowing for the accurate extraction and validation of region-specific radiomic features.

Approach 2: Feature Extraction from Automated Zone and Lesion Segmentations

In the first approach, radiomic features were extracted from manually segmented lesions, and the lesion's zone was incorporated as an additional feature to enhance PI-RADS classification. This method depended on the availability of ground-truth lesion masks and zone labels in the test dataset, which were used both for feature extraction and model evaluation.

In the second approach, designed to better replicate a real clinical workflow, the test dataset was processed through a fully automated pipeline utilizing previously developed segmentation models [37]. Specifically, the zonal segmentation model, published earlier [38, 39], was applied to automatically identify prostate zones on T2W images. Following this, the lesion segmentation model, trained on the same dataset and utilizing ADC images, was used to segment lesions automatically [40, 41]. To assign the zonal label to each lesion in an objective and reproducible manner, a custom Python script was implemented to calculate the spatial intersection between the automatically segmented lesion masks and zone masks. This procedure enabled automated, precise determination of the lesion's prostate zone without manual intervention.

Radiomic feature extraction was then performed in the same way as in the first approach, but now based on fully automated lesion and zone segmentations. Finally, the PI-RADS classification model was evaluated using this end-to-end automated pipeline, allowing assessment of classification performance in a realistic clinical scenario where both segmentation and zone assignment are performed without manual input.

Figure 2 depicts the step-by-step process of Approach 2 used for PI-RADS classification.

Feature Extraction Using Proposed CNN (Approach 3)

In another approach, we extracted features using a purposely trained Convolutional Neural Network (CNN) and then applied our ML models for PI-RADS classification.

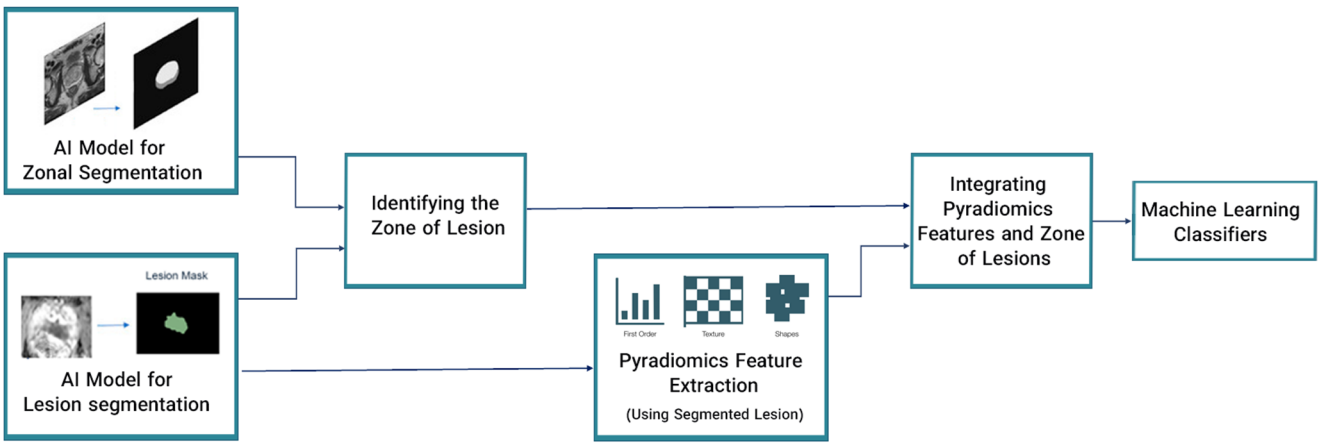


Fig. 2 Workflow of the fully automated pipeline (approach 2) for PI-RADS classification

For training the CNN, we used ADC images together with their corresponding masks as input, meaning each image and its mask were passed to the network simultaneously. This paired-input strategy allowed the network to learn both visual and spatial context, focusing more effectively on lesion regions [42, 43].

The CNN architecture consisted of three convolutional layers with dropout layers to mitigate overfitting. Before training, we resized both images and masks to 256 × 256 pixels, applied center cropping to retain only the prostate region, and then resized them to 128 × 128 pixels. To improve generalization, we applied data augmentation during training, including image rotations at various angles, horizontal and vertical flips, and small translations to simulate variations in patient positioning and acquisition.

The network was trained for 200 epochs using a learning rate of 1e-4 and a batch size of 32. The convolutional layers used a kernel size of 3 with ReLU activation functions, and the model was optimized using categorical cross-entropy loss. After training, features were extracted from the last dense layer and used as inputs to the ML classifiers for PI-RADS classification.

Our CNN-based approach is intended as a lightweight baseline for learned features rather than a full deep learning alternative to radiomics. We focused on ADC images, which consistently showed the highest performance in previous studies, and used relatively small, cropped images (128 × 128), which limited the effectiveness of very deep architectures. While deeper networks such as ResNet-50 were tested, they did not outperform our simple CNN. Additionally, to ensure a fair comparison with previous approaches, the same classical machine learning classifiers were used. This design allows our CNN to provide meaningful learned features while maintaining a consistent and comparable evaluation framework.

Machine Learning Classifiers

In all three approaches for PI-RADS classification, we evaluated multiple ML classifiers, including Random Forest (RF) [44], Support Vector Machine (SVM) [45], Extreme Gradient Boosting (XGBoost) [46], AdaBoost [47], Gradient Boosting [48], and LightGBM [49]. We also developed four ensemble models. Ensemble 1 utilized logistic regression as a meta-learner, combining the outputs of XGBoost and k-Nearest Neighbors (KNN). Ensemble 2 replaced the meta-learner with XGBClassifier, again combining XGBoost and KNN. Ensemble 3 implemented a voting strategy with XGBoost and KNN, while Ensemble 4 applied voting with XGBoost and RF. These combinations were selected after testing several model pairings and were found to achieve the best classification performance in our experiments.

All classifiers were trained using 5-fold cross-validation [50] to ensure robust performance estimation. Before training, feature values were standardized using StandardScaler [51] to remove the mean and scale variance to unit standard deviation, which is essential for distance-based methods like SVM and KNN. We also applied the LASSO feature selection [52] to reduce dimensionality by penalizing less informative features. Classifiers were trained both with and without feature selection; better results were obtained when LASSO was applied.

Briefly, RF is an ensemble of decision trees that reduces overfitting through bootstrap aggregation; SVM is a kernel-based classifier that maximizes the margin between classes; XGBoost, LightGBM, AdaBoost, and Gradient Boosting are gradient-boosting algorithms that iteratively build ensembles of weak learners to minimize prediction error, with XGBoost and LightGBM optimized for speed and regularization; and KNN is a distance-based method that classifies based on the majority class of the nearest neighbors. Logistic Regression was used in Ensemble 1 as a linear meta-learner to combine base classifier predictions, while XGBClassifier in Ensemble 2 provided a nonlinear combination. The voting ensembles (Ensemble 3 and Ensemble 4) aggregated classifier outputs through majority voting to improve generalization.

Results

In the following, we present the results obtained for all three proposed approaches. For Approach 1, radiomic features were extracted from all three MRI sequences (T2W, ADC, and DWI), and the features from these sequences were subsequently combined. This process yielded four sets of results. The extracted features were then used to train ML models for PI-RADS classification into three classes: PI-RADS 3, PI-RADS 4, and PI-RADS 5.

As described in detail in Approach 1, two distinct categories of features were extracted, and the results obtained using each feature category are reported separately. Model performance was evaluated using precision, recall, and AUC for each PI-RADS class. Additionally, the average accuracy and average AUC across all classes are reported.

For clarity, results for each approach are presented in separate sections.

Results of Approach 1 Using PyRadiomics Feature Set 1

The results obtained using Feature Set 1 for different input configurations are reported in Appendix A (Tables 8, 9, 10 and 11). As these results were lower than those obtained with Feature Set 2, they are presented in the appendix for completeness. For each configuration, performance is reported in terms of precision, recall, and AUC for each PI-RADS class, as well as average accuracy and average AUC across all classes.

According to the obtained results, the following trends were observed. Overall, models trained on ADC features achieved higher classification performance compared to those trained solely on DWI or T2W, with the highest accuracy observed for Ensemble 1 on ADC (0.77), accompanied by an AUC of 0.83. In contrast, DWI- and T2W-based models showed lower accuracy values, generally in the range of 0.49–0.63, although certain configurations achieved competitive AUC values above 0.80 (e.g., AdaBoost on DWI with 0.81).

Combining features from all sequences provided balanced performance, with several ensemble methods and gradient boosting models achieving accuracies above 0.65 and AUCs above 0.81. Notably, Ensemble 3 with combined features reached an accuracy of 0.69 and maintained strong per-class AUC values, particularly for PI-RADS 5 (AUC$\geq$0.94 in most cases).

Across all configurations, classification for PI-RADS 5 tended to be the most accurate and consistent (AUCs often above 0.94), whereas PI-RADS 3 classification was more challenging, with lower precision and recall, especially for DWI and T2W inputs. These trends highlight both the strength of ADC-derived features and the benefits of combining multi-sequence information for more robust PI-RADS classification.

