Received: December 5, 2025. Revised: January 3, 2026.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Synergistic Integration of Radiomics and Deep Features from Multiparametric MRI for Enhanced Prostate Cancer Classification

Nasser M. Al-Zidi[1]* **D. Vasumathi[1]**

[1]*Department of Computer Science and Engineering, University College of Engineering, Science & Technology Hyderabad, Jawaharlal Nehru Technological University Hyderabad, Hyderabad, India*
* Corresponding author's Email: alzidi.nasser@gmail.com

Abstract: Accurate classification of clinically significant (CS) versus clinically insignificant (CiS) prostate cancer is critical for treatment decisions. This study investigates the integration of radiomics and deep features from multiparametric MRI (mpMRI) for automated prostate cancer classification. A multimodal deep learning framework with Convolutional Block Attention Module processes three optimally selected MRI sequences (T2-weighted, high b-value DWI, and Ktrans) through parallel encoders. Deep features are extracted from the trained framework, while radiomics features are extracted from each sequence and then concatenated across sequences. Six feature configurations (radiomics-only, deep-only, radiomics-PCA, deep-PCA, combined, and combined-PCA) are evaluated using eleven machine learning classifiers on the official ProstateX challenge dataset (330 training, 208 test lesions). The highest performance is achieved by Voting 2 ensemble classifier with combined PCA features: 0.943 AUC (95% CI: 0.908-0.971), 89.9% accuracy, 81.2% sensitivity, and 92.5% specificity. Deep features substantially outperformed radiomics features (mean AUC: 0.899 vs 0.829, 8.44% improvement). Combined features with PCA significantly outperformed deep-only features (2.06% improvement, $p = 0.032$). Cross-modal correlation analysis (mean $|r| = 0.14 \pm 0.13$) provided theoretical validation that radiomics and deep features capture complementary information. This study demonstrates that systematic integration of radiomics and deep features with PCA-based dimensionality reduction achieves superior prostate cancer classification, offering a validated approach for clinical decision support.

Keywords: Prostate cancer, Multiparametric MRI, Radiomics, Deep features, Machine learning, Classification.

1. Introduction

Prostate cancer (PCa) represents the second most diagnosed malignancy and fifth leading cause of mortality among men, with 1.4 million new cases annually [1]. The clinical challenge lies in distinguishing clinically significant (CS, Gleason score ≥7) from clinically insignificant (CiS, Gleason score ≤6) disease, which determines treatment strategy. CS disease requires aggressive intervention to prevent progression, while CiS disease may be managed through active surveillance [2, 3]. Traditional diagnostic (PSA testing, digital rectal examination, TRUS-guided biopsy) exhibit limitations including low specificity, operator dependence, and procedural complications, contributing to both overdiagnosis and underdiagnosis. Multiparametric MRI (mpMRI) has emerged as a superior imaging modality, integrating T2-weighted, diffusion-weighted, and dynamic contrast-enhanced sequences [2, 4]. However, interpretation remains challenging with moderate inter-reader variability [5], necessitating automated diagnostic systems.

Deep learning, particularly convolutional neural networks, has demonstrated remarkable capability in medical image analysis through hierarchical feature learning [6-8]. Concurrently, radiomics—the high-throughput extraction of quantitative features from medical images—has shown promise in capturing texture, shape, and intensity patterns that correlate with disease characteristics [9, 10]. Both face limitations: deep learning requires large labeled datasets and lacks interpretability; radiomics relies

on handcrafted features missing complex hierarchical patterns. Critically, hybrid approaches lack theoretical validation of feature complementarity and systematic investigation of dimensionality reduction for integration. This study investigates radiomics features, deep features extracted from a multimodal attention-based framework, and their combinations with dimensionality reduction using Principal Component Analysis (PCA) [11] for distinguishing CS from CiS prostate cancer using the ProstateX dataset.

We evaluate six feature configurations (radiomics-only, deep-only, radiomics-PCA, deep-PCA, combined, and combined-PCA) across eleven machine learning classifiers. The main contributions include: (1) systematic comparison demonstrating that deep features substantially outperform radiomics features and that combined features with PCA achieve optimal performance, (2) theoretical validation through cross-modal correlation analysis proving that radiomics and deep features capture complementary rather than redundant information, and (3) rigorous statistical validation establishing that PCA-based integration significantly outperforms both individual feature types, providing evidence-based guidelines for feature integration in computer-aided diagnosis systems. The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the dataset, methodology, and evaluation protocol. Section 4 presents experimental results and comparisons. Section 5 discusses the findings and limitations. Section 6 concludes the paper.

2. Related work

Prostate cancer classification from mpMRI has evolved from traditional machine learning with handcrafted radiomics features to advanced deep learning architectures. Early approaches employed radiomics features—quantitative descriptors of image intensity, texture, and shape—with classical classifiers. Kitchen and Seah [12] achieved 0.82 AUC using SVM with radiomics on ProstateX, while Kwon et al. [13] and Sobecki et al. [14] reported AUCs of 0.63-0.82. Varan et al. [15] achieved 88% accuracy with fine-tuned linear SVM and key radiomics features. However, traditional radiomics approaches suffer from limitations including manual feature engineering requiring domain expertise, preprocessing sensitivity reducing robustness, restricted feature spaces missing complex patterns, and inability to capture hierarchical representations automatically.

Deep learning methods demonstrated progressive improvements on ProstateX. Mehrtash et al. [16] employed 3D CNNs (0.80 AUC), Seah et al. [17] developed auto-windowing CNNs (0.84 AUC), and Liu et al. [18] proposed XmasNet (0.84 AUC, ranked 2nd among 33 teams). Transfer learning approaches by Chen et al. [19] (VGG-16, 0.83 AUC, ranked 4th), Yuan et al. [20], Mehmood et al. [21], Abbasi et al. [22], and Yoo et al. [23] achieved 81-89% accuracy leveraging pre-trained networks. Wang and Wang [3] investigated optimal mpMRI sequence combinations using multi-input CNNs, achieving 0.89 AUC through systematic sequence selection. Recent advanced architectures include Santhirasekaram et al. [24] with multi-scale hybrid Transformers (0.94 AUC), Yang et al. [25] with deep learning ensembles (0.902 AUC), and explainable approaches by Hamm et al. [26] and Cai et al. [27]. Despite impressive performance, deep learning approaches face limitations including large labeled dataset requirements, black-box nature limiting clinical trust and interpretability, risk of overfitting with limited data, computational intensity, and potential to miss texture information routinely assessed by radiologists.

Hybrid approaches integrating radiomics and deep learning have emerged. Khanfari et al. [28] combined radiomics and deep features for prostate cancer grading using PROSTATEx-2, achieving 0.95 AUC and demonstrating deep features significantly outperformed radiomics alone. Castillo et al. [29] compared deep learning and radiomics models, finding radiomics demonstrated robust external validation (AUCs 0.88, 0.91, 0.65) versus deep learning (0.70, 0.73, 0.44), highlighting variability across datasets. Donisi et al. [30] achieved 80% accuracy (AUC < 0.80) using radiomics with tree-based algorithms. However, current hybrid approaches exhibit critical limitations: (1) limited investigation of PCA for dimensionality reduction and feature integration, (2) lack of theoretical justification that radiomics and deep features capture complementary rather than redundant information, (3) insufficient component-wise validation quantifying individual feature contributions, and (4) absence of systematic comparison across multiple integration strategies.

This study addresses these gaps through systematic investigation of radiomics and deep features extracted from mpMRI sequences for PCa classification, theoretical validation of feature complementarity via cross-modal correlation analysis, mutual information analysis, evaluation of six feature configurations including PCA-based dimensionality reduction, comprehensive comparison across eleven machine learning

classifiers, statistical validation using bootstrap confidence intervals and DeLong's test, and evaluation on official ProstateX challenge dataset enabling reproducible benchmark comparison.

3. Materials and methods

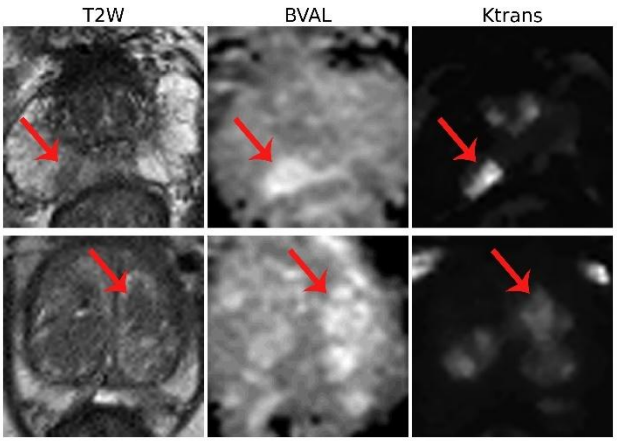
3.1 Dataset description

This investigation utilizes the ProstateX Challenge dataset [31], made publicly available through the SPIE-AAPM-NCI Prostate MR Classification Challenge. The dataset comprises multiparametric MRI scans from 346 patients acquired using 3T Siemens MAGNETOM Trio and Skyra scanners. The training cohort contains 330 lesions from 204 patients, while the test cohort comprises 208 lesions from 142 patients. We utilize the official challenge data splits with predefined patient-level split preventing data leakage. Three sequences were selected based on prior optimal combination analysis [32]: T2-weighted imaging for anatomical structure reference, high b-value diffusion-weighted imaging (BVAL), and (Ktrans) from DCE-MRI reflecting vascular permeability and angiogenesis. Each lesion is annotated with spatial coordinates (x,y,z), anatomical zone, and clinical significance designation. The test set was used only once for final evaluation. Fig. 1a illustrates representative MRI slices for CS and CiS cases across all sequences, while Fig. 1b shows the distribution of lesions across prostate zones in the training and test sets.

