D6.1

Vendor Specific AI Models

Related Work Package	WP6 – Development of Vendor-Specific and Vendor-Neutral AI Models
Related Tasks	T6.1, T6.2, T6.3
Lead Beneficiary	RADBOUDUMC
Contributing Beneficiaries	FCHAMPALIMAUD
Document version	v1.0
Deliverable Type	Report
Distribution level	PU
Contractual Date of Delivery	M32
Actual Date of Delivery	10 November 2023

Authors	Anindo Saha, Jasper Twilt, Maarten de Rooij, Jurgen Futterer, Henkjan Huisman
Contributors	FCHAMPALIMAUD José Guilherme de Almeida Nuno Rodrigues Ana Carolina Rodrigues Raquel Moreno Nickolas Papanikolaou CNR Rossana Buongiorno Claudia Caudai Giulio Del Corso Danila Germanese Eva Pachetti Maria Antonietta Pascali FORTH Grigorios Kalliatakis Dimitrios Zaridis Eugenia Mylona Avtantil Dimitriadis ADVANTIS Dimitris Agraniotis Zoi Giavri FPO Valentina Giannini Giovanni Maimone Simone Mazzetti Daniele Regge QUIBIM HULAFE Manuel Marfil Trujillo David Vallmanya Poch Leonor Cerdá Alberich
Reviewers	Nickolas Papanikolaou, Nikolaos Tachos, Manolis Tsiknakis, Sara Colantonio, Kostas Marias

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement **no 952159**

Contents

Executive Summary

Deliverable 6.1, led by partner RADBOUDUMC, titled "Development of Vendor-Specific AI Models," describes the work performed in Tasks 6.1, 6.2, and 6.3 in WP6. Task 6.1 describes the "Prospective data upload to the platform" (Leader: RADBOUDUMC, Participants: FPO, FCHAMPALIMAUD, HULAFE, UNIPI, IPC, HACETTEPE, GAONA St Savvas, RMH, QUIRONSALUD, IDIBGI, JCC, NCI)". Task 6.2 describes the "Deep learning methods for semi-automatic segmentation" (Leader: QUIBIM, Participants: ADVANTIS, FORTH, FCHAMPALIMAUD). Task 6.3 describes the "Development of vendor-specific models for diagnostics, prognosis, and treatment" (Leader: RADBOUDUMC, Participants: FCHAMPALIMAUD, FPO, HULAFE). The work performed significantly contributes to achieving especially three of seven objectives:

1 "Develop a comprehensive data resource related to prostate cancer for clinical care, research and innovation."

3 "Develop and Deploy Novel AI Models to Address the Unanswered Clinical Questions regarding Prostate Cancer Management across the Disease Continuum."

5 "Validate, verify and explain or interpret the performance of AI models in order to increase trust and render them applicable in clinical practice."

The WP6 concept is to collect and apply prospective data for two purposes. Firstly, prospective, per-device/center data allows for fine-tuning of trained AI master models (WP5) to a specific center or device that is hypothesized to optimize performance. Secondly, prospective data is new data that allows for robust validation of developed AI and segmentation models (WP5) on unseen data.

In this Deliverable 6.1, we report on the complete execution of the collection of prospective data (T6.1) and our strategy to mitigate earlier reported delays in data ingestion. We report the work prospectively validating previously developed (WP5) segmentation and detection algorithms for all eight use cases (T6.2). We report on our extensive scientific explorations of the unique vendor-specific concepts in the ProCAncer-I (T6.3). Vendor-specific fine-tuning is hypothesized to improve diagnostic performance over the master models developed in WP5. Elaborate experimentation work on radiomics AI shows that prospective fine-tuning indeed shows the expected improvement. We are also reporting ongoing extensive experiments with deep learning vendor-specific modeling. The results show a mix of benefits to deep learning AI master models that have successfully generalized over all vendors. We have discovered dependencies on the amount of data available, the varying case complexity, and the varying image quality ranges. These breakthrough observations allow for an AI technology-based adaptation of ProCAncer-I concepts and will enable the scientific community to choose strategies that lead to the best possible AI models for validation in WP7 to start *"addressing the unanswered clinical questions"*. The results are being submitted to scientific meetings and peer-reviewed journals.

Chapter 1

Prospective Data Collection

1.1 Chapter Summary

In this chapter, we show that we successfully collected prospective data, which enabled us to proceed with the tasks outlined in this Deliverable. We had to make slight adjustments due to data collection challenges that have been reported earlier. While data collection is a critical component of AI development, our experiences within this project, as well as various other EU initiatives, have revealed that it encompasses a broader range of considerations than initially anticipated. While the original retrospective data was successfully obtained (as detailed in Deliverable 5.1), the subsequent delay has impacted prospective data collection.

To address this, we have implemented a mitigation strategy involving modifying the criterion that initially distinguished data between retrospective and prospective cohorts. Per the project proposal, prospective data collection was defined as the enrollment of patients including their clinical and MRI data after January 2022. After internal deliberation during various meetings, it was determined that prospective cases should be defined as data satisfying the ground truth after January 2022. Consequently, this approach allows for including, for example, patients with negative MRI results from 2021 yet fulfilling the one-year follow-up criteria in 2022.

Moreover, in light of the delays and challenges faced by a few centers in enrolling participants and meeting the criteria for reference standards, a decision was made to permit the inclusion of additional retrospective patients in a separate data bucket, separate from the retrospective data as obtained and outlined during WP5 developments.

These measures have collectively enabled the acquisition of sufficient (prospective) data, facilitating the execution of all experiments within WP6.

1.2 Current Status of Prospective Data Collection

Enrolment of prospective patients started in each clinical center between May 2021 and July 2022 after each local Ethics Committee approved the study protocol. At the time being, the following figure coming from the ProstateNET monitoring service visualizes the current status of prospective data upload.

The expected total number of prospective patients to be collected and uploaded on the ProstateNET is about 4,500, corresponding to 8,700 data points. At the end of October 2023, the number of prospective cases available on the platform was 1,113 (25% of the total number of prospective cases) (Figure 1.1). All clinical data partners have started with the enrollment of prospective data, and almost all clinical data partners are in the final stages of uploading prospective cases.

In Figure 1.2, we observe the distribution of cases among the ProCAncer-I Use Cases (UC). Notably, the number of cases allocated to UC1 (detection of prostate cancer) and UC2 (characterization of prostate cancer) surpasses the number of cases in the other UCs. This distribution can be explained by the fact that the ground truth for UC1 and UC2 can be satisfied based on biopsy outcomes, while the remaining use cases necessitate a more extended patient follow-up period. For instance, confirming or excluding biochemical relapse after treatment requires a longer observation period, thereby delaying the inclusion of patient data

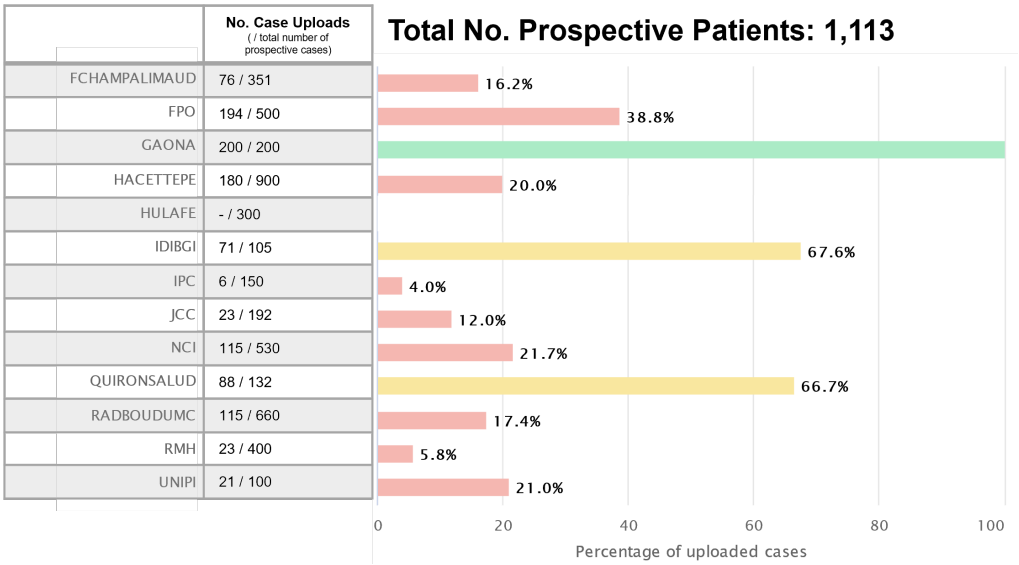


Figure 1.1: Prospective Data Upload Status (obtained end of October 2023)

for the corresponding use cases.

It is anticipated that the number of cases in the other use case categories will increase in the following months as the necessary extended follow-up periods to establish the ground truth for these use cases become available.

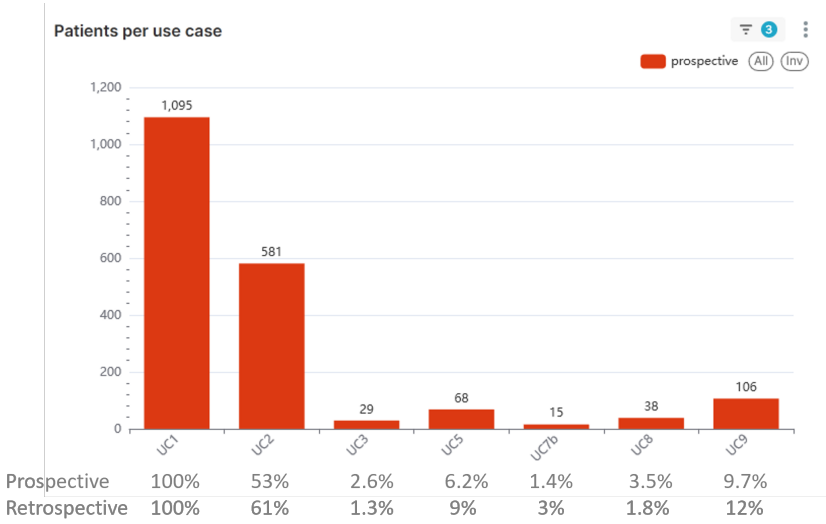


Figure 1.2: Prospective Data Distribution Among ProCancer-I Use Cases (obtained end of October 2023)

1.3 Additional Retrospective Data Upload for Validation

A few centers have raised challenges in prospective data enrollment, resulting in a reduced estimation of total patients available to upload as prospective data. Together with the consortium and in discussion with the Management Board, a decision was made to permit the inclusion of supplementary retrospective patients. These additional retrospective cases will be carefully compiled distinctively from the retrospective repository

compiled in WP5 and will be marked with the specific purpose of model validation. By incorporating this supplementary data, clinical partners can effectively meet their initial projections of data contribution, ensuring that this data is preserved for the same purpose as the prospective data.

In light of this decision, a new version of the electronic Case Report Form (eCRF) was created and subsequently made available to four out of the thirteen participating centers within the Consortium. As of the most recent update on October 30, 2023, these centers include FChampalimaud, IPC, RMH, and RadboudUMC.

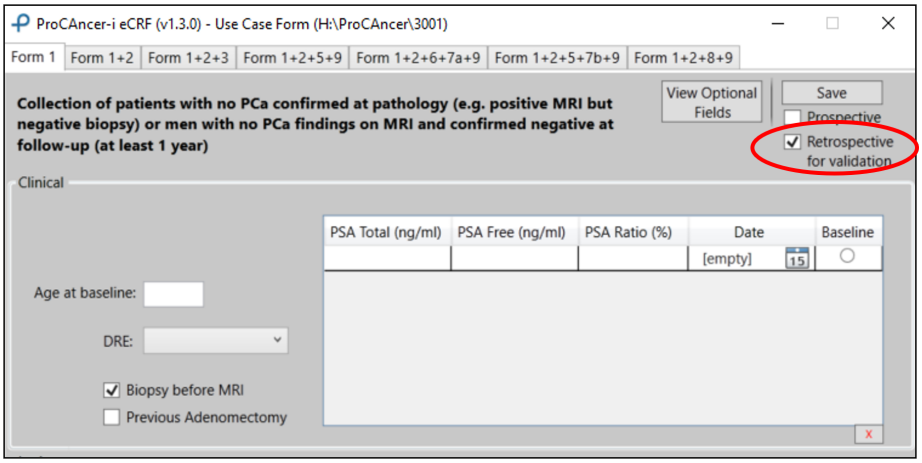


Figure 1.3: eCRF version with the possibility to tag retrospective cases for validation

1.4 Agreements on Ongoing Prospective Data Curation

The collection of prospective data will continue in the coming half year. The Management Board, in agreement with the whole Consortium, set the next deadlines for the prospective enrolment, i.e., 50% by the end of November 2023, 80% by the end of February 2024, with completion of the collection by May 2024.

Prospective						
#	Partner's name	no. of unique patients	now uploaded (20th Oct 2023)	Nov-23 50%	Feb-24 80%	May-24 100%
2	FCHAMPALIMAUD	351	71	176	281	351
3	RADBOUD	660	115	330	528	660
4	HULAFE	300	-	150	240	300
5	UNIPI	100	21	50	80	100
6	IPC (PCAL)	150	2	75	120	150
7	HACETTEPE	900	180	450	720	900
8	IDIBGI	105	71	53	84	105
9	JCC (VDC)	192	23	96	154	192
10	NCI	530	107	265	424	530
11	St. Savvas (GAONA)	200	200	100	160	200
12	RMH (ICR)	400	23	200	320	400
13	QUIRONSALUD	132	88	66	106	132
14	FPO	500	194	250	400	500
	Total	4.520	1.095	2.260	3.616	4.520

Figure 1.4: Overview of number prospective data uploaded and ongoing uploading agreements

Chapter 2

Prospective Validation of the Segmentation Master Models (Experiments Set 1)

2.1 Chapter Summary

In this chapter we present the evaluation of the segmentation models models developed in D5.3 by FCHAMPALIMAUD on prospective data. It can be noted that for both whole gland and zone segmentations, the performance is similar, while for lesions segmentation, we see a considerable improvement, which is most likely due to the small number of prospective annotated samples available. We do not provide a description of the training/data preparation since it was previously specified in D5.3.

2.2 Methods

Data Description

For prospective validation, cases were downloaded from the ProstateNet platform on October 11th 2023. An overview of the data, stratified by manufacturer, can be seen in Tab. 2.1. Since whole gland masks are generated by merging both Peripheral (PZ) and Transitional+Center (TZ) masks, they share the same data composition.

Analysis Description

For the prospective data validation presented in this chapter, we use only the FCHAMPALIMAUD ProstateAll models developed in D5.3. We test exclusively the full resolution T2W models as those were the best ones during retrospective evaluation. Additionally, we perform a fairness analysis regarding data provider.

	Total	Siemens	Philips	GE
Gland	211	29	176	6
Lesions	19	7	8	4

Table 2.1: Stratification of prospective samples by manufacturer for Gland and Lesion segmentation.

2.3 Results

In order to understand if our segmentation models were capable of generalizing to new, prospective data, we tested models that performed the best on the retrospective data - Full resolution ProstateAll nnUNets - on 211 cases, for both Whole Gland and Zone segmentations, and 19 cases for lesions segmentation.

Table 2.2 and Figure 2.1 show the obtained results. As it can be observed, the results are fairly similar to the ones obtained during retrospective evaluation, with Whole Gland showing the largest performance drop of $\approx 4\%$, Peripheral Zone dropping $\approx 2\%$, Transitional Zone gaining $\approx 1\%$, and lastly, Lesion segmentation showing a huge performance increase, of $\approx 27\%$. Analysing Fig. 2.1, it is clear that the performance of the lesion segmentation model is quite broad, both failing to segment anything in 3 out of 19 cases, and producing Dice scores above 0.8 for 9 out of 19 cases.

When comparing the performance for each of the providers (Fig. 2.2) it can be observed that overall, regardless of the segmentation task, all providers show similar scores, that is, apart from RMH, that shows considerably worse performance for all metrics regarding lesion segmentation. All zero and low scores obtained are in cases provided by that institution, which could indicate that there is some localized issue regarding the provided lesion annotation.

Additionally, it is also possible to know that the lesion segmentation models are capable of detecting the lesions, showing a very high Recall, of 0.86, for predictions above an IoU of 10%. In fact, if not for the outlier performance solely on RMH data, they would have a perfect Recall of 1.

	Whole Gland	Peripheral Zone	Transitional Zone	Lesions
Dice	0.88 ± 0.01	0.78 ± 0.01	0.87 ± 0.01	0.66 ± 0.06
HD	15.7 ± 1.36	19.19 ± 1.35	12.65 ± 0.88	30.15 ± 10.77
ASSD	0.57 ± 0.05	0.68 ± 0.04	0.79 ± 0.2	11.02 ± 8.17
RAVD	0.15 ± 0.02	0.21 ± 0.04	0.63 ± 0.33	0.09 ± 0.16
Recall	1	1	1	0.86

Table 2.2: Mean prospective results stratified by segmentation task.