Results of Approach 1 Using PyRadiomics Feature Set 2

Tables 2, 3, 4 and 5 summarize the results obtained with Feature Set 2 across different input configurations: ADC (Table 2), DWI (Table 3), T2W (Table 4), and the combination of all sequences (Table 5). For each configuration, model performance is reported in terms of precision, recall,

Table 2 Performance of PI-RADS Classification Models of Approach 1, Using PyRadiomics Feature Set 2 Across ADC Images

	Acc.	AUC.	PI-RADS 3			PI-RADS 4			PI-RADS 5		
			Prec.	Recall	AUC	Prec.	Recall	AUC	Prec.	Recall	AUC
RF	0.65	0.84	0.42	0.50	0.77	0.62	0.68	0.774	**1.00**	0.73	0.97
XGBoost	0.71	0.86	0.53	0.66	0.80	0.70	0.63	0.79	0.92	**0.86**	0.98
AdaBoost	0.69	0.82	0.55	**0.83**	0.79	0.66	0.63	0.69	**1.00**	0.66	0.98
Gradient Boosting	0.71	0.83	0.53	0.58	0.76	0.68	0.77	0.75	**1.00**	0.73	0.97
LightGBM	0.71	0.84	**0.58**	0.58	0.79	0.66	0.72	0.74	0.92	0.80	0.97
SVM	0.67	0.83	0.50	0.41	0.75	0.63	**0.86**	0.77	**1.00**	0.60	0.96
Ensemble 1	0.73	**0.87**	0.57	0.66	**0.83**	0.69	0.72	**0.81**	**1.00**	0.80	0.98
Ensemble 2	0.71	0.85	0.46	0.58	0.78	0.71	0.68	0.77	**1.00**	**0.86**	**0.99**
Ensemble 3	**0.75**	0.86	0.57	0.66	0.79	**0.72**	0.72	0.80	**1.00**	**0.86**	0.98
Ensemble 4	0.71	0.85	0.53	0.66	0.80	0.70	0.63	0.77	0.92	**0.86**	0.97

Table 3 Performance of PI-RADS Classification Models of Approach 1, Using PyRadiomics Feature Set 2 Across DWI Images

	Acc.	AUC.	PI-RADS 3			PI-RADS 4			PI-RADS 5		
			Prec.	Recall	AUC	Prec.	Recall	AUC	Prec.	Recall	AUC
RF	0.63	0.77	0.37	0.25	**0.69**	0.58	0.77	0.67	**0.91**	**0.73**	0.95
XGBoost	**0.65**	0.73	**0.50**	**0.41**	0.64	**0.60**	0.77	0.62	0.90	0.66	0.92
AdaBoost	0.49	0.72	0.26	**0.41**	0.62	0.44	0.36	0.56	**0.91**	**0.73**	**0.97**
Gradient Boosting	0.59	0.73	0.33	0.25	0.61	0.55	0.72	0.63	0.90	0.66	0.95
LightGBM	0.63	0.72	0.45	**0.41**	0.62	0.59	0.72	0.63	0.90	0.66	0.92
SVM	0.61	0.74	0.00	0.00	0.62	0.55	**0.90**	0.65	0.90	0.66	0.95
Ensemble 1	0.59	**0.75**	0.00	0.00	0.60	0.54	0.86	**0.70**	0.90	0.66	0.94
Ensemble 2	0.44	0.64	0.26	0.33	0.54	0.45	0.50	0.52	0.70	0.46	0.86
Ensemble 3	0.57	0.68	0.33	0.25	0.55	0.55	0.72	0.58	0.81	0.60	0.92
Ensemble 4	0.57	0.74	0.30	0.25	0.64	0.53	0.68	0.63	0.90	0.66	0.95

Table 4 Performance of PI-RADS Classification Models of Approach 1, Using PyRadiomics Feature Set 2 Across T2W Images

	Acc.	AUC.	PI-RADS 3			PI-RADS 4			PI-RADS 5		
			Prec.	Recall	AUC	Prec.	Recall	AUC	Prec.	Recall	AUC
RF	0.63	0.71	0.40	0.16	0.55	0.56	0.77	0.62	0.85	**0.80**	0.96
XGBoost	0.63	0.68	**0.42**	**0.25**	0.50	**0.57**	0.72	0.58	0.85	**0.80**	0.95
AdaBoost	0.63	**0.77**	0.40	0.16	**0.67**	0.56	0.77	0.66	0.85	**0.80**	0.97
Gradient Boosting	0.61	0.67	0.28	0.16	0.47	0.55	0.72	0.58	**0.92**	**0.80**	**0.98**
LightGBM	0.57	0.68	0.25	0.16	0.51	0.51	0.63	0.56	0.85	**0.80**	0.96
SVM	0.61	0.70	0.00	0.00	0.61	0.53	**0.95**	0.59	0.90	0.60	0.89
Ensemble 1	**0.65**	0.71	0.00	0.00	0.62	**0.57**	0.90	0.63	**0.92**	**0.80**	0.88
Ensemble 2	0.53	0.76	0.00	0.00	0.66	0.50	0.72	**0.67**	0.90	0.66	0.95
Ensemble 3	0.63	0.70	0.33	0.16	0.54	0.56	0.77	0.59	**0.92**	**0.80**	0.95
Ensemble 4	0.55	0.71	0.22	0.16	0.55	0.50	0.59	0.60	0.85	**0.80**	0.97

Table 5 Performance of PI-RADS Classification Models of Approach 1, Using PyRadiomics Feature Set 2 Across Combined Features of ADC, DWI, and T2W Images

	Acc.	AUC.	PI-RADS 3			PI-RADS 4			PI-RADS 5		
			Prec.	Recall	AUC	Prec.	Recall	AUC	Prec.	Recall	AUC
RF	0.69	0.83	0.50	0.66	0.76	**0.68**	0.59	0.75	0.92	**0.86**	0.98
XGBoost	0.67	0.80	0.47	0.66	0.73	0.66	0.54	0.72	0.92	**0.86**	0.95
AdaBoost	0.67	0.81	0.50	**0.75**	0.77	0.65	0.59	0.68	**1.00**	0.73	**0.98**
Gradient Boosting	0.67	0.78	0.46	0.58	0.71	0.65	0.59	0.70	0.92	**0.86**	0.94
LightGBM	0.67	0.79	0.50	0.50	0.71	0.62	0.68	0.70	0.92	0.80	0.95
SVM	0.59	0.78	0.50	0.25	0.73	0.52	**0.86**	0.65	**1.00**	0.46	0.97
Ensemble 1	0.69	**0.84**	**0.54**	0.50	**0.80**	0.64	0.72	**0.75**	0.92	0.80	0.98
Ensemble 2	**0.71**	0.83	0.53	0.58	0.79	0.68	0.68	0.73	0.92	**0.86**	0.97
Ensemble 3	0.69	0.83	**0.54**	0.50	0.79	0.64	0.72	0.74	0.92	0.80	0.98
Ensemble 4	0.67	0.82	0.46	0.58	0.75	0.65	0.59	0.73	0.92	**0.86**	0.97

and AUC for each PI-RADS class, along with the average accuracy and average AUC across all classes.

Evaluation of PyRadiomics Feature Set 2 across different input configurations revealed that ADC-based models generally outperformed those trained solely on DWI or T2W images. Among the ADC configurations, the highest overall accuracy was obtained with Ensemble 3 (0.75, AUC = 0.86), closely followed by Ensemble 1 and several gradient boosting variants. For these models, PI-RADS 5 classification remained consistently strong (AUC ≥ 0.97), while PI-RADS 3 continued to be the most challenging category, showing notably lower precision and recall.

DWI-based models showed moderate performance, with accuracy values often in the 0.57–0.65 range and fewer configurations exceeding an AUC of 0.75. AdaBoost and Ensemble 2 performed the weakest for this modality, whereas XGBoost and RF achieved relatively more balanced results. Similarly, T2W-based models tended to underperform compared to ADC and combined configurations, with most accuracies between 0.57 and 0.65, although certain ensemble methods maintained reasonable AUC values for PI-RADS 5.

When combining features from ADC, DWI, and T2W, performance improved over single-sequence DWI or T2W models and approached that of ADC-only configurations. Notably, Ensemble 2 on combined features achieved an accuracy of 0.71 with an AUC of 0.83, indicating the benefit of multi-sequence information for robust classification. Across all modalities and configurations, classification of PI-RADS 5 was consistently reliable, whereas PI-RADS 3 remained less accurately predicted, reflecting the inherent difficulty of differentiating intermediate cases.

Results of Approach 2 Using PyRadiomics Features

In the second approach, the entire workflow, from zonal and lesion segmentation to zone assignment, was performed automatically using previously developed segmentation models and a custom spatial overlap algorithm. Radiomic features were then extracted from the automated segmentations, and PI-RADS classification was carried out as in the first approach.

For Feature Set 1 in Approach 2, the results obtained using ADC images are reported in Appendix B (Table 12), as they are less relevant compared to the main results, while Table 6 shows the corresponding results for Feature Set 2.

Table 6 Performance of PI-RADS Classification Models of Approach 2, Using PyRadiomics Feature Set 2 Across ADC Images

	Acc.	AUC.	PI-RADS 3			PI-RADS 4			PI-RADS 5		
			Prec.	Recall	AUC	Prec.	Recall	AUC	Prec.	Recall	AUC
RF	0.60	0.81	0.16	0.16	0.71	0.63	0.50	**0.81**	0.75	**0.92**	0.91
XGBoost	0.66	0.82	**0.37**	**0.50**	0.74	0.70	0.50	0.78	0.80	**0.92**	0.93
AdaBoost	**0.72**	0.81	0.33	0.16	0.73	0.64	**0.78**	0.76	**0.92**	**0.92**	**0.93**
Gradient Boosting	0.63	0.80	0.28	0.33	0.72	0.77	0.50	0.74	0.70	**0.92**	0.92
LightGBM	0.69	0.81	**0.37**	**0.50**	0.72	**0.80**	0.57	0.77	0.80	**0.92**	0.92
SVM	0.57	0.78	0.16	0.16	0.66	0.57	0.57	0.77	0.76	0.76	0.91
Ensemble 1	0.63	**0.82**	0.33	0.16	**0.76**	0.55	0.71	0.79	0.83	0.76	0.91
Ensemble 2	0.48	0.64	0.00	0.00	0.58	0.46	0.50	0.58	0.64	0.69	0.76
Ensemble 3	0.60	0.81	0.28	0.33	0.74	0.53	0.50	0.77	0.84	0.84	0.92
Ensemble 4	0.66	0.81	0.33	**0.50**	0.71	0.77	0.50	0.79	0.80	**0.92**	0.91

Table 7 Performance of PI-RADS Classification Models of Approach 3, Using Extracted Features by Proposed CNN Across ADC Images

	Acc.	AUC.	PI-RADS 3			PI-RADS 4			PI-RADS 5		
			Prec.	Recall	AUC	Prec.	Recall	AUC	Prec.	Recall	AUC
RF	0.65	0.84	0.37	0.25	**0.78**	0.60	0.77	0.77	0.92	0.80	**0.96**
XGBoost	0.67	0.84	0.42	0.25	0.77	0.61	0.86	0.78	**1.00**	0.73	0.96
AdaBoost	0.53	0.82	**1.00**	0.08	0.76	0.48	**1.00**	0.74	**1.00**	0.20	0.96
Gradient Boosting	0.63	0.81	0.37	0.25	0.73	0.60	0.77	0.77	0.84	0.73	0.94
LightGBM	0.69	0.82	0.42	0.25	0.73	0.63	0.86	**0.79**	**1.00**	0.80	0.94
SVM	0.63	0.72	0.50	0.33	0.58	0.58	0.63	0.67	0.76	**0.86**	0.92
Ensemble 1	0.59	0.72	0.26	0.16	0.55	0.53	0.63	0.67	0.81	**0.86**	0.92
Ensemble 2	0.67	0.80	0.46	**0.50**	0.76	**0.66**	0.72	0.74	0.91	0.73	0.91
Ensemble 3	0.69	0.79	0.57	0.33	0.70	0.62	0.90	0.74	**1.00**	0.66	0.94
Ensemble 4	**0.73**	**0.84**	0.66	0.33	0.77	0.65	0.95	**0.79**	**1.00**	0.73	0.96

For Feature Set 1 (Table 12), the highest accuracy (0.69) was achieved by XGBoost, LightGBM, and Ensemble 3, while Gradient Boosting obtained the highest overall AUC (0.84). XGBoost also showed strong class-specific performance, particularly for PI-RADS 4 (precision = 0.66, recall = 0.57, AUC = 0.78) and PI-RADS 5 (precision = 0.85, recall = 0.92, AUC = 0.98).