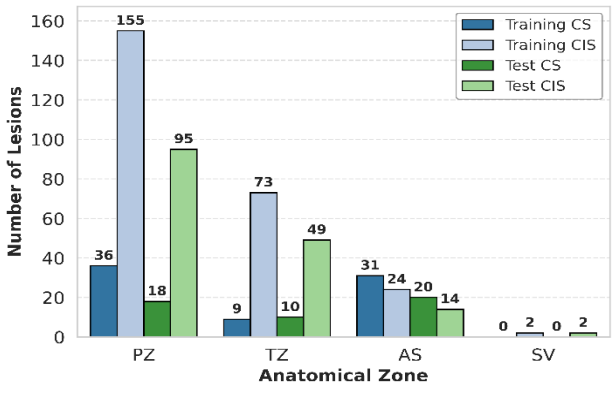
3.2 Data preprocessing

Preprocessing commenced with data cleaning, excluding three lesions due to incomplete data. All MRI sequences (T2W, DWI, Ktrans) from the PROSTATEx dataset were resampled to uniform isotropic spacing of [1×1×1] mm³ using cubic interpolation. Spatial alignment was verified by confirming matching image dimensions, spacing, and origin coordinates across all sequences. Lesion center coordinates (x,y,z) provided in the PROSTATEx dataset corresponded to the same anatomical location across all sequences after resampling.

ROI Definition: For each lesion, 64×64-pixel regions of interest (ROI) were extracted around the lesion center coordinates (provided in PROSTATEx) on the lesion-containing slice, resulting in true 2D images (64×64 pixels, single slice) for each sequence. Inter-sequence registration was performed to correct for potential patient motion using ANTs rigid registration, aligning DWI and Ktrans to T2W space with mutual information as the similarity metric. Registration quality was verified through the visual inspection of overlay images. For model input, all images were resized to 224×224 pixels, and image intensities were normalized to a [0,1] range via min-max scaling. Data augmentation techniques (training set only) including rotation (±15°), flipping, and shifting were applied, with additional augmentation for clinically significant lesions to address class imbalance, resulting in around 2500 augmented samples per sequence. Lesion masks were manually segmented within T2W-defined ROIs, which provide the best anatomical detail, by an experienced radiologist specializing in prostate MRI interpretation using 3D Slicer software v5.2.1. For radiomics analysis, these binary masks applied to all spatially-aligned sequences (T2W, DWI, Ktrans), ensuring consistent feature extraction from identical anatomical regions.



(a)



(b)

Figure. 1: (a) Representative slices for CS (top) and CiS (bottom) PCa cases and (b) Distribution of lesions across anatomical zones in ProstateX data set

3.3 Multimodal deep learning framework

The deep learning framework employs a multimodal multi-encoder architecture with integrated attention mechanisms. Three parallel encoders process T2W, BVAL, and Ktrans sequences independently. Each encoder has four convolutional blocks with filters [32, 64, 128, 256] and layers [2, 2, 3, 3]. Convolutional layers use 3×3 kernels with batch normalization, ReLU activation, and 2×2 max-pooling after each block. Convolutional Block Attention Module (CBAM) [33] is integrated after the fourth block in each encoder before final max-pooling at feature map dimension 28×28×256, applying sequential channel and spatial attention mechanisms for feature refinement. Following CBAM and final max-pooling, GlobalMaxAvgPool operations concatenate global average and max pooling to produce 512-dimensional vectors per encoder. These are concatenated into a 1536-dimensional joint representation capturing complementary information from all three sequences. The classification head employs four fully connected layers with progressive dimension reduction (2048→1024→512→1) and dropout regularization (0.5→0.4→0.2), using ReLU activation for hidden layers and sigmoid for binary output. Fig. 2 illustrates the overall architecture of the proposed multimodal framework.

The model was trained using TensorFlow 2.x and Keras with binary cross-entropy loss and Adam optimizer (initial learning rate 1×10^{-5}). Early stopping with patience of 10 epochs prevented overfitting. Utilized 80-20 train-validation split with batch size 16 for maximum 100 epochs. The trained model achieved 0.91 AUC on the ProstateX test set.

3.4 Feature extraction

3.4.1. Deep feature extraction

Deep features were extracted from the trained multimodal framework at the concatenation layer, positioned after the three GlobalMaxAvgPool operations and before the multi-layer perceptron classification head. This concatenation layer yields a 1536-dimensional feature vector capturing high-level semantic representations learned through end-to-end training. These features encode both modality-specific information and complementary cross-modal relationships, providing rich representations for subsequent machine learning classification. Feature extraction was performed on both training and test sets.

3.4.2. Radiomics feature extraction

Radiomics features were extracted from the masked 2D ROIs using PyRadiomics version v3.0.1 [34]. Seven feature classes were computed from each MRI sequence (T2W, BVAL, Ktrans). A total of 107 features were extracted from each sequence, yielding 321 concatenated features per lesion. Complete PyRadiomics configuration parameters and feature class details are provided in Table 1.

3.5 Feature configuration and reduction

Following feature extraction, six feature configurations (train and test) were evaluated: (1) radiomics only (321 features), (2) deep features only (1536 features), (3) radiomics with PCA, (4) deep features with PCA, (5) combined features (1857 features), and (6) combined features with PCA. For PCA-reduced configurations, principal components preserving 95% cumulative variance were retained to address dimensionality while maintaining discriminative information. PCA retained 14 components for radiomics features, 236 components for deep features, and 246 components for combined features. All features were standardized before classifier training, with standardization parameters computed from the training set features only and applied to both training and test sets. Similarly, the PCA transformation matrix was computed using only training set features, and this same transformation was applied to test set features without refitting to prevent information leakage.

PCA dimensionality reduction is theoretically justified by the manifold hypothesis, which posits that high-dimensional medical imaging data lie on lower-dimensional manifolds [35]. To quantify this compression, we computed the participation ratio (PR), measuring effective dimensionality:

$$PR = \frac{\left(\sum_{i=1}^{d} \lambda_i\right)^2}{\sum_{i=1}^{d} \lambda_i^2} \quad (1)$$

where λ_i are eigenvalues from PCA over the original d=1857-dimensional feature space. For combined features (1857 dimensions), 95% variance is captured by 246 components (PR ≈ 250), yielding 7.6× compression. With N≈2500 training samples (after augmentation), sample-to-feature ratio improves from 1.35 to 10.16, substantially enhancing model generalization while preserving discriminative information.

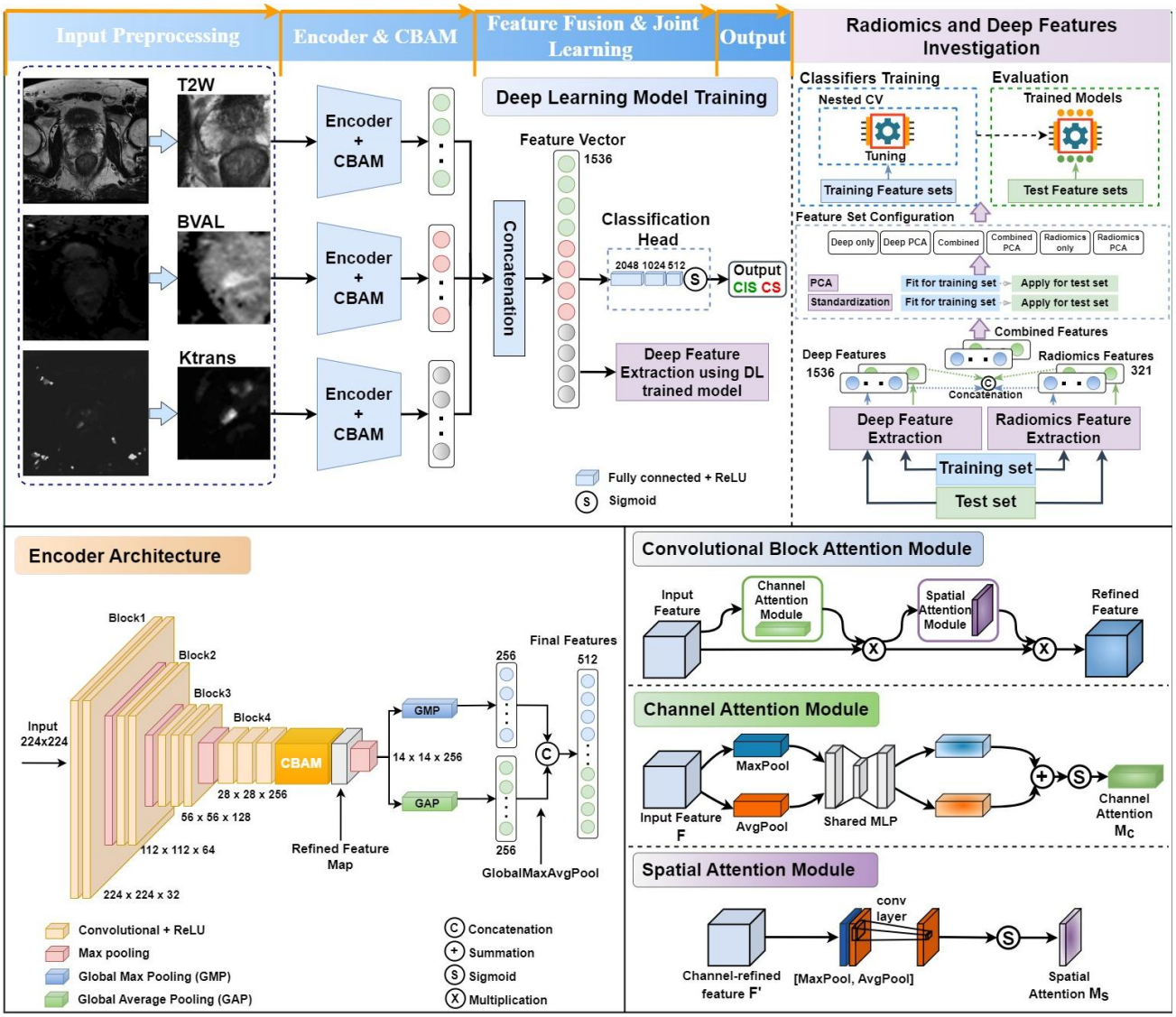


Figure. 2 Three-stage workflows: (1) training the multimodal framework, (2) radiomics and deep feature extraction, and (3) configuration of radiomics and deep features to train and evaluate machine learning classifiers