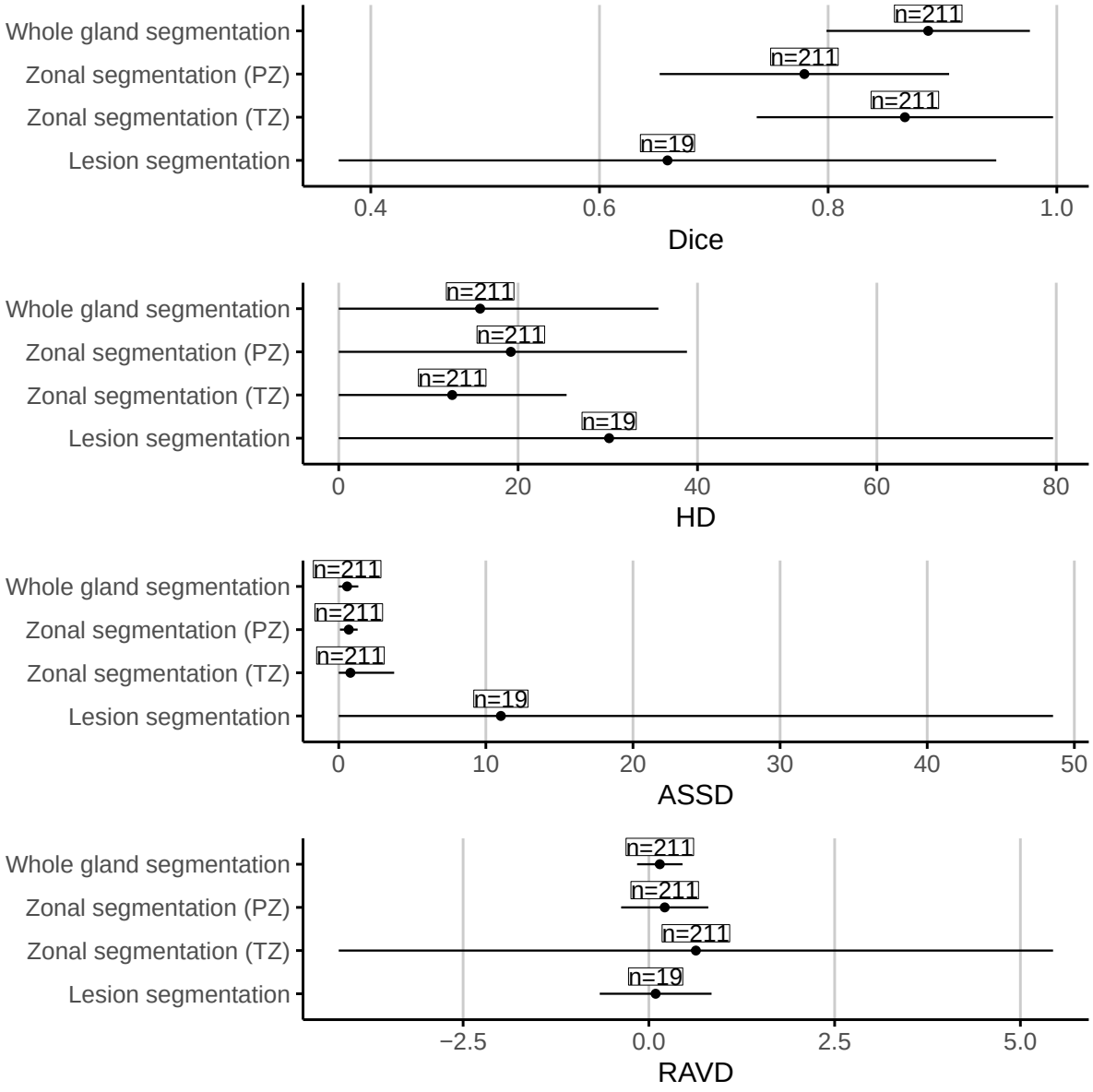


Figure 2.1: Distribution of the prospective results stratified by segmentation task.

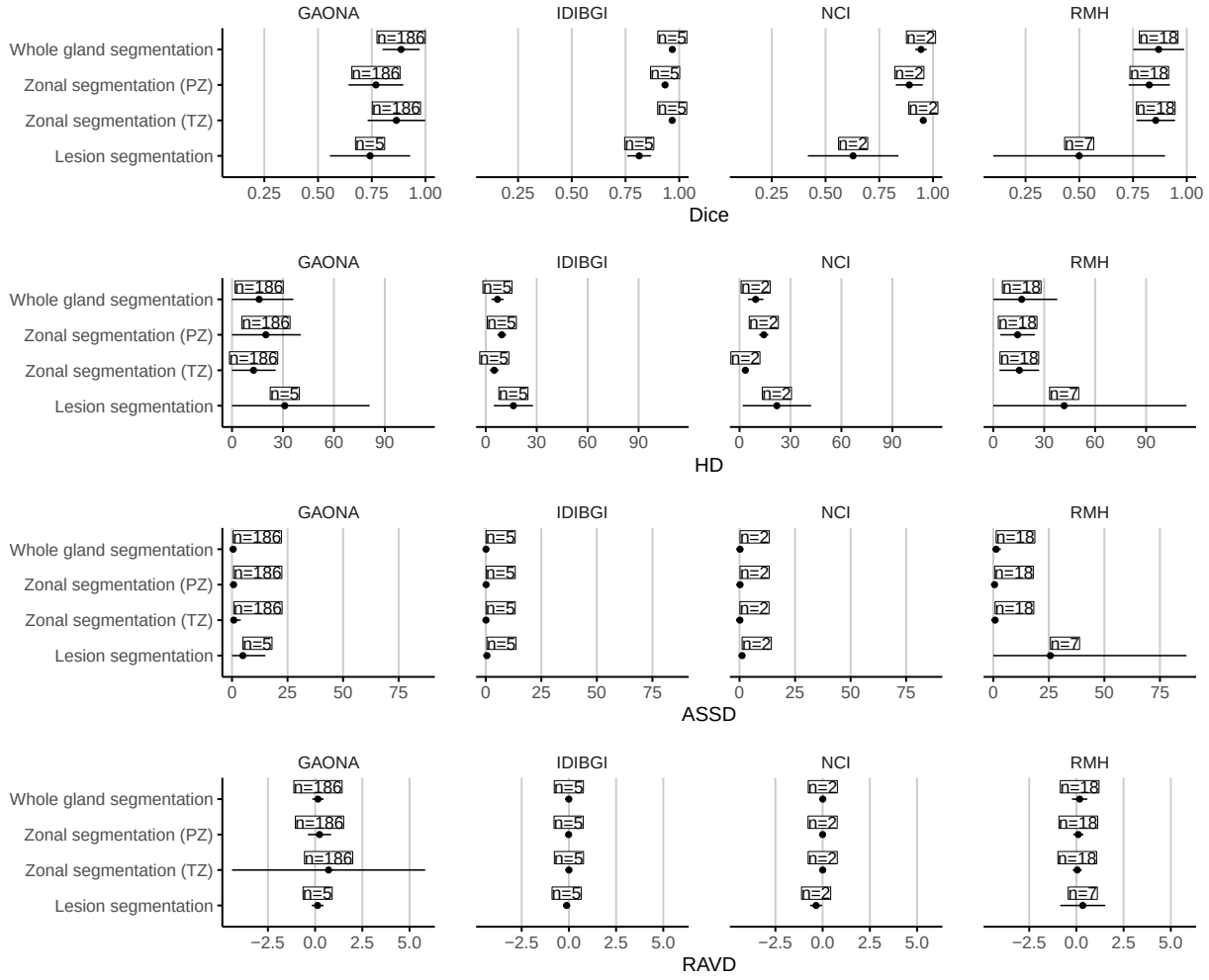


Figure 2.2: Distribution of the prospective results stratified by data provider.

2.4 Discussion

The results obtained from the prospective validation of the segmentation models show that there is no deterioration in performance, with whole gland and prostate zone models producing results similar to those obtained on retrospective data, and the lesion segmentation models showing a considerable performance improvement. Alas, this was done on a very small sample of prospective cases, which raises some concerns about the real applicability of the models.

Chapter 3

Prospective Validation of Radiomics Master Models (Experiments Set 1)

3.1 Chapter Summary

Each FCHAMPALIMAUD model developed in Deliverable 5.3 (Chapter 4: Radiomics Master Models) was validated with prospectively collected data. All prospective data preprocessing was exactly the same as in the retrospective dataset, so the preprocessing steps described in the following section (Methodology) are identical to those found in Deliverable 5.3 (Chapter 4: Radiomics Master Models). In this Chapter, we present the performance results on the prospective cases, which reflect the behaviour of these models on contemporary data.

3.2 Methods

Prospective Data Description

Our dataset consisted of T2W, DWI and ADC exams labelled as prospective in the ProstateNet image archive created under the scope of the ProCAncer-I project. The exams were acquired in the initial stages of the disease continuum. Ethics committee approval and patient consent were obtained by each clinical partner.

Feature Extraction

The extraction of radiomic and deep features was performed following the same methodology and parameters as described in Deliverable 5.3, including, for the radiomic features, the co-registration of T2W and DWI sequences and the automatic generation of the whole prostate gland segmentations. The quality of which, was assessed by an expert radiologist. Those results can be found in deliverable 5.3.

The clinical variables included for each use case are described in Deliverable 5.3. Cases with missing data were removed from the prospective validation cohort.

Model Validation

The master models developed and presented in deliverable 5.3 (Chapter 4: Radiomics Master Models) are summarized in Tables 3.1 and 3.2.

The master models were validated in a similar manner to the one described in Deliverable 5.3. We present the performance on the prospective cohort in terms of AUC, Sensitivity/Recall, Specificity, Precision, F1, F2 and Cohen's Kappa. Additionally, a fairness analysis is also conducted.

Large Use Cases	Target definition	Data type	MRI volumes	ML algorithm	Train size				Hold-out test set performance			
					Full	Siemens	Philips	GE	AUC	Sen	Spe	pdt
2	ISUP {1} vs ISUP {2,3,4,5}	radiomics	T2W, DWI and ADC	LightGBM	4107	2077	1365	665	0.7648	0.651	0.7647	0.7477
2	ISUP {1,2} vs ISUP {3,4,5}	radclin noERC	T2W, DWI and ADC	CatBoost	3899	2077	1365	457	0.7347	0.7931	0.4859	0.2158
2	ISUP {1,2,3} vs ISUP {4,5}	radiomics	T2W, DWI and ADC	LightGBM	4107	2077	1365	665	0.8427	0.5313	0.8802	0.2552
5	biochemical recurrence after RP at follow-up. Pre-surgery prediction	hybrid	T2W	CatBoost	709	429	182	98	0.6899	0.7143	0.5366	0.0038
5	biochemical recurrence after RP at follow-up. Post-surgery prediction	hybrid noERC	T2W	CatBoost	676	429	182	65	0.8188	0.8571	0.5854	0.0528

Table 3.1: Description of master models developed in workpackage 5.3. for each of the larger UCs

Small Use Cases	Target definition	Data type	MRI volumes	ML algorithm	Train size				CV performance			
					Full	Siemens	Philips	GE	AUC	Sen	Spe	pdt
3	metastasis at 6 months follow-up	hybrid noERC	DWI	SGD	62	31	22	9	0.8077	0.6155	1	0.5
6	biochemical recurrence after RT at follow-up	raddeep noERC	DWI	SGD	73	59	11	3	0.8393	0.8056	0.873	0.5
7a	urinary toxicity		T2W	sigmoid SVM	136				0.6988	0.7323	0.606	0.5?
7a	rectal toxicity		T2W	sigmoid SVM	136				0.7835	0.7654	0.6774	0.5?
8	left active surveillance	raddeep noERC	ADC	SGD	92	81	10	1	0.6905	0.6667	0.7143	0.5
7b	QoL (low, intermediate and high)	radclin	T2W	CatBoost	198	47	151	0	0.5489	-	-	-

Table 3.2: Description of master models developed in workpackage 5.3. for each of the smaller UCs

3.3 Results

Data Description

The total prospective dataset is composed of 530 patients (UC 2). Of these, 21 patients are also suitable for UC3, the prediction of metastasis, 47 patients for UC 5, the biochemical recurrence use case, and 30 patients for UC 8, the prediction of early withdrawal from the active surveillance program. The dataset size changes during the workflow are described in Table 3.3 for each use case.

	UC 2	UC 3	UC 5	UC 6	UC 7a	UC 7b	UC 8
Initial number of patients	530	21	47	0	0	0	30
T2 available	494	21	47	-	-	-	30
T2 segmentation	494	21	47	-	-	-	30
T2 extraction	492	21	46	-	-	-	30
Available ground truth	492	21	46	-	-	-	30
DWI exists	487	21	46	-	-	-	30
DWI exists and T2 mask available	487	21	46	-	-	-	30
DWI extraction	469	21	45	-	-	-	30
Available ground truth	469	21	45	-	-	-	30
ADC exists	489	21	46	-	-	-	30
ADC exists and T2 mask available	489	21	46	-	-	-	30
ADC extraction	466	21	45	-	-	-	30
Available ground truth	466	21	45	-	-	-	30

Table 3.3: Data workflow, specifying the number of prospective patients in each use case after each step.

For use cases 6, 7a and 7b, no prospective patients were available. For UC 8, as the model requires no endorectal coil in the exam, only 18 patients were available. All of these 18, had a negative ground truth and the model correctly predicted all of them, so no further results are shown. For UC 3, as the model requires no endorectal coil in the exam, only 12 patients were available. The performance on this small cohort is shown, however there were not enough patients to conduct a fairness analysis.

Use Case 2 - ISUP 1 vs ISUP 2,3,4,5

Model Performance

The prospective cohort for UC 2 included 466 patients. Defining the ground truth as ISUP 1 vs ISUP 2, 3, 4 and 5 resulted in 184 negative and 282 positive cases. The master model's predictions are shown in the confusion matrix in Figure 3.1 and the model's performance is described in Table 3.4, where the hold-out test set performance (previously shared in D5.3.) is included for comparison purposes. The results show the model was able to generalize to the prospective cohort, with minimal differences in performance. Overall, the performance found on the hold-out test set was reproduced and confirmed in the prospective cohort.

Fairness Analysis

Tables 3.5 to 3.11 show the master model's performance on different subsets of the prospective test set. Regarding scanner manufacturer (Table 3.5), the model is showing robustness, with its performance on the different cohorts not differing more than 10% across all performance metrics. For cases where an endorectal coil is used, there is a significant reduction in Precision (15.6% less precise), but this is followed by a 6.7% increase in Recall, which overall results in a very similar F-score.

Regarding lesion location, we see a significant drop in both precision and recall when an index lesion is not located in the peripheral zone (Table 3.7), which results in an F-score almost 16% lower when compared to a patient with an index lesion located in this area. This is not the case, however, for the transitional zone (Table 3.8), where the performance metrics for both subcohorts are not more than 5% apart, indicating strong robustness. Finally, regarding the central zone and anterior stroma (Tables 3.9 and 3.10), the relatively small number of cases prevent us from drawing definitive conclusions, however, the model does seem to perform better in the minority subcohorts, which might suggest robustness.

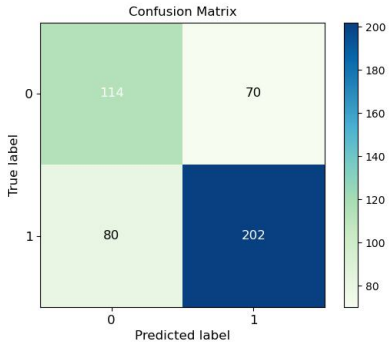


Figure 3.1: Confusion matrix of the radiomics_uc2_T2&DWI&ADC_LGBM model's predictions on the prospective cohort.

	Hold-out test set	**Prospective cohort**
AUC	0.7648	0.7364
Sensitivity/Recall/TPR	0.6510	0.7163
Specificity/TNR	0.7647	0.6196
Precision/PPV	0.8899	0.7426
F1	0.7519	0.7292
F2	0.6879	0.7214
CohensKappa	0.3305	0.3327

Table 3.4: Multi-metric performance of the radiomics_uc2_T2&DWI&ADC_LGBM model on the held-out test set and the prospective cohort.

manufacturer	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
SIEMENS	0.6931	0.7524	0.7619	0.7500	101	37	64	2077
PHILIPS	0.7163	0.6759	0.7604	0.6577	215	104	111	1365
GE MEDICAL SYSTEMS	0.6133	0.7486	0.7168	0.7570	150	43	107	665

Table 3.5: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by scanner manufacturer.

Endorectal coil	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Patients without ERC	0.7003	0.7186	0.7742	0.7059	397	159	238	3899
Patients with ERC	0.5507	0.7359	0.6182	0.7727	69	25	44	208

Table 3.6: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by presence/absence of endorectal coil.

index_lesion _location_PZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
1	0.6896	0.7401	0.7699	0.7331	393	142	251	3258
0	0.6164	0.5732	0.5455	0.5807	73	42	31	849

Table 3.7: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the peripheral zone.

index_lesion _location_TZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.6803	0.7274	0.7500	0.7220	391	150	241	3284
1	0.6667	0.6863	0.7000	0.6829	75	34	41	823

Table 3.8: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the transitional zone.