For Feature Set 2 (Table 6), AdaBoost achieved the highest accuracy (0.72) and a competitive AUC (0.81). LightGBM and XGBoost also performed consistently well, especially for PI-RADS 5 (AUC = 0.92 and 0.93, respectively). However, Ensemble 2 had the weakest performance, with both accuracy (0.48) and AUC (0.64) being the lowest.

Overall, switching from Feature Set 1 to Feature Set 2 led to improved accuracy in some models (e.g., AdaBoost) but not universally. XGBoost and LightGBM were consistently strong performers across both feature sets, while SVM and Ensemble 2 tended to lag in this fully automated pipeline.

Results of Approach 3

In Approach 3, we took a completely different direction by not using PyRadiomic features. Instead, we extracted features using the proposed CNN, trained on paired ADC images and masks, and then used these features as input to ML models for PI-RADS classification. Table 7 presents the results of this approach.

According to Table 7, Ensemble 4 achieved the best performance, with an accuracy of 0.73 and an AUC of 0.84, showing particularly strong results for PI-RADS 4 and PI-RADS 5. Other models, such as LightGBM and Ensemble 3, also performed competitively, although performance for PI-RADS 3 remained consistently lower across all models, indicating this class as the most challenging to classify.

Discussion

Multiparametric MRI enhances prostate tumor detection by combining anatomical information from T2W images with quantitative data from DWI and dynamic contrast-enhanced sequences [4].

PI-RADS v2, and its updated version PI-RADS v2.1, provide a standardized framework for assessing the risk of clinically significant prostate cancer based on lesion morphology and imaging appearance, enabling consistency across institutions [2]. However, assigning a PI-RADS score still depends on subjective evaluations of qualitative lesion

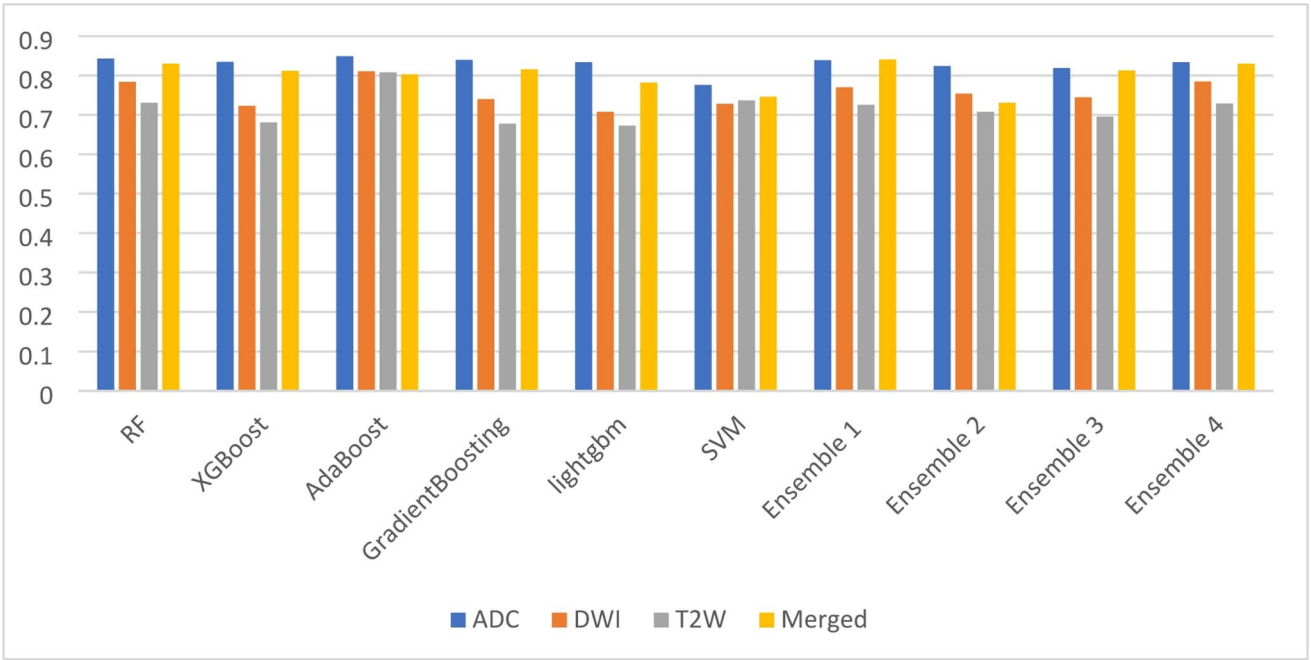


Fig. 3 Comparison of the AUCs achieved by all models in Approach 1 using Feature Set 1 for PI-RADS classification

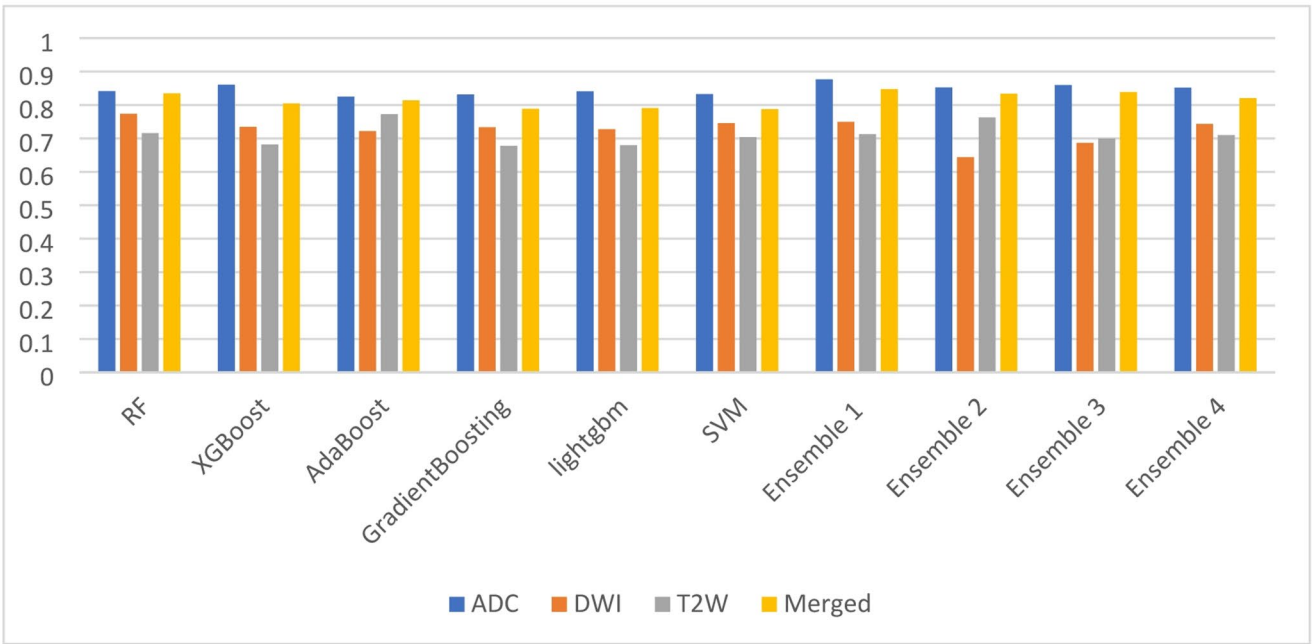


Fig. 4 Comparison of the AUCs achieved by all models in Approach 1 using Feature Set 2 for PI-RADS classification

characteristics, such as distinguishing heterogeneous T2W lesions (PI-RADS 3) from homogeneous ones (PI-RADS 4). This reliance on visual interpretation introduces variability and potential disagreement among clinicians [7, 53].

In this study, we evaluated multiple ML approaches for automated PI-RADS classification using radiomic features extracted from multiparametric MRI sequences, including T2W, DWI, and ADC maps. Across the three approaches, our results consistently highlighted the superior performance of ADC-based models compared to DWI or T2W alone, with Ensemble and gradient boosting variants achieving the highest overall accuracy and AUC values.

