Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Explainable and Uncertainty Aware Hybrid Deep Learning Framework for Prostate MRI Image Classification

Kallol Acharjee
JIS College of Engineering
Kaushik Adhikary
JIS University
Sumit Das
JIS College of Engineering
Madhu Bala
Lovely Professional University
Shifa Shah
shifa.shah@sharda.ac.in
Sharda University

Research Article

Keywords: prostate MRI, stacking ensemble, explainable artificial intelligence, uncertainty estimation, SHAP, Monte-Carlo dropout

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11213475/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An Explainable and Uncertainty Aware Hybrid Deep Learning Framework for Prostate MRI Image Classification

Kallol Acharjee[1*], Kaushik Adhikary[2], Sumit Das[3], Madhu Bala[4], Shifa Shah[5]

[1,2]Department of Computer Science and Engineering, JIS University, Kolkata, India
[1,3]Department of Information Technology, JIS College of Engineering, Kalyani, India
[4]School of Computing and Artificial Intelligence, Lovely Professional University, Punjab, India
[5]School of Computer Science and Engineering, Sharda University, Greater Noida,Uttar Pradesh, India

[1]kallol.acharjee@jiscollege.ac.in, [2]kaushik.adhikary@jisuniversity.ac.in, [3]sumit.das@jiscollege.ac.in, [4]madhucse0105@gmail.com, [5]shifa.shah@sharda.ac.in

ORCID: K. Acharjee, 0009-0006-9126-0925; K. Adhikary, 0000-0001-6514-0395; S. Das, 0000-0002-8018-9151; M.Bala, 0000-0003-2478-3532; S. Shah,0000-0003-4042-3289

Abstract

Background: Artificial Intelligence systems for prostate Magnetic Resonance Imaging (MRI) require not only strong discrimination but also interpretable outputs and an indication of predictive uncertainty. This study develops an image-level framework that combines engineered MRI features, complementary neural learners, post-hoc explanations, and Monte-Carlo (MC) dropout. **Methods:** DICOM images were represented by 265 features: a 256-bin normalized intensity histogram, five statistical descriptors, and four Gray Level Co-occurrence Matrix (GLCM) descriptors. A Multilayer Perceptron (MLP), one-dimensional Convolutional Neural Network (1D-CNN), and residual dense network produced class probabilities. These six probabilities were concatenated with the original features and processed by a 271-input stacking meta-learner. The supplied experiment used a stratified 60/20/20 training/validation/test split. Performance was assessed with accuracy, precision, sensitivity, F1-score, and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC). LIME and SHAP offered both local and global explanations, and 50 stochastic forward passes estimated MC-dropout uncertainty. **Results:** On the reported 1,200-image test partition, 98.25% of accuracy, 98.23% of F1-score, and 0.9907 of ROC-AUC were achieved by the hybrid model, exceeding the best base-model accuracy by 2.58 percentage points. The confusion matrix contained 596 true-negatives, 4 false-positives, 17 false-negatives, and 583 true-positives, corresponding to 99.32% precision and 97.17% sensitivity. **Conclusion:** Probability-level fusion of heterogeneous learners can improve image-level discrimination while supporting feature attribution and uncertainty inspection. However, the current labels are derived from DICOM study descriptions rather than pathology-confirmed cancer status. The reported performance is therefore methodological and must not be interpreted as clinical cancer-detection performance until the model is rerun with patient-grouped partitions, training-only preprocessing and resampling, verified cancer labels, and independent external validation.

Keywords: prostate MRI; stacking ensemble; explainable artificial intelligence; uncertainty estimation; SHAP; Monte-Carlo dropout

1. Introduction:

The incidence, development, and death from prostate cancer, the second most frequent cancer in males globally, vary significantly depending on age, race, genetic predisposition, and environmental variables. For the diagnosis, risk assessment, and treatment planning of prostate cancer, Magnetic Resonance Imaging (MRI), particularly multi-parametric MRI (mpMRI), has emerged as a non-invasive, radiation-free, and highly informative technique [1]. In contemporary radiotherapy, the most popular methods for treating prostate cancer are Volumetric-Modulated Arc Therapy(VMAT) and Intensity-Modulated RadioTherapy (IMRT). Compared to 3D conformal radiation therapy, these proven methods enable the delivery of larger doses to the target with lower doses to the Organs-At-Risk (OARs). By adjusting Dose-Volume Histogram (DVH) parameters and their weights to meet clinical requirements, inverse planning can provide patient-specific dose distributions. However, this process of trial and error takes time and the

quality of the planner's availability, experience, and competence. Thus, an effective method of producing 3D dosage distributions with high planning quality is needed [2]. Treatment planning for External Beam Radiation Therapy (EBRT) involves numerous manual procedures, including choosing the appropriate beam angles, energy, and forms. A trial-and-error approach of choosing and adjusting dose-volume criteria is used to improve a treatment plan for volumetric modulated arc therapy (VMAT) and intensity modulated radiation therapy (IMRT). Until most or all of the predetermined clinical goals are met, this manual iteration keeps going. Depending on the patient's anatomy and the planner's experience, this could take many hours. Furthermore, the growing number of cancer patients and more sophisticated treatment methods result in a greater burden and less time for EBRT treatment planning. This could be a constraint that causes variations in the quality of the plans [3]. This study investigates a hybrid framework that combines a compact engineered representation of DICOM images with three neural base learners: an MLP, a 1D-CNN, and a residual dense network. Their posterior probabilities are appended to the original feature vector and supplied to a deep meta-learner. SHAP and LIME are used for feature-level interpretation, while MC dropout characterizes stochastic predictive variation. The contributions are threefold:

1. a unified feature-and-probability fusion architecture for image-level binary classification;
2. integration of global/local explanations and stochastic uncertainty estimates into the same inference workflow; and
3. a transparent methodological audit separating the reported computational result from the evidence required for pathology-confirmed, patient-level cancer detection.

The third contribution is particularly important. In the supplied implementation, the target is constructed from the DICOM StudyDescription field. Consequently, the experiment distinguishes records associated with a prostate MRI study description from other records; it does not establish histopathology-confirmed malignancy. The manuscript therefore uses the reported data to evaluate the computational framework while explicitly limiting the clinical interpretation.