Regarding country of origin (Table 3.11), the model seems to perform in a similar manner, overall. It is worth pointing out the lowest performance, which was achieved with Lithuanian data. Here, the model achieved a perfect precision, but an extremely low recall. The highest performance was reached with Portuguese data, however the small size of the subcohort prevents us from drawing definitive conclusions. Finally, the model exceeded expectations with data from Greece, which was almost completely absent during

index_lesion _location_CZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.6733	0.7138	0.7356	0.7085	453	182	271	3919
1	0.8462	0.9091	0.9091	0.9091	13	2	11	188

Table 3.9: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the central zone.

index_lesion _location_AS	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.6741	0.7132	0.7375	0.7074	451	181	270	3859
1	0.8000	0.9016	0.8462	0.9167	15	3	12	248

Table 3.10: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the anterior stroma.

training. Despite this, on the Greek subcohort the model achieved the second highest F-score (after Portugal) of 0.8781.

country	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Netherlands	0.6833	0.7143	0.6969	0.7188	60	28	32	1622
Portugal	0.8182	0.9000	0.9000	0.9000	11	1	10	576
Lithuania	0.7922	0.1351	1.0000	0.1111	77	59	18	466
UK	0.6000	0.7229	0.8000	0.7059	20	3	17	447
Turkey	0.6346	0.6965	0.8485	0.6667	52	10	42	359
Italy	0.5541	0.7317	0.6207	0.7659	74	27	47	281
Spain	0.6309	0.7026	0.7963	0.6825	84	21	63	243
Greece	0.7500	0.8781	0.7313	0.9245	88	35	53	18

Table 3.11: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by country of origin of the data.

Use Case 2 - ISUP 1,2 vs ISUP 3,4,5

Model Performance

The prospective cohort for UC 2 included 395 patients, after exclusion of 71 cases where the MRI examination was performed with an endorectal coil. Defining the ground truth as ISUP 1 or 2 vs ISUP 3, 4 and 5 resulted in 292 negative and 103 positive cases. The master model's predictions are shown in the confusion matrix in Figure 3.2 and the model's performance is described in Table 3.12, where the hold-out test set performance (previously shared in D5.3.) is included for comparison purposes.

Overall, the model was able to generalize to the prospective cohort. In terms of precision, F1, F2 and Cohen's Kappa the performance on the prospective data was no more than 3 percentual points away from the hold-out test set performance. Additionally, for the same probability decision threshold, we observe a minor reduction in specificity (6% lower), but a significant rise in sensitivity (10% higher).

Fairness Analysis

Tables 3.13 to 3.18 show the master model's performance on different subsets of the prospective test set. Regarding scanner manufacturer (Table 3.13), the model shows similar recall scores in the three subcohorts, however, Phillips cases stand out in terms of precision, scoring 10% higher than Siemens and GE.

Regarding lesion location, we see a consistent rise in performance on the minority sub-cohorts: index lesion located in the TZ, CZ or AS (Tables 3.8, 3.9 and 3.10, respectively). Though for the CZ and AS the relatively small number of cases prevents us from drawing definitive conclusions. In terms of PZ lesion location we see a reduction in precision and an increase in recall, when the index lesion is not located in the PZ, but, overall, this results in a similar F-score.

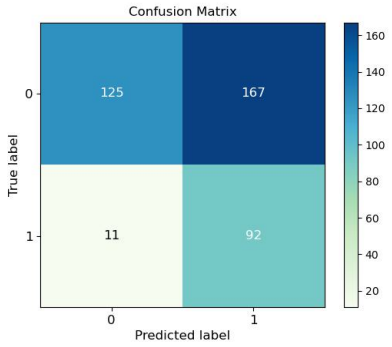


Figure 3.2: Confusion matrix of the radclin_uc2_T2&DWI&ADC_noERC _CatBoost model's predictions on the prospective cohort.

	Hold-out test set	**Prospective cohort**
AUC	0.7347	0.7801
Sensitivity/Recall/TPR	0.7931	0.8932
Specificity/TNR	0.4859	0.4281
Precision/PPV	0.3866	0.3552
F1	0.5198	0.5083
F2	0.6553	0.6855
CohensKappa	0.2128	0.2156

Table 3.12: Multi-metric performance of the radclin_uc2_T2&DWI&ADC_noERC_CatBoost model on the held-out test set and the prospective cohort.

manufacturer	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
SIEMENS	0.5149	0.6667	0.3188	0.9167	101	77	24	2077
PHILIPS	0.6093	0.7162	0.4077	0.8833	215	155	60	1365
GE MEDICAL SYSTEMS	0.4304	0.6250	0.2833	0.8947	79	60	19	457

Table 3.13: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by scanner manufacturer.

index_lesion _location_PZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
1	0.5375	0.6894	0.3673	0.8829	333	239	94	3077
0	0.6129	0.6522	0.2727	1	62	53	9	822

Table 3.14: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the peripheral zone.

index_lesion _location_TZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.5333	0.6643	0.3410	0.8706	330	245	85	31134
1	0.6308	0.7895	0.4286	1	65	47	18	786

Table 3.15: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the transitional zone.

index_lesion _location_CZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.5431	0.6688	0.3387	0.8842	383	288	95	3716
1	0.7500	0.9302	0.7273	1	12	4	8	183

Table 3.16: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the central zone.

Regarding country of origin (Table 3.18), the model shows consistently high recall, but fluctuating precision. The highest performances are reached with data from Italy, followed by UK, Turkey and Greece, all displaying F-scores above 0.7. The worst performances were surprisingly found with cases from Lithuania and Portugal, which were respectively the third and second most represented cohorts seen during training.

index_lesion _location_AS	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.5474	0.6746	0.3455	0.8854	380	284	96	3653
1	0.6000	0.8537	0.5385	1	15	8	7	246

Table 3.17: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the anterior stroma.

country	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Netherlands	0.4833	0.6771	0.2955	1	60	47	13	1622
Portugal	0.4546	0.4546	0.1429	1	11	10	1	576
Lithuania	0.7792	0.3704	0.1053	1	77	75	2	466
UK	0.7000	0.7692	0.5455	0.8571	20	13	7	447
Turkey	0.5385	0.7500	0.4500	0.9000	52	32	20	359
Spain	0.4405	0.6522	0.3231	0.8750	84	60	24	243
Italy	0.6667	0.8333	0.5000	1	3	2	1	73
Greece	0.4773	0.7109	0.4225	0.8571	88	53	35	18

Table 3.18: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by country of origin of the data.

Use Case 2 - ISUP 1,2,3 vs ISUP 4,5

Model Performance

The prospective cohort for UC 2 included 466 patients. Defining the ground truth as ISUP 1, 2 or 3 vs ISUP 4 and 5 resulted in 397 negative and 68 positive cases. The master model's predictions are shown in the confusion matrix in Figure 3.3 and the model's performance is described in Table 3.19, where the hold-out test set performance (previously shared in D5.3) is included for comparison purposes.

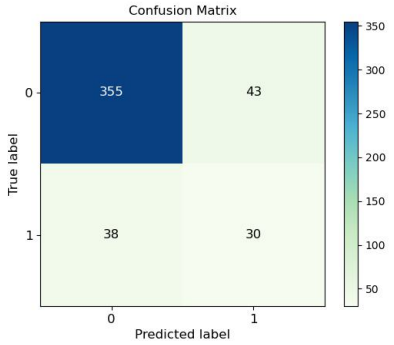


Figure 3.3: Confusion matrix of the radiomics_uc2_T2&DWI&ADC_LGBM model's predictions on the prospective cohort.

	Hold-out test set	Prospective cohort
AUC	0.8427	0.7564
Sensitivity/Recall/TPR	0.5313	0.4412
Specificity/TNR	0.8802	0.8919
Precision/PPV	0.4595	0.4109
F1	0.4928	0.4255
F2	0.5152	0.4348
CohensKappa	0.3870	0.3233

Table 3.19: Multi-metric performance of the radiomics_uc2_T2&DWI&ADC_LGBM model on the held-out test set and the prospective cohort.

A slight reduction is observable in all performance metrics (no larger than 9%) with the exception of Specificity, which shows an increase of 1%.

Fairness Analysis

Tables 3.20 to 3.26 show the master model's performance on different subsets of the prospective test set. Regarding scanner manufacturer (Table 3.20), there are clear differences across subcohorts. The lowest performance is obtained with GE cases, which can be partly explained by the lowest performance on ERC exams (Table 3.21) since most GE exams were acquired with endorectal coil.

manufacturer	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
SIEMENS	0.7921	0.6000	0.3333	0.7500	101	89	12	2077
PHILIPS	0.8930	0.4676	0.6842	0.4333	215	185	30	1365
GE MEDICAL SYSTEMS	0.7533	0.3053	0.2963	0.3077	150	124	26	650

Table 3.20: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by scanner manufacturer.

Endorectal coil	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Patients without ERC	0.8363	0.4856	0.4091	0.5094	397	344	53	3897
Patients with ERC	0.7681	0.2239	0.4286	0.2000	69	54	15	194

Table 3.21: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by presence/absence of endorectal coil.

Regarding lesion location, the model seems relatively robust with variations of F-score no greater than 10%. On both PZ and TZ (Tables 3.22 and 3.23) we see a slight increase in recall in the minority sub-cohorts. Regarding the central zone and anterior stroma (Tables 3.24 and 3.25), the relatively small number of cases prevent us from drawing definitive conclusions.

index_lesion _location_PZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
1	0.8219	0.4355	0.4355	0.4355	393	331	62	3245
0	0.8493	0.4286	0.2727	0.5000	73	67	6	846

Table 3.22: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the peripheral zone.

index_lesion _location_TZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.8312	0.4152	0.4035	0.4182	391	336	55	3271
1	0.8000	0.5147	0.4375	0.5385	75	62	13	820

Table 3.23: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the transitional zone.

index_lesion _location_CZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.8344	0.4235	0.3881	0.4333	453	393	60	3903
1	0.5385	0.5263	0.6667	0.5000	13	5	8	188

Table 3.24: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the central zone.

index_lesion _location_AS	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.8359	0.4299	0.4091	0.4355	451	389	62	3844
1	0.5333	0.4839	0.4286	0.5000	15	9	6	247

Table 3.25: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the anterior stroma.

The highest performance (0.8333 F-score) is found on cases from Lithuania and UK, however these

subcohorts contain one single positive case each, which immediately results in a perfect recall score. Thus, this is not a reliable performance estimate as it is highly influenced by the perfect recall obtained. Following these, are the Netherlands and Greece (third and fourth highest performances), which represent respectively, the most and the least represented subsets during training.

country	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Netherlands	0.7167	0.6034	0.3182	0.7778	60	51	9	1622
Portugal	0.9091	0	0	0	11	10	1	575
Lithuania	0.9870	0.8333	0.5000	1.0000	77	76	1	466
UK	0.9500	0.8333	0.5000	1.0000	20	19	1	447
Turkey	0.6923	0.5000	0.5000	0.5000	52	36	16	362
Italy	0.7568	0.2083	0.3750	0.1875	74	58	16	266
Spain	0.7857	0.2174	0.1429	0.2500	84	76	8	243
Greece	0.8977	0.5479	0.8889	0.5000	88	72	16	15

Table 3.26: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the prospective test set, divided by country of origin of the data.

Use Case 3

Model Performance

The prospective cohort for UC 3 included 22 patients, however, only 12 of them had exams taken without endorectal coil. The master model's predictions on this small cohort are shown in the confusion matrix in Figure 3.4 and the model's performance is described in Table 3.27, where the hold-out test set performance (previously shared in D5.3) is included for comparison purposes.

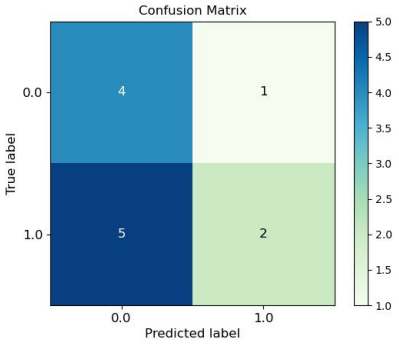


Figure 3.4: Confusion matrix of the hybrid_uc3_DWI_noERC_SGD model's predictions on the prospective cohort.

	Hold-out test set	**Prospective cohort**
AUC	0.8077	0.6286
Sensitivity/Recall/TPR	0.6155	0.2857
Specificity/TNR	1	0.8000
Precision/PPV	1	0.6667
F1	0.7595	0.4000
F2	0.6657	0.3226
CohensKappa	0.3466	0.0769

Table 3.27: Multi-metric performance of the hybrid_uc3_DWI_noERC_SGD model on the held-out test set and the prospective cohort.

A consistent and significant reduction is observable in all performance metrics.

Use Case 5 - Pre-surgery

Model Performance

The prospective cohort for UC 5 included 46 patients. The master model's predictions on this cohort are shown in the confusion matrix in Figure 3.5 and the model's performance is described in Table 3.28, where the hold-out test set performance (previously shared in D5.3) is included for comparison purposes.

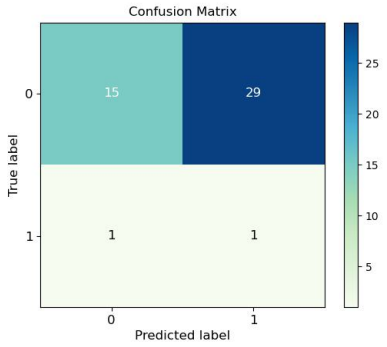


Figure 3.5: Confusion matrix of the hybrid_uc5_T2_CatBoost model's predictions on the prospective cohort.

	Hold-out test set	Prospective cohort
AUC	0.6899	0.3182
Sensitivity/Recall/TPR	0.7143	0.5
Specificity/TNR	0.5366	0.3409
Precision/PPV	0.2083	0.0333
F1	0.3226	0.0625
F2	0.4808	0.1316
CohensKappa	0.1250	-0.0207

Table 3.28: Multi-metric performance of the hybrid_uc5_T2_CatBoost model on the held-out test set and the prospective cohort.

Fairness Analysis

Tables 3.29 to 3.34 show the master model's performance on different subsets of the prospective test set. The small number of positive cases in the prospective cohort (2 cases) makes it very difficult to draw definitive conclusions. Most performance metrics easily go to zero when a sub-cohort does not include any positive cases or fails in the one or two that it has. For this reason, we will focus on accuracy for this part of the analysis. Looking at this metric, the model seems relatively robust across all sensitive attributes, with accuracy always between 0.32 and 0.44, with the exception of Greek cases, where the model performed with less than 1% accuracy.

manufacturer	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
PHILIPS	0.3200	0	0	0	25	25	0	429
SIEMENS	0	0	0	0	1	0	1	182
GE MEDICAL SYSTEMS	0.4000	0.2941	0.0769	1	20	19	1	98

Table 3.29: hybrid_uc5_T2_CatBoost model performance on sub cohorts of the prospective test set, divided by scanner manufacturer.

index_lesion _location_PZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
1	0.3500	0.1471	0.0385	0.5000	40	38	2	633
0	0.3333	0	0	0	6	6	0	76

Table 3.30: hybrid_uc5_T2_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the peripheral zone.

index_lesion _location_TZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.3421	0.1515	0.0400	0.5000	38	36	2	592
1	0.3750	0	0	0	8	8	0	117

Table 3.31: hybrid_uc5_T2_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the transitional zone.

index_lesion _location_CZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.3488	0.1389	0.0357	0.5000	43	41	2	709
1	0.3333	0	0	0	3	3	0	0

Table 3.32: hybrid_uc5_T2_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the central zone.

index_lesion _location_AS	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.3636	0.1389	0.0357	0.5000	44	42	2	649
1	0	0	0	0	2	2	0	60

Table 3.33: hybrid_uc5_T2_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the anterior stroma.

country	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Spain	0.4286	0	0	0	7	6	1	96
Turkey	0.4400	0.2632	0.0667	1	25	24	1	82
Lithuania	0.3333	0	0	0	3	3	0	88
Greece	0.0909	0	0	0	11	11	0	0

Table 3.34: hybrid_uc5_T2_CatBoost model performance on sub cohorts of the prospective test set, divided by country of origin of the data.

Use Case 5 - Post-surgery

Model Performance

The prospective cohort for UC 5 included 46 patients, all of them with no endorectal coil. The master model's predictions on this cohort are shown in the confusion matrix in Figure 3.6 and the model's performance is described in Table 3.35, where the hold-out test set performance (previously shared in D5.3) is included for comparison purposes.

Fairness Analysis

Tables 3.36 to 3.41 show the master model's performance on different subsets of the prospective test set. In concordance with the pre-surgical context, the small number of positive cases in the prospective cohort (2 cases) makes it very difficult to draw definitive conclusions. Most performance metrics easily go to zero when a sub-cohort does not include any positive cases or fails in the one or two that it has. For this reason, we will focus on accuracy for this part of the analysis. Additionally, the extremely small size of some sub-cohorts prevents us from accurately assessing the performance (this was the case for all index lesion locations, Tables 3.37 - 3.40).

Regarding scanner manufacturer, the model achieves the highest performance with GE cases (accuracy 0.55), while it drops significantly for Philips (accuracy 0.16). In terms of country of origin (Table 3.41) the highest performance is achieved with Turkish cases (accuracy 0.52).