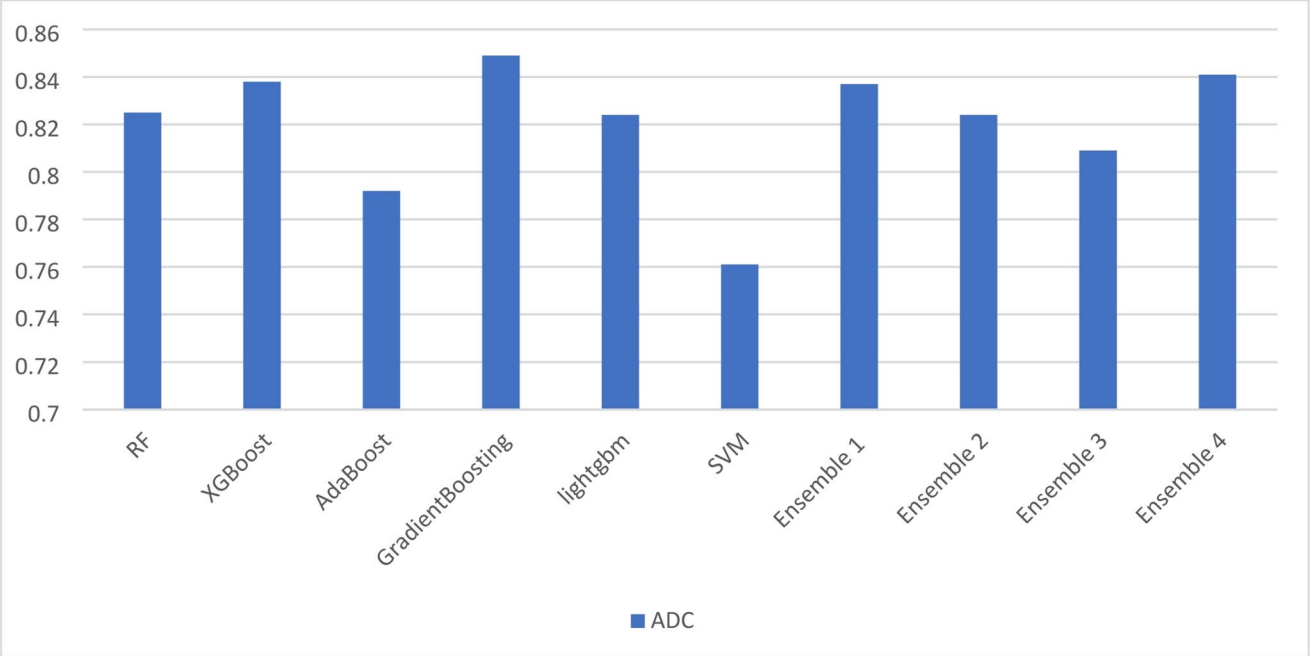


Fig. 5 Comparison of the AUCs achieved by all models in Approach 2 using Feature Set 1 for PI-RADS classification

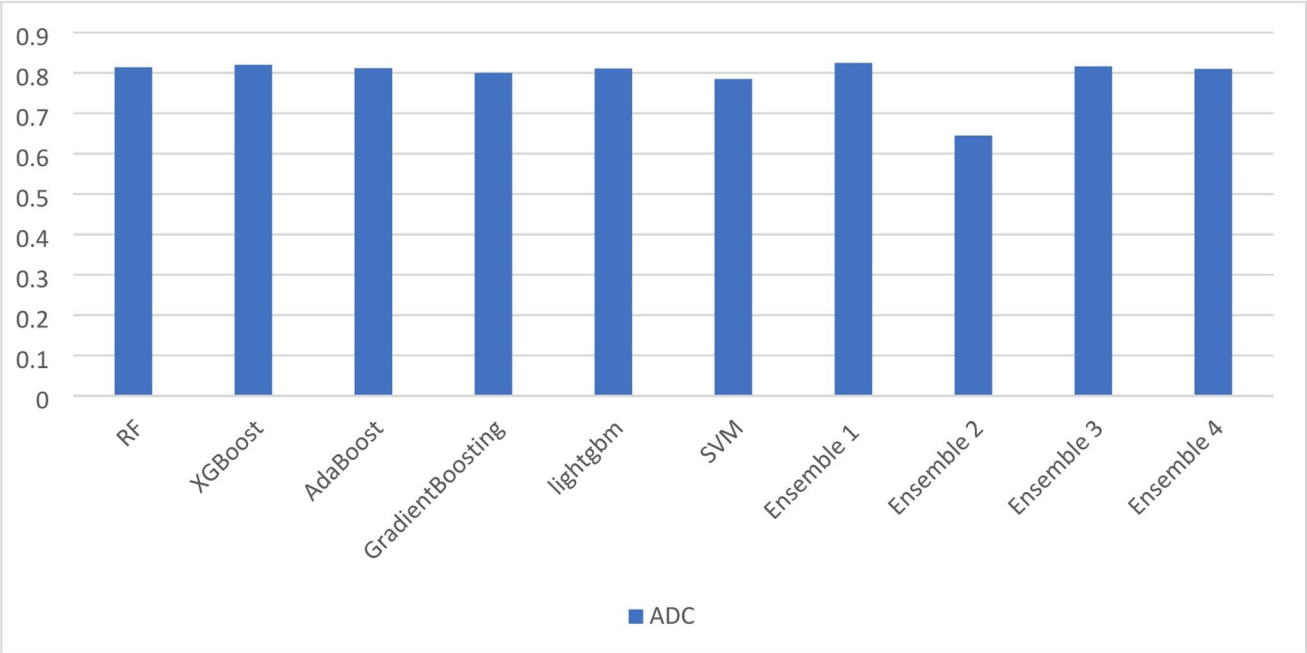


Fig. 6 Comparison of the AUCs achieved by all models in Approach 2 using Feature Set 2 for PI-RADS classification

Notably, PI-RADS 5 lesions were classified reliably across all models, whereas PI-RADS 3 remained the most challenging category, reflecting the inherent difficulty in distinguishing intermediate-risk lesions.

Figure 3 compares AUCs of ML models for PI-RADS classification using Feature Set 1. ADC features consistently achieve the highest AUC across most models, while T2W features generally perform the worst. Merged features improve performance over individual sequences, particularly for XGBoost, GradientBoosting, and the ensemble models. Ensemble models show competitive AUCs, often surpassing single classifiers when using merged features.

According to Fig. 4, which shows the comparison of AUCs achieved by all models in Approach 1 using Feature

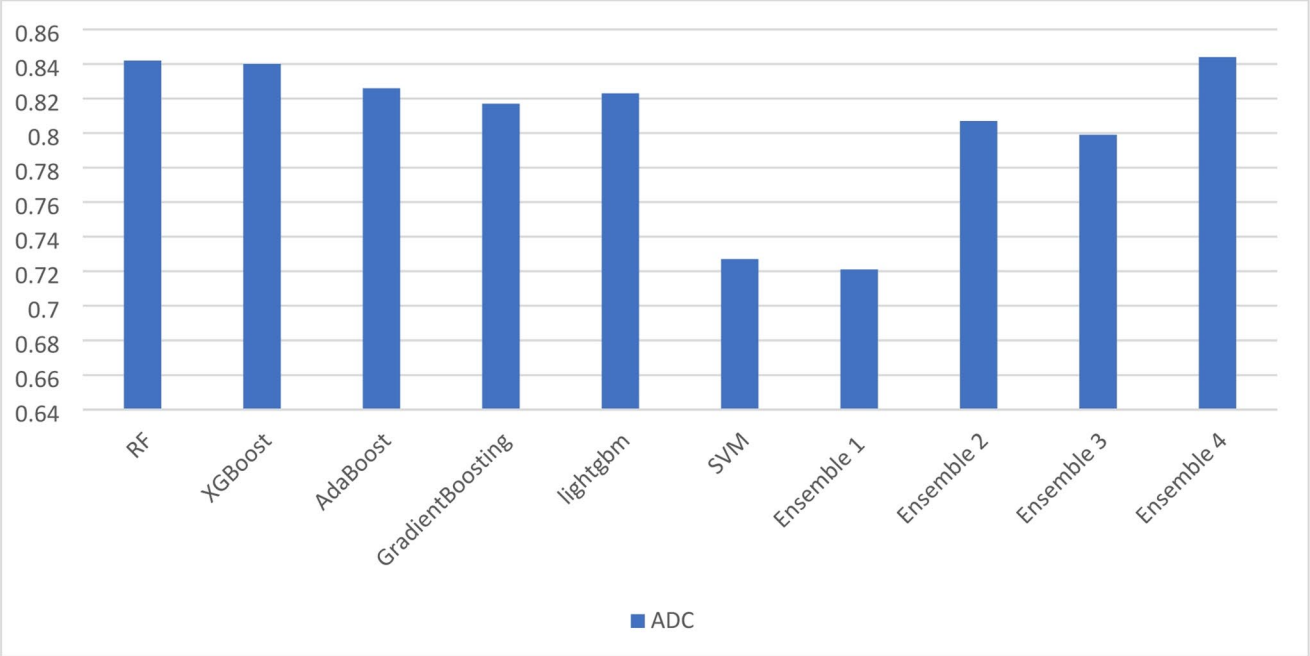


Fig. 7 Comparison of the AUCs achieved by all models in Approach 3

Set 2 for PI-RADS classification, ADC consistently delivers the highest performance across most models. Merged features also perform well, particularly in ensemble models like Ensemble 1 and Ensemble 4. DWI and T2W exhibit lower and more variable AUCs, indicating weaker discriminative power. Ensemble approaches tend to improve model robustness and classification accuracy. Overall, ADC derived features turn out to be the most important.