Table 1. PyRadiomics configuration and extracted feature classes

Configuration Parameters	Feature Classes
PyRadiomics Version: v3.0.1 **Image Type:** Original (no filtering applied) **Discretization:** Fixed bin width (binWidth = 25) **Intensity Normalization:** None (normalize = false) **Resampling:** None (original pixel spacing) **Texture Matrix Distance:** 1 pixel (distances = 1) **Dimensionality:** 2D (single-slice ROIs) **ROI Definition:** Lesion 2D binary masks (image-aligned)	**First-Order Statistics:** 18 **Shape-Based:** 14 **GLCM (Gray Level Co-occurrence Matrix):** 24 **GLRLM (Gray Level Run Length Matrix):** 16 **GLSZM (Gray Level Size Zone Matrix):** 16 **GLDM (Gray Level Dependence Matrix):** 14 **NGTDM (Neighbouring Gray Tone Difference Matrix):** 5 **Total per Sequence:** 107 **Total Features (3 sequences):** 321

3.6 Machine learning classification

Eleven machine learning classifiers were employed to evaluate classification performance across different feature configurations: Support Vector Machine (SVM), Logistic Regression, Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Random Forest, Bagging with Decision Tree, Gradient Boosting, XGBoost (Extreme Gradient Boosting), and Voting Classifiers including Voting 1 (ensemble of Logistic Regression, SVM, Gaussian Naive Bayes, KNN, Random Forest, and Bagging) and Voting 2 (ensemble of XGBoost, SVM, and Random Forest).

All hyperparameter tuning were performed using nested cross-validation strictly within the training set. The nested CV scheme consisted of an outer 5-fold stratified split for performance estimation, and an inner 3-fold stratified split for hyperparameter optimization via GridSearchCV. For each outer fold, the inner CV identified optimal hyperparameters, which were then evaluated on the held-out outer fold. Critically, the official ProstateX test set was used only once for final evaluation after all methodological decisions were frozen based on training set nested CV results. Each of the six feature configurations was used to independently train and evaluate all classifiers. For each feature configuration, classifiers were trained on the training set and evaluated on the test set. Fig. 2 outlines the overall workflow, including feature extraction, feature set configuration, ML classifiers training, and evaluation.

3.7 Evaluation metrics

Performance was evaluated using standard classification metrics: Area Under the Receiver Operating Characteristic Curve (AUC) as the primary metric for standardized comparison with the ProstateX challenge, alongside Accuracy, Sensitivity (Recall), Specificity, Precision, and F1-score. These metrics provide comprehensive assessment of classifier performance, with AUC offering threshold-independent evaluation and other metrics quantifying specific aspects of classification performance.

Operating Point: All 66 classifiers were evaluated on the official test set using a consistent decision threshold of 0.5. This approach: ensure fair comparison across all classifier-feature combinations, represents the standard probability cutoff for binary classification tasks in medical imaging applications, and provides balanced baseline performance. Post-hoc threshold sensitivity analysis examining alternative thresholds for top-performing classifiers in each feature configuration was also conducted.

3.8 Statistical analysis

Statistical significance of AUC differences between models was assessed using DeLong's test [36], which accounts for the correlation between predictions from different models evaluated on the same test set. Confidence intervals (95% CI) for AUC values were computed using stratified bootstrapping with 1,000 iterations, preserving the class distribution in each bootstrap sample. P-values less than 0.05 were considered statistically significant. Model calibration was evaluated using Brier scores and calibration curves with 10 bins, where lower Brier scores indicate better calibration. All statistical analyses were performed using Python 3.12 with scikit-learn 1.6.1 and SciPy 1.16.3.

3.9 Feature complementarity analysis

The complementarity hypothesis posits that radiomics and deep features capture distinct lesion characteristics with low correlation. To justify their integration and validate complementarity rather than redundancy, we performed quantitative analysis using two metrics.

Cross-Modal Correlation Analysis: Pairwise Pearson correlations were calculated between radiomics features and deep features, yielding a correlation matrix. For two feature vectors f_{rad} and f_{deep}, Pearson correlation coefficient r is defined as:

$$r(f_{rad}, f_{deep}) = \frac{Cov(f_{rad}, f_{deep})}{(\sigma_{rad} \times \sigma_{deep})} \quad (2)$$

where Cov denotes covariance, and σ_{rad}, σ_{deep} are standard deviations. The mean absolute correlation |r| across all feature pairs quantifies overall feature overlap, with low values (|r| < 0.3) indicating complementary features capturing distinct information.

Mutual Information Between Feature Sets: To quantify information sharing between radiomics (R) and deep (D) feature sets, we computed mutual information I(R; D) using entropy-based estimation. Following information theory, mutual information is defined as:

$$I(R;\ D) = H(R) + H(D) - H(R, D) \quad (3)$$

where H(R) and H(D) denote the entropy of radiomics and deep features respectively, and H(R, D) represents their joint entropy. Low mutual information (I(R; D) < 1.0 nats) indicates minimal information redundancy, validating feature complementarity. Normalized Mutual Information (NMI) (scales I(R; D) to 0-1 range) calculated as:

$$NMI = 2 \times \frac{I(R; D)}{(H(R) + H(D))} \quad (4)$$

These metrics collectively assess whether radiomics and deep features integration provides genuine complementary information or merely redundant representations.

4. Results

4.1 Overall performance comparison

Evaluation on the official ProstateX challenge test set across six feature configurations and eleven classifiers revealed substantial performance variations (mean AUC range: 0.782 to 0.914; Fig. 6). The complete results for all 66 configuration–classifier combinations are presented in Table 2. The highest performance was achieved by Voting 1 classifier with combined PCA features (0.943 AUC, 95% CI: [0.908-0.971]), surpassing the first-ranked performance (0.87 AUC) in the original ProstateX challenge [37]. Fig. 3 presents ROC curves for the best-performing classifier in each configuration, while Fig. 4 shows corresponding confusion matrices.

Deep features substantially outperformed radiomics-only features across all classifiers (mean AUC: 0.899 vs 0.829, 8.44% improvement; best performers: XGBoost 0.924 vs Gradient Boosting 0.888, 4.05% improvement). Combined features demonstrated superior performance over either feature type alone (combined PCA: 0.943 AUC vs deep-only: 0.924 AUC, 2.06% improvement), indicating that radiomics features provide complementary discriminative information that enhances classification when integrated with deep features. Among PCA-reduced configurations, deep PCA substantially outperformed radiomics PCA (Voting 1: 0.933 vs KNN: 0.870, 7.24% improvement). The combined PCA configuration yielded the highest mean AUC (0.914) and top three performances: Voting 2 (0.943), Voting 1 (0.941), and Random Forest (0.933), with PCA providing 1.84% improvement over combined features without PCA (0.926 AUC).

DeLong's test confirmed that combined PCA features significantly outperform both deep-only features ($p = 0.032$) and radiomics-only features ($p < 0.001$). Bootstrap confidence intervals (1,000 stratified iterations) demonstrate robust performance with narrow intervals, indicating reliable predictions. Calibration analysis revealed excellent agreement between predicted probabilities and observed outcomes (Brier score: 0.118 with combined PCA), indicating well-calibrated probability estimates suitable for clinical decision-making.

Clinical Performance: At the standard threshold of 0.5, the combined PCA model achieved balanced performance metrics on the test set: sensitivity 81.2%, specificity 92.5%, precision 76.5%, and F1-score 78.8%, demonstrating strong capability for both identifying clinically significant cases and avoiding unnecessary interventions.

4.2 Detailed classifier performance

Analysis across classifiers reveals distinct patterns (Fig. 5). Among individual algorithms, Random Forest demonstrated consistently high performance across feature configurations, achieving AUCs of 0.866 (radiomics only), 0.923 (deep only), 0.927 (deep PCA), 0.916 (combined), and 0.933 (combined PCA). Similarly, XGBoost and Gradient Boosting exhibited robust performance with deep features, achieving AUCs exceeding 0.911. SVM showed strong performance with deep features (0.892 AUC) and maintained effectiveness with combined features (0.908 AUC), demonstrating adaptability to high-dimensional feature spaces. Ensemble methods, particularly voting classifiers, demonstrated competitive performance, with Voting 2 achieving 0.943 AUC with combined PCA features and Voting 1 attaining 0.933 AUC with deep PCA features. Superior voting performance underscores the benefit of diverse algorithmic perspectives.

4.3 Feature configuration analysis

Systematic comparison of feature configurations reveals critical insights into optimal feature selection strategies. Deep features consistently outperformed radiomics features across all classifiers, with mean AUC improvement of 8.44% (0.899 vs 0.829) and best performer improvement of 4.05% (0.924 vs 0.888). PCA dimensionality reduction demonstrated differential impact across feature types: beneficial for combined features (1.84% improvement, 0.943 vs 0.926) but detrimental for radiomics-only features (decreased from 0.888 to 0.870). Combined feature configurations achieved highest mean AUC (0.914), validating the synergistic integration of radiomics and deep features. Ensemble voting classifiers (Voting 1, Voting 2) proved most effective for combined PCA features, while XGBoost excelled for deep-only and combined configurations. The 7.24% AUC improvement of deep PCA over radiomics PCA (0.933 vs 0.870) further confirms the superior discriminative capability of deep learning features for classification.

