Feature Extraction Strategies for Clinically Significant Prostate Cancer Detection on Biparametric MRI: A Systematic Ablation of the Local-Global MIL Framework

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Detecting clinically significant prostate cancer (csPCa) on biparametric MRI (bpMRI) under patient-level weak supervision is challenging because lesion-level annotations are rarely available at scale.This study present a systematic ablation of the Local–Global Multiple Instance Learning (LoGoMIL) architecture introduced by Redekop et al. (2022) on the public PI-CAI 2022 benchmark (1,476 patients), evaluating eleven configurations across six orthogonal design axes: local-only vs local-global context, encoder pretraining strategy (pretrained MedicalNet vs. patch-level pretraining), attention pooling variant in the local branch (plain vs. gated), classifier head (linear vs. MLP), encoder freeze/unfreeze policy, and input modality count (T2W, T2W+ADC, T2W+ADC+HBV). All experiments share a five-fold cross-validation protocol with patient-level AUROC reporting. Two findings stand out. First, adding the global gland-level branch improves the local only baseline significantly. Second, the single largest gain comes from the number of input modalities: adding ADC to T2W improves the T2W only model trained from-scratch by approximately 10 percentage points, and including the high b-value DWI sequence yields the best overall configuration (mean AUROC 0.777 ± 0.031).

Keywords: prostate cancer · biparametric MRI · multiple instance learning · weakly supervised learning · feature extraction

1 Introduction

Prostate cancer is the second most frequently diagnosed cancer in men worldwide, yet a large proportion of detected tumours are clinically insignificant and would remain indolent for the patient's lifetime [1, 2]. Reliably distinguishing clinically significant prostate cancer (csPCa, defined as ISUP Grade Group ≥ 2) from indolent disease is therefore a central objective of modern urological practice [3].

Over the last decade, biparametric MRI (bpMRI) combining T2-weighted (T2W) anatomical imaging with diffusion-weighted imaging (DWI) and the derived apparent diffusion coefficient (ADC) maps has become a cornerstone of

pre-biopsy risk stratification [4, 5]. The PI-CAI 2022 study demonstrated that AI systems can achieve performance non-inferior to expert uroradiologists on the same bpMRI data [5], validating automated detection as clinically viable.

A fundamental constraint of large-scale prostate MRI datasets is the scarcity of voxel-level lesion annotations. Although the PI-CAI dataset includes voxel-level lesion annotations with ISUP grades for a subset of patients, these annotations are not available across the entire cohort. Consequently, models trained on all available data must rely on patient-level binary labels, placing the problem in the weakly supervised setting settings [7]. Multiple Instance Learning (MIL) [9, 10] addresses this by treating the patient as a "bag" of patch instances with unknown labels and learning from bag-level labels alone.

A recurring architectural tension is between local lesion-level detail and global gland-level anatomical context. The Local-Global MIL (LoGoMIL) framework of Redekop et al. [11] resolves this by running parallel local and global encoder branches whose feature vectors are fused by element-wise addition. On a private cohort of 986 examinations under three-fold cross-validation, LoGoMIL achieves a mean AUROC of approximately 0.75. However, several design questions remain open: the relative merits of domain-specific pretraining strategies, attention pooling variants, classifier head depth, and multi-modal input fusion have not been systematically characterised within this framework on a public benchmark.

We close this gap with an eleven-experiment ablation study on the public PI-CAI 2022 training and development dataset. Our main contributions are: (i) the first systematic evaluation of LoGoMIL on a large, heterogeneous public benchmark using five-fold cross-validation; (ii) a controlled comparison of six design axes varied one at a time.

2 Related Work

Local–Global MIL architecture. Redekop et al. [11] proposed LoGoMIL, combining a local branch that encodes small 3D patches via attention-based MIL pooling with a global branch that encodes the whole prostate gland volume. The two 256-dimensional feature vectors are fused by element-wise addition before classification, allowing the model to jointly leverage fine-grained lesion-level cues and coarse anatomical context under patient-level weak supervision.

Pretraining strategies. Chen et al. [13] introduced Med3D, pre-training a family of 3D ResNets on eight public volumetric medical datasets via a multi-task segmentation objective. The resulting MedicalNet backbones provide domain-specific weight initialisation for 3D medical imaging tasks, reducing data requirements compared to training from scratch. We evaluate both this generic 3D pretraining strategy and a task-specific patch-classification pretraining of the local encoder within the LoGoMIL framework.