2. Literature Review:

Recent studies have highlighted a few shot multimodal-deep-learning model, CL-MGNET, integrating PET/CT scan data, mpMRI data, and clinical data, which demonstrated strong performance in predicting aggressive prostate-cancer features. With AUCs of 0.877 and 0.872 in internal and external validation, the model showed excellent generalizability despite limited training data. [4].This research work shows the PRAD-Hybrid CNN (PRADHC), combining EfficientNet and Residual Blocks, achieved 99.34% accuracy and AUC in automated prostate cancer detection from MRI. Its enhanced feature extraction and robustness suggest strong potential as a supportive tool for radiologists in early diagnosis and clinical decision-making.[5].Recent studies have explored deep learning-based automated segmentation of prostate cancer using PSMA PET/CT to decrease observer variability and increase diagnostic accuracy. A proposed framework combining deep learning-based prostate contouring, watershed segmentation, and SUV-based lesion filtering achieved strong agreement with expert annotations, with a median Dice similarity coefficient of 0.93 and low false-negative rate of 2.9%. These findings demonstrate the potential of automated PSMA PET/CT segmentation as a reliable tool for prostate cancer delineation and clinical assessment [6].This research has investigated deep learning-based prostate cancer detection using micro-ultrasound (micro-US) imaging. The suggested PSNet model showed great promise for real-time csPCa detection and biopsy guiding, achieving a patient-level ROC-AUC of 91% with 99% specificity after being trained on raw micro-US power spectra and merged with PSA measurements [7]. The effect of MRI image quality on deep learning-based prostate cancer detection has been emphasised in recent studies. While radiologists maintained more consistent accuracy, a nnaunet model performed worse on low-quality bpMRI images (AUC 0.63 vs. 0.71), highlighting the necessity of standardised MRI methods and quality assurance for trustworthy AI-assisted diagnosis [8]. In order to enhance feature extraction while lowering complexity and overfitting, this study has proposed hybrid deep-learning architectures for automated prostate cancer detection from MRI by integrating VGG16, LeNet, and MobileNetV2. With 96.3% accuracy, 96.7% recall, and 96.2% F1-score, the suggested model outperformed ResNet50 and

EfficientNetB0 and showed great promise for precise and effective prostate cancer diagnosis [9].With the aim to enhance adaptive radiotherapy planning for prostate cancer, recent research has looked into deep learning-based deformable image registration (DIR). During MR-guided adaptive radiation therapy, segmentation-regularized models like ProRSeg showed better contour propagation, which may enhance dose accumulation, target coverage, and organ sparing [10]. The advantages of combining imaging and clinical data for AI-based prostate cancer detection have been shown by this study. The potential for dependable clinical decision assistance was demonstrated by a bimodal deep learning model that combined bpMRI with clinical data, achieving an AUC of 0.82 and sensitivity of 0.73. while uncertainty analysis increased prediction confidence (accuracy 0.85) [11]. For better prostate cancer diagnosis from multimodal MRI, recent research has suggested hybrid deep learning frameworks that include evolutionary optimisation, graph-based learning, and attention processes. The proposed method showed great promise for precise tumour segmentation, feature representation, and trustworthy prostate cancer classification, achieving 97.82% classification accuracy and 97.8% Dice score [12].This study investigated the use of deep learning and sophisticated MRI biomarkers in conjunction with radiologist evaluation to enhance the detection of csPCa. The value of AI-assisted multimodal assessment was demonstrated by a multicenter study that found that combining RSI and DL models with PI-RADS improved diagnostic performance, obtaining AUCs of 0.78 and 0.80, respectively, compared with 0.75 for PI-RADS alone [13]. To lessen human labour and interobserver variability, this research has examined deep learning based automated-segmentation of urinary organs at risk (OARs) in prostate cancer radiation therapy. With a mean DSC of 0.95 for the bladder, a multicenter nnU-Net framework showed excellent segmentation performance, indicating its potential to increase efficiency and consistency in radiation planning [14].This study has demonstrated how metastatic bone lesions in prostate cancer CT scan may be automatically identified and segmented using deep learning. AI-assisted lesion assessment and enhanced clinical workflow efficiency were supported by EfficientNet's better detection performance (F1-score 0.82) and nnUNetv2's superior segmentation (DSC 0.74) [15]. Deep learning and computational pathology have been investigated in this research to predict Ki-67 proliferation score from H&E-stained histopathology images in metastatic prostate cancer. The greatest results were obtained using a foundation model and TransMIL, with an AUROC of 0.841 for high Ki-67 expression, indicating the potential of AI-based pathology for non-invasive biomarker prediction [16]. The promise of deep learning for automated and quantitative Gleason grading in prostate cancer histopathology has been demonstrated by this study. With concordance indices of 0.69 for RFS, 0.72 for MFS, and 0.64 for OS, a MObileNetV3 and ResNet50-based system outperformed traditional ISUP grading in terms of grading generalisabisability and prognostic prediction [17]. When compared to PI-RADS evaluation, this research work has indicated that ensemble deep learning models using numerous mpMRI sequences can enhance the diagnosis of clinically significant prostate cancer (csPCa). The suggested framework showed better diagnostic accuracy, especially in patients with PSA levels<= 10ng/ml, with an AUC of 0.902 compared to 0.759 for PI-RADS [18].

3. Materials and Methods

3.1. Study-design and scope

This provided implementation description and result figures are used to retrospectively recreate this computational investigation. The experiment described used a DICOM picture record as the unit of analysis. The information that is currently provided does not specify the number of distinct patients, the distribution of scanners, the dates of acquisition, the pathological reference standard, or an external test cohort. All numerical findings below are therefore reported as image-level results and have not been independently reproduced from the original dataset in this manuscript preparation step.

3.2 DICOM ingestion and proxy-label construction

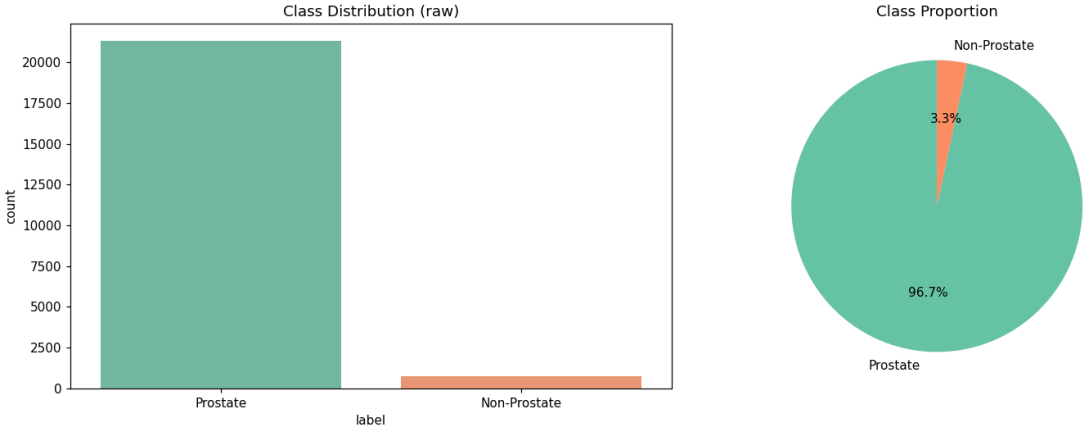


Figure 1: Raw image-record distribution before balancing, as reported by the supplied implementation. Labels are derived from DICOM

The implementation recursively identifies DICOM files and records each file path, PatientID, StudyDescription, and SeriesDescription. A binary proxy target is derived from StudyDescription:

$$y_i = \begin{cases} 1, & \textit{if StudyDescription} = \textit{“MRI Prostate ERC”}, \\ 0, & \textit{otherwise.} \end{cases} \tag{1}$$

Thus, class 1 denotes membership in a named prostate-MRI study category, not a verified malignant lesion. The raw records were highly imbalanced: the supplied class-distribution figure reports approximately 96.7% prostate-labelled and 3.3% non-prostate-labelled images (Figure 1).