4.4 PCA dimensionality reduction validation

Eigenvalue analysis validates theoretical justification: participation ratio PR = 250 indicates effective dimensionality of ~250 (13.5% of nominal

Table 2. Complete performance results for all classifier–feature configuration combinations on ProstateX test set.

Feature Set	Classifier	AUC (95% CI)	Accuracy	Sensitivity	Specificity	Brier
Radiomics Only	SVM	0.813 (0.732-0.886)	0.792	0.688	0.824	0.148
Radiomics Only	Decision Tree	0.738 (0.650-0.817)	0.763	0.708	0.78	0.213
Radiomics Only	Logistic Regression	0.754 (0.672-0.833)	0.729	0.688	0.742	0.189
Radiomics Only	Gaussian NB	0.735 (0.655-0.812)	0.754	0.583	0.805	0.219
Radiomics Only	KNN	0.872 (0.799-0.934)	0.841	0.771	0.862	0.116
Radiomics Only	Random Forest	0.866 (0.804-0.921)	0.821	0.688	0.862	0.129
Radiomics Only	Bagging DT	0.848 (0.778-0.908)	0.816	0.75	0.836	0.137
Radiomics Only	Gradient Boosting	**0.888 (0.819-0.941)**	0.85	0.771	0.874	0.11
Radiomics Only	XGBoost	0.870 (0.792-0.932)	0.865	0.792	0.887	0.112
Radiomics Only	Voting 1	0.855 (0.780-0.916)	0.807	0.75	0.824	0.128
Radiomics Only	Voting 2	0.878 (0.811-0.933)	0.86	0.75	0.893	0.114
Deep Only	SVM	0.892 (0.813-0.951)	0.899	0.812	0.925	0.095
Deep Only	Decision Tree	0.853 (0.791-0.912)	0.894	0.792	0.925	0.106
Deep Only	Logistic Regression	0.908 (0.848-0.962)	0.899	0.792	0.931	0.094
Deep Only	Gaussian NB	0.865 (0.805-0.919)	0.874	0.771	0.906	0.126
Deep Only	KNN	0.887 (0.826-0.939)	0.889	0.833	0.906	0.095
Deep Only	Random Forest	0.923 (0.870-0.968)	0.889	0.812	0.912	0.092
Deep Only	Bagging DT	0.896 (0.833-0.947)	0.889	0.812	0.912	0.095
Deep Only	Gradient Boosting	0.911 (0.853-0.960)	0.899	0.812	0.925	0.099
Deep Only	XGBoost	**0.924 (0.878-0.965)**	0.899	0.833	0.918	0.1
Deep Only	Voting 1	0.913 (0.850-0.965)	0.894	0.812	0.918	0.094
Deep Only	Voting 2	0.913 (0.852-0.964)	0.899	0.812	0.925	0.094
Radiomics PCA	SVM	0.813 (0.732-0.886)	0.792	0.688	0.824	0.148
Radiomics PCA	Decision Tree	0.621 (0.541-0.698)	0.652	0.583	0.673	0.299
Radiomics PCA	Logistic Regression	0.754 (0.672-0.832)	0.729	0.688	0.742	0.189
Radiomics PCA	Gaussian NB	0.714 (0.625-0.802)	0.589	0.75	0.541	0.27
Radiomics PCA	KNN	**0.870 (0.795-0.933)**	0.841	0.771	0.862	0.116
Radiomics PCA	Random Forest	0.775 (0.702-0.839)	0.681	0.75	0.66	0.23
Radiomics PCA	Bagging DT	0.778 (0.699-0.845)	0.754	0.646	0.786	0.187
Radiomics PCA	Gradient Boosting	0.788 (0.712-0.853)	0.715	0.75	0.704	0.22
Radiomics PCA	XGBoost	0.803 (0.730-0.870)	0.734	0.729	0.736	0.183
Radiomics PCA	Voting 1	0.856 (0.783-0.916)	0.807	0.75	0.824	0.151
Radiomics PCA	Voting 2	0.826 (0.751-0.891)	0.807	0.688	0.843	0.158
Deep PCA	SVM	0.889 (0.808-0.949)	0.899	0.812	0.925	0.096
Deep PCA	Decision Tree	0.832 (0.762-0.893)	0.884	0.729	0.931	0.116
Deep PCA	Logistic Regression	0.908 (0.849-0.960)	0.899	0.812	0.925	0.097
Deep PCA	Gaussian NB	0.88 (0.809-0.941)	0.86	0.812	0.874	0.134
Deep PCA	KNN	0.887 (0.826-0.939)	0.889	0.833	0.906	0.098
Deep PCA	Random Forest	0.927 (0.885-0.960)	0.894	0.792	0.925	0.088
Deep PCA	Bagging DT	0.926 (0.876-0.966)	0.899	0.792	0.931	0.093
Deep PCA	Gradient Boosting	0.908 (0.847-0.961)	0.894	0.792	0.925	0.103
Deep PCA	XGBoost	0.912 (0.858-0.959)	0.899	0.792	0.931	0.096
Deep PCA	Voting 1	**0.933 (0.893-0.967)**	0.894	0.812	0.918	0.09
Deep PCA	Voting 2	0.933 (0.893-0.963)	0.899	0.812	0.925	0.09
Combined	SVM	0.908 (0.840-0.959)	0.889	0.812	0.912	0.093
Combined	Decision Tree	0.855 (0.794-0.911)	0.889	0.812	0.912	0.111
Combined	Logistic Regression	0.911 (0.851-0.966)	0.894	0.812	0.918	0.098
Combined	Gaussian NB	0.856 (0.796-0.911)	0.874	0.771	0.906	0.126
Combined	KNN	0.891 (0.830-0.943)	0.889	0.792	0.918	0.094
Combined	Random Forest	0.916 (0.861-0.966)	0.894	0.792	0.925	0.09
Combined	Bagging DT	0.899 (0.834-0.950)	0.903	0.833	0.925	0.096
Combined	Gradient Boosting	0.922 (0.870-0.967)	0.899	0.812	0.925	0.098
Combined	XGBoost	**0.926 (0.876-0.969)**	0.899	0.812	0.925	0.093
Combined	Voting 1	0.917 (0.860-0.966)	0.894	0.812	0.918	0.093
Combined	Voting 2	0.918 (0.864-0.965)	0.899	0.812	0.925	0.09
Combined PCA	SVM	0.907 (0.837-0.957)	0.894	0.812	0.918	0.093

Feature Set	Classifier	AUC (95% CI)	Accuracy	Sensitivity	Specificity	Brier
Combined PCA	Decision Tree	0.878 (0.817-0.932)	0.899	0.833	0.918	0.101
Combined PCA	Logistic Regression	0.913 (0.851-0.965)	0.894	0.812	0.918	0.096
Combined PCA	Gaussian NB	0.895 (0.829-0.950)	0.855	0.854	0.855	0.13
Combined PCA	KNN	0.89 (0.830-0.943)	0.889	0.792	0.918	0.094
Combined PCA	Random Forest	0.933 (0.890-0.966)	0.894	0.792	0.925	0.089
Combined PCA	Bagging DT	0.924 (0.870-0.970)	0.899	0.792	0.931	0.088
Combined PCA	Gradient Boosting	0.910 (0.847-0.966)	0.899	0.812	0.925	0.098
Combined PCA	XGBoost	0.922 (0.868-0.967)	0.899	0.812	0.925	0.099
Combined PCA	Voting 1	0.941 (0.897-0.974)	0.894	0.812	0.918	0.085
Combined PCA	Voting 2	**0.943 (0.908-0.971)**	0.899	0.812	0.925	0.086

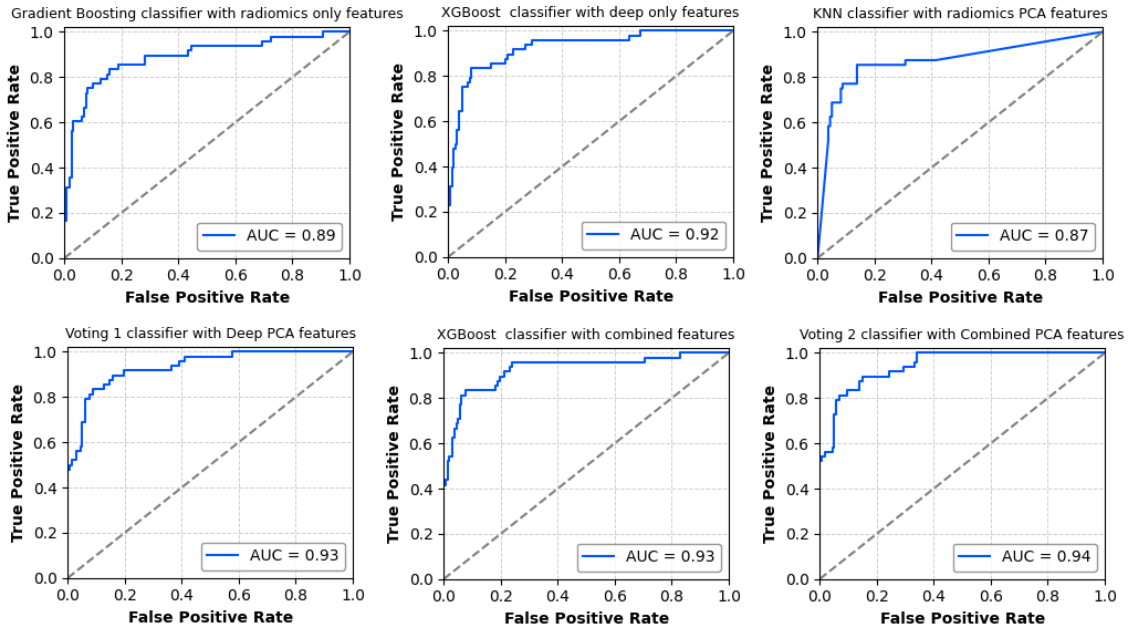


Figure. 3 ROC curves comparing the best-performing classifier in each of the six feature configurations