Multi-modal fusion strategies. Gunashekar et al. [15] systematically compared early, intermediate, and late fusion strategies for multi-parametric prostate MRI

lesion detection. They reported that late fusion stacking modality channels at the network input is the most robust strategy when modalities differ substantially in dynamic range and noise characteristics, as is the case for T2W, ADC, and HBV. This motivates our channel-stacking approach for multi-modal experiments.

Attention pooling variants. Ilse et al. [10] introduced Attention-Based MIL (AB-MIL), replacing fixed pooling operators with a learned attention distribution over instances. Their plain variant computes attention weights via a softmax over a single tanh-gated projection, while the gated variant introduces a sigmoid gate multiplied element-wise with the attention head, increasing expressivity at the cost of additional parameters. Both variants are systematically evaluated in this work.

3 Materials and Methods

3.1 Dataset and Preprocessing

All experiments use the public PI-CAI 2022 training and development dataset [6], comprising 1,500 bpMRI examinations from 1,476 patients acquired at three Dutch centres. Each examination provides T2W, ADC, and high b-value DWI (HBV) sequences together with a binary patient-level csPCa label (ISUP Grade Group ≥ 2). The cohort contains 425 positive (28.3%) and 1,075 negative cases. We use the official five-fold cross-validation split released with the challenge.

Preprocessing steps include: (1) resampling the T2W scan to a common voxel spacing and co-registering ADC/HBV to the T2W reference grid; (2) cropping to the prostate bounding box using the publicly released semi-supervised nnU-Net gland segmentation of Bosma et al. [16]; (3) robust intensity normalisation with IQR clipping and per-patient in-gland standardisation; (4) patch extraction for the local branch via a sliding window of $8\times8\times8$ voxels with 50% overlap, retaining patches with at least 50% gland coverage; and (5) resampling the cropped gland to $32\times64\times64$ for the global branch. For each fold k, the remaining four folds form a train-plus-validation pool from which a stratified 10% validation split is held out for checkpoint selection.

3.2 Model Architecture

We implement the LoGoMIL architecture of Redekop et al. [11] with extensions. The model processes each patient examination through two parallel branches whose outputs are fused before classification (see Fig. 1).

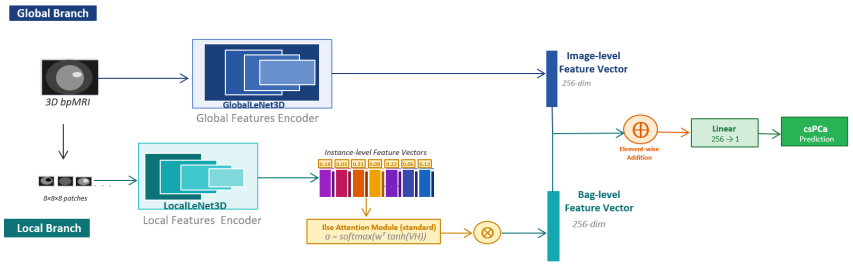

Fig. 1. The LoGoMIL architecture adopted from Redekop et. al [11]: local patch branch with attention-based MIL pooling, global gland branch, and element-wise fusion prior to classification.

Local branch. The local branch operates on a bag of K small 3D patches sampled within the prostate gland mask. Each patch has shape $C{\times}8{\times}8{\times}8$, where $C \in \{1, 2, 3\}$ is the number of input modality channels. A shared 3D CNN encoder: a LeNet-style network with two convolutional blocks followed by a fully connected layer maps each patch independently to a 256-dimensional embedding $\mathbf{h}_k$. Because the encoder is shared across all instances in the bag, the number of patches K can vary between patients without architectural changes.

The patch embeddings are aggregated into a single bag-level feature vector $\mathbf{z}_{\text{loc}} \in \mathbb{R}^{256}$ using Attention-Based MIL pooling [10]:

$$\mathbf{z}_{\text{loc}} = \sum_{k=1}^{K} a_k \mathbf{h}_k, \qquad a_k = \frac{\exp\big(\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_k)\big)}{\sum_{j=1}^{K} \exp(\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_j))}, \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{128\times256}$ and $\mathbf{w} \in \mathbb{R}^{128}$ are trainable parameters. A gated variant [10] is also evaluated, where the attention weights are additionally modulated by a sigmoid gate: $a_k \propto \exp\big(\mathbf{w}^\top[\tanh(\mathbf{V}\mathbf{h}_k) \odot \sigma(\mathbf{U}\mathbf{h}_k)]\big)$, with $\mathbf{U} \in \mathbb{R}^{128\times256}$ a second learnable matrix.