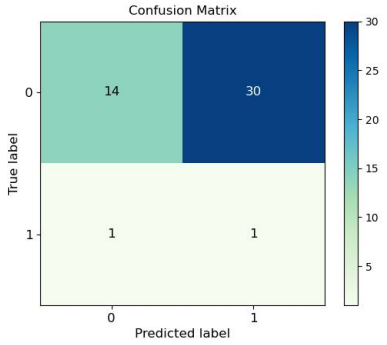


Figure 3.6: Confusion matrix of the hybrid_uc5_T2_noERC_CatBoost model's predictions on the prospective cohort.

	Hold-out test set	Prospective cohort
AUC	0.8188	0.2727
Sensitivity/Recall/TPR	0.8571	0.5
Specificity/TNR	0.5854	0.3182
Precision/PPV	0.2609	0.0323
F1	0.4	0.0606
F2	0.5882	0.1282
CohensKappa	0.2272	-0.02297

Table 3.35: Multi-metric performance of the hybrid_uc5_T2_noERC_CatBoost model on the held-out test set and the prospective cohort.

manufacturer	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
PHILIPS	0.1600	0	0	0	25	25	0	429
SIEMENS	0	0	0	0	1	0	1	182
GE MEDICAL SYSTEMS	0.5500	0.3571	0.1000	1	20	19	1	65

Table 3.36: hybrid_uc5_T2_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by scanner manufacturer.

index_lesion _location_PZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
1	0.3500	0.1471	0.0385	0.5000	40	38	2	602
0	0.1667	0	0	0	6	6	0	74

Table 3.37: hybrid_uc5_T2_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the peripheral zone.

index_lesion _location_TZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.3684	0.1563	0.0417	0.5000	38	36	2	561
1	0.1250	0	0	0	8	8	0	115

Table 3.38: hybrid_uc5_T2_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the transitional zone.

index_lesion _location_CZ	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.3256	0.1351	0.0345	0.5000	43	41	2	709
1	0.3333	0	0	0	3	3	0	0

Table 3.39: hybrid_uc5_T2_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the central zone.

index_lesion _location_AS	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.3409	0.1351	0.0345	0.5000	44	42	2	616
1	0	0	0	0	2	2	0	60

Table 3.40: hybrid_uc5_T2_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by presence/absence of lesion in the anterior stroma.

country	Accuracy	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Spain	0	0	0	0	7	6	1	96
Turkey	0.5200	0.2941	0.0769	1	25	24	1	82
Lithuania	0.3333	0	0	0	3	3	0	88
Greece	0.0909	0	0	0	11	11	0	0

Table 3.41: hybrid_uc5_T2_noERC_CatBoost model performance on sub cohorts of the prospective test set, divided by country of origin of the data.

3.4 Discussion

The prospective validation of the master models developed in Deliverable 5.3. is of extreme relevance for their translation to the clinical setting, as it should, when done correctly, give the user an estimate of the model's "real-world" performance. To do this type of analysis in a relevant manner, we need a significant amount of data, hopefully, representative of the heterogeneity found in the clinic, whether in terms of image quality, clinician expertise, patient socio-economic status, and so on. Unfortunately, the consortium was not able to generate these large volumes of data in the prospective arm of the project for most use cases, making it impossible to accurately prospectively validate them and keeping them as proof-of-concept type studies.

The exception to this, was use case 2, where over 400 cases were included in the prospective cohort. For all ground truth definitions (ISUP 1 vs 2345, ISUP 12 vs 345 and ISUP 123 vs 45), the models generalized extremely well to the prospective cases, showing only minor variations in the performance metrics when compared with the hold-out test set performance reported in Deliverable 5.3.

Contrary to what was found for use case 2, the performance of models from use cases 3 and 5 dropped significantly in the prospective cohort. Though it is relevant to take into account the rather small size of the prospective dataset (12 and 46 cases, for use cases 3 and 5, respectively) and, in the case of use case 5, its extreme imbalance (only 2 positive cases). For these reasons, it would be imprudent to assume that this is the model's actual performance out in the "real-world".

Despite this, several reasons could be behind the high and low generalizability of models trained for the different use cases. Firstly, use case 2 models rely mainly on radiomics-only or combined with clinical variables, while use case 3/5 models were trained with hybrid data, comprising radiomics, clinical variables and deep features, the latter of which are known to easily overfit. Secondly, use case 3/5 models were trained using only one MRI volume, namely T2W and DWI for use case 3 and 5, respectively, while use case 2 models were trained with information from the three MRI sequences available (T2W, DWI and ADC). The latter process resembles in a closer way the clinician's overall practice when assessing a prostate cancer patient, since the three volumes provide distinct and relevant information for the clinical decisions made. The final reason, and the most likely suspect in the differences of generalization power, is the size of the dataset used for training. While use case 2 models were trained with around 4000 cases, the datasets used in use cases 3 and 5 comprised only around 60 and 700 cases, respectively. Even though, for use case 5, this is a significant number of cases, it is still five times less than in use case 2, which is likely to later reflect in the generalization power of the final models.

Additionally, there is also a higher degree of difficulty when trying to predict biochemical recurrence or metastatic development (both future events) as opposed to the clinical aggressiveness already present in the image. Not to mention the confounding variables that also highly influence recurrence-free survival after radical prostatectomy and that we are unable to control for, such as, for example, the surgeon's expertise.

To conclude, despite the challenges faced, namely prospective dataset size and heterogeneity in the different use cases, this was a successful validation of the master models presented in Deliverable 5.3, especially for use case 2 where all three models were able to maintain their performance in the prospective cohort.

Chapter 4

Prospective Validation of Deep Learning Master Models (Experiments Set 1)

4.1 Chapter Summary

For this chapter, the evaluation of deep-learning models developed in D5.3 by FCHAMPALIMAUD on prospective data are presented. Here, a particular aspect becomes evident — there is a non-negligible drop in performance for most mpMRI models across targets. We demonstrate that the likely driver of this is the shift in scanner/vendor composition.

In this section we validate the models developed in D5.3 with prospective data. We will not provide a description of training/data preparation at this stage as it had been previously specified in D5.3. Nonetheless, a clear methodological description of training and data preparation is provided ahead in chapter 10 as models were retrained and the reader is referred there if any details on the training are necessary.

4.2 Methods

Data Description

For prospective validation, cases were downloaded from the ProstateNet platform on October 11th 2023. After excluding all retrospective cases and keeping only cases where all three sequences were available, a total of 466 cases remained (Table 4.1).

Analysis Description

For the validation in this section, only the FCHAMPALIMAUD VGG models developed in D5.3. Both T2W and mpMRI (T2W+DWI+ADC) models were tested to understand how using different sequences affects generalisation. A fairness (subgroup) analysis is also performed in terms of dataset provider, lesion PI-RADS, manufacturer (vendor), age, PSA, and lesion location. Finally, an enrichment analysis, described ahead in the results, is also presented to help make sense of differences in performance.

Manufacturer	**ISUP=1**	**ISUP=2**	**ISUP=3**	**ISUP=4**	**ISUP=5**
GE (ERC)	26	20	10	11	4
GE (no ERC)	19	47	15	2	11
Philips	103	50	30	14	17
Siemens	22	40	13	6	6

Table 4.1: Prospective dataset composition stratified by ISUP and vendor (manufacturer).

4.3 Results

Prospective Validation

To understand how performance generalises to new, prospective data, the sequence-only VGG models detailed above — which performed the best out of other architectures — were tested on a prospective test set with 466 cases.

Prospective set composition. Firstly, it should be noted that the relative proportions of cases changed both in terms of data provider (Figure 4.1) and ISUP distribution (Figure 4.2). Of note is the fact that studies acquired using Siemens were considerably more prevalent in the retrospective cohort, whereas Philips are now the most prevalent, and the fact that the manufacturer composition also significantly changed.

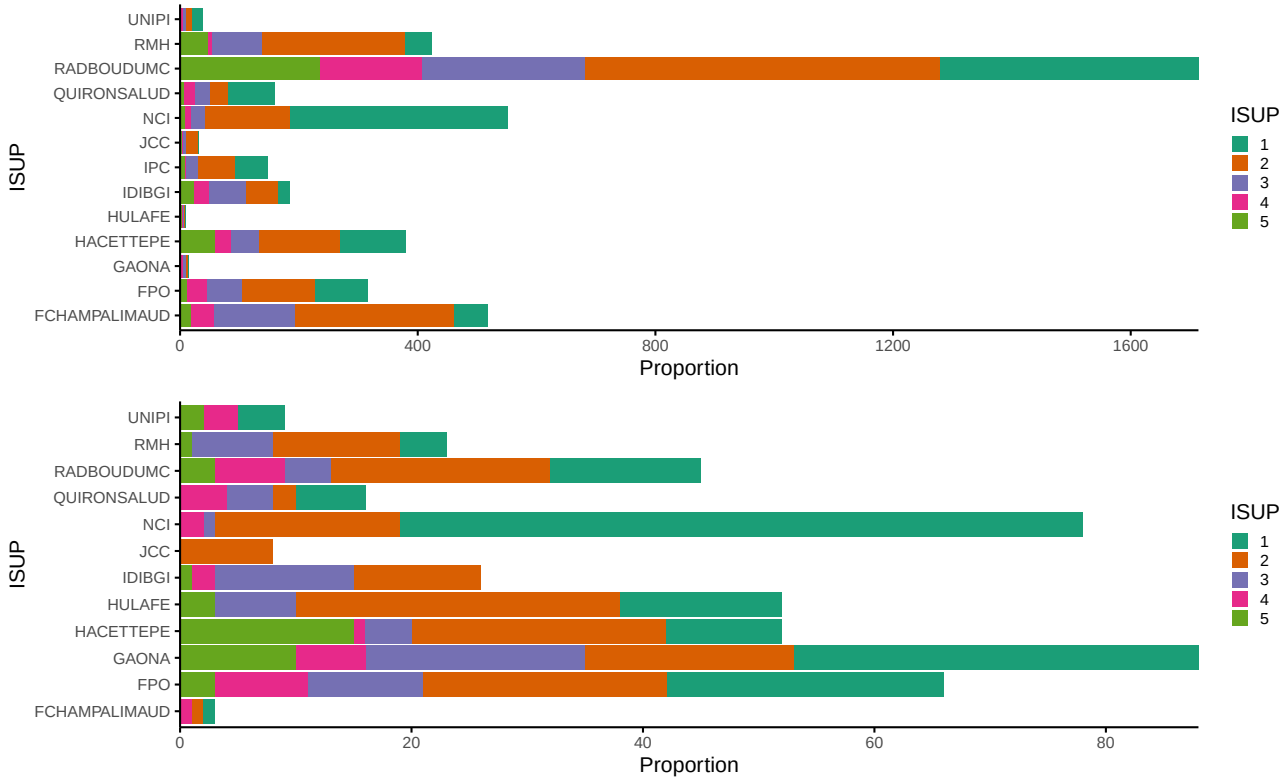


Figure 4.1: Case count stratified by ISUP and data provider for the retrospective (top) and prospective (bottom) data.

Performance analysis Firstly, it should be noted that poor generalization is observed for mpMRI models but not for T2W-only models (Figure 4.3; Figure 4.4), with the exception of the intermediate vs. high risk models where poor generalization is observed for both T2W-only and mpMRI models.

These results are not necessarily unexpected — indeed, as what is presented later in the feature visualization and concerning dataset distances (Figure 10.31; Figure 10.32; Figure 10.15; Figure 10.16; Figure 10.47; Figure 10.48) hints that the differences in composition observable in Figure 4.1 and **??** could have a significant impact on performance. In other words, shifts in the prevalence of different data providers and manufacturers could lead to shifts in performance.

Fairness analysis. To gain a better grasp on the failure of these models a fairness analysis was performed to better understand in which cases they showed a decrease in performance (Figure 4.5). In general, trends

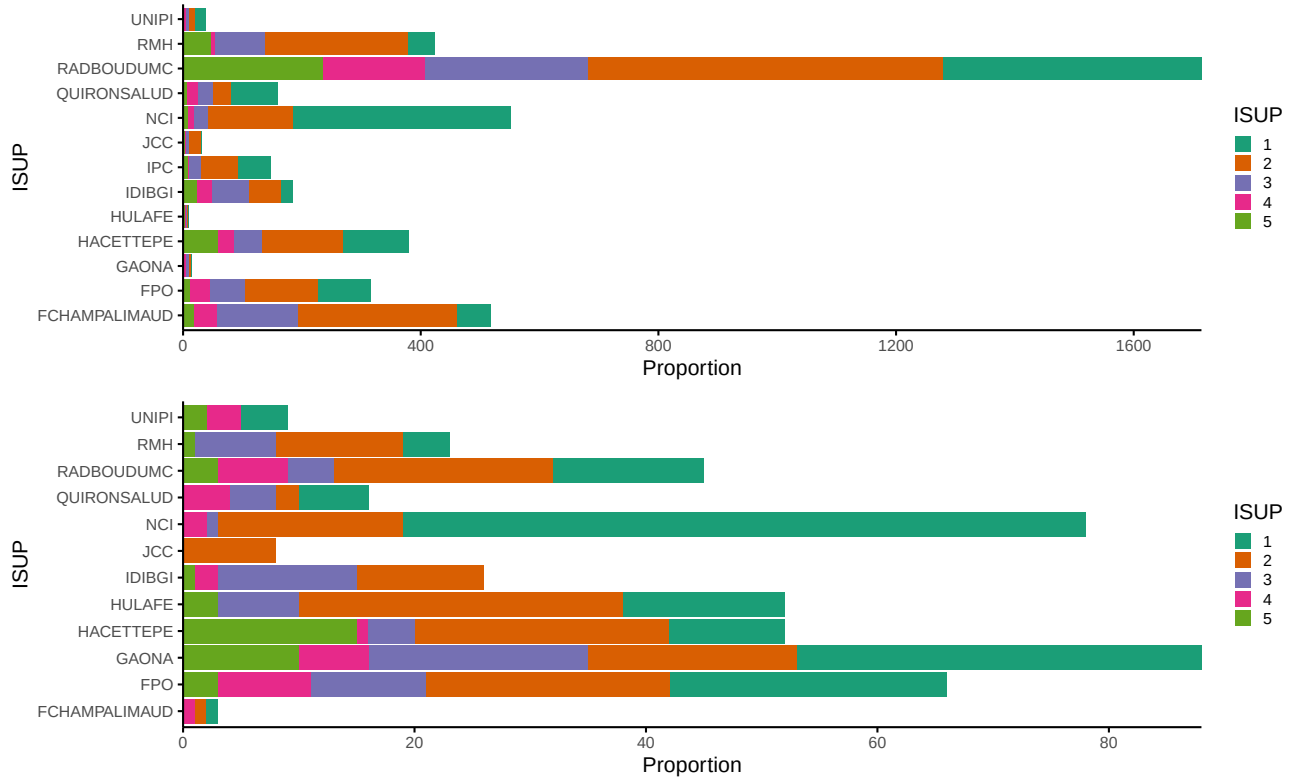


Figure 4.2: Case count stratified by ISUP and manufacturer for the retrospective (top) and prospective (bottom) data.

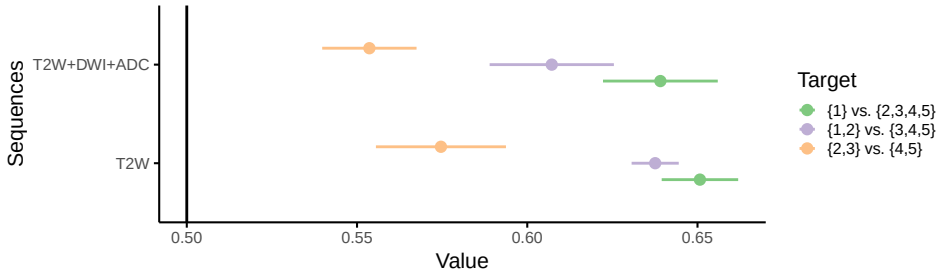


Figure 4.3: Performance of sequence-only VGG models trained on all manufacturers on the prospective test set. Circles represent the mean of the 5-folds and the horizontal lines represent the standard error around the mean.

are similar to those described in D5.3, with performance being largely determined by dataset size and manufacturer (data providers using Philips scanners such as FCHAMPALIMAUD, RADBOUD or GAONA show considerably better performance than other dataset providers). A negative trend is observed for both the low vs. possibly high and possibly low vs. high when looking at age, which provides a relevant subset where the application of these models can be of critical importance — indeed, at lower ages the performance is consistently higher than at higher ages. For the remaining subgroups, no relevant trend was detected.