3.3. Image normalization and feature engineering

Pixel intensities are rescaled to the integer range $[0,255]$. Each image is then represented by a 265-dimensional vector

$$\boldsymbol{x} = \left[\boldsymbol{h}_{256}, H, \mu_w, C_{RMS}, \gamma_1, \gamma_2, g_{con}, g_{eng}, g_{hom}, g_{cor}\right] \in \mathbb{R}^{265}, \tag{2}$$

where $\boldsymbol{h}_{256}$is a normalized 256-bin intensity histogram; H, μ_w, C_{RMS}, γ_1, and $\gamma_2$are entropy, weighted mean, normalized RMS contrast, skewness, and excess kurtosis; and the final four terms are GLCM contrast, energy, homogeneity, and correlation. For normalized bin probabilities p_i and histogram counts h_i, the principal statistical descriptors are

$$H = -\sum_i p_i \, log_2(p_i + \varepsilon), \tag{3}$$

$$\mu_w = \frac{\sum_i i \, h_i}{\sum_i h_i}, \tag{4}$$

$$\sigma = \sqrt{\frac{\sum_i h_i \, (i - \mu_w)^2}{\sum_i h_i}}, \quad C_{RMS} = \frac{\sigma}{255}. \tag{5}$$

Skewness and excess kurtosis are computed from the standardized third and fourth central moments. For texture analysis, intensities are quantized to 64 levels and GLCMs are evaluated at one-pixel distance for $\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$; each texture descriptor is averaged over the four directions.

3.4. Balancing, standardization, and partitioning

The supplied implementation resamples each class to a maximum of 3,000 image records, shuffles the result, standardizes features using

$$z = (x - \mu)/\sigma, \tag{6}$$

and applies a stratified 60/20/20 training/validation/test split. The resulting confusion matrix indicates a balanced test set of 1,200 images (600 per class).

One of the main issues with validity is the sequence in which these procedures are performed. Resampling and scaler fitting take place prior to partitioning, according to the source description. This can transmit duplicate records and test-distribution information into model development. A publication-grade rerun should instead (i) group records by patient before splitting, (ii) fit the scaler on the training fold only, (iii) resample only the training fold, and (iv) preserve untouched validation, internal-test, and external-test cohorts.

3.5. Base learners

Table 1: Reported architecture of the base learners

Model	Core architecture	Intended representation
MLP	$265 \rightarrow 128 \rightarrow 64 \rightarrow 2$ with batch normalization, ReLU, dropout, and softmax	Global nonlinear interactions among engineered features
1D-CNN	Three one-dimensional convolution/pooling stages, global average pooling, and dense softmax output	Local patterns among adjacent feature positions
Residual dense	64-unit dense projection followed by a residual block and softmax output	Transformed features with shortcut-preserved information

Three models process the same 265 features but impose different inductive biases. Their reported structures are summarized in Table 1.

For logits zk, the two-class probability is

$$P(y = k \mid \boldsymbol{x}) = \frac{exp(z_k)}{\sum_j exp(z_j)} \tag{7}$$

The residual learner uses $y = F(x) + S(x)$, where F is the learned transformation and S is the shortcut projection.

3.6. Hybrid stacking meta-learner

Each base model returns probabilities for the two classes. The original features and six probabilities are concatenated:

$$\boldsymbol{x}_H = [\boldsymbol{x}, \boldsymbol{p}_{MLP}, \boldsymbol{p}_{CNN}, \boldsymbol{p}_{Res}] \in \mathbb{R}^{271} \tag{8}$$

The meta-learner follows $271 \rightarrow 512 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 2$. Training is reported in two phases: first the base models are frozen while the meta-learner is trained; then all layers are unfrozen for joint fine-tuning. The complete workflow is shown in Figure 2.

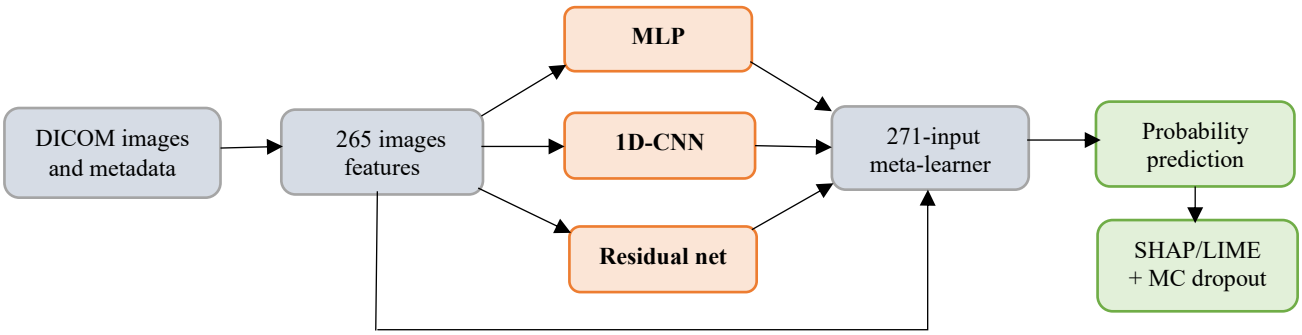


Figure 2: Proposed feature-and-probability fusion workflow. The original features bypass the base learners and are concatenated with their probability outputs.

For strict stacking evaluation, the meta-learner should receive out-of-fold base-model predictions generated without exposure to the corresponding training target. The supplied material does not state that out-of-fold construction was used; nested cross-validation or a fixed hold-out stacking fold is therefore required in the rerun.

3.7. Evaluation measures

Let true positives, true negatives, false positives, and false negatives be represented by TP, TN, FP, and FN. The reported measures are

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{9}$$

$$Precision = \frac{TP}{TP + FP}, \quad Sensitivity = \frac{TP}{TP + FN}, \tag{10}$$

$$F_1 = \frac{2 \cdot Precision \cdot Sensitivity}{Precision + Sensitivity}. \tag{11}$$

ROC-AUC summarizes sensitivity against the false-positive rate over decision thresholds. Specificity, negative predictive value(NPV), balanced accuracy, and Matthews correlation coefficient (MCC) are additionally derived from the supplied confusion matrix. No confidence intervals, repeated-split estimates, or hypothesis tests were provided.