Global branch. The global branch encodes the entire prostate gland in a single forward pass. The cropped gland volume is resampled to a fixed shape of $C{\times}32{\times}64{\times}64$ voxels, regardless of the original gland size, ensuring a consistent input resolution. A deeper 3D CNN with three convolutional blocks interleaved with max-pooling layers, followed by a fully connected layer maps this volume to a 256-dimensional global feature vector $\mathbf{z}_{\text{glo}} \in \mathbb{R}^{256}$. Unlike the local branch, which attends to fine-grained patch-level cues, the global branch captures coarse anatomical context such as zonal structure and gland morphology that individual patches cannot recover.

Fusion and classification. The local and global representations are combined by element-wise addition: $\mathbf{z} = \mathbf{z}_{\text{loc}} + \mathbf{z}_{\text{glo}}$. This fused vector is passed to a classifier

head that produces a scalar patient-level csPCa score via sigmoid activation. We evaluate two classifier variants: a single linear layer and a small MLP with one hidden layer of width 256, GELU activation, and dropout layer.

3.3 Experimental Design

Eleven experiments were perfomed to assess the design choices

Local branch. Prostate patches are passed through a 3D CNN encoder, and the 256 dimensional patch embeddings are combined via attention pooling. Three encoder. Three encoder variants are investigated: (i) a 3D LeNet-style network trained from scratch (**LeNet$^{\mathbf{S}}$**), (ii) the same network initialised from a patch-classification pretraining checkpoint (**LeNet$^{\mathbf{P}}$**), and (iii) a frozen MedicalNet 3D ResNet-18 backbone (**MedNet-RN18**).Both the plain and gated variant of the attention based MIL pooling are evaluated [10]

Global branch. Two encoder strategies are investigated; training from scratch vs. a frozen MedicalNet 3D ResNet-18 backbone..

Fusion and classification. The fused representation $\mathbf{z} = \mathbf{z}_{\text{loc}} + \mathbf{z}_{\text{glo}}$ is passed to either a linear head or a small MLP (one hidden layer of width 256, GELU activation, dropout) to produce a patient-level csPCa score.

Table 1 summarises the eleven configurations. Each experiment isolates one design axis while holding all preprocessing, architecture components, training protocol, and cross-validation split fixed. Experiment 1 is a preliminary single-fold local-only baseline; Experiments 2–11 use the full five-fold protocol.

Table 1. Ablation configurations. **LeNet$^{\mathbf{S}}$**: 3D LeNet trained from scratch; **LeNet$^{\mathbf{P}}$**: patch-pretrained 3D LeNet; **MedNet**: frozen MedicalNet ResNet-18. **(2ch)**/**(3ch)**: multi-channel input. Local pretrain state: Frz = Frozen, UFrz = Unfrozen, F→U = Staged.

#	Modality	Local enc.	Global enc.	Pretrain state	Attn	Head	CV
1	T2W	LeNet^{S}	—	n/a	Plain	Linear	1-fold
2	T2W	LeNet^{S}	LeNet^{S}	n/a	Plain	Linear	5-fold
3	T2W	MedNet	MedNet	Frz	Plain	Linear	5-fold
4	T2W	LeNet^{P}	LeNet^{S}	UFrz	Plain	Linear	5-fold
5	T2W	LeNet^{P}	LeNet^{S}	UFrz	Gated	Linear	5-fold
6	T2W	LeNet^{P}	LeNet^{S}	Frz	Plain	MLP	5-fold
7	T2W	LeNet^{P}	LeNet^{S}	UFrz	Plain	MLP	5-fold
8	T2W+ADC	LeNet^{S} (2ch)	LeNet^{S} (2ch)	n/a	Plain	Linear	5-fold
9	T2W+ADC	LeNet^{P} (2ch)	LeNet^{S} (2ch)	Frz	Plain	Linear	5-fold
10	T2W+ADC	LeNet^{P} (2ch)	LeNet^{S} (2ch)	F→U	Plain	Linear	5-fold
11	T2W+ADC+HBV	LeNet^{S} (3ch)	LeNet^{S} (3ch)	n/a	Plain	Linear	5-fold

3.4 Training Protocol

All models are trained using the Adam optimiser with an initial learning rate of 1×10^{-4} and a ReduceLROnPlateau scheduler (factor 0.1, patience 5 epochs). The training objective is binary cross-entropy; the classifier bias is initialised negatively to compensate for class imbalance ($\sim$28% positive cases). Models are trained for up to 50 epochs per fold; the checkpoint with the highest validation AUROC is retained for evaluation. Held-out test folds are never used during training or model selection.