Model enrichment analysis. To better understand how shifts in scanner prevalence between retrospective and prospective sets affect AUC, an enrichment metric was defined as the ratio between the prevalence of a model in the prospective set for a given provider and the prevalence of a model in the retrospective set

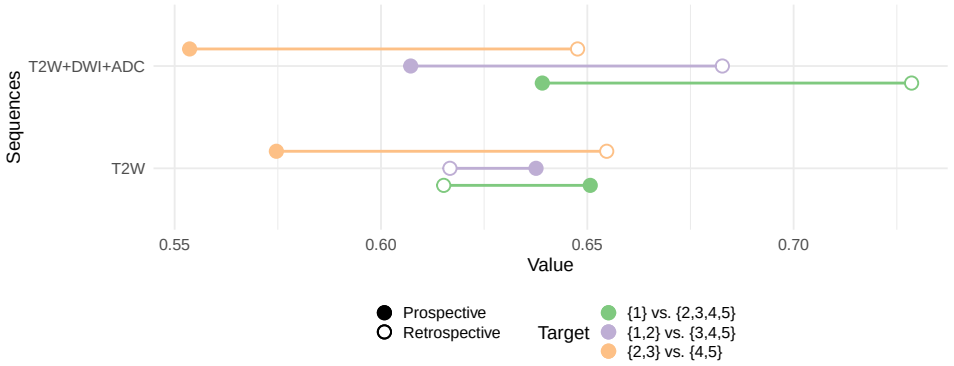


Figure 4.4: Comparison of sequence-only VGG models trained on the all manufacturers on the hold-out and prospective test set. Circles represent the mean of the 5-folds and the horizontal lines represent the standard error around the mean.

for the same provider. More formally, let $n_{\text{model,data}}$ and n_{data} be the number of studies acquired with a given scanner model in a given data provider and the number of studies acquired in a given data provider, respectively, with a r or p superscript denoting retrospective or prospective studies. The enrichment is thus calculated as enrichment $= \frac{n^p_{\text{model,data}}/n^p_{\text{data}}}{n^r_{\text{model,data}}/n^r_{\text{data}}}$. When enrichment $= 1$, this implies that a scanner is equally prevalent for a given data provider. We say that data is "retrospective biased" when enrichment < 1.2 (the model/data provider combination was more common in the retrospective set). and "prospective biased" otherwise (the model/data provider combination was more common in the prospective dataset).

In Figure 4.6 and Figure 4.7 it is easily observable that there are relatively large shifts in how scanner models are distributed between retrospective and prospective dataset. Considering HACETTEPE as an example, the most prevalent scanner in the retrospective set (Philips Ingenia; the same happens with QUIRONSALUD with Philips Achieva) is no longer used for the prospective cases, whereas a GE Signa Architect becomes the most prevalent in the prospective test set after being only the third most prevalent during training with retrospective data. An important aspect of this analysis is that models were exposed to significantly higher amounts of variability during training — most scanners that were used to acquire images during the retrospective set were no longer used during the prospective data collection.

Using the heuristic classification of retrospective and prospective biased data, the AUC was calculated separately for both; this showed that a large part of the poor generalisation for the low vs. possibly high and possibly low vs. high targets can be attributed to shifts to the prevalence of different models across data providing institutions (Figure 4.8). This is only the case for T2-only models with the intermediate vs. high target definition — indeed, mpMRI models for this case are no better than random when considering the retrospective biased data, suggesting that these models are likely to not be applicable in a clinical setting. Finally, it should be noted that this significantly reduces the number of applicable instances for some data centres, while also guaranteeing that these models are harder to transfer to new settings Figure 4.9.

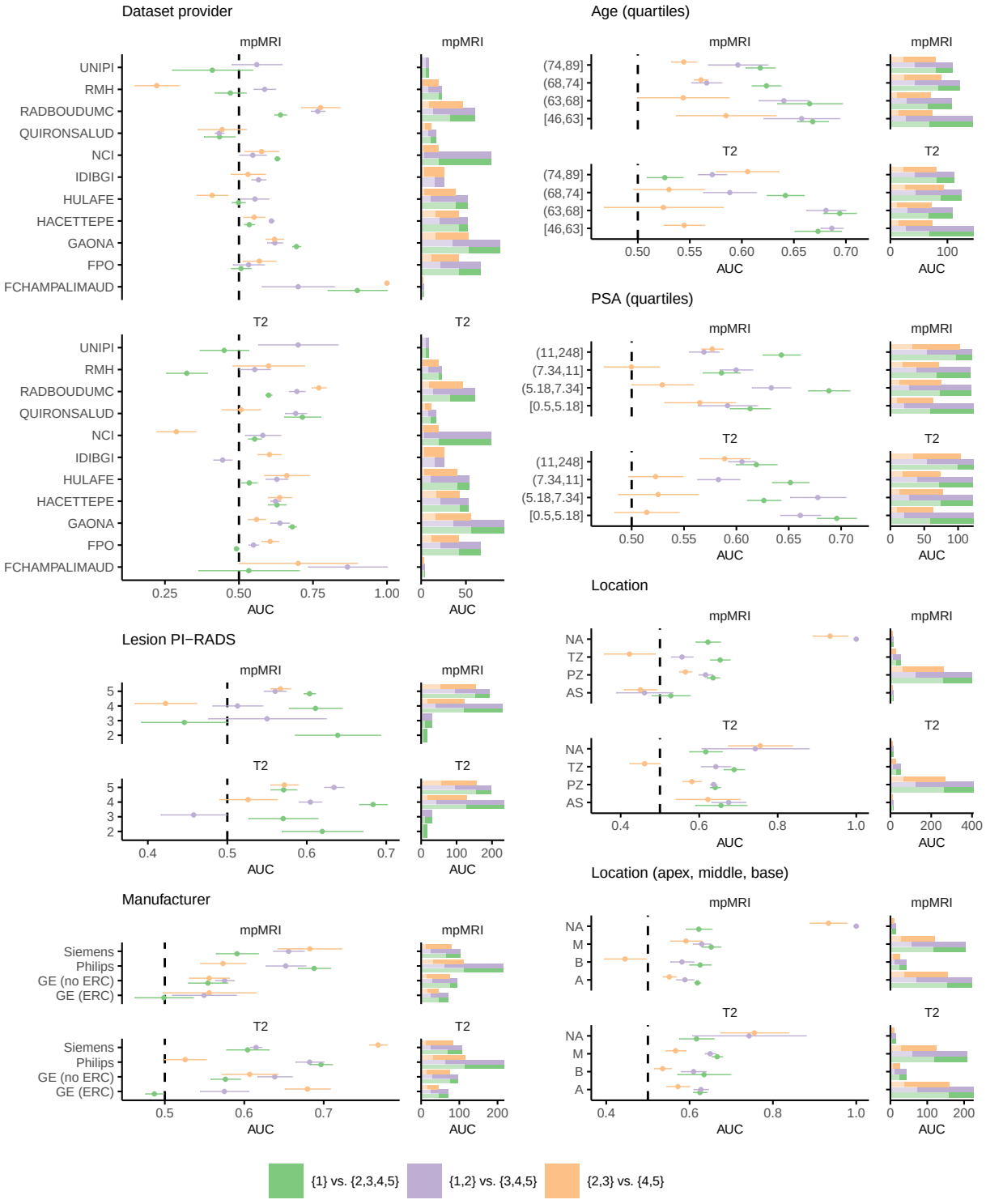
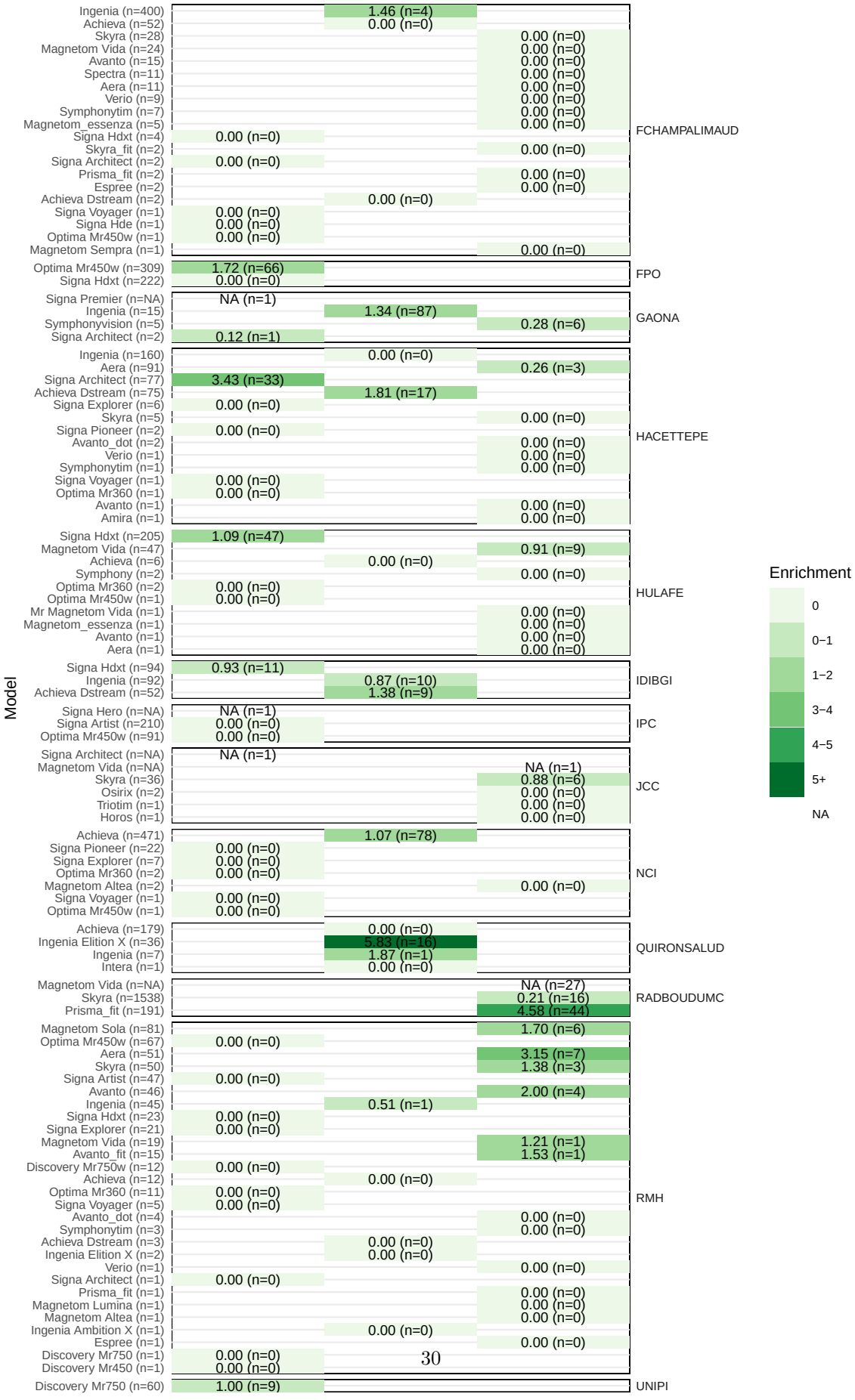


Figure 4.5: Fairness analysis for all binary target definitions for the deep-learning models. Point-and-range plots represent the performance (mean and standard error of the mean) and the coloured vertical lines represent the expected performance on the whole dataset. Horizontal bar plots represent the counts in each target and stratum for the testing set. The lighter fraction of the bars represents the positive cases.


Model
Ingenia (n=400)
Achieva (n=52)
Skyra (n=28)
Magnetom Vida (n=24)
Avanto (n=15)
Spectra (n=11)
Aera (n=11)
Verio (n=9)
Symphonytim (n=7)
Magnetom_essenza (n=5)
Signa Hdxt (n=4)
Skyra_fit (n=2)
Signa Architect (n=2)
Prisma_fit (n=2)
Espree (n=2)
Achieva Dstream (n=2)
Signa Voyager (n=1)
Signa Hde (n=1)
Optima Mr450w (n=1)
Magnetom Sempra (n=1)
FCHAMPALIMAUD
Optima Mr450w (n=309)
Signa Hdxt (n=222)
FPO
Signa Premier (n=NA)
Ingenia (n=15)
Symphonyvision (n=5)
Signa Architect (n=2)
GAONA
Ingenia (n=160)
Aera (n=91)
Signa Architect (n=77)
Achieva Dstream (n=75)
Signa Explorer (n=6)
Skyra (n=5)
Signa Pioneer (n=2)
Avanto_dot (n=2)
Verio (n=1)
Symphonytim (n=1)
Signa Voyager (n=1)
Optima Mr360 (n=1)
Avanto (n=1)
Amira (n=1)
HACETTEPE
Signa Hdxt (n=205)
Magnetom Vida (n=47)
Achieva (n=6)
Symphony (n=2)
Optima Mr360 (n=2)
Optima Mr450w (n=1)
Mr Magnetom Vida (n=1)
Magnetom_essenza (n=1)
Avanto (n=1)
Aera (n=1)
HULAFE
Signa Hdxt (n=94)
Ingenia (n=92)
Achieva Dstream (n=52)
IDIBGI
Signa Hero (n=NA)
Signa Artist (n=210)
Optima Mr450w (n=91)
IPC
Signa Architect (n=NA)
Magnetom Vida (n=NA)
Skyra (n=36)
Osirix (n=2)
Triotim (n=1)
Horos (n=1)
JCC
Achieva (n=471)
Signa Pioneer (n=22)
Signa Explorer (n=7)
Optima Mr360 (n=2)
Magnetom Altea (n=2)
Signa Voyager (n=1)
Optima Mr450w (n=1)
NCI
Achieva (n=179)
Ingenia Elition X (n=36)
Ingenia (n=7)
Intera (n=1)
QUIRONSALUD
Magnetom Vida (n=NA)
Skyra (n=1538)
Prisma_fit (n=191)
RADBOUDUMC
Magnetom Sola (n=81)
Optima Mr450w (n=67)
Aera (n=51)
Skyra (n=50)
Signa Artist (n=47)
Avanto (n=46)
Ingenia (n=45)
Signa Hdxt (n=23)
Signa Explorer (n=21)
Magnetom Vida (n=19)
Avanto_fit (n=15)
Discovery Mr750w (n=12)
Achieva (n=12)
Optima Mr360 (n=11)
Signa Voyager (n=5)
Avanto_dot (n=4)
Symphonytim (n=3)
Achieva Dstream (n=3)
Ingenia Elition X (n=2)
Verio (n=1)
Signa Architect (n=1)
Prisma_fit (n=1)
Magnetom Lumina (n=1)
Magnetom Altea (n=1)
Ingenia Ambition X (n=1)
Espree (n=1)
Discovery Mr750 (n=1)
Discovery Mr450 (n=1)
RMH
Discovery Mr750 (n=60)
UNIPI
ge
philips
siemens
Enrichment
0
0–1
1–2
3–4
4–5
5+
NA

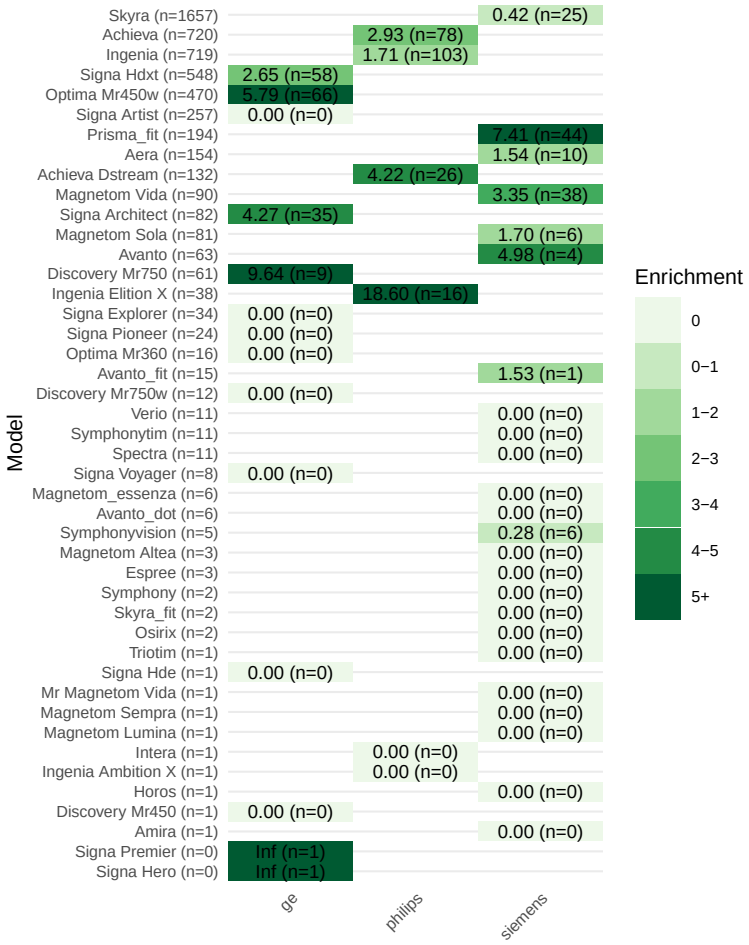


Figure 4.7: Enrichment values for different model/manufacturer combinations. The colour represents the enrichment values, whereas the text specifies the enrichment and the number of cases in the prospective set. The number after each model name specifies the number of data points in the retrospective set.