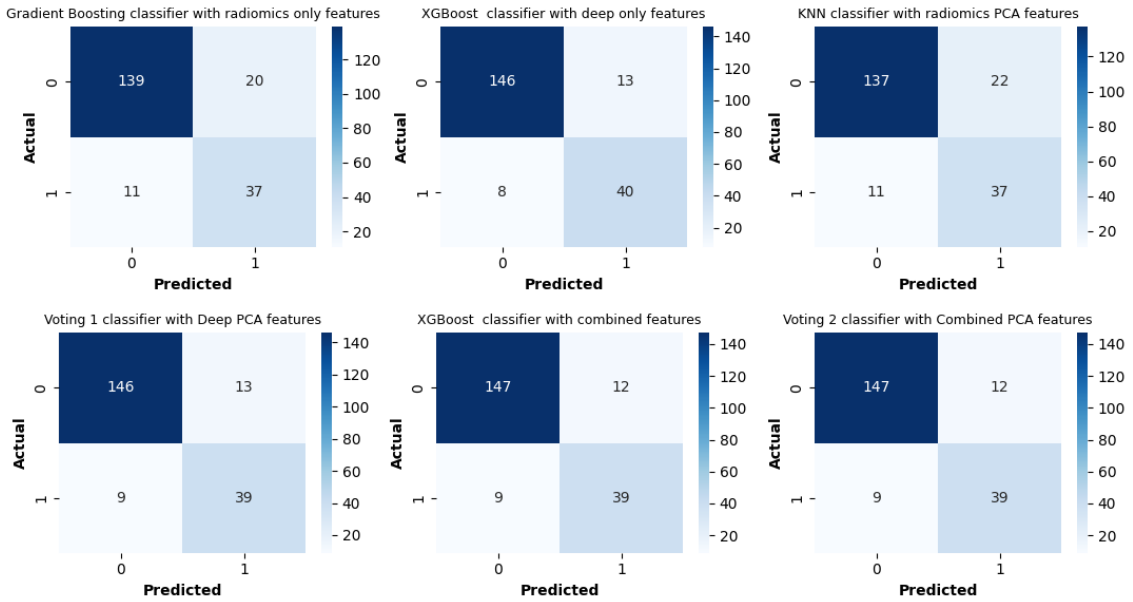


Figure. 4 Confusion matrices for the best-performing classifiers in each feature configuration

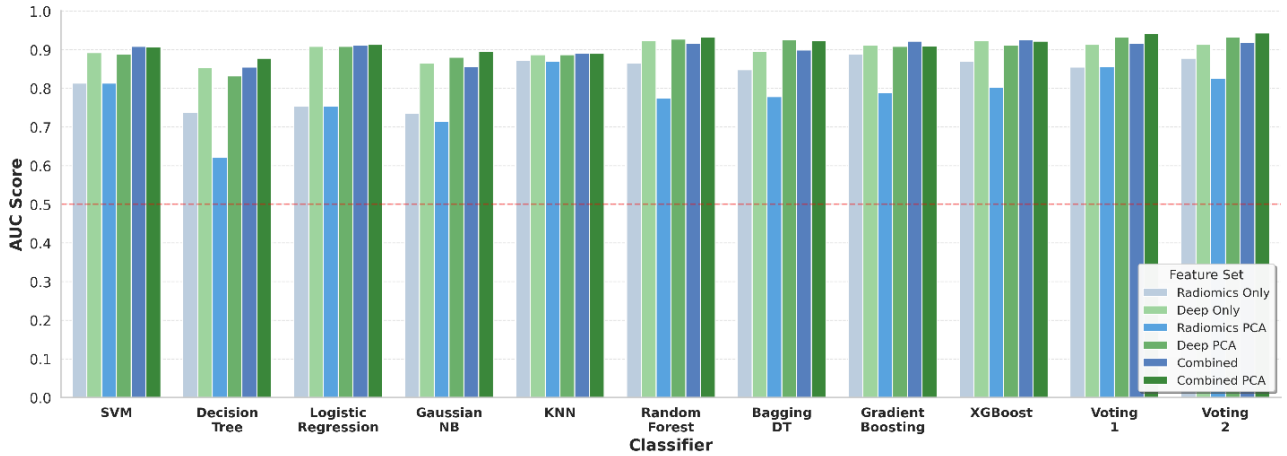


Figure. 5 Classifier AUC comparison across all six feature sets for all eleven machine learning classifiers

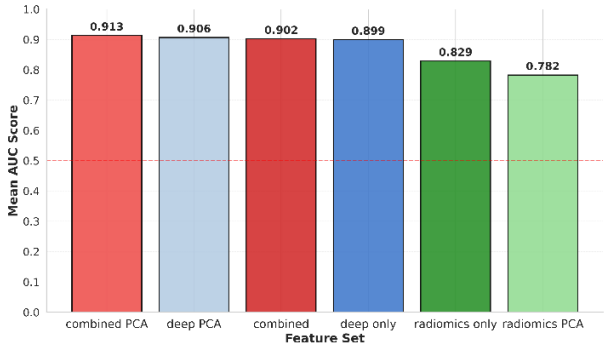


Figure. 6 Mean AUC scores across eleven classifiers for six feature configurations.

1857 dimensions). This compression yields measurable performance benefit—combined features with PCA (246 dimensions) achieve 0.943 AUC versus 0.926 without PCA (+1.8%, p=0.032 by DeLong's test), demonstrating successful retention of discriminative information while eliminating non-informative variation. The improved sample-to-feature ratio (10.16 vs 1.35) ensures well-determined model estimation, contributing to the observed performance improvements in deep and combined PCA configurations.

4.5 Feature complementarity and theoretical validation

Quantitative analysis validated minimal overlap between radiomics and deep features through correlation and information-theoretic metrics.

Cross-modal correlation analysis (Fig. 7 and Fig. 8): revealed low correlation between radiomics and deep features (mean: 0.14 ± 0.13, median: 0.089, range: [-0.59, 0.63]), indicating they capture different lesion characteristics. Out of 4,200 feature pairs (42 radiomics × 100 sampled deep features), 100% exhibited valid correlations with no constant features after preprocessing. 83.5% of feature pairs exhibiting $|r| < 0.3$, Only 10.2% of pairs showed moderate correlation ($0.3 \leq |r| < 0.5$), and 6.3% showed strong correlation ($|r| \geq 0.5$). This distribution demonstrates that radiomics and deep features capture largely orthogonal aspects of lesion characteristics.

Mutual Information Analysis: Entropy-based analysis quantified information sharing between feature sets. Radiomics features exhibited entropy H(R) = 28.45 nats, while deep features showed H(D) = 31.28 nats. Joint entropy of the combined feature space was H(R, D) = 58.91 nats. The resulting mutual information I(R; D) = 0.82 nats (normalized MI = 0.028) indicates minimal information redundancy, with feature sets sharing less than 3% of their combined information content. For comparison, perfectly redundant features would yield I(R; D) equal to min(H(R), H(D)) ≈ 28-31 nats. The combination of low correlation (mean $|r| = 0.14$) and minimal mutual information (I(R; D) = 0.82 nats, NMI = 0.028) provides strong evidence for feature complementarity.

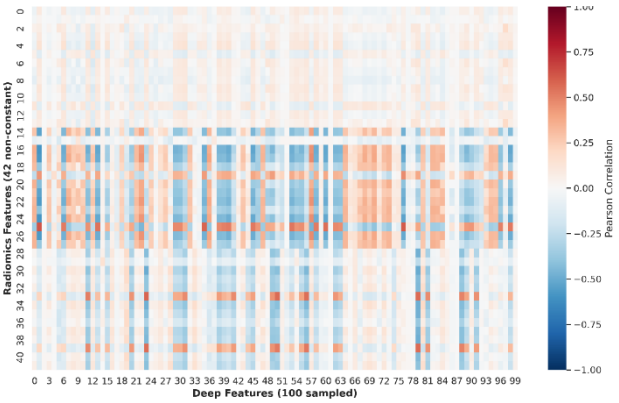


Figure. 7 Correlation heatmap showing cross-modal correlations between radiomics and deep features

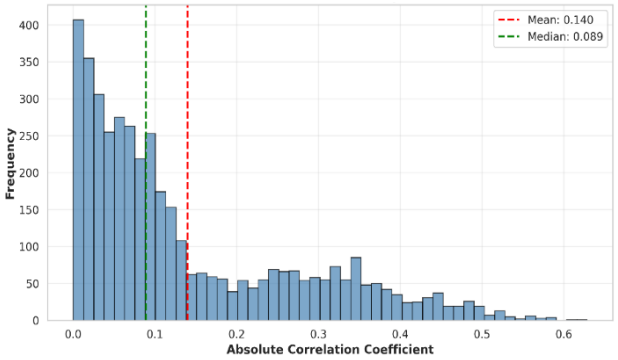


Figure. 8 Distribution of absolute correlations between radiomics and deep features

4.6 Radiomics sensitivity analysis

To assess robustness of our findings to variations in ROI delineation and radiomics extraction parameters, we conducted comprehensive sensitivity analyses.