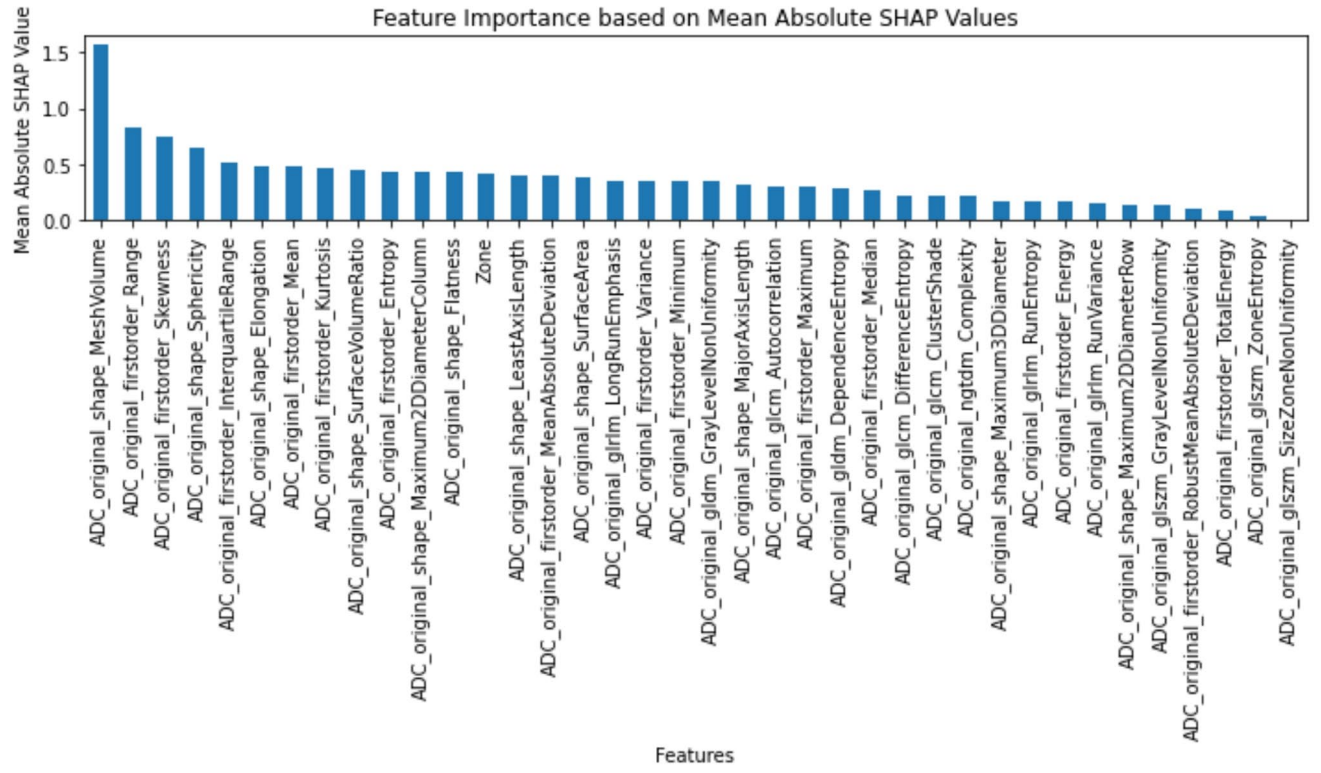
Figures 5 and 6 present the AUCs achieved by all models in Approach 2 using Feature Set 1 and Feature Set 2, respectively. Across both feature sets, model performance is relatively stable, with most AUCs clustering around the 0.80 mark. In Fig. 5 (Feature Set 1), GradientBoosting and Ensemble 4 deliver the highest AUCs, whereas in Fig. 6 (Feature Set 2), the performance is more uniform across models, though Ensemble 2 shows a noticeable drop. SVM underperforms in both cases, suggesting limited compatibility with ADC-based input. Overall, the results indicate that while both feature sets are effective, Feature Set 1 slightly enhances model separation and peak performance, particularly for ensemble and boosting methods.

Figure 7 compares the AUCs achieved by various models in Approach 3 using ADC features for PI-RADS classification. Tree-based models like RF and XGBoost maintain strong performance, with AUCs above 0.83. Ensemble 4 achieved the better results, indicating the strength of its model combination in this approach. In contrast, SVM and Ensemble 1 show the weakest performance, both dropping below 0.74. These results suggest that while some ensemble strategies may fail to generalize, others (like Ensemble 4) can significantly boost predictive power.

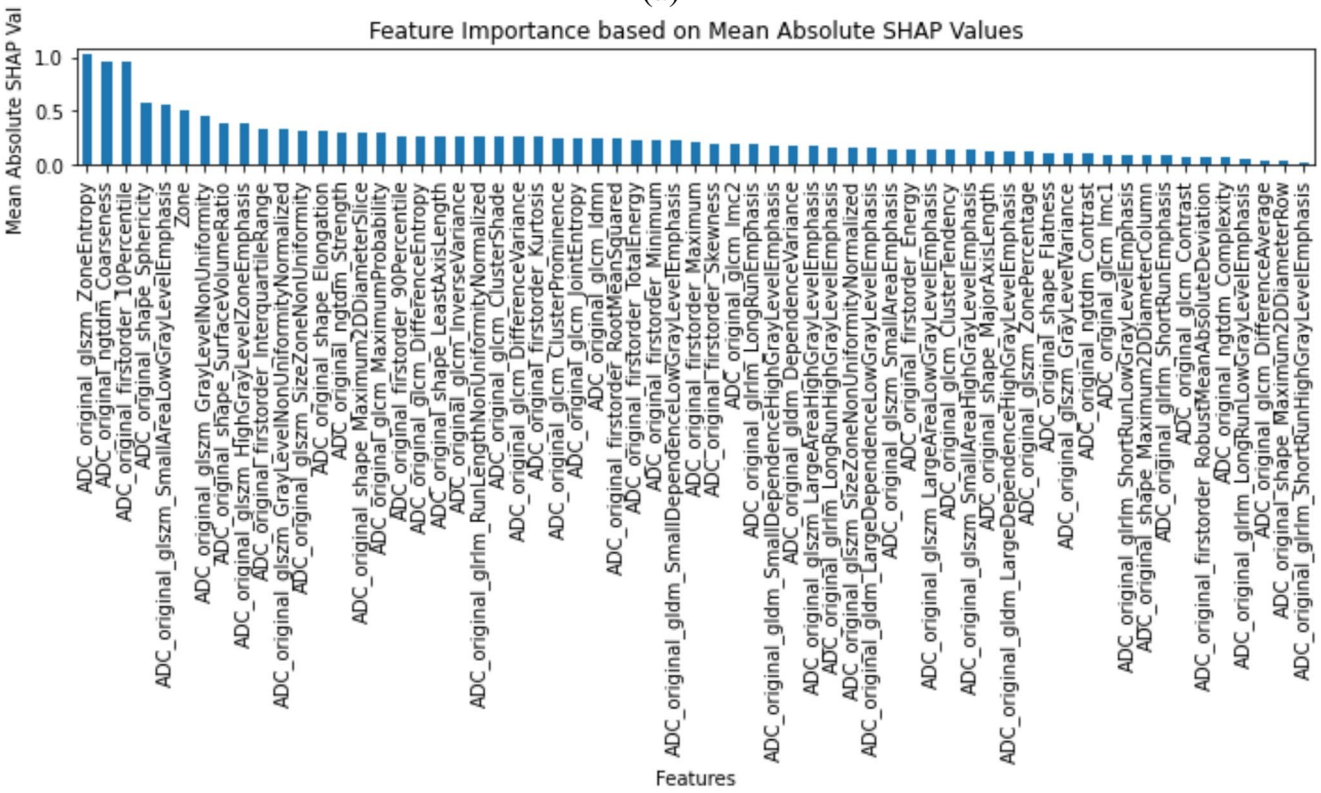
To gain insights into feature contributions and model interpretability, SHAP (SHapley Additive exPlanations) was employed. SHAP is a unified framework for interpreting predictions, grounded in cooperative game theory, that assigns each feature an importance value for a particular prediction [54]. It has become a standard method for model explainability due to its consistency and local accuracy. However, applying SHAP to complex ensemble models, especially stacked ensembles, is often computationally intensive or infeasible due to their structural complexity and lack of direct SHAP integration [55]. Therefore, XGBoost, which consistently achieved high AUCs across experiments and was a core component in the ensemble models, was selected for SHAP analysis.

Figures 8 and 9 illustrate the SHAP-based interpretability results obtained from XGBoost using ADC features in Approach 1. Figure 8 presents the feature importance plots based on the mean absolute SHAP values for both Feature Set 1 and Feature Set 2, highlighting which features contribute most on average to the model's predictions. Figure 9 shows the mean SHAP value plots, capturing the average impact (both direction and magnitude) of each feature on the model output across all classes for the same two feature sets.

These visualizations display the average impact of each feature on the model output across all classes, based on the mean absolute SHAP values, effectively summarizing overall feature importance in the classification task.



(a)



(b)

Fig. 8 Feature importance plots based on the mean absolute SHAP values, in **a** approach 1 using feature set 1, and **b** approach 1 using feature set 2

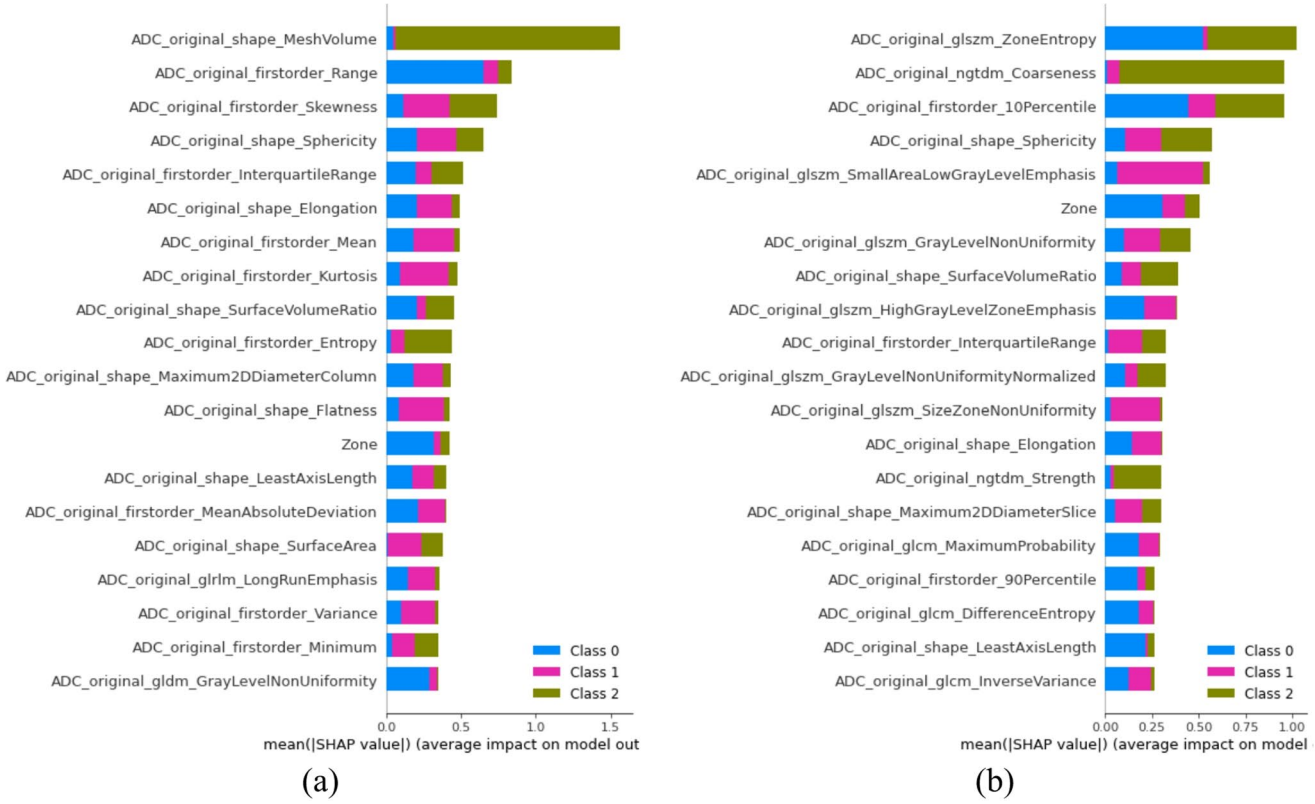


Fig. 9 Feature importance plots based on the mean SHAP value based on the classes (Class 0: PI-RADS 3, Class 1: PI-RADS 4, and Class 3: PI-RADS 5), in **a** approach 1 using feature set 1, and **b** approach 1 using feature set 2

Based on Fig. 9. a, for Class 0 (PI-RADS 3), MeshVolume and Range have the highest positive impact, indicating their strong role in identifying this class. For Class 1 (PI-RADS 4), features like Skewness, Sphericity, and Elongation show notable influence, suggesting the model uses shape irregularities to differentiate this group.