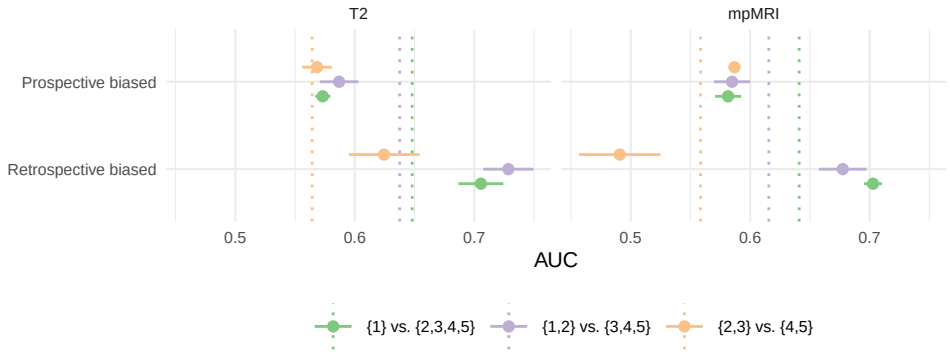


Figure 4.8: Performance stratified by target, enrichment classification and sequence used for modelling.

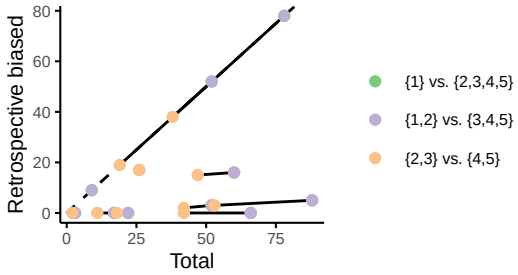


Figure 4.9: Change in the number of cases when evaluating on all instances and only in instances which are retrospective biased. Lines connecting separate points signify that both points belong to same dataset but have different target definitions.

4.4 Discussion

The results presented here concerning the prospective validation of deep-learning models allow the outline of some core conclusions which create significant issues with the applicability of these models:

1. Models are likely to require additional finetuning with new data to guarantee that their deployment does not deteriorate. This entails a framework that not only sustains predictive approaches, but also implements real-time or continuous evaluation of these models, requiring constant feedback from medical doctors

2. Generalisation using mpMRI data appears to be significantly harder than with T2W data (even when considering retrospective data in the prospective evaluation, T2W models are comparable to mpMRI models). This is possibly associated with a number of issues — while T2W images are relatively stable at least in qualitative terms (excluding studies acquired with ERC), diffusion or functional imaging is significantly more variable and the patterns associated with it are harder to learn when using such cohorts. Indeed, the results outlined for radiomics models point in two concrete directions forward — in the absence of proper lesion annotations, the best approach forward (which guarantees the best generalisation) is likely to be a combination of automatic prostate segmentation and radiomics.

Chapter 5

Prospective Validation of Deep Learning Master Models (Experiments Set 2)

5.1 Chapter Summary

In Deliverable 5.3, FORTH presented four different deep learning-based models to help us determine if a tumor is present or not in a patient's MRI examination. The VGG model was chosen because it had the best ACC, AUC, and F1 score results. In this chapter we are testing the generalisation ability of our VGG model on prospective data. Similarly to Deliverable 5.3, two subsets were used for training the models: i) entire dataset and ii) data from the left branch only for the negative PCa class, leaving out data that had positive MRI results but negative biopsy findings.

5.2 Methods

Data Description

We validated our model (trained on retrospective data) using ProstateNET's 425 prospective cases. This subset included 207 positive PCa cases and 218 negative PCa cases (see Table 8.1). Details on the retrospective training data and its preparation can be found in D5.3.

Case Type	Number of Cases
Positive PCa	**207**
Negative PCa	**218**

Table 5.1: Prospective dataset breakdown for Use Case 1.

5.3 Results

Prospective Validation

In this section, we present the outcomes from models trained on retrospective data when tested on prospective data across two distinct scenarios. Additionally, we showcase the findings from D5.3 to evaluate how these models generalize to unseen prospective data.

Table 8.2 displays results from the model subjected to two testing scenarios: Initially, it is trained and evaluated using hold-out retrospective data, and subsequently, the same model, trained on retrospective

	retrospective data	*prospective data*
ACC	**0.7479**	0.7254
AUC	**0.8121**	0.8013
F1 score	**0.7860**	0.7335

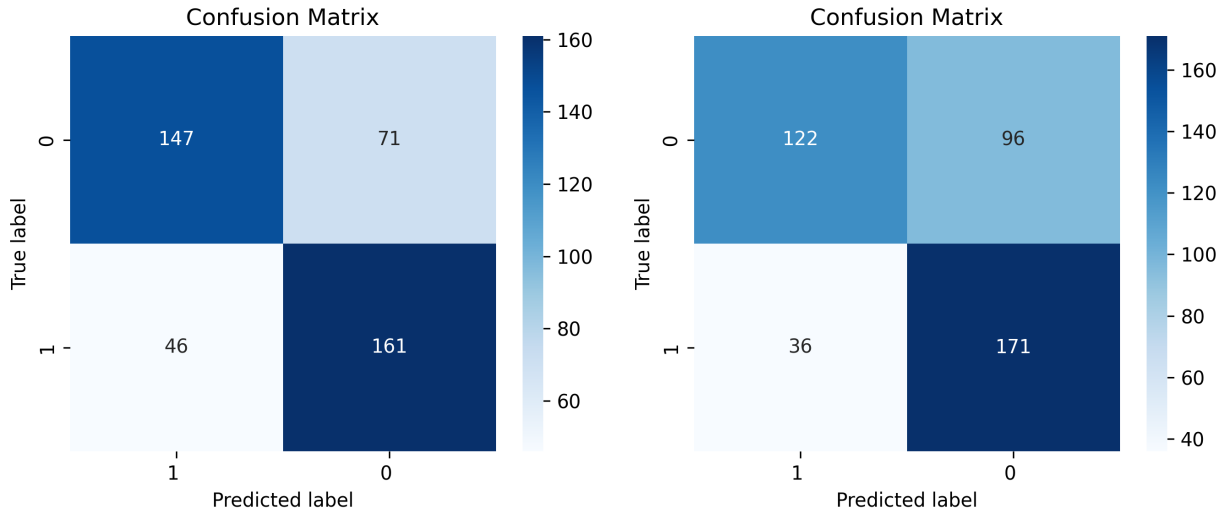
Table 5.2: Classification results for Use Case 1 using our best performing VGG model trained on both branches and tested on hold-out retrospective and prospective data respectively.

	retrospective data	*prospective data*
ACC	**0.7956**	0.6878
AUC	**0.8242**	0.7812
F1 score	**0.8454**	0.7215

Table 5.3: Classification results for Use Case 1 using our best performing VGG trained exclusively on left branch data, and tested on hold-out retrospective and prospective data respectively.

data, is tested on prospective data. The data in both scenarios are derived from both branches (see section 5.3 of Deliverable 5.3). On the other hand, Table 8.3 shows results for the other scenario where training data has derived from the left branch only.

The confusion matrices showed in Figure 8.1 reveal a tendency of the model to missclassify normal (negative PCa) cases as cancerous more frequently. This observation was also a driving factor for our experiments with the UC1-T2w-LeftBranchRaw dataset. Removing cases that were initially labeled as MRI-positive helps the model make better predictions. The challenging cases, which even confused clinicians, seem to have features that make classification difficult, leading to more errors.



(a) *UC1-T2w* dataset

(b) *UC1-T2w-LeftBranchRaw* dataset

Figure 5.1: Confusion matrices for our best performing model trained on both branches and tested on prospective data (left) and trained on left branch and tested on the same prospective data (right)

5.4 Discussion

When we tested our models on new unseen data, we notice a decrease in their performance for both scenarios. The performance dropped by 2% in the first scenario and by a notable 10% in the second. This significant drop for the second scenario is mainly down to the fact that prospective data originate from the right branch of the flowchart presented in Deliverable 5.3. Essentially, in our second scenario, we're training on the left branch and testing on the right. Results in both deliverables 5.3 and 6.1 indicated that these cases are quite challenging.

Chapter 6

Prospective Validation of Deep Learning Master Models (Experiments Set 3)

6.1 Chapter Summary

In this chapter, among the three models that FPO presented in Deliverable 5.3 to detect and segment prostate cancer (PCa) using multiparametric input, FPO selected the one that reached the highest performances in terms of detection rate (DR) and true negative rate (TNR). The selected model was the one that used a Unet in which the encoder was replaced with a Resnet50 and that received as input a 3-channel 2D image in which T2w, ADC map, and DWI images were concatenated. Since no tumoral masks were provided for prospective data, we were able to assess only detection performances.

6.2 Methods

Prospective Data Description

The dataset consisted of T2W, DWI, and ADC exams labeled as prospective in the multi-center ProstateNet image archive created under the scope of the ProCAncer-I project. A total of 269 cases were used (169 positive and 100 negative patients). Figure 6.1 shows the distribution of positive and negative patients across centers and vendors.

Pre-processing

Before feeding the networks, some pre-processing steps were applied. First, in case T2w and hbDWI/ADC didn't have the same slice thickness, they were co-registered with the T2w image, using an elastic transformation and the mutual information as metric. Then, all sequences were cropped and resampled in order to have the same resolution and field of view (FOV), and the N4 bias correction filter was applied to the T2w image to correct inhomogeneities due to the coil. Finally, a *in-house* developed algorithm to automatically segment the prostate was applied and each sequence was cropped around the automatically segmented prostate area using a bounding box of 224x224 pixels to ease the network training and reduce the computational cost. Then, a pixel standardization using the z-score technique was applied at the patient level. Pixel intensities values were rescaled between 0 and 1, and all voxels outside the prostate area were set to 0. Finally, 2D slices were transformed into RGB images in which each RGB channel is represented by a different sequence (T2w, ADC, and hbDWI). Once the output images were generated, a binary threshold filter was applied to the probability maps returned by the networks to obtain the automatic masks of the tumors. Then, connected areas smaller than 50 voxels were discarded.

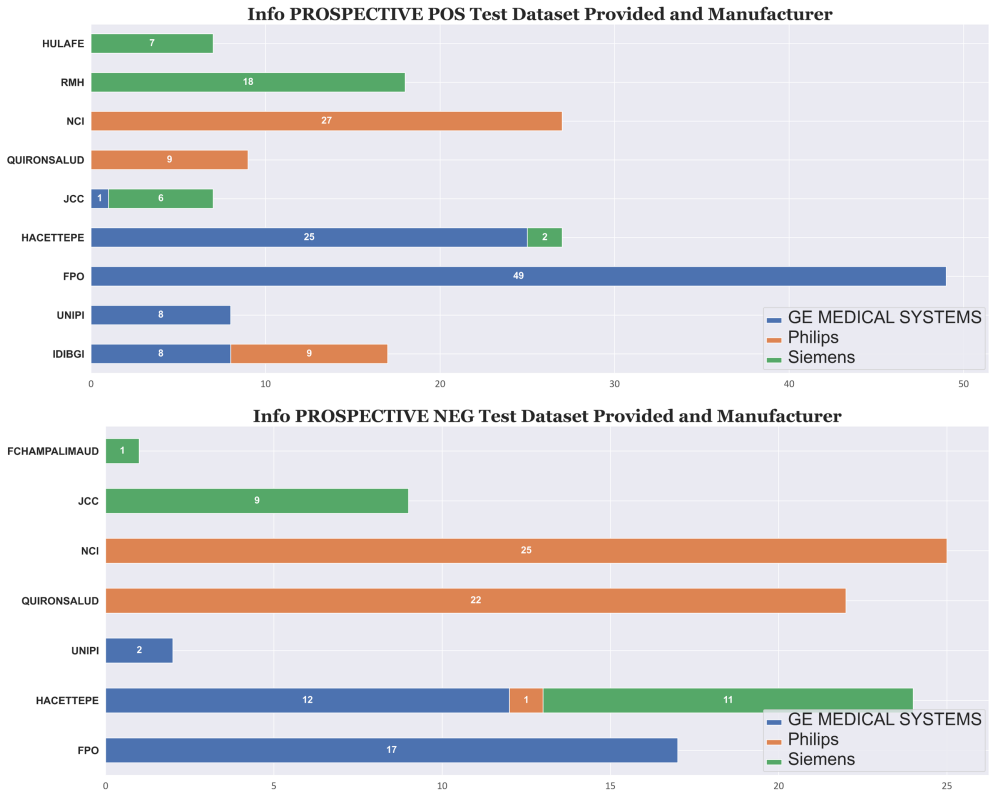


Figure 6.1: Distribution of positive and negative patients across centers and vendors

6.3 Results

On positive patients, the model reached good performances in detecting PCas, reaching a DR of 76% (129/169). Viceversa, considering negative patients, the model didn't reach high accuracy in detecting negative patients, showing a TNR of 40%, i.e., 40 out of 100 patients were correctly considered as negative. Table 6.1 and 6.2 show the performances of the master model stratified per vendors, respectively on positive and negative patients.

Vendor	Detection rate (%)	Detection rate (rate)
Master model on GE	86	79/91
Master model on Philips	60	27/45
Master model on Siemens	70	23/33

Table 6.1: Results of the master model on positive cases stratified per vendor

Vendor	True Negative Rate (%)	True Negative Rate (rate)
Master model on GE	29	9/31
Master model on Philips	40	19/48
Master model on Siemens	57	12/21

Table 6.2: Results of the master model on negative cases stratified per vendor

6.4 Discussion

The prospective validation of the master models developed in Deliverable 5.3 for the detection of PCa gave us a hint about the performance that could be obtained in clinical practice. Results obtained in terms of detection rate were promising, however, differences between vendors were observed: results obtained on GE data were statistically higher than those obtained on both Philips ($p \leq 0.001$). This might be due to the fact that most Philips cases did not have the same slice thickness between T2w and ADC/hbDWI images, therefore image registration might have introduced some biases. On the other side, we observed that negative patients were often misclassified as positive, meaning that the network is not able to discard suspected areas mimicking tumoral regions. This can be due either to the fact that no negative patients were used to train the network or to the low number of patients with a manually segmented mask available for training.
These results prompted us to start developing a post-processing method to discard false positive candidates and increase TNR (ongoing research).

Chapter 7

Prospective Validation of Deep Learning Master Models (Experiments Set 4)

7.1 Chapter Summary

Within ProCAncer-I, *master models* are defined as artificial intelligence (AI) algorithms trained on all available prostate MRI data, regardless of their use cases, vendors, and parent centers. While *prospective validation* is defined as testing any AI model on held-out contemporary data. In this chapter, and very early on in the Procancer-I project, partner Radboudumc explored an alternative strategy for the prospective validation of deep learning master models, particularly for UC1 and UC2. Partner Radboudumc has been active in the prostate MRI field for decades and used its infrastructure and data as leverage in the Procancer-i project. As part of performing its WP6 tasks in the Procancer-i project it hypothesized that state-of-the-art AI models, trained using thousands of patient cases, are non-inferior to radiologists at clinically significant prostate cancer (csPCa; ISUP $\geq$ 2) detection using MRI. To test this extended prospective validation hypothesis, Radboudumc designed an international comparative study: the PI-CAI (Prostate Imaging—Cancer Artificial Intelligence) challenge (`https://pi-cai.grand-challenge.org/`). Partner RADBOUDUMC was ideally positioned to swiftly start this type of study early in the Procancer-i project, because of their AI experimentation platform (Grand-Challenge) and organized data availability. Radboudumc's data plus newer data has now all been contributed to Procancer-I. The AI experimentation platform is now also incorporated in Procancer-i to help expedite its WP7 Clinical Validation. In the PI-CAI study, we investigated AI systems (*master models*) that were independently developed, trained, tuned, and tested at detecting ISUP $\geq$ 2 cancers using a large multi-center cohort of 10K patient examinations, in comparison to international radiologists participating in a multi-reader, multi-case observer study. This is inline with the UC1 and UC2 use cases as defined in the Procancer-i project. A key Grand-Challenge concept allowed PI-CAI to use secret, sequestered test data to validate internationally contributed AI models. The same test data was used to run reader experiments with radiologists. This setting allows a prospective validation and ranking of AI as well as a comparison to clinical performance. The PI-CAI challenge is visible as a work that is partly supported by Procancer-i. The PI-CAI models are contributed to the Procancer-i model repository. PI-CAI will remain visible and contribute to the sustainability of Procancer-i work. Our preliminary findings from this study have been published as a peer-reviewed paper at the `Medical Imaging in Deep Learning Conference` [26] and have been presented in oral presentations at the `108th Scientific Assembly and Annual Meeting of the Radiological Society of North America (RSNA 2022)`, `2023 European Congress of Radiology (ECR 2023)`, `38th Annual Congress of the European Association of Urology (EAU 23)` and at the `Annual Meeting of the Society for Advanced Body Imaging (SABI 2023)`.

