Fine-tuning Vision Foundation Models for Multi-Modal Prostate MR Sequence Classification

Stefan Denner[1] STEFAN.DENNER@DKFZ-HEIDELBERG.DE
Bálint Kovács[1]
David Zimmerer[1]
Deepa Krishnaswamy[2]
Dimitrios Bounias[1]
Raphael Stock[1]
Markus Bujotzek[1]
Fergus Imrie[3]
Andrey Fedorov[2]
Klaus H. Maier-Hein[1]
[1] *Division of Medical Image Computing, German Cancer Research Center, Heidelberg, Germany*
[2] *Brigham and Women's Hospital, Boston, MA, USA*
[3] *Department of Statistics, University of Oxford, Oxford, UK*

Editors: Under Review for MIDL 2025

Abstract

Assigning MRI sequence types is essential yet remains a tedious, manual step in prostate imaging workflows. Current automated approaches relying solely on images or DICOM metadata often struggle with protocol variability and metadata inaccuracies, limiting their generalizability. We propose fine-tuning vision foundation models within different fusion strategies integrating image and metadata. We achieve state-of-the-art F1-score of 1.00 and 0.98 on internal and external test sets, respectively, demonstrating robust generalization.

Keywords: MRI, Prostate Cancer, Sequence Classification, Vision Foundation Models

1. Introduction

Magnetic resonance imaging (MRI) plays a crucial role in tissue characterization by providing complementary information through multiple sequences with distinct tissue contrasts, making it particularly valuable in the diagnosis of prostate cancer (PCa) and the guidance of subsequent interventions (Turkbey and Choyke, 2018). The PI-RADS guidelines (Weinreb et al., 2016; Turkbey et al., 2019) recommend acquiring prostate MRI in a multiparametric fashion, including T2-weighted (T2w), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced (DCE) sequences. However, machine learning algorithms often require only a subset of the available prostate MRI scans (Bhattacharya et al., 2022), including primary sequences or derived images like the apparent diffusion coefficient (ADC) map. In this context, manual data curation remains a tedious and time-consuming task.

Automatic methods utilizing either DICOM metadata (Gauriau et al., 2020; Cluceru et al., 2023) or imaging data (Kasmanoff et al., 2023; Salome et al., 2023) have been introduced but face limited generalizability due to variability in acquisition protocols and frequent metadata inaccuracies. To enhance robustness, Krishnaswamy et al. (2024) combined

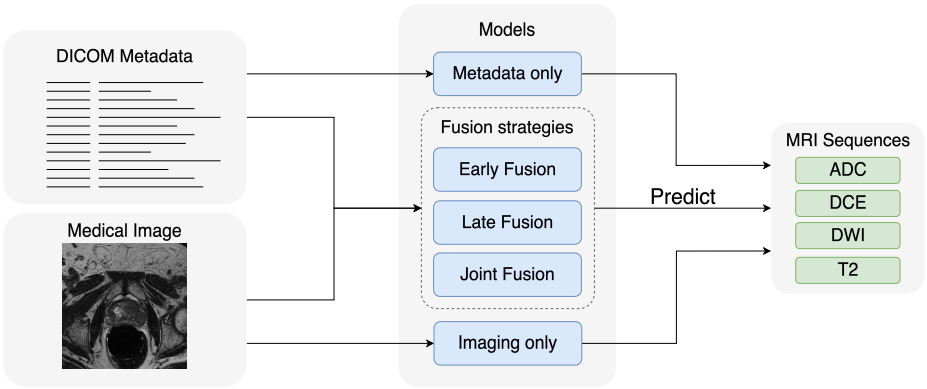

Figure 1: Overview of the proposed multi-modal MRI sequence classification approach.

image features with DICOM metadata, achieving improved yet still limited generalization due to the necessity of training from scratch on relatively small, annotated prostate MRI datasets.

To overcome the reliance on extensive labeled datasets, we propose fine-tuning pretrained vision foundation models for multi-modal MRI sequence classification. Specifically, we investigate the impact of different pretraining strategies (supervised, self-supervised, weakly supervised), model sizes, and various fusion strategies integrating image data with metadata. We validate our approach on internal and external datasets, achieving state-of-the-art F1 scores of 1.00 and 0.98, respectively, demonstrating superior generalization capability and robustness.