3.8. Explainability and uncertainty

SHAP is used for global rankings and individual feature attributions. Conceptually, a prediction is decomposed as

$$f(\boldsymbol{x}) = \phi_0 + \sum_{i=1}^{265} \phi_i, \tag{12}$$

where positive ϕ_i values move the output toward the positive proxy class and negative values move it away. LIME fits a locally weighted surrogate around an individual feature vector to identify features supporting or opposing that prediction.

For MC dropout, dropout remains active during $N = 50$ stochastic forward passes. If p^j is the positive-class probability on pass j, the reported summary is

$$\bar{p} = \frac{1}{N}\sum_{j=1}^{N} p^j, \quad s_p^2 = \frac{1}{N}\sum_{j=1}^{N}\left(p^j - \bar{p}\right)^{2.} \tag{13}$$

The variance measures stochastic dispersion but is not, by itself, a calibrated probability of error.

3.9. Reporting information required before submission

The following items were not available in the supplied manuscript and should be completed for a journal submission: dataset name and persistent identifier; inclusion/exclusion criteria; unique-patient count; MRI field strength, sequences, and acquisition parameters; ground-truth adjudication; ethics approval and consent basis; software/library versions; optimizer, learning rate, batch size, dropout rate, stopping rule, and random seed; patient-level partition code; missing-data handling; and external-validation protocol.

4. Results

4.1. Learning behaviour

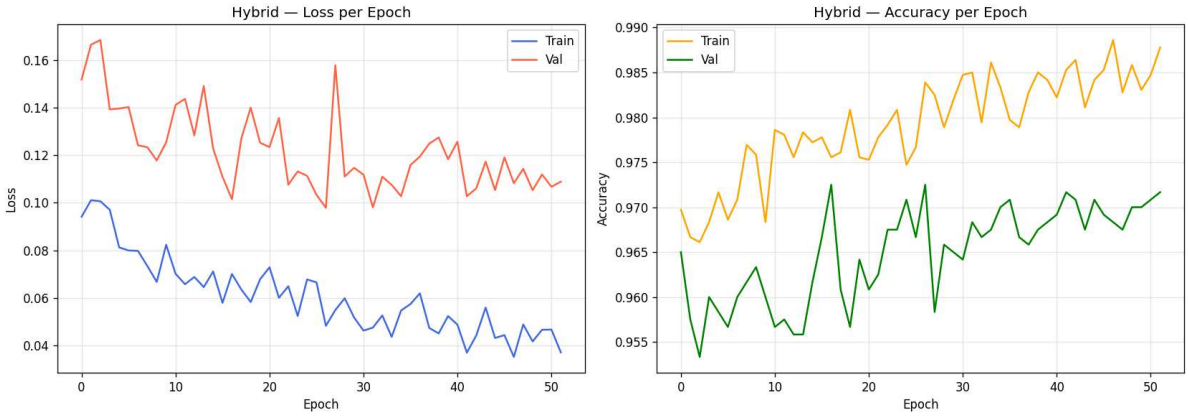


Figure 3: Reported accuracy (right) and training and validation loss (left) for hybrid models

The hybrid training curves span approximately 52 epochs (Figure 3). Training loss decreased while training accuracy approached 0.99. Validation accuracy fluctuated around 0.95–0.97 and the validation loss remained higher than training loss. Rather than depending solely on the final epoch, the continuing generalisation, and patient-grouped validation should be investigated in the rerun.

4.2. Comparison with base learners

Table 2: Reported image-level performance

Model	Accuracy (%)	F1 (%)	ROC-AUC
MLP	95.67	95.50	0.9896
1D-CNN	95.00	94.98	0.9894
Residual network	95.67	95.53	0.9829
Hybrid stacking	**98.25**	**98.23**	**0.9907**

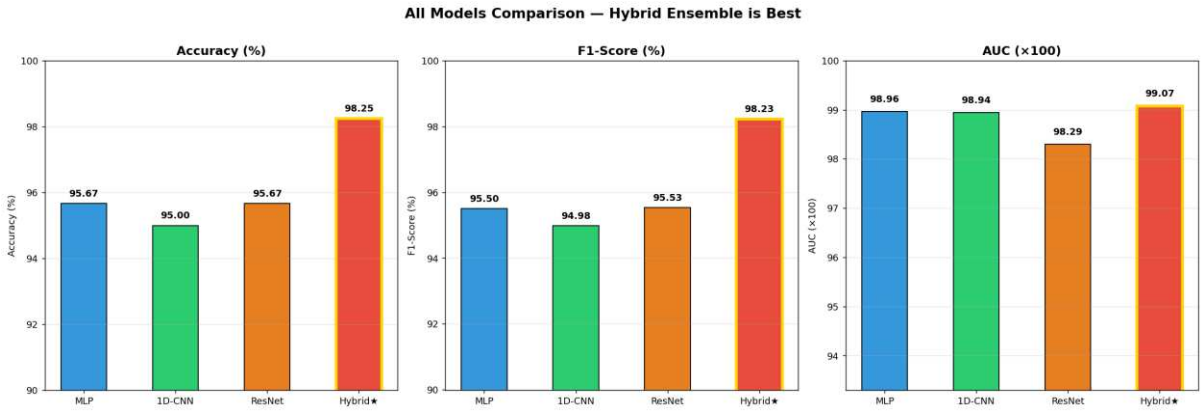


Figure 4: Reported ROC-AUC comparison, F1-score, and accuracy. The source figure plots AUC as a percentage.

According to Table 2 and Figure 4, the hybrid model had the greatest reported accuracy, F1-score, and ROC-AUC. The best individual accuracy of 95.67%, attained by both the MLP and residual learner, was 2.58 percentage points lower than its 98.25% accuracy. Since every model had already demonstrated excellent ranking performance, the ROC-AUC gain was lower.