In Class 2 (PI-RADS 5), Skewness and Mean contribute most, reflecting intensity-related characteristics as key indicators. Some features, such as Skewness and Range, contribute meaningfully across all classes but with varying magnitudes. Overall, the model leverages both shape and intensity-based features differently across classes to support PI-RADS classification.

According to Fig. 9 b, features related to Class 2 (PI-RADS 5) have the highest SHAP impact, with texture metrics like ZoneEntropy and Coarseness being most influential. Class 1 (PI-RADS 4) features, including first-order and small area emphasis metrics, show moderate importance. Class 0 (PI-RADS 3) features have lower but consistent effects across variables. Overall, texture features from ADC images dominate model predictions, while shape and first-order features contribute less. This highlights the key role of texture heterogeneity in class differentiation.

When compared with prior research, our results reveal both agreement with established trends and meaningful improvements in certain aspects of PI-RADS classification performance.

Zhu et al. [9] conducted one of the largest studies on automated PI-RADS scoring using a multicenter dataset of 927 patients and 1,186 lesions, reporting AUCs between 0.85 and 0.87 for PI-RADS 3–5 classification. Our best-performing ADC-based ensemble models achieved comparable overall discrimination (AUC = 0.83–0.86) despite being trained on a smaller dataset, suggesting that well-optimized radiomic and ensemble learning frameworks can generalize effectively even in limited-data settings. Similarly, Winkel et al. [11] achieved an AUROC of 0.828 using bi-parametric MRI and a DL model, with sensitivity and specificity of 76.7% and 85.9%, respectively; our ensemble models demonstrated comparable performance, particularly in distinguishing high-risk lesions (PI-RADS 5).

Study [10] reported an overall accuracy of 58% for direct PI-RADS category prediction using DL, which increased to 86% when allowing ± 1 score flexibility. In contrast, our best ensemble model achieved accuracies up to 0.77 without any tolerance margin, highlighting a higher precision in strict category prediction, particularly for ADC-derived features.

Radiomic studies focusing on ADC maps [12–15] consistently demonstrated that ADC-based features yield

superior classification accuracy compared to DWI or T2W. For example, reference [13] achieved 79% overall accuracy (AUC ≈ 0.85), and [14] reported AUC = 0.85, sensitivity = 83%, and specificity = 78%. Our results corroborate these findings, with ADC-based ensemble models achieving the highest accuracy (up to 0.77) and AUC values exceeding 0.83. These consistent outcomes reaffirm the pivotal role of ADC radiomics in characterizing tumor cellularity and discriminating between benign and malignant prostate tissue.

Several studies demonstrated that combining multiparametric MRI sequences enhances classification robustness. In [15], integrating T2W, DWI, ADC, and DCE features achieved a validation AUC = 0.931. Likewise, our combined-sequence models reached AUC = 0.83 and accuracy = 0.71, approaching the performance of larger multicenter datasets, thus confirming the added value of feature fusion. Similarly [13], reported 86% sensitivity for peripheral zone lesions using combined T2W and ADC features, which aligns with our observation that multi-sequence integration improves lesion-level stability across zones.

Deep radiomics and ensemble-based methods have also shown competitive performance across the literature. For example [16], combined nnU-Net and XGBoost, achieving a patient-level AUROC of 0.91, while [21] reported up to 88% accuracy and an AUROC of 0.94 using a deep ensemble model. Although our dataset was more limited, our ensemble configurations (e.g., Ensemble 3 and Ensemble 4) achieved accuracies between 0.71 and 0.77 and AUCs up to 0.86, which are within the range of these advanced architectures. This suggests that ensemble learning applied to handcrafted radiomic features remains a powerful and computationally efficient alternative to deep end-to-end frameworks, especially for smaller datasets.

A persistent challenge across nearly all studies, including [12, 17, 18], is the accurate classification of PI-RADS 3 lesions. These intermediate cases exhibit overlapping imaging and radiomic characteristics between benign and malignant tissue, leading to reduced sensitivity and precision. Our results confirmed this difficulty, as PI-RADS 3 classification consistently showed the lowest recall and precision values across all feature sets and MRI modalities. Conversely, high-risk lesions (PI-RADS 5) maintained high AUCs (≥ 0.94), consistent with prior literature.

Finally, several studies [19–22] have emphasized the value of incorporating clinical parameters such as PSA density and patient age to enhance classification reliability. Although our work focused exclusively on imaging-based radiomics, the comparable performance achieved suggests that radiomic signatures can effectively capture much of the underlying biological variability represented by these clinical indicators. Nonetheless, future research that integrates clinical, molecular, and imaging data could further strengthen diagnostic precision and bring model performance closer to expert-level assessment.

A consistent challenge across both prior studies and our own results lies in the differentiation of PI-RADS 3 lesions. Their inherently ambiguous imaging characteristics, which overlap with both lower and higher-grade categories, continue to limit classification accuracy even in advanced models. Addressing this issue will require improved feature extraction methods, more sophisticated model training strategies, and the inclusion of relevant clinical or biochemical markers to better capture subtle disease patterns.

Furthermore, dataset size and heterogeneity remain key factors affecting model generalizability. Many existing studies, and to some extent, ours, are constrained by limited sample sizes, which restrict external validity and fail to encompass the full variability encountered in clinical practice. Future research should therefore prioritize multicenter collaborations and standardized imaging protocols to build more diverse and representative datasets.

In summary, our findings are consistent with and, in several respects, extend existing literature. ADC-based radiomic features and ensemble learning approaches demonstrated robust and stable performance, achieving results comparable to those reported in larger deep learning studies. While PI-RADS 5 lesions are classified with high reliability, the persistent difficulty of accurately identifying PI-RADS 3 cases underscores the ongoing need for refined feature design, multi-sequence fusion, and broader, standardized datasets to achieve clinically reliable and generalizable AI-driven PI-RADS assessment.

Conclusions

This study highlights the value of multiparametric MRI, particularly ADC imaging, for automated PI-RADS classification of prostate lesions. Among the evaluated approaches, combining radiomics and deep learning features provided the most robust performance, with consistent improvements when integrating multiple MRI sequences. While classification of intermediate-risk lesions (PI-RADS 3) remains challenging, the proposed framework supports more objective and reproducible assessment. Overall, these findings reinforce the potential of AI-assisted analysis to enhance risk stratification and clinical decision-making in prostate cancer.

Appendix A

See Tables 8, 9, 10 and 11.

Table 8 Performance of PI-RADS Classification Models of Approach 1, Using PyRadiomics Feature Set 1 Across ADC Images

	Acc.	AUC.	PI-RADS 3			PI-RADS 4			PI-RADS 5		
			Prec.	Recall	AUC	Prec.	Recall	AUC	Prec.	Recall	AUC
RF	0.65	0.84	0.46	0.58	0.79	0.60	0.63	0.76	1.00	0.73	0.97
XGBoost	0.67	0.83	0.50	0.66	**0.79**	0.63	0.63	0.75	1.00	0.73	0.94
AdaBoost	0.67	**0.84**	0.57	**0.66**	0.75	0.63	0.77	**0.81**	1.00	0.53	**0.97**
Gradient Boosting	0.67	0.84	0.50	0.58	0.79	0.62	0.68	0.76	1.00	0.73	0.96
LightGBM	0.65	0.83	0.47	**0.66**	0.78	0.65	0.68	0.77	1.00	0.60	0.94
SVM	0.61	0.77	0.46	0.58	0.72	0.59	0.72	0.65	1.00	0.46	0.94
Ensemble 1	**0.77**	0.83	**0.66**	**0.66**	0.77	**0.72**	**0.81**	0.77	1.00	**0.80**	0.96
Ensemble 2	0.65	0.82	0.50	**0.66**	0.78	0.63	0.63	0.75	0.90	0.66	0.93
Ensemble 3	0.69	0.81	0.50	**0.66**	0.75	0.66	0.63	0.75	1.00	**0.80**	0.94
Ensemble 4	0.67	0.83	0.50	**0.66**	0.78	0.63	0.63	0.75	1.00	0.73	0.96

Table 9 Performance of PI-RADS Classification Models of Approach 1, Using PyRadiomics Feature Set 1 Across DWI Images