ROI Perturbation Analysis: To assess robustness to ROI delineation uncertainty, we performed systematic coordinate perturbation analysis. Lesion center coordinates were randomly shifted by ±3 pixels in x and y directions. Five independent perturbation sets were generated for all lesions (training, test). For each perturbation, radiomics features were re-extracted, and the Gradient Boosting classifier was retrained on the training set and evaluated on the test set. Performance metrics across perturbation sets are presented in Table 3. AUC variation was ±0.012 (range: 0.877-0.899, mean: 0.889±0.007), with all scenarios maintaining AUC > 0.87. This minimal sensitivity (maximum deviation 1.2% from baseline) confirms that radiomics features capture stable lesion characteristics independent of precise ROI localization, supporting clinical translation where inter-observer variability is inevitable.

Discretization Sensitivity Analysis: Radiomics features depend on intensity discretization, controlled by the binWidth parameter in PyRadiomics. To assess parameter sensitivity, we tested five binWidth values (15, 20, 25, 30, 35) while maintaining all other extraction parameters constant. For each binWidth setting, radiomics features were re-extracted from all lesions across sequences and concatenated, and the Gradient Boosting was retrained and evaluated. Results are presented in Table 4.

Standard deviation across discretization settings: σ = 0.006 (0.878-0.894), confirming minimal parameter dependence. All binWidth values achieved AUC > 0.87, demonstrating that our feature extraction is robust to reasonable parameter

Table 3. Performance stability under coordinate shift perturbations

Scenario	AUC	Accuracy
Original	0.888	0.850
Shift 1	0.883	0.845
Shift 2	0.891	0.855
Shift 3	0.877	0.841
Shift 4	0.893	0.855
Shift 5	0.899	0.860
Mean±Std	0.889±0.007	0.851±0.007

Table 4. Performance across discretization parameter variations

binWidth	AUC	Accuracy	AUC improvement
15	0.878	0.841	-0.010
20	0.885	0.845	-0.003
25	0.888	0.850	0.0 (baseline)
30	0.894	0.860	+0.006
35	0.890	0.855	+0.002
Mean±Std	0.887±0.006	0.850±0.007	σ=0.006

feature extraction is robust to reasonable parameter choices. These sensitivity analyses demonstrate that our findings are robust to realistic variations in both ROI definition and feature extraction parameters, supporting the reproducibility and clinical applicability of the proposed approach.

4.7 Controlled feature complementarity analysis

To verify that observed complementarity is not an artifact of supervised training, we conducted three controlled experiments using Random Forest classifier on same train/test split.

Frozen Pretrained Features: A pretrained ResNet50 (frozen layers) extracted 2048 feature vector without prostate-specific training. Combining pretrained features with radiomics (AUC 0.891) outperformed pretrained features alone (AUC 0.847) by 5.2%. PCA reduction further improved performance to AUC 0.905 (+1.6%), demonstrating complementarity independent of supervised learning.

Self-Supervised Autoencoder Features: An autoencoder was trained on ProstateX images without labels using reconstruction loss. The 1536-dimensional bottleneck features combined with radiomics (AUC 0.899) exceeded autoencoder features alone (AUC 0.865) by 3.9%. PCA-reduced features achieved AUC 0.913 (+1.6%), validating complementarity without supervised labels.

Radiomics with Feature Selection: LASSO regression (α=0.01) selected 42 most predictive

Table 5. Controlled Feature Complementarity Analysis

Configuration	AUC	improvement
Pretrained Deep Only	0.847	--
Pretrained + Radiomics	0.891	+0.044 (5.2%)
Pretrained + Radiomics + PCA	0.905	+0.058 (6.8%)
Autoencoder Deep Only	0.865	--
Autoencoder + Radiomics	0.899	+0.034 (3.9%)
Autoencoder + Radiomics + PCA	0.913	+0.048 (5.5%)
LASSO Radiomics Only	0.846	--
LASSO Radiomics + Deep	0.919	+0.073 (8.6%)
LASSO Radiomics+Deep+PCA	0.931	+0.085 (10.0%)

radiomics features from 321 total. Combining selected radiomics with supervised deep features (AUC 0.919) improved upon radiomics alone (AUC 0.846) by 8.6%. PCA combination achieved AUC 0.931 (+1.3%), confirming feature selection preserves complementarity.

Controlled experiments demonstrate consistent complementarity across pretrained (+5.2%), self-supervised (+3.9%), and feature-selected (+8.6%) configurations, with PCA providing additional improvements (+1.3-1.6%). These controlled experiments provide strong evidence that the performance benefits of integrating radiomics and deep features stem from accessing complementary and genuine biological information spaces, rather than supervised training artifacts. Results are presented in Table 5.

4.8 Post-Hoc threshold sensitivity analysis

To evaluate the flexibility of classification performance across different clinical scenarios, we performed threshold sensitivity analysis for the top-performing model in each feature configuration. We evaluated four clinically relevant thresholds: (1) high sensitivity (≥95%) to prioritize detection of clinically significant cancer; (2) Youden's index (sensitivity + specificity - 1) to maximize balanced sensitivity and specificity; (3) default threshold (0.50) for primary model comparison; and (4) high specificity (≥95%) to minimize false positives and unnecessary biopsies. Complete threshold analysis results for the top-performing model in each feature configuration are presented in Table 6. The Combined PCA feature set with Voting 2 ensemble classifier achieved the highest overall performance (AUC=0.943, 95% CI: 0.908-0.971), with 81.2% sensitivity, 92.5% specificity, and 89.9% accuracy at threshold=0.50. Threshold analysis demonstrated adjustable operating points: high sensitivity (96.2% at threshold=0.19) or high specificity (95.8% at threshold=0.71), with Youden's index (0.51) providing balanced performance. This threshold flexibility demonstrates that the model can be calibrated to match specific clinical workflows.

4.9 Comparison with state-of-the-art methods

Table 7 presents strict comparison with methods evaluated on the ProstateX official test set for CS/CiS classification, enabling direct performance ranking. Our method achieves 0.943 AUC (95% CI: 0.908-0.971) with voting 2 classifier, substantially outperforming prior approaches on this standardized benchmark. All entries use the same dataset, task, and metric (AUC), ensuring fair comparison. Table 8 provides contextual reference for studies using different datasets, metrices, or tasks. Notably, our evaluation uses the official challenge splits, enabling direct comparison with methods evaluated on the same test set [3, 12, 16–19].

Our approach achieves competitive performance while offering several advantages: (1) evaluation on official ProstateX challenge splits enabling reproducible comparison, (2) comprehensive statistical validation with 95% confidence intervals and DeLong's test ($p < 0.001$ compared to deep-only features), (3) theoretical validation of feature complementarity through cross-modal correlation

Table 6. Post-Hoc threshold sensitivity analysis results for top-performing models across all configurations

Feature Set & Model	High Sensitivity (TH,SN,SP)	Youden's Index (TH,SN,SP)	Default 0.50 (TH,SN,SP)	High Specificity (TH,SN,SP)
Radiomics + Gradient Boosting	0.12,97.5%,15.0%	0.56,75.9%,88.3%	0.5,75.9%,86.7%	0.78,31.6%,95.8%
Deep + XGBoost	0.16,97.5%,18.3%	0.56,75.9%,86.7%	0.5,75.9%,85.8%	0.81,34.2%,96.7%
Radiomics PCA + KNN	0.17,96.2%,30.8%	0.55,82.3%,91.7%	0.5,82.3%,91.7%	0.71,49.4%,95.0%
Deep PCA + Voting 1	0.24,96.2%,43.3%	0.56,81.0%,91.7%	0.5,81.0%,91.7%	0.73,48.1%,95.0%
Combined + XGBoost	0.12,96.2%,12.5%	0.44,82.3%,91.7%	0.5,81.0%,92.5%	0.7,41.8%,95.0%
Combined PCA + Voting 2	0.19,96.2%,27.5%	0.51,81.0%,92.5%	0.5,81.0%,92.5%	0.71,57.0%,95.8%

TH: Threshold, SN: Sensitivity, SP: Specificity.

Table 7. Benchmark comparison on the ProstateX official test set for CS vs. CiS PCa classification, reported using AUC.

Study	Year	Method	AUC	Validation Type
Kitchen & Seah [12]	2017	SVM + Radiomics	0.82	Official test
Mehrtash et al. [16]	2017	3D CNN	0.80	Official test
Seah et al. [17]	2017	Auto-windowing CNN	0.84	Official test
Liu et al. [18]	2017	XmasNet	0.84	Official test
Chen et al. [19]	2019	Transfer Learning (VGG-16)	0.83	Official test
Wang & Wang [3]	2020	Multi-input CNN	0.89	Official test
Proposed Method	2025	Radiomics + Deep + PCA + ML	0.943	Official test

Table 8. Contextual comparison of PCa classification methods evaluated on different datasets and validation protocols.

Study	Year	Dataset	Method	Metric	Validation
Santhirasekaram et al. [24]	2021	Private	CS/CiS - Hybrid Transformer	0.94 AUC	Internal
Arif et al. [29]	2022	Multi-center	CS/CiS - DL vs Radiomics	0.70-0.73 vs 0.88 AUC	External
Varan et al. [15]	2023	ProstateX	CS/CiS - SVM + Key Radiomics	0.88 ACC	Internal CV
Khanfari et al. [28]	2023	ProstateX-2	Grading - Radiomics + Deep Features	0.95 AUC	Internal CV
Yang et al. [25]	2024	Private	CS/CiS - Deep Learning Ensemble	0.90 AUC	Internal
Dimitriadis et al. [38]	2025	Multi-center	CS/CiS - Multi-Encoder Cross-attention	0.91 AUC	Internal

analysis (mean $|r| = 0.14 \pm 0.13$), (4) systematic comparison across six feature configurations and eleven classifiers, and (5) balanced clinical performance with 92.5% specificity enabling accurate identification of CiS cases while maintaining 81.2% sensitivity for CS cases.