4.3. Confusion-matrix and ROC analysis

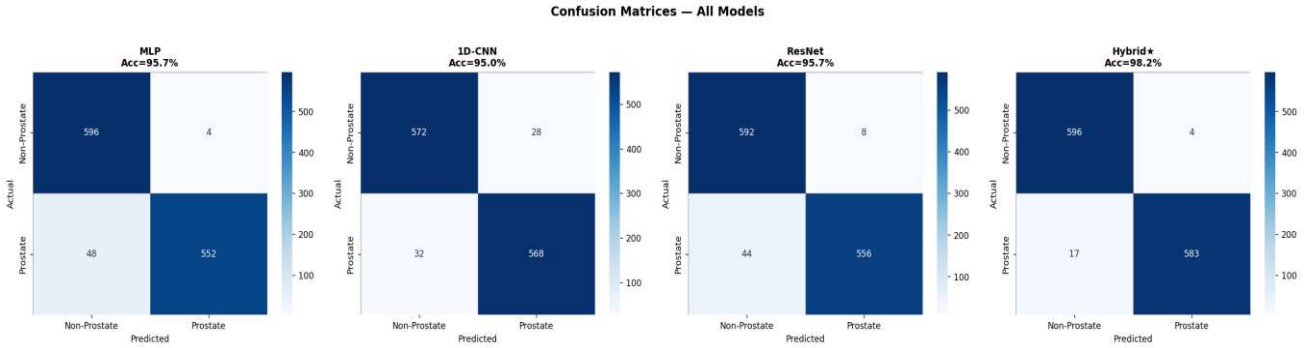


Figure 5: Confusion matrices for the residual network, hybrid model, MLP, and 1D-CNN were reported

Table 3: Measures of hybrid models obtained using the provided confusion matrix.

Measure	Value (%)
Accuracy	98.25
Precision (positive predictive value)	99.32
Sensitivity (recall)	97.17
Specificity	99.33
F1-score	98.23
Negative predictive value	97.23
Balanced accuracy	98.25
Matthews correlation coefficient	96.52

There is 596 TN, 4 FP, 17 FN, and 583 TP in the hybrid confusion matrix (Figure 5). As a result, the test partition has 1,200 image records that are split equally between the two proxy classes. Measures derived directly from these four counts are shown in Table 3. While the 17 false negatives lower sensitivity to 97.17%, the low false-positive count yields 99.32% accuracy and 99.33% specificity.

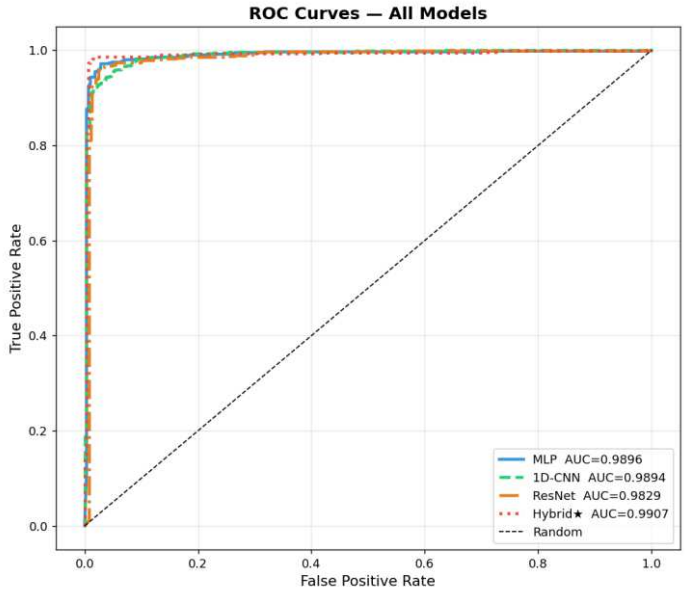
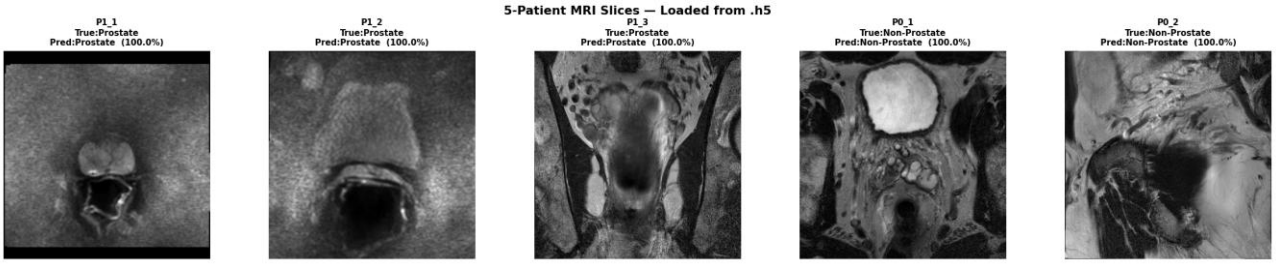


Figure 6: ROC curves for the hybrid model and the three basic models were reported

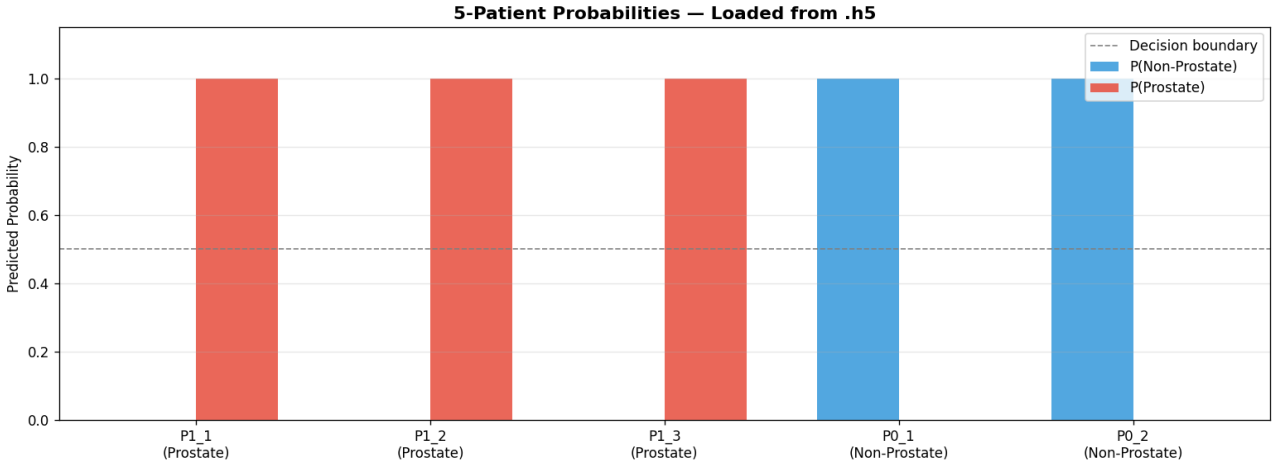
All four ROC curves may be found in the upper-left corner of Figure 6, with the hybrid model achieving the highest recorded ROC-AUC of 0.9907. The apparent variations in ROC-AUC values should be regarded as descriptive rather than statistically proven since there is no patient-grouped external cohort or confidence interval available.

4.4. Illustrative image-level predictions

Three positive-proxy and two negative-proxy photos were among the five test instances for which the saved hybrid model was reloaded. The samples on display had a high degree of confidence in their classification (Figure 7). These examples show the inference interface, but clinical performance, error modes, and patient-level resilience cannot be estimated using a manually chosen five-image panel.