2. Methodology

Data We use the prostate MRI dataset collection introduced by Krishnaswamy et al. (2024), comprising two internal datasets (used for training, validation, and testing) and five external datasets (used exclusively for testing). Following the original preprocessing, we use the extracted center slices from 3D volumes along with the metadata predefined splits.

Methods We evaluate multiple vision foundation models across different fusion strategies. Specifically, we fine-tune three pretrained vision encoders: supervised Vision Transformer (ViT) (Dosovitskiy et al., 2020), self-supervised DINOv2 (Oquab et al., 2023) (both pretrained on natural images), and weakly-supervised BiomedCLIP (Zhang et al., 2023), pretrained on scientific biomedical image-text pairs. Additionally, we assess the impact of varying model sizes.

As baselines, we include image-only and metadata-only approaches. To systematically study multi-modal fusion, we adopt the strategies described in Imrie et al. (2025): (1) Early fusion concatenates extracted image features from the vision encoders with metadata before classification via a multilayer perceptron (MLP); (2) Late fusion ensembles predictions from separate image-only and metadata-only models (using Autoprognosis (Imrie et al., 2025) for metadata); and (3) Joint fusion concatenates image and metadata features and trains the entire model end-to-end. We perform an ablation across all vision encoders and fusion strategies. Full training details are provided in appendix A.

Table 1: F1 scores for our proposed multi-modal MRI sequence classification approach. Avg. best in bold. †= Results taken from (Krishnaswamy et al., 2024)

		Internal					External				
		ADC	DCE	DWI	T2w	**Avg.**	ADC	DCE	DWI	T2w	**Avg.**
	Images †	0.99	0.99	0.99	0.99	0.99	0.99	0.89	0.59	0.93	0.85
	Metadata †	1.00	1.00	1.00	1.00	1.00	0.91	1.00	0.60	0.98	0.87
	Images + metadata †	0.99	0.99	1.00	1.00	1.00	0.99	0.99	0.72	0.98	0.92
	Autoprognosis	1.00	1.00	1.00	1.00	1.00	0.95	0.99	0.14	0.94	0.76
Imaging only	ViT-B/16	1.00	1.00	1.00	1.00	1.00	0.96	0.94	0.82	0.93	0.91
	ViT-L/16	1.00	1.00	1.00	1.00	1.00	0.98	0.92	0.88	0.87	0.91
	DINOv2 ViT-S/14	1.00	1.00	1.00	1.00	1.00	0.93	0.94	0.88	0.93	0.92
	DINOv2 ViT-B/14	1.00	1.00	1.00	1.00	1.00	0.98	0.91	0.92	0.85	0.92
	DINOv2 ViT-L/14	1.00	1.00	1.00	1.00	1.00	0.94	0.92	0.90	0.90	0.91
	BiomedCLIP ViT-B/16	1.00	1.00	1.00	1.00	1.00	0.98	0.96	0.93	0.92	0.95
Early Fusion	ViT-B/16	1.00	1.00	1.00	1.00	1.00	0.96	0.96	0.82	0.97	0.93
	ViT-L/16	1.00	1.00	1.00	1.00	1.00	0.81	0.97	0.89	0.67	0.84
	DINOv2 ViT-S/14	0.96	1.00	1.00	0.97	0.98	0.87	0.99	0.87	0.88	0.90
	DINOv2 ViT-B/14	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.90	0.99	0.97
	DINOv2 ViT-L/14	1.00	1.00	0.98	0.96	0.98	0.68	1.00	0.84	0.35	0.72
	BiomedCLIP ViT-B/16	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.88	0.98	0.96
Late fusion	ViT-B/16	1.00	1.00	1.00	1.00	1.00	0.98	0.99	0.89	0.99	0.96
	ViT-L/16	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.92	0.97	0.96
	DINOv2 ViT-S/14	1.00	1.00	1.00	1.00	1.00	0.95	0.98	0.81	0.98	0.93
	DINOv2 ViT-B/14	1.00	1.00	1.00	1.00	1.00	0.97	0.94	0.88	0.89	0.92
	DINOv2 ViT-L/14	1.00	1.00	1.00	1.00	1.00	0.95	0.98	0.85	0.97	0.94
	BiomedCLIP ViT-B/16	1.00	1.00	1.00	1.00	1.00	0.97	0.98	0.91	0.97	0.96
Joint fusion	ViT-B/16	1.00	1.00	1.00	1.00	1.00	0.97	0.88	0.91	0.57	0.83
	ViT-L/16	1.00	1.00	1.00	1.00	1.00	0.98	0.95	0.94	0.86	0.93
	DINOv2 ViT-S/14	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.86	1.00	0.95
	DINOv2 ViT-B/14	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.93	0.99	**0.98**
	DINOv2 ViT-L/14	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.93	0.99	0.97
	BiomedCLIP ViT-B/16	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.93	1.00	**0.98**