4 Results

4.1 Model performance

Table 2 reports validation AUROC for all configurations. Experiment 1 (local-only) reaches a single-fold validation AUROC of 0.580, near chance, confirming that patch-level features alone carry limited discriminative power. Adding the global branch in Experiment 2 raises the AUROC to 0.694 on fold 0 (+7.5 pp). Within the T2W-only regime, the best configuration is Experiment 6 (frozen patch-pretrained encoder, plain attention, MLP head) at 0.747 ± 0.036; gated attention (Exp. 5) consistently underperforms its plain counterpart.

The most impactful single intervention is the addition of ADC: Experiment 8 (T2W+ADC framework trained from scratch) reaches 0.752 ± 0.037, a gain of +9.7 pp over Experiment 2. Adding HBV as a third channel yields the best overall result in Experiment 11 (0.777 ± 0.031).

Table 2. Model performance in terms of validation AUROC. $\mu\pm\sigma$ denotes mean and standard deviation across the five cross-validation folds. Experiment 1 reports single-fold results only. Best mean per modality regime in **bold**.

#	Configuration	Modality	F_0	F_1	F_2	F_3	F_4	$\mu\pm\sigma$
1	Local-only (preliminary)	T2W	0.580	–	–	–	–	0.580 (1 fold)
2	LoGoMIL trained from scratch	T2W	0.694	0.718	0.642	0.599	0.621	0.655 ± 0.044
3	MedicalNet, both branches frozen	T2W	0.662	0.704	0.647	0.715	0.699	0.685 ± 0.026
4	Patch-pretrained local + plain attn.	T2W	0.746	0.774	0.642	0.620	0.720	0.701 ± 0.060
5	Patch-pretrained local + gated attn.	T2W	0.699	0.709	0.683	0.632	0.672	0.679 ± 0.027
6	Frozen pretrained + MLP	T2W	0.786	0.783	0.748	0.692	0.725	**0.747 ± 0.036**
7	Unfrozen pretrained + MLP	T2W	0.749	0.750	0.741	0.708	0.714	0.732 ± 0.018
8	End-to-end trained from scratch	T2W+ADC	0.748	0.821	0.741	0.706	0.744	**0.752 ± 0.037**
9	Frozen pretrained	T2W+ADC	0.730	0.736	0.725	0.717	0.724	0.727 ± 0.006
10	Freeze-then-unfreeze	T2W+ADC	0.734	0.729	0.726	0.708	0.720	0.723 ± 0.009
11	End-to-end trained from scratch	T2W+ADC+HBV	0.771	0.817	0.727	0.802	0.769	**0.777 ± 0.031**

4.2 Statistical Analysis

A Friedman test across Experiments 2-11 using per-fold AUCs as paired observations yields $\chi^2_F(9)$=29.43, p=0.0005, rejecting the null hypothesis of equal

performance. Nemenyi post-hoc analysis at α=0.05 (critical difference CD=6.06) identifies four significant pairwise differences: The models in experiments 11, 6, and 8 are each significantly better than the models in Experiment 2 trained from-scratch using T2W only (Exp. 2, rank 9.30 vs. ranks 2.00, 3.00, 3.10 respectively);The model in experiment 11 is additionally superior to the gated-attention configuration (Exp. 5, rank 8.20). All other pairwise differences fall below CD.

5 Discussion and Conclusion

This ablation study yields four headline findings for LoGoMIL-based csPCa detection on bpMRI.

Global context is essential. Purely local patch-based MIL achieves near-chance performance on T2W (AUROC 0.580). Adding the global gland-level branch raises this by 7.5 pp, confirming that whole-gland context complements patch-level features and cannot be recovered from patches alone.

Input modality dominates all other design choices. Adding ADC to T2W improves the LoGoMIL performance by ∼10 pp, the largest single-step gain of the entire programme. Adding HBV as a third modality adds a further ∼2.5 pp, yielding the best overall configuration (Experiment 11, mean AUROC 0.777). The lower bootstrap CI of Experiment 11 (0.751) matches the ∼0.75 AUROC reported by Redekop et al. [11] on a smaller private cohort, and does so in a fully public, larger-scale setting.