(a) Five chosen MRI slices with anticipated labelling



(b) Two-class probabilities that correspond

Figure 7: Forecasts from the provided result set that serve as examples. In this case, the proxy classes obtained from metadata are denoted by "Prostate" and "Non-Prostate".

4.5. Uncertainty and explanations

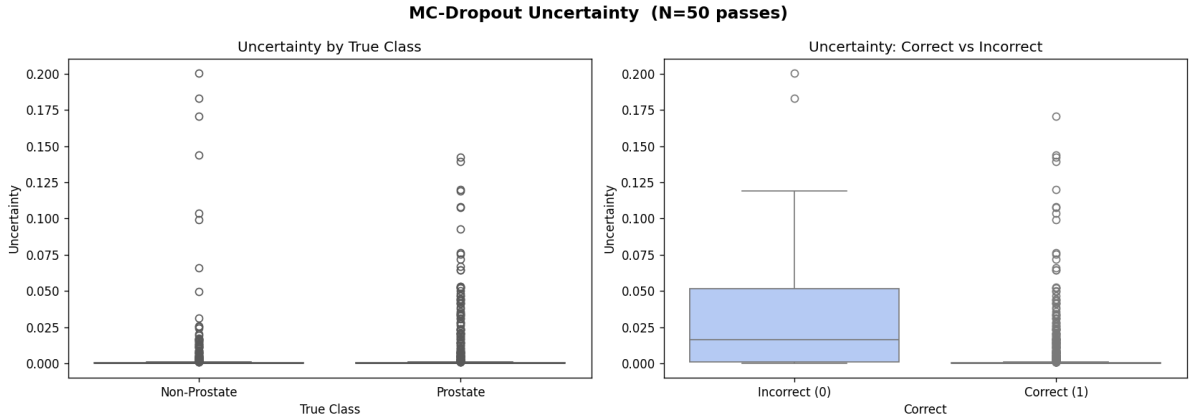
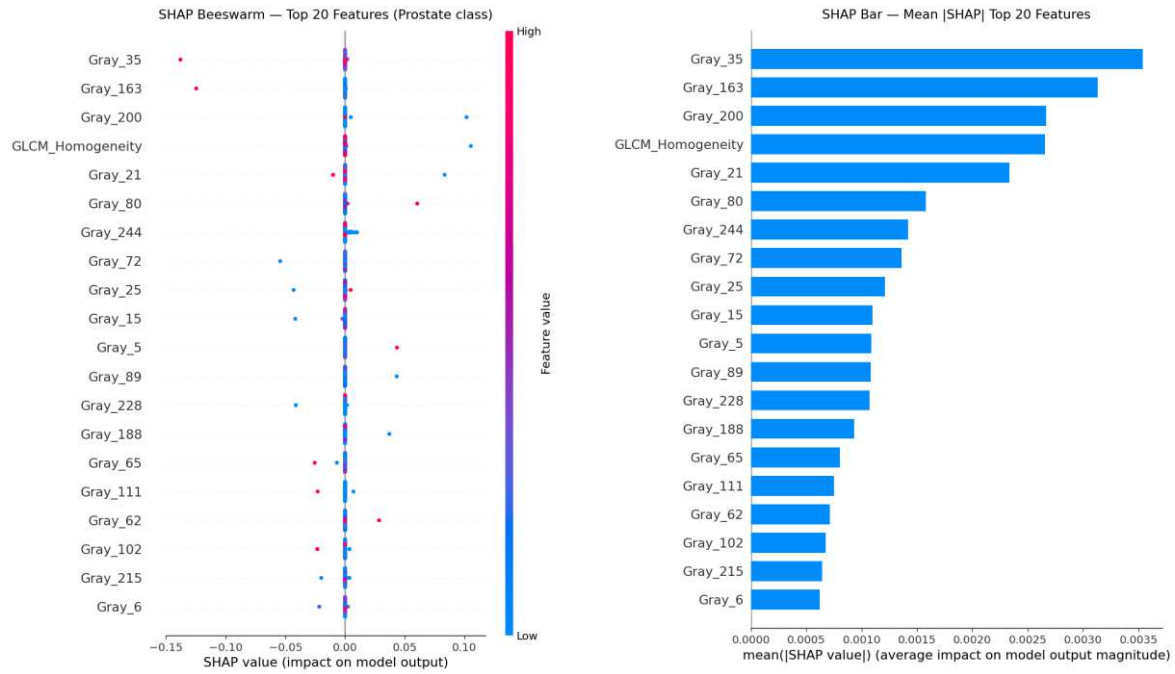


Figure 8: MC-dropout distributions from fifty random forward passes were reported

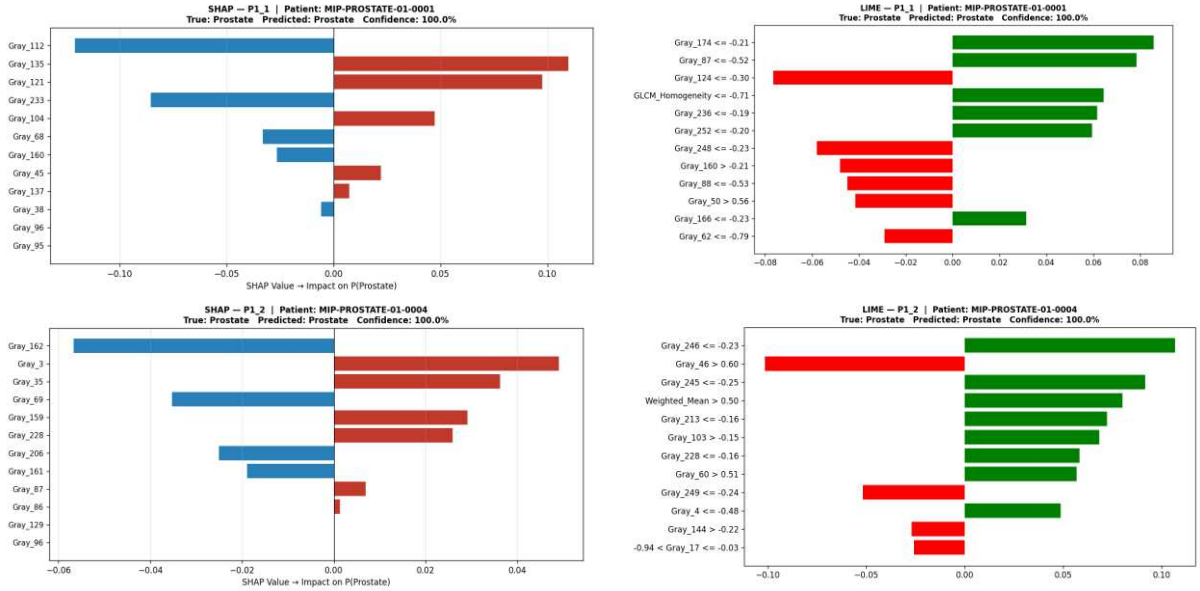
Predictive dispersion is compared by proxy class and correctness using the MC-dropout technique (Figure 8). There are outliers and overlaps in the displayed distributions. The image only enables qualitative inspection in the absence of numerical summary, calibration error, risk-coverage curves, or a predetermined uncertainty threshold; it does not prove that the variance accurately finds problems.