3. Results and Discussion

Our results (table 1) show perfect performance (F1 = 1.00) on the internal test dataset for image-only, late fusion, and joint fusion strategies across all evaluated sequence types. On the external dataset, joint fusion using either BiomedCLIP ViT-B/16 or DINOv2 ViT-B/14 achieves the best overall performance (F1 = 0.98). Specifically, T2w and DCE sequences consistently reach perfect scores (F1 = 1.00), while performance slightly drops for ADC (F1 = 0.98) and notably for DWI (F1 = 0.93). Notably, DWI classification substantially benefits from image information compared to metadata alone (F1 = 0.93 vs. 0.60), underscoring the importance of visual data for this sequence. Across all strategies, BiomedCLIP ViT-B/16 consistently yields the highest overall performance, likely due to its extensive weakly-supervised pretraining on biomedical image-text pairs. Additionally, self-supervised DINOv2 models marginally outperform supervised ViT models on average. Overall, our approach demonstrates that fine-tuned vision foundation models with joint image-metadata fusion effectively generalize across datasets, significantly reducing dependence on large-scale annotated datasets.

References

Indrani Bhattacharya, Yash S Khandwala, Sulaiman Vesal, Wei Shao, Qianye Yang, Simon JC Soerensen, Richard E Fan, Pejman Ghanouni, Christian A Kunder, James D Brooks, et al. A review of artificial intelligence in prostate cancer detection on imaging. *Therapeutic advances in urology*, 14:17562872221128791, 2022.

BN Bloch, A Jain, and Jaffe CC. Data from PROSTATE-DIAGNOSIS [dataset], 2015. URL https://wiki.cancerimagingarchive.net/display/Public/PROSTATE-DIAGNOSIS.

P Choyke, B Turkbey, P Pinto, M Merino, and B Wood. Data from PROSTATE-MRI, 2016. URL https://www.cancerimagingarchive.net/collection/prostate-{MRI}/.

J Cluceru, JM Lupo, Y Interian, R Bove, and JC Crane. Improving the automatic classification of brain MRI acquisition contrast with machine learning. *Journal of Digital Imaging*, 36(1):289–305, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

A Fedorov, M Schwier, D Clunie, C Herz, S Pieper, R Kikinis, C Tempany, and F Fennessy. An annotated test-retest collection of prostate multiparametric MRI. *Scientific data*, 5 (1):1–3, 2018.

R Gauriau, C Bridge, L Chen, F Kitamura, NA Tenenholtz, JE Kirsch, KP Andriole, MH Michalski, and BC Bizzo. Using DICOM metadata for radiological image series categorization: a feasibility study on large clinical brain MRI datasets. *Journal of digital imaging*, 33:747–62, 2020.

Fergus Imrie, Bogdan Cebere, Eoin F McKinney, and Mihaela van der Schaar. Autoprognosis 2.0: Democratizing diagnostic and prognostic modeling in healthcare with automated machine learning. *PLOS digital health*, 2(6):e0000276, 2023.

Fergus Imrie, Stefan Denner, Lucas S Brunschwig, Klaus Maier-Hein, and Mihaela Van Der Schaar. Automated ensemble multimodal machine learning for healthcare. *IEEE Journal of Biomedical and Health Informatics*, 2025.

N Kasmanoff, MD Lee, N Razavian, and YW Lui. Deep multi-task learning and random forest for series classification by pulse sequence type and orientation. *Neuroradiology*, 65: 77–87, 2023.

Deepa Krishnaswamy, Bálint Kovács, Stefan Denner, Steve Pieper, David Clunie, Christopher P Bridge, Tina Kapur, Klaus H Maier-Hein, and Andrey Fedorov. Automatic classification of prostate mr series type using image content and metadata. *arXiv preprint arXiv:2404.10892*, 2024.

G Litjens, O Debats, J Barentsz, N Karssemeijer, and H Huisman. Computer-aided detection of prostate cancer in MRI. *IEEE transactions on medical imaging*, 33(5):1083–92, 2014.

G Litjens, J Futterer, and H Huisman. Data from prostate-3T, 2016. URL https://wiki.cancerimagingarchive.net/display/Public/Prostate-3T.