7.2 Methods

The PI-CAI study protocol was established with 16 experts across prostate radiology, urology, and AI [29]. This retrospective study included 10207 prostate MRI exams (9129 patients) curated from four European tertiary care centers based in the Netherlands and Norway between 2012 and 2021. All patients were men suspected of harboring prostate cancer without a history of treatment or prior csPCa findings. Imaging was acquired using commercial 1.5 or 3T MRI scanners equipped with surface coils. In the first phase of this study, algorithm developers worldwide were invited to design AI models for detecting csPCa in biparametric MRI (bpMRI), using 1500 training cases that were made publicly available. For a given bpMRI exam, AI models were required to complete two tasks: localize all csPCa lesions (if any) and predict the case-level likelihood of csPCa diagnosis. To this end, AI models could use imaging data and several variables (PSA, patient age, prostate volume, scanner model) to inform their predictions. Once developed, these algorithms were independently tested using a hidden cohort of 1000 patient cases (including external data from an unseen center) in a fully blinded setting, where histopathology and a follow-up period of ≥ 3 years were used to establish the reference standard. For more details, please refer to [29].

7.3 Results

Between June and November 2022, >830 AI developers (>50 countries) opted-in and >310 algorithm submissions were made. Parallel to this, 79 radiologists (55 centers, 22 countries) were enlisted in a multi-reader, multi-case observer study, whose primary objective was to estimate clinicians' performance at this same task. The distribution of AI developers and radiologists has been illustrated in Fig. 7.1. When trained on 1500 cases, the top five most performant prostate-AI models reached 0.88±0.01 AUROC in case-level diagnosis and 76.38±0.74% sensitivity at 0.5 false positives per case in lesion detection (as shown in Table 7.1), which is comparable to that of radiologists' performance reported in the literature. When ensembled with equal weighting, diagnostic performance increased substantially to 0.912 AUROC, indicating notable diversity among the top five methods.

Model	AUROC	AP	Sens @ 0.5 FP/Patient
Y. Yuan et al. (Australia)	0.881	0.633	77.64%
C. A. Nader et al. (France)	0.889	0.615	76.63%
A. Karagöz et al. (Turkey)	0.889	0.614	75.38%
X. Li, S. Vesal, S. Saunders et al. (USA)	0.871	0.612	76.13%
H. Kan et al. (China)	0.886	0.593	76.13%
Ensemble of Top Five Models (Global)	0.912	–	–

Table 7.1: Case-level diagnostic performance, as estimated by the *Area Under Receiver Operating Characteristic* (AUROC) metric, and lesion-level detection performance, as estimated by the *Average Precision* (AP) and the detection sensitivity at 0.5 false positives per patient metrics, across 1000 testing cases.

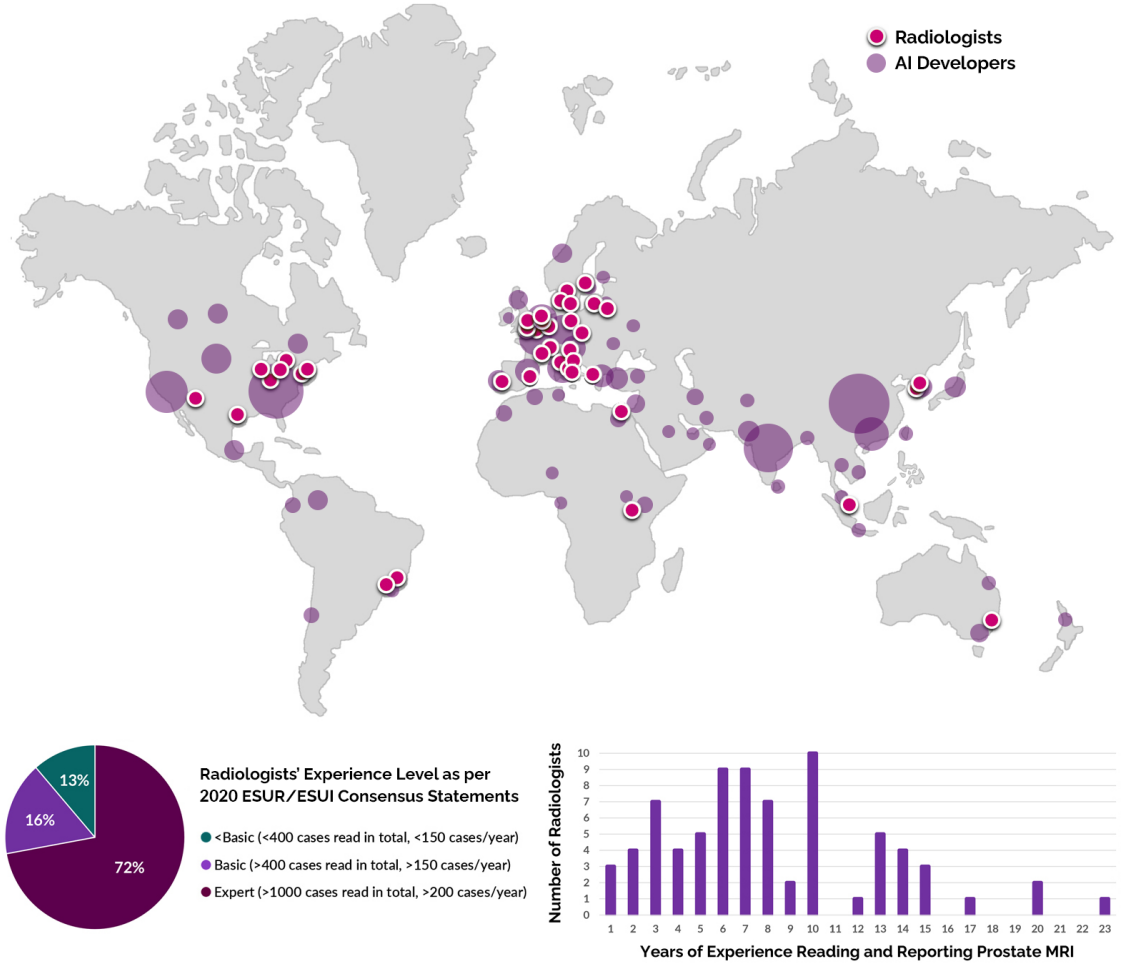


Figure 7.1: Distribution of >830 AI developers (>50 countries) and 79 radiologists (55 centers, 22 countries) participating in the PI-CAI challenge as of 10 November, 2022. Radiologists' experience varies between 1 and 23 years (median seven years), where 72% (57) of readers can be categorized as "expert" based on the 2020 ESUR/ESUI consensus statements [5].

7.4 Discussion

We observed that well-engineered deep learning systems, trained using a curated set of 1500 cases with a strong reference standard, can match or potentially outperform radiologists at detecting ISUP ≥ 2 cancers in prostate MRI (supported by the fact that five independent teams of developers around the world were able to reach similar levels of high diagnostic performance as reported in literature). In the next phase of this study, these AI models will be re-trained using a private dataset of 9107 cases, performance will be re-evaluated across 1000 testing cases, and the ensembled AI system will be benchmarked against the radiologists participating in the reader study. The historical reads made during routine practice to conclude non-inferiority or superiority (if applicable).

Chapter 8

Vendor Specific Segmentation Models (Experiments Set 1)

8.1 Chapter Summary

In this chapter, FCHAMPALIMAUD expand the analysis presented in D5.3, by training the same whole prostate and lesion segmentation nnUNet models on different subsets of data stratified by manufacturer (vendor) — GE, Philips and Siemens. We perform both a CV and hold-out test set analysis of these models, and compare the results between manufacturer, and to those obtained by the master models outlier in deliverable 5.3.

8.2 Methods

Data Description

We used the same set of retrospective data used in deliverable 5.3, which consists of T2W cases downloaded from the ProstateNet platform on March 13th 2023. An overview of the data, stratified by manufacturer, can be seen in Tab. 8.1. Since during the development of the D5.3 master models the data was not stratified by vendor, guaranteeing equal proportions of vendor-specific data throughout the different partitions, we define new CV folds and hold-out test partitions following the same procedures defined in D5.3. This introduces a minor caveat since $1:1$ comparisons between master models and vendor-specific models may not be entirely comparable.

Target	Total	Siemens	Philips	GE
Whole prostate gland	638	152	245	239
Lesions	461	136	184	136

Table 8.1: Stratification of the ProstateNet samples by target (whole prostate gland and lesions) and manufacturer for all segmentation tasks. Whole gland masks are generated by merging both Peripheral (PZ) and Transitional+Center (TZ) masks and share the same data composition.

Deep learning model specification

We train 3D full-resolution nnUNet models [12] (nnUNet) for each vendor (GE, Philips and Siemens). This framework makes use of stochastic gradient descent with Nesterov momentum ($\mu = 0.99$), a maximum initial learning rate of 0.01, and polynomial [4] learning rate decay policy which reduces the learning rate by a factor of $(1 - \frac{epoch}{epoch_{max}})^{0.9}$ on each epoch. The objective function combines a cross-entropy loss and a generalized Dice loss function.

Model evaluation

The cross-validation performance of each model was summarised as the average validation Dice score calculated using the parameters corresponding to the highest Dice score observed during training. The hold-out test set performance was calculated as the Dice score (DS), Hausdorff Distance (HD), Average Symmetric Surface distance (ASSD), and Relative Absolute Volume Difference (RAVD) [43] (both distance metrics were calculated using MedPy [18]).

8.3 Results

Cross-validation results

As shown in Tab. 8.2, two distinct scenarios can be highlighted for the cross-validation results: For whole gland segmentation, we can see that the performance is very similar across all metrics, between all manufacturers, and that these are also similar to the ones obtained by the master models detailed in D5.3; On the other hand, the lesion segmentation models show large differences between manufacturer, with the Siemens model clearly outperforming the other two, and being the only model producing results similar to those obtained by the master models detailed in D5.3. This is particularly interesting as all three vendor-specific data subsets have similar sizes (as highlighted in Table 8.1).

	Dice	HD	RAVD	ASSD	Recall
			Gland		
Siemens	0.9 ± 0.01	12.37 ± 1.49	0.04 ± 0.02	0.52 ± 0.11	1.0 ± 0.0
Philips	0.9 ± 0.01	12.43 ± 1.13	0.04 ± 0.02	0.47 ± 0.07	1.0 ± 0.0
GE	0.91 ± 0.01	11.23 ± 1.33	0.21 ± 0.21	0.42 ± 0.03	1.0 ± 0.0
			Lesions		
Siemens	0.36 ± 0.03	74.32 ± 6.17	0.08 ± 0.11	15.11 ± 2.4	0.7 ± 0.04
Philips	0.24 ± 0.02	64.66 ± 5.32	0.82 ± 0.37	22.62 ± 3.37	0.5 ± 0.04
GE	0.29 ± 0.03	74.39 ± 6.7	1.17 ± 0.45	17.92 ± 2.67	0.58 ± 0.05

Table 8.2: nnUNet CV results for all segmentation tasks, stratified by manufacturer. For each dataset, the average Dice, Hausdorf, RAVD, ASSD and Recall performance, along with their respective standard deviation, are presented.

Hold-out test set results

To analyse the hold-out test set results and assess the generalization across datasets and models, models trained on specific vendors were tested on data from other manufacturers as this allows us to assess how feasible is the transference of these models to concrete instances and applications where there may exist relevant alterations to image acquisition.

For whole gland segmentation models (Tab. 8.3), all manufacturer models show the same generalization degree to Siemens and Philips data, whilst GE shows a 2/3% improvement over the others on GE data. It can also be noted that these results match those of the ProstateNet master models detailed in deliverable 5.3, which highlights how this task may be relatively simple from a modelling perspective and unaffected by shifts to the domain. The consistently high recall (1.0) highlights that these models capture a relevant portion of the prostate under all scenarios.

Regarding the lesion segmentation models (Tab. 8.4), some interesting aspects can be outlined. Similar to what was observed during cross-validation, Siemens models are the best performing ones. However, it is on Philips data that models tend to perform the best on average — this includes GE models, which perform relatively poorly on GE data but in similar ways to other models on Philips data. When evaluating the lesion detection capabilities (Recall) of each models, we can see that the Siemens model produces very good results, in particular when tested on in-distribution data (on Siemens data). It should be highlighted here that there is a distinctive difference between data acquired using GE scanners and data acquired with Philips or Siemens scanners — the former has some instances where data was acquired with the use of an endorectal coil. As shown in chapter 10, this has a significant impact on the performance of classification models and

it is likely that this equally impacts the performance of models trained for lesion segmentation/detection tasks.

To compare results with those obtained for the model trained on the entire dataset, assessed how models compare to one another (Table 8.5 with Table 8.6, the last of which includes results already reported in D5.3). It becomes relatively clear that Dice scores have a wide overlap between models trained on specific subsets and those trained on the larger ProstateNet dataset (it should still be highlighted that Dice scores are consistently lower for vendor-specific models). When considering the recall (i.e. performance as lesion detection model), performance suffers considerable alterations — indeed, Siemens models appear to retain their performance when compared with models trained on the entire ProstateNet dataset, but the recall for GE and Philips models is considerably lower than that of the ProstateNet-trained models.

Trained on		Tested on: Siemens	Tested on: Philips	Tested on: GE	
Siemens		0.92 ± 0.03	0.91 ± 0.03	0.88 ± 0.09	Dice
		8.81 ± 5.8	21.59 ± 43.48	17.87 ± 45.61	HD
		0.02 ± 0.09	0.01 ± 0.07	0.01 ± 0.12	RAVD
		0.32 ± 0.11	0.81 ± 2.38	0.62 ± 0.64	ASSD
		1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	Recall
Philips		0.92 ± 0.03	0.92 ± 0.02	0.89 ± 0.1	
		8.28 ± 5.38	14.12 ± 24.47	18.21 ± 45.59	
		0.01 ± 0.08	0.01 ± 0.07	-0.01 ± 0.11	
		0.33 ± 0.11	0.36 ± 0.31	0.6 ± 0.65	
		1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	
GE		0.92 ± 0.03	0.92 ± 0.03	0.91 ± 0.09	
		9.08 ± 5.56	17.39 ± 27.73	17.12 ± 46.21	
		-0.0 ± 0.08	-0.01 ± 0.07	0.0 ± 0.09	
		0.33 ± 0.1	0.45 ± 0.49	0.52 ± 0.65	
		1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	

Table 8.3: nnUNet whole gland segmentation hold-out test set results. For each pairwise evaluation, the average Dice, Hausdorf, RAVD, ASSD and Recall performance, along with their respective standard deviations, are presented.

Trained on		Tested on: Siemens	Tested on: Philips	Tested on: GE	
Siemens		0.33 ± 0.28	0.33 ± 0.31	0.19 ± 0.26	Dice
		40.84 ± 50.38	79.66 ± 69.08	63.53 ± 71.39	HD
		-0.3 ± 0.42	-0.14 ± 1.28	0.21 ± 2.17	RAVD
		11.23 ± 32.49	28.32 ± 44.85	20.31 ± 34.75	ASSD
		0.7 ± 0.1	0.63 ± 0.09	0.4 ± 0.11	Recall
Philips		0.15 ± 0.22	0.34 ± 0.33	0.23 ± 0.3	
		34.68 ± 51.57	47.94 ± 51.18	26.72 ± 49.08	
		-0.37 ± 0.4	-0.04 ± 1.15	0.05 ± 0.92	
		5.78 ± 10.95	12.45 ± 27.82	3.42 ± 8.13	
		0.4 ± 0.11	0.56 ± 0.1	0.4 ± 0.11	
GE		0.22 ± 0.29	0.29 ± 0.31	0.21 ± 0.28	
		45.14 ± 65.89	49.02 ± 53.59	51.54 ± 59.9	
		-0.31 ± 0.36	0.09 ± 1.97	-0.26 ± 0.71	
		8.7 ± 15.1	12.53 ± 31.32	7.56 ± 9.24	
		0.4 ± 0.11	0.56 ± 0.1	0.45 ± 0.11	

Table 8.4: nnUNet lesion segmentation hold-out test set results stratified by vendor. For each pairwise evaluation, the average Dice, Hausdorf, RAVD, ASSD and Recall performance, along with their respective standard deviation, are presented.

Vendor	Dice	HD	RAVD	ASSD	Recall
GE	0.25 ± 0.29	48.61 ± 59.42	-0.14 ± 1.33	9.90 ± 22.22	0.48 ± 0.17
Philips	0.25 ± 0.30	37.65 ± 51.47	-0.11 ± 0.93	7.76 ± 19.57	0.46 ± 0.17
Siemens	0.29 ± 0.30	63.26 ± 66.79	-0.08 ± 1.47	20.83 ± 39.20	0.58 ± 0.17

Table 8.5: nnUNet lesion segmentation hold-out test set results for models trained on specific vendors and tested on the entire ProstateNet dataset. For each pairwise evaluation, the average Dice, Hausdorf, RAVD, ASSD and Recall performance, along with their respective standard deviation, are presented.

Dice	HD	RAVD	ASSD	Recall
0.33 ± 0.3	54.52 ± 62.22	0.26 ± 1.59	14.32 ± 35.78	0.62 ± 0.06

Table 8.6: nnUNet lesion segmentation hold-out test set results for models trained on the entire ProstateNet dataset. For each pairwise evaluation, the average Dice, Hausdorf, RAVD, ASSD and Recall performance, along with their respective standard deviation, are presented.

8.4 Discussion

These results highlight some important trends in model performance which are significantly different for whole gland segmentation models and for lesion segmentation models.