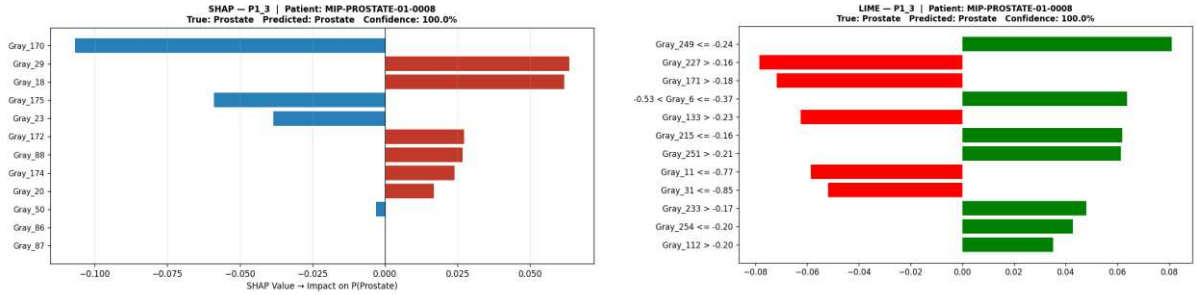


(a) SHAP beeswarm (b) Mean absolute SHAP values

Figure 9: Summaries of global SHAP for the positive proxy class.

For the positive proxy class, global SHAP outputs rank feature contributions (Figure 9).While the mean absolute SHAP plot summarises overall importance, the beeswarm plot displays both direction and dispersion. The ensemble employs textural information in addition to the intensity histogram, as evidenced by the appearance of several GLCM-derived features among the strongest contributors. The ranking should be understood as model attribution rather than a biological explanation because features are artificial summaries and may be associated.





(a) Local SHAP attribution (b) Local LIME attribution

Figure 10: Examples of justifications for a single chosen positive-proxy prediction

Figure 10 displays representative local SHAP and LIME explanations. Although they employ distinct surrogate assumptions and scales, both approaches reveal both positive and negative feature contributions. Descriptive confidence may rise when approaches agree; disagreement should lead to a careful examination rather than the automatic acceptance of either answer.

5. Discussion

5.1. Principal findings

According to proposed experiment, on the same image-level split, feature-and-probability fusion can do better than any of the three individual learners. The accuracy boost of 2.58 percentage points is in line with the anticipated function of stacking, which gives the meta-learner access to both the original representation and possibly complementary model judgements. The confusion matrix also shows that the gain was obtained on the given split with a somewhat higher number of false negatives and few false positives. The integrated output of the framework is it wider value. A class probability, stochastic dispersion, SHAP decomposition, and LIME surrogate can all be obtained from a single inference. This combination is aligned with recent work emphasizing multimodal evidence, uncertainty, and expert/model complementarity [11, 13]. Nevertheless, the additional outputs are informative only if validated. Explanations can reveal the model's statistical dependencies but cannot show that those dependencies are clinically appropriate, and MC-dropout variance can be miscalibrated under distribution shift.

5.2. Comparison with related methods

Recent prostate-imaging studies report strong results from residual hybrids, ensemble MRI models, attention mechanisms, and multimodal fusion [4, 5, 12, 18]. Direct numerical ranking against those studies would be inappropriate because the cohorts, prediction targets, units of analysis, and validation designs differ. The present 0.9907 ROC-AUC is measured for a metadata-derived image-classification task, whereas other studies may predict clinically significant cancer, grade, invasion, or segmentation outcomes using pathology, biopsy, or expert annotation. The most defensible comparison is therefore architectural: heterogeneous model fusion can improve representation, but external and clinically anchored validation determines whether that gain transfers to care.

5.3. Clinical interpretation and intended use

The framework may be assessed as a second-reader or triage component instead of an independent diagnostic system if it were retrained with appropriate labels. The anticipated probability, calibration range, uncertainty flag, and a geographically or feature-linked explanation should all be reported by a clinically focused interface. Additionally, it should maintain radiologist monitoring and provide an abstention rule for ambiguous or out-of-distribution exams. Patient-level sensitivity for clinically significant cancer, specificity, calibration, decision-curve net benefit, and performance across sites, scanners, prostate zones, lesion sizes, and demographic subgroups should all be prioritised in the evaluation.

5.4. Limitations and required validation

The current result cannot be clinically interpreted due to a number of constraints.

1. **Proxy target:** Instead of biopsy, prostatectomy, or expert lesion adjudication, labels are derived from StudyDescription. As a result, the task is not proven to detect cancer.

2. **Unit-of-analysis ambiguity:** The number of patients and the number of images per patient are unclear, however outcomes for images are published. Correlated slices from a single patient can be divided into several partitions using image-level splitting.
3. **Preprocessing leakage:** Before splitting, the source implementation fits the scaler. Only training data can be used to estimate all learnt preprocessing parameters.
4. **Resampling leakage:** Duplicate or closely related entries may be spread across folds by balancing before partitioning. Only the training fold may be used for resampling.
5. **Stacking leakage:** Out-of-fold base predictions for meta-learning training are not documented in the source. Construction must be nested or out-of-fold.
6. **Incomplete provenance:** The provided material does not specify the dataset identification, inclusion criteria, scanner/acquisition parameters, patient demographics, or refence-standard time.
7. **Single-split evaluation:** There were no reports of an external cohort, repeated cross-validation, confidence intervals, calibration statistics, or formal model-comparison tests.
8. **Explanation scope:** Instead of using localised anatomy, SHAP and LIME use tailored summary features. They were not evaluated for stability, clinical plausibility, or sensitivity to associated characteristics.

Before fitting the model, a rigorous rerun should lock an analysis plan; use pathology or expert-confirmed patient labels; split by patient and ideally by institution; fit preprocessing and resampling within each training fold; generate out-of-fold stacking probabilities; tune only on validation folds; maintain an unaltered external test set; and report confidence intervals, and uncertainty based selective performance. These adjustments should result in a more cautious but reliable estimate.