	Acc.	AUC.	PI-RADS 3			PI-RADS 4			PI-RADS 5		
			Prec.	Recall	AUC	Prec.	Recall	AUC	Prec.	Recall	AUC
RF	**0.63**	0.78	0.37	0.25	0.70	0.58	0.77	0.67	**0.91**	0.73	0.97
XGBoost	0.59	0.72	0.37	0.25	0.63	0.54	0.77	0.63	0.90	0.60	0.89
AdaBoost	0.59	**0.81**	0.38	**0.41**	**0.71**	0.56	0.63	**0.73**	0.90	0.66	0.97
Gradient Boosting	0.57	0.74	0.30	0.25	0.61	0.53	0.68	0.63	0.90	0.66	0.97
LightGBM	0.49	0.70	0.20	0.16	0.59	0.46	0.63	0.59	0.88	0.53	0.93
SVM	**0.63**	0.72	0.00	0.00	0.61	0.56	**0.95**	0.61	0.90	0.66	0.95
Ensemble 1	0.59	0.77	0.00	0.00	0.66	0.54	0.90	0.70	0.90	0.60	0.93
Ensemble 2	**0.63**	0.75	**0.66**	0.16	0.67	0.56	**0.95**	0.64	0.88	0.53	0.93
Ensemble 3	0.61	0.74	0.44	0.33	0.65	0.56	0.77	0.66	0.90	0.60	0.91
Ensemble 4	**0.63**	0.78	0.45	**0.41**	**0.71**	**0.59**	0.72	0.67	0.90	0.66	0.96

Table 10 Performance of PI-RADS Classification Models of Approach 1, Using PyRadiomics Feature Set 1 Across T2W Images

	Acc.	AUC.	PI-RADS 3			PI-RADS 4			PI-RADS 5		
			Prec.	Recall	AUC	Prec.	Recall	AUC	Prec.	Recall	AUC
RF	0.61	0.73	0.00	0.00	0.61	0.54	0.77	0.61	0.86	**0.86**	0.96
XGBoost	0.57	0.68	0.16	0.20	0.50	0.51	0.63	0.60	0.85	0.80	0.93
AdaBoost	0.61	**0.80**	**0.37**	0.25	**0.71**	**0.57**	0.72	**0.73**	0.84	0.73	0.93
Gradient Boosting	0.59	0.67	0.16	0.08	0.47	0.53	0.68	0.59	0.86	**0.86**	0.96
LightGBM	0.61	0.67	0.33	0.16	0.48	0.55	0.72	0.57	0.85	0.80	0.96
SVM	**0.63**	0.73	0.00	0.00	0.63	0.55	**0.95**	0.60	**0.90**	0.66	**0.96**
Ensemble 1	**0.63**	0.72	0.00	0.00	0.58	0.55	0.90	0.65	0.84	0.73	0.93
Ensemble 2	0.46	0.70	0.23	**0.33**	0.53	0.47	0.50	0.63	0.88	0.53	**0.95**
Ensemble 3	0.57	0.69	0.00	0.00	0.53	0.51	0.77	0.58	0.84	0.73	**0.96**
Ensemble 4	**0.63**	0.72	0.33	0.16	0.61	**0.57**	0.72	0.61	0.86	**0.86**	0.96

Table 11 Performance of PI-RADS Classification Models of Approach 1, Using PyRadiomics Feature Set 1 Across Combined Features of ADC, DWI, and T2W Images

	Acc.	AUC.	PI-RADS 3			PI-RADS 4			PI-RADS 5		
			Prec.	Recall	AUC	Prec.	Recall	AUC	Prec.	Recall	AUC
RF	0.65	**0.83**	0.47	**0.66**	0.76	0.63	0.54	**0.74**	0.92	**0.80**	0.98
XGBoost	0.65	0.81	0.47	**0.66**	0.74	0.63	0.54	0.71	0.92	**0.80**	0.97
AdaBoost	0.59	0.80	0.35	0.41	0.74	0.56	0.63	0.67	**1.00**	0.66	**0.98**
Gradient Boosting	0.65	0.81	0.47	**0.66**	0.76	0.63	0.54	0.70	0.92	**0.80**	0.98
LightGBM	0.59	0.78	0.43	0.58	0.72	0.54	0.54	0.67	0.90	0.66	0.95
SVM	0.53	0.74	0.14	0.08	0.70	0.48	**0.72**	0.58	**1.00**	0.60	0.95
Ensemble 1	0.65	0.84	**0.53**	**0.66**	**0.81**	0.60	0.63	0.73	0.90	0.66	0.97
Ensemble 2	0.51	0.73	0.28	0.33	0.67	0.48	0.54	0.58	0.90	0.60	0.93
Ensemble 3	**0.69**	0.81	**0.53**	**0.66**	0.79	**0.66**	0.63	0.70	0.92	**0.80**	0.94
Ensemble 4	0.63	**0.83**	0.44	**0.66**	0.78	0.61	0.50	0.73	0.92	**0.80**	0.97

Appendix B

See Table 12,

Table 12 Performance of PI-RADS Classification Models of Approach 2, Using PyRadiomics Feature Set 1 Across ADC Images

	Acc.	AUC.	PI-RADS 3			PI-RADS 4			PI-RADS 5		
			Prec.	Recall	AUC	Prec.	Recall	AUC	Prec.	Recall	AUC
RF	0.60	0.82	0.16	0.16	0.74	0.61	0.57	0.78	0.78	0.84	0.95
XGBoost	**0.69**	0.83	**0.42**	**0.50**	0.74	**0.66**	0.57	0.78	0.85	0.92	0.98
AdaBoost	0.63	0.79	0.25	0.16	0.71	0.56	**0.64**	0.71	0.84	0.84	0.94
Gradient Boosting	0.63	**0.84**	0.28	0.33	0.78	0.63	0.50	0.78	0.80	0.92	0.98
LightGBM	**0.69**	0.82	**0.42**	**0.50**	0.77	**0.66**	0.57	0.74	0.85	0.92	0.95
SVM	0.51	0.76	0.00	0.00	0.66	0.50	0.42	0.72	0.64	0.84	0.89
Ensemble 1	0.66	0.83	0.20	0.16	0.79	0.61	0.57	0.74	**0.86**	**1.00**	0.97
Ensemble 2	0.57	0.82	0.40	0.33	**0.83**	0.05	0.42	0.70	0.68	0.84	0.93
Ensemble 3	**0.69**	0.80	0.33	0.33	0.71	**0.66**	0.57	0.74	**0.86**	**1.00**	0.96
Ensemble 4	0.66	0.84	0.33	0.33	0.73	0.61	0.57	**0.81**	0.85	0.92	**0.97**

Author Contributions All the authors contributed to Methodology and to Investigation, S.M. and M.A. contributed to Conceptualization and Writing the original draft, S.M., I.B.Z., F.D. R.D.M., L.D.P., E.C., A.L., M.L. and D.F. contributed to Data Curation. S.F., L.D.P., and M.A. contributed to Software, A.M., E.D., G.G. and M.A. contributed to Funding Acquisition and to Supervision, E.D., C.M., G.G. and M.A. reviewed and edited the manuscript.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under the CISC project (Marie Skłodowska-Curie grant agreement no. 955901 https:/ /www.ciscproject.eu/, accessed on 18 March 2025). This work was partially supported by the MUSA (Multilayered Urban Sustainability Action) project, funded by the European UnionNextGenerationEU, under the Mission 4 Component 2 Investment Line of the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R\&D "innovation ecosystems", set up of "territorial leaders in R&D" (CUP G43C22001370007, Code ECS00000037); Program "piano sostegno alla ricerca" PSR and the PSR-GSA-Linea 6; Project ReGAInS (code 2023-NAZ-0207/DIP-ECC-DISCO23), funded by the Italian University and Research Ministry, within the Excellence Departments program 2023–2027 (law 232 / 2016); and FAIR-Future Artificial Intelligence Research-Spoke 4-PE00000013-D53C22002380006, funded by the European Union-Next Generation EU within the project NRPP M4C2, Investment 1.,3 DD. 341, 15 March 2022.

Data availability The Data are not made publicly available.

Declarations

Conflict of Interests On behalf of all authors, the corresponding author states that there are no conflicts of interest.

Ethical Approval Approval for this study was obtained following a comprehensive review process, ensuring that all ethical considerations and privacy concerns were adequately addressed to protect participants' rights. Approval on 20/11/2024 by the Ethical Committee "Comitato Etico Territoriale Lombardia 3", Study ID: 5105, code "PI-RADSv2", title "Predizione della malignità delle lesioni prostatiche mediante analisi AI di immagini RM multiparametriche con mezzo di contrasto".

Research Involving Human and/or Animals – This research did not involve humans nor animals.

Informed Consent Consent to Participate declarations not applicable.

References

1. Ahmed HU, El-Shater Bosaily A, Brown LC, Gabe R, Kaplan R, Parmar MK, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. Lancet. 2017;389:815–22. https://doi.org/10.1016/S0140-6736(16)32401-1.
2. Barentsz JO, Weinreb JC, Verma S, Thoeny HC, Tempany CM, Shtern F, et al. Synopsis of the PI-RADS v2 Guidelines for Multiparametric Prostate Magnetic Resonance Imaging and Recommendations for Use. Eur Urol. 2016;69:41–9. https://doi.org/10.1016/j.eururo.2015.08.038.
3. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2015;278:563–77. https://doi.org/10.1148/radiol.2015151169.
4. Turkbey B, Mani H, Shah V, Rastinehad AR, Bernardo M, Pohida T, et al. Multiparametric 3T prostate magnetic resonance imaging