G Litjens, O Debats, J Barentsz, N Karssemeijer, and H Huisman. SPIE-AAPM PROSTATEx challenge data (version 2) [dataset], 2017. URL https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=23691656.

A Madabhushi and M Feldman. Fused radiology-pathology prostate dataset (prostate fused-MRI-pathology), 2016. URL https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=23691514.

S Natarajan, A Priester, D Margolis, J Huang, and L Marks. Prostate MRI and ultrasound with pathology and coordinates of tracked biopsy (Prostate-MRI-US-Biopsy) (version 2) [data set], 2013. URL https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=68550661.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

P Salome, F Sforazzini, G Brugnara, A Kudak, M Dostal, C Herold-Mende, S Heiland, J Debus, A Abdollahi, and M Knoll. MR-Class: A python tool for brain mr image classification utilizing one-vs-all dcnns to deal with the open-set recognition problem. *Cancers*, 15:1820, 2023.

A Singanamalli, M Rusu, RE Sparks, NN Shih, A Ziober, L Wang, J Tomaszewski, M Rosen, M Feldman, and A Madabhushi. Identifying in vivo DCE MRI markers associated with microvessel architecture and gleason grades of prostate cancer. *Journal of Magnetic Resonance Imaging*, 43:149–158, 2016.

GA Sonn, S Natarajan, DJ Margolis, M MacAiran, P Lieu, J Huang, FJ Dorey, and LS Marks. Targeted biopsy in the detection of prostate cancer using an office based magnetic resonance ultrasound fusion device. *Journal of Urology*, 189(1):86–91, 2013.

Baris Turkbey and Peter L Choyke. Future perspectives and challenges of prostate mr imaging. *Radiologic Clinics*, 56(2):327–337, 2018.

Baris Turkbey, Andrew B Rosenkrantz, Masoom A Haider, Anwar R Padhani, Geert Villeirs, Katarzyna J Macura, Clare M Tempany, Peter L Choyke, Francois Cornud, Daniel J Margolis, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European urology*, 76 (3):340–351, 2019.

Jeffrey C Weinreb, Jelle O Barentsz, Peter L Choyke, Francois Cornud, Masoom A Haider, Katarzyna J Macura, Daniel Margolis, Mitchell D Schnall, Faina Shtern, Clare M Tempany, et al. PI-RADS prostate imaging–reporting and data system: 2015, version 2. *European urology*, 69(1):16–40, 2016.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Table 2: Dataset collection. Number of MR series (patients in parentheses) included for the analysis. ERC = endorectal coil used, †=multiple manufacturers, ‡=multiple magnetic field strengths. Table adapted with minor modifications from (Krishnaswamy et al., 2024)

	Dataset	ERC	T2W	DWI	ADC	DCE	Train	Val	Test
Internal	QIN-Prostate-Repeatability (Fedorov et al., 2018)	✓	30 (15)	30 (15)	30 (15)	30 (15)	✓	✓	✓
	ProstateX† (Litjens et al., 2014, 2017)	–	431 (346)	357 (346)	356 (346)	15456 (346)	✓	✓	✓
External	Prostate-MRI (Choyke et al., 2016)	✓	26 (26)	52 (26)	–	51 (26)	–	–	✓
	Prostate-3T† (Litjens et al., 2016)	–	64 (64)	–	–	–	–	–	✓
	Prostate-Diagnosis (Bloch et al., 2015)	✓	93 (91)	–	–	–	–	–	✓
	Prostate-MRI-US-Biopsy†‡ (Natarajan et al., 2013) (Sonn et al., 2013)	✓	958 (792)	110 (108)	1019 (836)	–	–	–	✓
	Prostate-Fused-MRI-Pathology (Singanamalli et al., 2016) (Madabhushi and Feldman, 2016)	✓	46 (27)	13 (12)	12 (12)	102 (28)	–	–	✓

Appendix A. Trainings details

We fine-tune the vision foundation models using a two step approach. We first freeze the backbone and only fine-tune the last layer with a learning rate of 0.0001 for 5 epochs. Then, we unfreeze the whole model and continue fine-tuning with a learning rate of 1e-6 until convergence of the weighted cross entropy validation loss. We ensemble the four cross-fold validation trained models to generate the final prediction. For autoprognosis (Imrie et al., 2023), used in late fusion, we utilize the default configurations with the number of folds set to four.