6. Conclusion

This study work introduces a hybrid framework that integrates probability outputs from MLP, 1D-CNN, and residual base learners with 265 intensity, statistical, and textural features in an explainable and uncertainty-aware manner. 98.25% image-level accuracy, 98.23% F1-score, and 0.9907 ROC-AUC were obtained by the disclosed hybrid, while supplemental information regarding feature contributions and stochastic variation was provided by SHAP, LIME, and MC dropout. The outcome confirms that heterogeneous stacking for prostate MRI analysis should be further studied. Because the target is obtained from metadata and the disclosed preprocessing, resampling, and splitting sequence may introduce leakage, it does not yet support a claim of clinical prostate cancer diagnosis. The required next step is multicenter, pathology confirmed, leakage controlled, patient grouped validation.

Data Availability Statement

The data that support the finding of this study are available from The Cancer Imaging Archive (TCIA) under the PROSTATE-MRI collection https://www.cancerimagingarchive.net/collection/prostate-mri/ [19]. The dataset is available subject to TCIA's data usage policies and restrictions. We acknowledge the original data contributors and cite the dataset in the reference list.

Funding Declaration

There is no funding

Clinical Trial

Not Applicable

Consent To Publish

Not Applicable

Ethics and Consent

Not Applicable

References:

[1] K. V. Ghag, J. Jagdale, and M. Vyawahare, "MR ProGradeNet A Multi-Resolution Deep Learning Framework for Automated Prostate Cancer Grading from MRI," *Semin. Ultrasound CT MRI*, Mar. 2026, doi: 10.1053/j.sult.2026.03.003.

[2] Y. Koike, H. Takegawa, Y. Anetai, S. Ohira, S. Nakamura, and N. Tanigawa, “Patient-specific three-dimensional dose distribution prediction via deep learning for prostate cancer therapy: Improvement with the structure loss,” *Phys. Med.*, vol. 107, p. 102544, Mar. 2023, doi: 10.1016/j.ejmp.2023.102544.

[3] M. Lempart *et al.*, “Volumetric modulated arc therapy dose prediction and deliverable treatment plan generation for prostate cancer patients using a densely connected deep learning model,” *Phys. Imaging Radiat. Oncol.*, vol. 19, pp. 112–119, Jul. 2021, doi: 10.1016/j.phro.2021.07.008.

[4] C. Miao *et al.*, “Generalizable Deep Learning for Prostate Cancer Risk Stratification: Multicenter Study Integrating 18F-PSMA-1007 PET/CT and mpMRI,” *Acad. Radiol.*, vol. 33, no. 3, pp. 1107–1120, Mar. 2026, doi: 10.1016/j.acra.2025.12.050.

[5] J. Liu, L. Hou, Y. Xu, Y. Zhang, and H. Zong, “PRAD-Hybrid CNN (PRADHC): A Deep Learning Model for Assisted Diagnosis of Prostate Cancer on MRI,” *http://www.eurekaselect.com*, Mar. 2026, doi: 10.2174/0115734056413897251202130301.

[6] R. Alipour *et al.*, “Fully automated deep learning-based algorithm for contouring primary malignancy on PSMA PET/CT: results from the proPSMA trial in men with intermediate to high-risk prostate cancer,” *EANM J.*, vol. 3, p. 100219, Jun. 2026, doi: 10.1016/j.eanmj.2026.100219.

[7] A. El Kaffas *et al.*, “Prostate Cancer Detection on Micro-Ultrasound Raw Data Using a Deep Learning Neural Network,” *Ultrasound Med. Biol.*, May 2026, doi: 10.1016/j.ultrasmedbio.2026.04.002.

[8] E. H. P. Pooch *et al.*, “In prostate cancer diagnosis, a deep learning model shows increased susceptibility to MRI quality variations compared to radiologists,” *Eur. J. Radiol. Artif. Intell.*, vol. 6, Jun. 2026, doi: 10.1016/j.ejrai.2026.100084.

[9] N. T j, “An enhanced deep learning framework for prostate cancer detection using modified VGG16 and LeNet-MobileNetV2 integration,” *Results Eng.*, vol. 27, p. 106918, Sep. 2025, doi: 10.1016/j.rineng.2025.106918.

[10] S. Madhavan *et al.*, “Segmentation regularized training for multi-domain deep learning registration applied to magnetic resonance-guided prostate cancer radiotherapy,” *Phys. Imaging Radiat. Oncol.*, vol. 39, p. 100989, May 2026, doi: 10.1016/j.phro.2026.100989.

[11] K. Taguelmimt *et al.*, “Towards more reliable prostate cancer detection: Incorporating clinical data and uncertainty in MRI deep learning,” *Comput. Biol. Med.*, vol. 194, p. 110440, Aug. 2025, doi: 10.1016/j.compbiomed.2025.110440.

[12] M. Palanisamy and K. Sennidurai, “Multi-scale attention and hypergraph neural network based deep learning framework for prostate cancer classification and segmentation,” *Biomed. Signal Process. Control*, vol. 113, p. 108925, Mar. 2026, doi: 10.1016/j.bspc.2025.108925.

[13] Y. Song *et al.*, “Deep Learning Artificial Intelligence and Restriction Spectrum Imaging for Patient-level Detection of Clinically Significant Prostate Cancer on Biparametric Magnetic Resonance Imaging,” *Eur. Urol. Open Sci.*, vol. 85, pp. 49–59, Mar. 2026, doi: 10.1016/j.euros.2026.01.014.

[14] J. Le Guévelou *et al.*, “Geometric evaluation of a deep learning method for segmentation of urinary OARs on magnetic resonance imaging for prostate cancer radiotherapy,” *Clin. Transl. Radiat. Oncol.*, vol. 56, p. 101091, Jan. 2026, doi: 10.1016/j.ctro.2025.101091.

[15] S. J. Pawan *et al.*, “Deep learning-based detection and segmentation of osseous metastatic prostate cancer lesions on computed tomography,” *Eur. J. Radiol. Artif. Intell.*, vol. 2, p. 100005, Jun. 2025, doi: 10.1016/j.ejrai.2025.100005.

[16] Z. Chen *et al.*, “Deep learning for predicting Ki-67 scores from H&E-stained prostate cancer images,” *J. Pathol. Inform.*, vol. 21, p. 100633, Apr. 2026, doi: 10.1016/j.jpi.2026.100633.

[17] S. Patkar *et al.*, “A selective CutMix approach improves generalizability of deep learning-based grading and risk assessment of prostate cancer,” *J. Pathol. Inform.*, vol. 15, p. 100381, Dec. 2024, doi: 10.1016/j.jpi.2024.100381.

[18] C. Yang *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.*, vol. 42, no. 5, p. 158.e17-158.e27, May 2024, doi: 10.1016/j.urolonc.2024.01.021.

[19] “PROSTATE-MRI,” The Cancer Imaging Archive (TCIA), doi: http://doi.org/10.7937/K9/TCIA.2016.6046GUDv