to detect cancer: histopathological correlation using prostatectomy specimens processed in customized magnetic resonance imaging based molds. J Urol. 2011;186:1818–24. https://doi.org/10.1016/j.juro.2011.07.013.
5. Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, et al. Eur Urol. 2019;76:340–51. https://doi.org/10.1016/j.eururo.2019.02.033. Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2.
6. Eklund M, Jäderling F, Discacciati A, Bergman M, Annerstedt M, Aly M, et al. MRI-Targeted or Standard Biopsy in Prostate Cancer Screening. N Engl J Med. 2021;385:908–20. https://doi.org/10.1056/NEJMoa2100852.
7. Rosenkrantz AB, Ginocchio LA, Cornfeld D, Froemming AT, Gupta RT, Turkbey B, et al. Interobserver Reproducibility of the PI-RADS Version 2 Lexicon: A Multicenter Study of Six Experienced Prostate Radiologists. Radiology. 2016;280:793–804. https://doi.org/10.1148/radiol.2016152542.
8. Saha A, Hosseinzadeh M, Huisman H. End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. Med Image Anal. 2021;73:102155. https://doi.org/10.1016/j.media.2021.102155.
9. Bao J, Qiao X, Song Y, Su Y, Ji L, Shen J, et al. Prediction of clinically significant prostate cancer using radiomics models in real-world clinical practice: a retrospective multicenter study. Insights Imaging. 2024;15:68. https://doi.org/10.1186/s13244-024-01631-w.
10. Sanford T, Harmon SA, Turkbey EB, Kesani D, Tuncer S, Madariaga M, et al. Deep-Learning-Based Artificial Intelligence for PI-RADS Classification to Assist Multiparametric Prostate MRI Interpretation: A Development Study. J Magn Reson Imaging. 2020;52:1499–507. https://doi.org/10.1002/jmri.27204.
11. Youn SY, Choi MH, Kim DH, Lee YJ, Huisman H, Johnson E, et al. Detection and PI-RADS classification of focal lesions in prostate MRI: Performance comparison between a deep learning-based algorithm (DLA) and radiologists with various levels of experience. Eur J Radiol. 2021;142:109894. https://doi.org/10.1016/j.ejrad.2021.109894.
12. Zhang KS, Schelb P, Kohl S, Radtke JP, Wiesenfarth M, Schimmöller L, et al. Improvement of PI-RADS-dependent prostate cancer classification by quantitative image assessment using radiomics or mean ADC. Magn Reson Imaging. 2021;82:9–17. https://doi.org/10.1016/j.mri.2021.06.013.
13. Bonaffini PA, De Bernardi E, Corsi A, Franco PN, Nicoletta D, Muglia R, et al. Towards the Definition of Radiomic Features and Clinical Indices to Enhance the Diagnosis of Clinically Significant Cancers in PI-RADS 4 and 5 Lesions. Cancers (Basel). 2023;15. https://doi.org/10.3390/cancers15204963.
14. Wang J, Wu C-J, Bao M-L, Zhang J, Wang X-N, Zhang Y-D. Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer. Eur Radiol. 2017;27:4082–90. https://doi.org/10.1007/s00330-017-4800-5.
15. Li M, Yang L, Yue Y, Xu J, Huang C, Song B. Use of Radiomics to Improve Diagnostic Performance of PI-RADS v2.1 in Prostate Cancer. Front Oncol. 2020;10:631831. https://doi.org/10.3389/fonc.2020.631831.
16. Nketiah G, Sunoqrot MRS, Sandsmark E, Langørgen S, Selnæs K, Bertilsson H et al. Deep Radiomics Detection of Clinically Significant Prostate Cancer on Multicenter MRI: Initial Comparison to PI-RADS Assessment. 2024. https://doi.org/10.48550/arXiv.2410.16238
17. Brancato V, Aiello M, Basso L, Monti S, Palumbo L, Di Costanzo G, et al. Evaluation of a multiparametric MRI radiomic-based approach for stratification of equivocal PI-RADS 3 and upgraded PI-RADS 4 prostatic lesions. Sci Rep. 2021;11:643. https://doi.org/10.1038/s41598-020-80749-5.
18. Jaen-Lorites JM, Ruiz-Espana S, Pineiro-Vidal T, Santabarbara JM, Maceira AM, Moratal D. Multiclass Classification of Prostate Tumors Following an MR Image Analysis-Based Radiomics Approach. Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf. 2022;2022:1436–9. https://doi.org/10.1109/EMBC48229.2022.9871746.
19. Singh D, Kumar V, Das CJ, Singh A, Mehndiratta A. Machine learning-based analysis of a semi-automated PI-RADS v2.1 scoring for prostate cancer. Front Oncol. 2022;12:961985. https://doi.org/10.3389/fonc.2022.961985.
20. Oerther B, Engel H, Wilpert C, Nedelcu A, Sigle A, Grimm R, et al. Score Assignment and Lesion Detection in Prostate MRI. Cancers (Basel). 2025;17. https://doi.org/10.3390/cancers17050815. Multi-Center Benchmarking of a Commercially Available Artificial Intelligence Algorithm for Prostate Imaging Reporting and Data System (PI-RADS).
21. Yang C, Li B, Luan Y, Wang S, Bian Y, Zhang J, et al. Deep learning model for the detection of prostate cancer and classification of clinically significant disease using multiparametric MRI in comparison to PI-RADs score. Urol Oncol Semin Orig Investig. 2024;42. https://doi.org/10.1016/j.urolonc.2024.01.021. :158.e17-158.e27.
22. Wei X, Xu J, Zhong S, Zou J, Cheng Z, Ding Z, et al. Diagnostic value of combining PI-RADS v2.1 with PSAD in clinically significant prostate cancer. Abdom Radiol. 2022;47:3574–82. https://doi.org/10.1007/s00261-022-03592-4.
23. Kan Y, Zhang Q, Hao J, Wang W, Zhuang J, Gao J, et al. Clinico-radiological characteristic-based machine learning in reducing unnecessary prostate biopsies of PI-RADS 3 lesions with dual validation. Eur Radiol. 2020;30:6274–84. https://doi.org/10.1007/s00330-020-06958-8.
24. Hectors SJ, Chen C, Chen J, Wang J, Gordon S, Yu M, et al. Magnetic Resonance Imaging Radiomics-Based Machine Learning Prediction of Clinically Significant Prostate Cancer in Equivocal PI-RADS 3 Lesions. J Magn Reson Imaging. 2021;54:1466–73. https://doi.org/10.1002/jmri.27692.
25. Koh D-M, Collins DJ. Diffusion-weighted MRI in the body: applications and challenges in oncology. AJR Am J Roentgenol. 2007;188:1622–35. https://doi.org/10.2214/AJR.06.1403.
26. DeLano MC, Cooper TG, Siebert JE, Potchen MJ, Kuppusamy K. High-b-value diffusion-weighted MR imaging of adult brain: image contrast and apparent diffusion coefficient map features. AJNR Am J Neuroradiol. 2000;21:1830–6.
27. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.
28. Maintz JB, Viergever MA. A survey of medical image registration. Med Image Anal. 1998;2:1–36. https://doi.org/10.1016/s1361-8415(01)80026-8.
29. Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The Design of SimpleITK. Front Neuroinform. 2013;7:45. https://doi.org/10.3389/fninf.2013.00045.
30. Rueckert D, Sonoda LI, Hayes C, Hill DLG, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging. 1999;18:712–21. https://doi.org/10.1109/42.796284.
31. Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. IEEE Trans Med Imaging. 2000;19:143–50. https://doi.org/10.1109/42.836373.
32. Shinohara RT, Sweeney EM, Goldsmith J, Shiee N, Mateen FJ, Calabresi PA, et al. Statistical normalization techniques for magnetic resonance imaging. NeuroImage Clin. 2014;6:9–19. https://doi.org/10.1016/j.nicl.2014.08.008.

33. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77:e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.
34. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441–6. https://doi.org/10.1016/j.ejca.2011.11.036.
35. Barentsz JO, Richenberg J, Clements R, Choyke P, Verma S, Villeirs G, et al. ESUR prostate MR guidelines 2012. Eur Radiol. 2012;22:746–57. https://doi.org/10.1007/s00330-011-2377-y.
36. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine Learning methods for Quantitative Radiomic Biomarkers. Sci Rep. 2015;5:13087. https://doi.org/10.1038/srep13087.
37. Fouladi S et al. Exploring UNet-Based Models for Prostate Lesion Segmentation from Multi-Sequence MRI (T2W, ADC, DWI). Submitted to World Wide Web. 2025.
38. Fouladi S, Di Palma L, Darvizeh F, Fazzini D, Maiocchi A, Papa S, Gianini G, Alì M. Neural Network Models for Prostate Zones Segmentation in Magnetic Resonance Imaging. Information. 2025;16. https://doi.org/10.3390/info16030186.
39. Bovio A, Barile M, Pallotta F, Pede L, Maiocchi A, Alì M, Darvizeh F, Fazzini D, Lacavalla F, Banzi M, Gianini G, Mio C, Berto F, Bondaruc R, Damiani E, Fouladi S. A Federated Learning Architecture for Prostate MRI Image Segmentation. To appear in Proceedings of the 4th Italian Conference on Big Data and Data Science, Torino, Italy, September 2025 (ITADATA2025).
40. Fouladi S et al. Advanced Prostate MRI Analysis: UNET-Based Models for Zonal and Lesion Segmentation. International Conference on Management of Digital. Cham: Springer Nature Switzerland, 2024, pp. 174–187.
41. Fouladi S, Darvizeh F, Di Meo R, Bossi Zanetti I, Gianini G, Damiani E, Cambie E, Licata A, Maiocchi A, Ali M. Fazzini D.A hybrid LSTM-UNet architecture for segmentation of MRI prostatic lesion images, To appear in: Proceedings of International Conference on Management of Digital. Cham: Springer Nature Switzerland, 2025.
42. Shen D, Wu G, Suk H-I. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng. 2017;19:221–48. https://doi.org/10.1146/annurev-bioeng-071516-044442.
43. Ronneberger O, Fischer P, Brox T. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. U-Net: Convolutional Networks for Biomedical Image Segmentation BT - Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing; 2015. pp. 234–41.
44. Breiman L. Random Forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010950718922.
45. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97. https://doi.org/10.1007/BF00994018.
46. Chen T, Guestrin C, XGBoost:. A Scalable Tree Boosting System. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. New York, NY, USA: Association for Computing Machinery; 2016. pp. 785–794. https://doi.org/10.1145/2939672.2939785
47. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci. 1997;55:119–39. https://doi.org/10.1006/jcss.1997.1504.
48. Friedman J. Greedy Function Approximation: A Gradient Boosting Machine. Ann Stat. 2000;29. https://doi.org/10.1214/aos/1013203451.
49. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017.
50. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 2001;14.
51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
52. Tibshirani R. Regression shrinkage selection via the LASSO. J R Stat Soc Ser B. 2011;73:273–82. https://doi.org/10.2307/41262671.
53. Muller BG, Shih JH, Sankineni S, Marko J, Rais-Bahrami S, George AK, et al. Prostate Cancer: Interobserver Agreement and Accuracy with the Revised Prostate Imaging Reporting and Data System at Multiparametric MR Imaging. Radiology. 2015;277:741–50. https://doi.org/10.1148/radiol.2015142818.
54. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proc. 31st Int. Conf. Neural Inf. Process. Syst. Red Hook, NY, USA: Curran Associates Inc.; 2017. pp. 4768–4777.
55. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif Intell. 2021;298:103502. https://doi.org/10.1016/j.artint.2021.103502.