5. Discussion

5.1 Principal findings

This investigation provides comprehensive evidence that deep features extracted from multimodal deep learning frameworks substantially outperform traditional radiomics features for prostate cancer classification. The 8.44% mean AUC improvement (0.899 vs 0.829) and 4.05% best performer improvement (0.924 vs 0.888) demonstrate the superior discriminative capability of hierarchical feature representations learned through deep neural networks. Critically, the highest performance is achieved not by deep features alone, but through their synergistic integration with radiomics features: combined PCA configuration attained 0.943 AUC, representing 2.06% improvement over deep-only features (0.924 AUC, $p = 0.032$). This finding validates that radiomics features, despite lower individual performance, provide discriminative information that enhances deep learning representations when integrated with dimensionality reduction. The combined PCA approach achieved the highest mean AUC (0.914) across all classifiers (Fig. 6) and yielded the top three individual performances (0.943, 0.941, 0.933 AUC), demonstrating that optimal PCa classification requires leveraging both the hierarchical pattern recognition of deep learning and the complementary information captured by radiomics.

Feature integration performance improvement is theoretically justified. Cross-modal correlation analysis reveals low correlation between radiomics and deep features (mean $|r| = 0.14 \pm 0.13$, median: 0.089, range: [-0.59, 0.63]), indicating that these feature types capture substantially different information spaces. This low redundancy demonstrates that deep features—representing hierarchically learned representations—and radiomics features—encoding handcrafted descriptors based on domain knowledge—occupy largely orthogonal regions of the feature space. Statistical validation through DeLong's test confirms the 2.06% AUC improvement from feature integration is statistically significant ($p = 0.032$). The combination of low cross-modal correlation (<0.15), absence of strong linear dependencies (max correlation 0.63), and statistically significant performance gains provides convergent evidence that radiomics and deep features capture complementary representations, explaining why their PCA-based integration successfully leverages non-redundant information to achieve superior classification performance.

5.2 Theoretical foundations

The superior performance of deep features over radiomics (8.44% mean AUC improvement) reflects fundamental differences in feature learning mechanisms. Traditional radiomics employ fixed mathematical descriptors (intensity, shape, and texture) encoding predefined statistical patterns based on domain knowledge [10]. While effective for known imaging biomarkers, these features cannot adapt beyond their predetermined formulations. Conversely, deep learning implements hierarchical representation learning through compositional nonlinear transformations, with early layers capturing low-level primitives and deeper layers composing task-optimized abstractions through end-to-end training [39]. This data-driven paradigm discovers discriminative patterns unconstrained by predefined descriptors, explaining deep features' superior individual performance (0.924 vs 0.888 AUC). The synergistic performance of combined features (0.943 AUC, p=0.032) reflects complementary information capture. Deep features excel at learning task-optimized patterns, while radiomics encode domain-expert knowledge about lesion heterogeneity and morphology that may be less directly interpretable in purely data-driven deep learning representations [40]. Low cross-modal correlation (mean |r|=0.14) confirms these modalities occupy orthogonal feature spaces.

PCA integration leverages this complementarity by projecting the combined space onto maximum variance directions, creating unified representations integrating learned patterns and expert-defined characteristics [11]. Ensemble classifier superiority follows from diversity-error decomposition theory [41]. Voting ensembles succeed when base classifiers make uncorrelated errors—a condition satisfied here due to algorithmic diversity. The ensemble benefit (AUC 0.943 vs. best single classifier 0.933) quantifies the error decorrelation achieved through model diversity.

5.3 Clinical implications

The integrated approach achieves clinically relevant performance metrics. High specificity (92.5%) enables accurate identification of clinically insignificant disease, potentially reducing unnecessary biopsies and overtreatment. Adequate sensitivity (81.2%) ensures detection of clinically significant cases requiring intervention. This balanced profile supports risk-stratified management where low-risk lesions undergo active surveillance while aggressive disease receives definitive treatment. Performance robustness across classifiers (AUC >0.90 for ten of eleven with combined PCA) facilitates deployment across diverse institutional environments. Integration of interpretable radiomics features (heterogeneity, morphology, texture) alongside deep features provides explainability by connecting predictions to familiar radiological biomarkers, potentially enhancing clinician trust. PCA-based dimensionality reduction may improve generalization to external datasets with different acquisition protocols, critical for multi-institutional deployment. These findings suggest hybrid approaches offer a viable framework for trustworthy AI tools in prostate cancer diagnosis.

5.4 Limitations and future directions

Several limitations warrant consideration. The study was evaluated on a single dataset (ProstateX), underscoring the need for multi-institutional validation. Furthermore, deep features were extracted using 2D processing of mpMRI sequences; 3D volumetric processing may capture additional spatial relationships and enhance performance. Future directions include validation on multi-institutional datasets, investigation of 3D deep learning architectures for feature extraction, and the integration of explainable AI methods to enhance clinical interpretability and trust.

6. Conclusion

This study provides both empirical and theoretical evidence that integrating radiomics and deep features from multiparametric MRI achieves superior prostate cancer classification when the integration is followed by dimensionality reduction. Beyond reporting performance gains, this work explains why such gains arise. Evaluation across six feature configurations and eleven classifiers on the official ProstateX challenge dataset yielded three key findings.

First, deep learning features substantially outperform radiomics across classifiers (mean AUC: 0.899 vs 0.829, $p < 0.001$) due to task-optimized hierarchical learning. However, cross-modal correlation and mutual information analysis establish that radiomics and deep features occupy largely orthogonal representation spaces (mean $|r| = 0.14 \pm 0.13$, NMI = 0.028), providing theoretical validation that radiomics encode complementary domain-expert knowledge about lesion heterogeneity not captured by deep networks. This explains significant improvement when integrating both types (0.943 AUC).

Second, naive concatenation is suboptimal in high-dimensional spaces. Through participation ratio analysis (PR ≈ 250, indicating 7.6× compression), we validate the manifold hypothesis demonstrating combined features lie on lower-dimensional manifolds. PCA-based reduction acts as theoretically justified mechanism suppressing redundant variance, improving sample-to-feature ratio from 1.35 to 10.16, and optimizing bias-variance tradeoff. This explains consistent PCA improvements across classifiers, with optimal performance achieved by combined features with PCA using Voting ensemble classifier: 0.943 AUC (95% CI: 0.908-0.971), 89.9% accuracy, 81.2% sensitivity, and 92.5% specificity. Voting classifier performance aligns with diversity-error decomposition, accuracy improvements stem from uncorrelated error patterns across base classifiers.

Third, controlled experiments using frozen pretrained models (0.905 AUC), self-supervised features (0.913 AUC), and LASSO-selected radiomics (0.931 AUC) confirm improvements arise from genuine complementarity, not supervised training artifacts. Consistent gains validate that hybrid representations leverage distinct biological cues.

The scientific contribution lies not in higher performance (0.943 AUC, surpassing ProstateX 0.87 AUC benchmark), but in providing a theoretically grounded framework for integrating heterogeneous features. Clinically balanced performance (81.2% sensitivity, 92.5% specificity) enables reducing unnecessary biopsies while detecting clinically significant disease.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization; methodology; software; validation; formal analysis; investigation; resources; data curation; writing—original draft preparation, N. M. Al-Zidi; writing—review and editing, D. Vasumathi; visualization, N. M. Al-Zidi; supervision, D. Vasumathi; project administration, D. Vasumathi; funding acquisition, N. M. Al-Zidi.

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", *CA: A Cancer Journal for Clinicians*, Vol. 74, No. 3, pp. 229–263, 2024.

[2] H. U. Ahmed, A. El-Shater Bosaily, L. C. Brown, R. Gabe, R. Kaplan, M. K. Parmar, Y. Collaco-Moraes, K. Ward, R. G. Hindley, A. Freeman, A. P. Kirkham, R. Oldroyd, C. Parker, and M. Emberton, "Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study", *The Lancet*, Vol. 389, No. 10071, pp. 815–822, Feb. 2017.

[3] Y. Wang and M. Wang, "Selecting proper combination of mpMRI sequences for prostate cancer classification using multi-input convolutional neuronal network", *Physica Medica*, Vol. 80, pp. 92–100, 2020.

[4] N. Mottet et al., "EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer—2020 Update. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent", *European Urology*, Vol. 79, No. 2, pp. 243–262, 2021.

[5] A. B. Rosenkrantz, L. A. Ginocchio, D. Cornfeld, A. T. Froemming, R. T. Gupta, B. Turkbey, A. C. Westphalen, J. S. Babb, and D. J. Margolis, "Interobserver Reproducibility of the PI-RADS Version 2 Lexicon: A Multicenter Study of Six Experienced Prostate Radiologists", *Radiology*, Vol. 280, No. 3, pp. 793–804, Sep. 2016.

[6] N. M. Al-Zidi and D. Vasumathi, "Magnetic Resonance Imaging-based Prostate Cancer Detection and Classification using Machine Learning and Deep Learning Techniques: A Survey", In: *Proc. of 2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, pp. 304–311, 2024.

[7] G. Valizadeh, M. Morafegh, F. Fatemi, M. Ghafoori, and H. Saligheh Rad, "Enhancing Prostate Cancer Classification: A Comprehensive Review of Multiparametric MRI and Deep Learning Integration", *Journal of Magnetic Resonance Imaging,* Vol. 62, No. 6, pp. 1603–1648, 2025.