Pretraining helps within each modality regime, but is superseded by adding a modality. The best T2W-only configuration implements a patch pretrained encoder.In the dual modality regime, a frozen pretrained encoder is actually dominated by training from scratch, suggesting that end-to-end adaptation of both encoders to the joint T2W+ADC statistics is preferable.

Simpler components generalise better. Plain Ilse attention in experiment 4 outperforms the gated variant in experiment 5. When the pretrained encoder is kept frozen, performance is slightly lower than when it is fine-tuned, but the results are more consistent across folds. Overall, training the full model from scratch using all three MRI sequences (T2W, ADC, and HBV) gives the best and most reliable results, it achieves the highest mean AUC of 0.777.

Limitations. All evaluations use the PI-CAI 2022 training and development cohort; the hidden test set is unavailable outside the challenge infrastructure. Validation AUCs are optimistically biased by model selection on the same split.Only LoGoMIL with element-wise feature fusion is examined; alternative fusion strategies (intermediate or attention-weighted) and alternative patch sizes remain unexplored.

Conclusion. We recommend LoGoMIL trained end-to-end from scratch on three-channel bpMRI (T2W+ADC+HBV) with plain attention and a linear classifier head as the best reproducible configuration on PI-CAI 2022, with statistically

confirmed superiority over the T2W-only models trained from-scratch. Our findings highlight input modality count as the primary lever for csPCa detection under patient-level weak supervision, and provide a rigorous public-benchmark baseline for future work on prostate bpMRI analysis.

References

1. Sung, H., Ferlay, J., Siegel, R.L., et al.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **71**(3), 209–249 (2021)
2. Ahmed, H.U., El-Shater Bosaily, A., Brown, L.C., et al.: Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. Lancet **389**(10071), 815–822 (2017)
3. Kasivisvanathan, V., Rannikko, A.S., Borghi, M., et al.: MRI-targeted or standard biopsy for prostate-cancer diagnosis. N. Engl. J. Med. **378**(19), 1767–1777 (2018)
4. Turkbey, B., Rosenkrantz, A.B., Haider, M.A., et al.: Prostate imaging reporting and data system version 2.1: 2019 update. Eur. Urol. **76**(3), 340–351 (2019)
5. Saha, A., Bosma, J.S., Twilt, J.J., et al.: Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. Lancet Oncol. **25**(7), 879–887 (2024)
6. Saha, A., Twilt, J.J., Bosma, J.S., et al.: The PI-CAI challenge: public training and development dataset. https://pi-cai.grand-challenge.org (2022)
7. Trombetta, R., Rouvière, O., Lartizien, C.: Weakly supervised deep learning model with size constraint for prostate cancer detection in multiparametric MRI and generalization to unseen domains. In: Proc. Medical Imaging with Deep Learning (MIDL). PMLR (2024)
8. Zheng, Y., Zhang, J., Huang, D., et al.: Detecting MRI-invisible prostate cancers using a weakly supervised deep learning model. Int. J. Biomed. Imaging **2024**, 2741986 (2024)
9. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1–2), 31–71 (1997)
10. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: Proc. ICML, PMLR, vol. 80, pp. 2127–2136 (2018)
11. Redekop, E., Sarma, K.V., Kinnaird, A., et al.: Attention-guided prostate lesion localization and grade group classification with multiple instance learning. In: Proc. Medical Imaging with Deep Learning (MIDL). PMLR (2022)
12. Redekop, E., Pleasure, M., Wang, Z., et al.: Codebook VQ-VAE approach for prostate cancer diagnosis using multiparametric MRI. In: Proc. CVPR Workshops (2024)
13. Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
14. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. In: Advances in NeurIPS, vol. 32 (2019)
15. Gunashekar, D.D., Bielak, L., Oerther, B., et al.: Comparison of data fusion strategies for automated prostate lesion detection using mpMRI correlated with whole mount histology. Radiat. Oncol. (2024)
16. Bosma, J.S., Saha, A., Hosseinzadeh, M., et al.: Semi-supervised learning with report-guided pseudo labels for deep learning-based prostate cancer detection using biparametric MRI. Radiology: AI **5**(5), e230031 (2023)

17. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
18. Isensee, F., Jaeger, P.F., Kohl, S.A.A., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
19. Gaudiano, C., Mottola, M., Bianchi, L., et al.: Beyond multiparametric MRI and towards radiomics to detect prostate cancer. Cancers **14**(24), 6156 (2022)
20. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: 3rd Int. Conf. on Learning Representations (ICLR) (2015)