[8] M. He, Y. Cao, C. Chi, X. Yang, R. Ramin, S. Wang, G. Yang, O. Mukhtorov, L. Zhang, A. Kazantsev, M. Enikeev, and K. Hu, "Research progress on deep learning in magnetic resonance imaging–based diagnosis and treatment of prostate cancer: a review on the current status and perspectives", *Frontiers* in Oncology, Vol. 13, 2023.

[9] Z. Liu, S. Wang, D. Dong, J. Wei, C. Fang, X. Zhou, K. Sun, L. Li, B. Li, M. Wang, and J.

Tian, "The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges", *Theranostics*, Vol. 9, No. 5, pp. 1303–1322, 2019.

[10] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data", *Radiology*, Vol. 278, No. 2, pp. 563–577, Feb. 2016.

[11] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments", Philosophical Transactions of the Royal Society A: Mathematical, *Physical and Engineering Sciences*, Vol. 374, No. 2065, p. 20150202, Apr. 2016.

[12] A. Kitchen and J. Seah, "Support vector machines for prostate lesion classification", In: *Proc. of SPIE*, p. 1013427, 2017.

[13] D. Kwon, I. M. Reis, A. L. Breto, Y. Tschudi, N. Gautney, O. Zavala-Romero, C. Lopez, J. C. Ford, S. Punnen, A. Pollack, and R. Stoyanova, "Classification of suspicious lesions on prostate multiparametric MRI using machine learning", *Journal of medical imaging*, Vol. 5, No. 3, p. 34502, 2018.

[14] P. Sobecki, D. Życka-Malesa, I. Mykhalevych, A. Gora, K. Sklinda, and A. Przelaskowski, "Feature Extraction Optimized For Prostate Lesion Classification", In: Proceedings of the 9th International Conference on Bioinformatics and Biomedical Technology, In: *Proc. of ICBBT '17*. New York, NY, USA: Association for Computing Machinery, pp. 22–27, 2017.

[15] M. Varan, J. Azimjonov, and B. Maçal, "Enhancing Prostate Cancer Classification by Leveraging Key Radiomics Features and Using the Fine-Tuned Linear SVM Algorithm", *IEEE Access*, Vol. 11, pp. 88025–88039, 2023.

[16] A. Mehrtash, A. Sedghi, M. Ghafoorian, M. Taghipour, C. M. Tempany, W. M. Wells, T. Kapur, P. Mousavi, P. Abolmaesumi, and A. Fedorov, "Classification of Clinical Significance of MRI Prostate Findings Using 3D Convolutional Neural Networks", *Proceedings of SPIE--the International Society for Optical Engineering*, Vol. 10134, 2017.

[17] J. C. Y. Seah, J. S. N. Tang, and A. Kitchen, "Detection of prostate cancer on multiparametric MRI", In: *Proc. of Medical Imaging 2017: Computer-Aided Diagnosis*, SPIE, pp. 585–588, 2017.

[18] S. Liu, H. Zheng, Y. Feng, and W. Li, "Prostate Cancer Diagnosis using Deep Learning with 3D Multiparametric MRI", In: *Proc. of Medical Imaging 2017: Computer-Aided Diagnosis*, p. 1013428, 2017.

[19] Q. Chen, S. Hu, P. Long, F. Lu, Y. Shi, and Y. Li, "A Transfer Learning Approach for Malignant Prostate Lesion Detection on Multiparametric MRI", *Technology in Cancer Research & Treatment*, Vol. 18, 2019.

[20] Y. Yuan, W. Qin, M. Buyyounouski, B. Ibragimov, S. Hancock, B. Han, and L. Xing, "Prostate cancer classification with multiparametric MRI transfer learning model", *Medical Physics*, Vol. 46, No. 2, pp. 756–765, 2019.

[21] M. Mehmood, S. H. Abbasi, K. Aurangzeb, M. F. Majeed, M. S. Anwar, and M. Alhussein, "A classifier model for prostate cancer diagnosis using CNNs and transfer learning with multi-parametric MRI", *Frontiers in Oncology*, Vol. 13, 2023.

[22] A. A. Abbasi, L. Hussain, I. A. Awan, I. Abbasi, A. Majid, M. S. A. Nadeem, and Q.-A. Chaudhary, "Detecting prostate cancer using deep learning convolution neural network with transfer learning approach", *Cognitive Neurodynamics*, Vol. 14, No. 4, pp. 523–533, 2020.

[23] S. Yoo, I. Gujrathi, M. A. Haider, and F. Khalvati, "Prostate Cancer Detection using Deep Convolutional Neural Networks", *Scientific Reports*, Vol. 9, No. 1, pp. 1–10, 2019.

[24] A. Santhirasekaram, K. Pinto, M. Winkler, E. Aboagye, B. Glocker, and A. Rockall, "Multi-scale Hybrid Transformer Networks: Application to Prostate Disease Classification", In: *Multimodal Learning for Clinical Decision Support*, pp. 12–21, 2021.

[25] C. Yang, B. Li, Y. Luan, S. Wang, Y. Bian, J. Zhang, Z. Wang, B. Liu, X. Chen, M. Hacker, Z. Li, X. Li, and Z. Wang, "Deep learning model for the detection of prostate cancer and classification of clinically significant disease using multiparametric MRI in comparison to PI-RADs score", *Urologic Oncology: Seminars and Original Investigations*, Vol. 42, No. 5, pp. 158.e17-158.e27, 2024.

[26] C. A. Hamm, G. L. Baumgärtner, F. Biessmann, N. L. Beetz, A. Hartenstein, L. J. Savic, K. Froböse, F. Dräger, S. Schallenberg, M. Rudolph, A. D. J. Baur, B. Hamm, M. Haas, S. Hofbauer, H. Cash, and T. Penzkofer, "Interactive Explainable Deep Learning Model Informs Prostate Cancer Diagnosis at MRI", *Radiology*, Vol. 307, No. 4, 2023.

[27] J. C. Cai, H. Nakai, S. Kuanar, A. T. Froemming, C. W. Bolan, A. Kawashima, H.

Takahashi, L. A. Mynderse, C. D. Dora, M. R. Humphreys, P. Korfiatis, P. Rouzrokh, A. K. Bratt, G. M. Conte, B. J. Erickson, and N. Takahashi, "Fully Automated Deep Learning Model to Detect Clinically Significant Prostate Cancer at MRI", *Radiology*, Vol. 312, No. 2, 2024.

[28] H. Khanfari, S. Mehranfar, M. Cheki, M. Mohammadi Sadr, S. Moniri, S. Heydarheydari, and S. M. Rezaeijo, "Exploring the efficacy of multi-flavored feature extraction with radiomics and deep features for prostate cancer grading on mpMRI", *BMC Medical Imaging*, Vol. 23, No. 1, p. 195, 2023.

[29] J. M. Castillo T., M. Arif, M. P. A. Starmans, W. J. Niessen, C. H. Bangma, I. G. Schoots, and J. F. Veenland, "Classification of Clinically Significant Prostate Cancer on Multi-Parametric MRI: A Validation Study Comparing Deep Learning and Radiomics", *Cancers*, Vol. 14, No. 1, p. 12, 2021.

[30] L. Donisi, G. Cesarelli, A. Castaldo, D. R. De Lucia, F. Nessuno, G. Spadarella, and C. Ricciardi, "A Combined Radiomics and Machine Learning Approach to Distinguish Clinically Significant Prostate Lesions on a Publicly Available MRI Dataset", *Journal of Imaging*, Vol. 7, No. 10, p. 215, 2021.

[31] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, "ProstateX Challenge data [dataset]", *The Cancer Imaging Archive*, 2017, doi: 10.7937/K9TCIA.2017.MURS5CL.

[32] N. M. Al-Zidi and D. Vasumathi, "Identifying Optimal Multiparametric MRI Sequence Combinations for Prostate Cancer Classification: An Integrated Deep Feature and Machine Learning Approach", In: *Proc. of 2024 5th IEEE Global Conference for Advancement in Technology (GCAT)*, pp. 1–7, 2024.

[33] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module", In: *Proc. of Computer Vision – ECCV 2018*, pp. 3–19, 2018.

[34] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype", *Cancer Research*, Vol. 77, No. 21, pp. e104–e107, 2017.

[35] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 8, pp. 1798–1828, 2013.

[36] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach", *Biometrics*, Vol. 44, No. 3, p. 837, 1988.

[37] S. G. Armato, H. Huisman, K. Drukker, L. Hadjiiski, J. S. Kirby, N. Petrick, G. Redmond, M. L. Giger, K. Cha, A. Mamonov, J. Kalpathy-Cramer, and K. Farahani, "PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images", *Journal of Medical Imaging*, Vol. 5, No. 4, p. 44501, 2018.

[38] A. Dimitriadis, G. Kalliatakis, R. Osuala, D. Kessler, S. Mazzetti, D. Regge, O. Diaz, K. Lekadir, D. Fotiadis, M. Tsiknakis, N. Papanikolaou, and K. Marias, "Assessing Cancer Presence in Prostate MRI Using Multi-Encoder Cross-Attention Networks", *Journal of Imaging*, Vol. 11, No. 4, p. 98, 2025.

[39] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, Vol. 521, No. 7553, pp. 436–444, 2015.

[40] L. Rundo and C. Militello, "Image biomarkers and explainable AI: handcrafted features versus deep learned features", *European Radiology Experimental*, Vol. 8, No. 1, p. 130, 2024.

[41] T. G. Dietterich, "Ensemble Methods in Machine Learning", *Multiple Classifier Systems*, pp. 1–15, 2000.