Whole gland segmentation models. Whole gland segmentation models show no relevant alterations when models are applied across datasets stratified by the vendor used to acquire each study. This entails a much easier deployment and ease of transfer to new settings and centres — the relatively high performance of these models leads to the ideal case scenario where little is necessary to transfer them to new circumstances.

Lesion segmentation models. On the other hand, lesion segmentation models have show significant drops in performance upon transferral to datasets acquired using scanners from vendors different from those used during training. This may lead to complications in the deployment of these models which should make end-users be relatively careful about their transferal — while Siemens models can be transfered to both Siemens and Philips data, their transferal to GE data can lead to relatively poor performance. This is particularly true in scenarios where such data has been partly acquired using endorectal coils. For this reason, the benefit of training these models using a dataset comprising multiple vendors and scenarios (as is the case of ProstateNet) becomes much more evident.

Limitations. While the comparisons drawn between models trained on specific subsets of data and models trained on the entire ProstateNet data are relevant, those drawn here can lead to difference estimates which may not be the most correct — indeed, the subsets used in each case are relatively different as testing sets are different between both cases as new CV folds and hold-out test sets had to be calculated differently to ensure equal vendor representation for D6.1.

Chapter 9

Vendor Specific Radiomics Models (Experiments Set 1)

9.1 Chapter Summary

Here, FCHAMPALIMAUD extended the radiomics work developed by FCHAMPALIMAUD in D5.3 to develop models specific to each vendor. More concretely, we apply a consistent machine-learning pipeline to use cases 2, 3, 5, 7, 7b and 8 after stratifying by vendor/manufacturer considering Siemens, Philips and GE Healthcare. Furthermore, given that a persistent endorectal coil has been identified, we also stratify by endorectal coil. We use this work to highlight performance differences between each model and study how the inclusion/exclusion of different feature types (clinical/radiological, deep features, radiomics features) can alter performance.

9.2 Methods

Data Description

Our dataset consisted of T2W, DWI and ADC exams from the ProstateNet image archive created under the scope of the ProCAncer-I project. The exams were acquired in the initial stages of the disease continuum by 13 different clinical partners, 3 scanner manufacturers and 27 scanner models. Ethics committee approval and patient consent were obtained by each clinical partner.

Segmentation

Automatic segmentation of the whole prostate gland was performed on T2W sequences using a segmentation model developed in-house. The full details of this model are shown in chapter 2.

The generated masks were post-processed in two stages. Firstly, the largest object was selected. An object was defined as a group of connected voxels. Here, it was assumed that the largest object would have the highest probability of covering the actual gland. Secondly, so as to smooth mask borders, a Delaunay triangulation was calculated on the convex hull of the selected object.

Sequence Co-registration

Due to the absence of segmentation masks for the diffusion sequences, T2W sequences (moving image) were co-registered to the DWI sequences' space (fixed image), and the calculated transformation matrix was then applied to the segmentation mask generated previously. The co-registration algorithm was a 3-resolution pyramid of rigid registrations. The transformed mask was then used for the radiomics extraction of the diffusion sequences. The co-registration parameters file can be found in Table 9.1. For wide field-of-view DWI sequences, a center crop was applied to facilitate the co-registration.

```
// Components
(Registration "MultiResolutionRegistration")
(FixedImagePyramid "FixedRecursiveImagePyramid")
(MovingImagePyramid "MovingRecursiveImagePyramid")
(Interpolator "LinearInterpolator")
(Metric "AdvancedMattesMutualInformation")
(Optimizer "AdaptiveStochasticGradientDescent")
(ResampleInterpolator "FinalBSplineInterpolator")
(Resampler "DefaultResampler")
(Transform "EulerTransform")

// **********Pyramid
(NumberOfResolutions 3)

// **********Transform
(AutomaticTransformInitializationMethod "GeometricCenter")
(AutomaticScalesEstimation "true")

// **********Optimizer
(MaximumNumberOfIterations 300)
(AutomaticParameterEstimation "true")

// **********Several
(WriteTransformParametersEachIteration "false")
(WriteTransformParametersEachResolution "false")
(WriteIterationInfo "false")
(WriteResultImage "true")
(ShowExactMetricValue "false")
(ResultImageFormat "nii")

// **********ImageSampler
(ImageSampler "RandomCoordinate")
(CheckNumberOfSamples "true")
(NewSamplesEveryIteration "true")
(MaximumNumberOfSamplingAttempts 8)
(NumberOfSpatialSamples 2048)
(NumberOfSamplesForExactGradient 4096)

// **********Interpolator and Resampler
// Order of B-Spline interpolation used for applying the final deformation:
(FinalBSplineInterpolationOrder 3)

// Default pixel value for pixels that come from outside the picture:
(DefaultPixelValue 0)
```

Table 9.1: Co-registration parameter file.

Radiomic Features Extraction

Bias field correction was performed on T2W sequences using the N4 Bias Field Correction algorithm [36] and the Python package Simple ITK (version 2.0.0) [42]. First, each image's x-, y- and z-spacing were checked for discrepancies. Since x- and y-spacings differed from z-spacing, feature extraction was later performed in 2D. Additionally, images' x- and y-spacings differed within and between patients, so T2W sequences were resampled to the 95th quantile value of 0.6875, and DWI and ADC were resampled to the 95th quantile value of 2.0. Image intensities were normalized. The bin width was selected for each image filter to produce discretized images with between 30 and 130 bins. The full description of extraction parameters for each modality can be found in Table 9.2.

Radiomic features were extracted from the whole gland segmentation using the Pyradiomics package (version 3.0) [38] in Python (version 3.7.9) [39]. All the pre-processing steps mentioned before were performed as parameters of the extractor function, except for the bias field correction, which was performed prior to the extraction. All image filters and feature classes were enabled, resulting in a total of 1223 features calculated per sequence. The mathematical expressions and semantic meanings of the features extracted can be found at https://pyradiomics.readthedocs.io/en/latest/.

T2W extraction parameters	DWI extraction parameters	ADC extraction parameters
imageType:	imageType:	imageType:
Original:	Original:	Original:
binWidth: 5	binWidth: 12	binWidth: 5
Wavelet:	Wavelet:	Wavelet:
binWidth: 3	binWidth: 8	binWidth: 4
Square:	Square:	Square:
binWidth: 3	binWidth: 8	binWidth: 3
SquareRoot:	SquareRoot:	SquareRoot:
binWidth: 8	binWidth: 16	binWidth: 8
Logarithm:	Logarithm:	Logarithm:
binWidth: 16	binWidth: 25	binWidth: 12
Exponential:	Exponential:	Exponential:
binWidth: 0.5	binWidth: 3	binWidth: 1
Gradient:	Gradient:	Gradient:
binWidth: 5	binWidth: 4	binWidth: 3
LBP2D:	LBP2D:	LBP2D:
binWidth: 0.1	binWidth: 0.1	binWidth: 0.1
LoG: 'sigma' : [1.0, 3.0]	LoG: 'sigma' : [1.0, 3.0]	LoG: 'sigma' : [1.0, 3.0]
featureClass:	featureClass:	featureClass:
firstorder:	firstorder:	firstorder:
glcm:	glcm:	glcm:
glrlm:	glrlm:	glrlm:
glszm:	glszm:	glszm:
gldm:	gldm:	gldm:
ngtdm:	ngtdm:	ngtdm:
shape:	shape:	shape:
setting:	setting:	setting:
binWidth: 5	binWidth: 5	binWidth: 5
normalize: True	normalize: True	normalize: True
normalizeScale: 100	normalizeScale: 100	normalizeScale: 100
force2D: True	force2D: True	force2D: True
voxelArrayShift: 300	voxelArrayShift: 300	voxelArrayShift: 300
resampledPixelSpacing:	resampledPixelSpacing:	resampledPixelSpacing:
[0.6875, 0.6875, 0]	[2, 2, 0]	[2, 2, 0]
geometryTolerance: 0.00001	geometryTolerance: 0.00001	geometryTolerance: 0.00001

Table 9.2: Radiomics Extraction parameters.

Deep Features

To generate deep features for each instance, we used the bottleneck of a U-Net model pre-trained on segmenting the whole prostate gland using T2W volumes. To calculate a segmentation prediction, the U-Net model first encodes the image into a low resolution volume with high semantic information (320 features in our case) and uses this information to obtain a segmentation map for a given object (whole prostate gland in our case). We encode each T2W volume and extract the maximum value of each feature, obtaining a 320-sized vector characterizing each image.

Clinical Features

The clinical variables included for each use case can be found in Table 9.3. Missing numerical values were imputed with a KNNImputer. Missing categorical values in the variables perineural_invasion, extra_prostatic_extension, seminal_vesical_invasion and resection_margin_status were set to "Not Assessed", while the remaining missing categorical values were imputed to the most frequent category.

For UC 5, two contexts were considered: presurgery and postsurgery. For the latter, the clinical variables included are the ones listed in Table 9.3. While, for the former, all information reported during or immediately after the surgery was removed, namely the variables prostatectomy_method, resection_margin_status, extraprostatic_extension, perineural_invasion, seminal_vesicle_invasion, previous_adenectomy and prostatectomy_nerve_sparing were excluded.

Use Cases	2, 3 and 8	5 (post-surgery) and 7b	5 (pre-surgery) and 6
Clinical variables	age_at_baseline (num) baseline_psa_total (num) index_lesion_pirads (cat) lesion_location_PZ (bool) lesion_location_TZ (bool) lesion_location_CZ (bool) lesion_location_AS (bool)	age_at_baseline (num) baseline_psa_total (num) index_lesion_pirads (cat) lesion_location_PZ (bool) lesion_location_TZ (bool) lesion_location_CZ (bool) lesion_location_AS (bool) prostatectomy_method (cat) resection_margin_status (cat) extraprostatic_extension (cat) perineural_invasion (cat) seminal_vesicle_invasion (cat) gleason1 (num) gleason2 (num) ISUP grade (num) previous_adenomectomy (bool) prostatectomy_nerve_sparing (bool)	age_at_baseline (num) baseline_psa_total (num) index_lesion_pirads (cat) lesion_location_PZ (bool) lesion_location_TZ (bool) lesion_location_CZ (bool) lesion_location_AS (bool) gleason1 (num) gleason2 (num) ISUP grade (num)

Table 9.3: Clinical variables included for each use case. "num" indicates a numerical variable; "bool" indicates a binary variable; "cat" indicates a categorical variable.

Dataset Construction

The train/test split was performed for the larger use cases at patient level with the Python scikit-learn package (version 0.23.2) [7]. The hold-out test sets consisted of 200 randomly selected patients for UC2 and 50 for UCs 5 and 7b. The split was stratified so that both train and test sets have the same label distribution. The train and test sets label distribution can be found in tables 9.4 and 9.5, for binary and multiclass tasks, respectively. For the smaller use cases, namely 3, 6 and 8, only the cross-validation performance is reported.

Use Cases	Target (binary)	Train Set		Test Set	
		0	1	0	1
2	ISUP 1 VS 2345	1360	3603	51	148
	ISUP 12 VS 345	3288	1675	141	58
	ISUP 123 VS 45	4145	818	167	32
3	no metastasis in 6 months VS metastasis developed	15	63	-	-
5	no biochemical recurrence after RP at follow-up VS biochemical recurrence	612	101	43	7
6	no biochemical recurrence after RT at follow-up VS biochemical recurrence	120	16	-	-
8	stayed in active surveillance VS left active surveillance	128	10	-	-

Table 9.4: Label distribution in the train and test sets for each binary classification problem.

Use Cases	Target (multiclass)	Train Set			Test Set		
		0	1	2	0	1	2
2	ISUP 123 VS 45	1360	2785	818	51	116	32
7b	epic 26 [0, 71] vs]71, 84] vs]84, 100]	71	75	62	14	20	15

Table 9.5: Label distribution in the train and test sets for each multiclass classification problem.

Different data subsets were tested for their training ability. Pure radiomics datasets were appended clinical and/or deep features and their performance was compared. Training with the full dataset was compared to training with patients whose exams had been taken in each scanner manufacturer independently, with the exception of TOSHIBA, whose patients were removed due to low representability. The exclusion of patients where an endorectal coil had been used was also tested. And, finally, we compared training with the full MRI sequence set to training with each sequence independently. From the final 64 training combinations, the subsets with less than 30 cases were discarded. The discriminated data sizes of the training set are shown in Tables 9.6 - 9.11 for each use case.

UC 2 - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
scanner	SIEMENS	2180	2086	2077	2077
	PHILIPS	1399	1374	1365	1365
	GE	1404	1145	665	665
	GE_noERC	906	659	457	457

Table 9.6: Discriminated data sizes of the training sets for UC2.

UC 3 - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
scanner	SIEMENS	20	18	18	18
	PHILIPS	23	22	22	22
	GE	19	10	7	7
	GE_noERC	16	7	7	7

Table 9.7: Discriminated data sizes of the training sets for UC3.

UC 5 - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
scanner	SIEMENS	182	173	173	173
	PHILIPS	429	425	420	420
	GE	105	85	64	64
	GE_noERC	65	45	30	30

Table 9.8: Discriminated data sizes of the training sets for UC5.

UC 6 - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
scanner	SIEMENS	-	-	-	-
	PHILIPS	48	48	48	48
	GE	65	64	27	27
	GE_noERC	-	-	-	-

Table 9.9: Discriminated data sizes of the training sets for UC6.

For UC 7b, there were no endorectal coil cases, so this setting was removed. Lastly, to minimize training time, a first initial evaluation of all MRI sequences was done for Radiomics only, as well as Radiomics + Clinical variables and, given the results consistently showed that DWI features provided the best outcome, all further models were trained only using DWI data, for a total of 8 models. The discriminated data sizes of the training set are shown in Table 9.10.

UC 7b - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
scanner	SIEMENS	53	51	51	51
	PHILIPS	179	179	175	175
	GE	2	2	2	2
	GE_noERC	-	-	-	-

Table 9.10: Discriminated data sizes of the training sets for UC7b.

UC 8 - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
scanner	SIEMENS	64	64	64	64
	PHILIPS	11	9	8	8
	GE	43	39	3	3
	GE_noERC	38	35	0	0

Table 9.11: Discriminated data sizes of the training sets for UC8.

Preprocessing Pipeline

All the steps described in this section were performed exclusively on the train set and only on the numerical variables. Features were scaled to have zero mean and standard deviation equal to 1 (Python package scikit-learn version 1.0.2). Features with low variance were identified and excluded. Here, a threshold of 0.01 was considered. Finally, feature correlation was assessed. Feature pairs were considered correlated if their Spearman correlation was higher than 0.8. Out of the two, the feature with the highest average correlation across all features was eliminated.

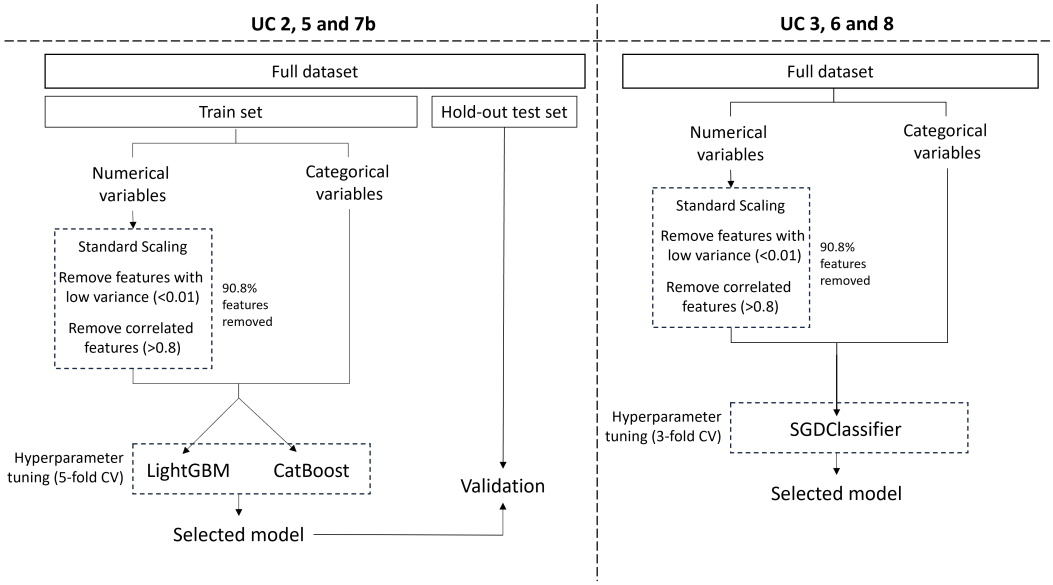


Figure 9.1: Radiomics model development pipeline.

Training

For radiomics and raddeep models, a light gradient boosting machine (LGBM) [13] was trained, while for radclin or hybrid models, which may include categorical data, the CatBoost [24] algorithm was used. Regarding the smaller UCs, a support vector machine (SVM) classifier was selected for UC 7b and Stochastic gradient descent algorithm (SGD) was prefered for UCs 3, 6 and 8. Hyperparameter tuning was performed for each algorithm and each parameter combination was evaluated through cross-validation (5 folds for UCs 2 and 5; 3 folds for UCs 3, 6 and 8). For UCs 3, 6, 7b and 8 a random search approach was selected, as less data is available so a less biased optimization is preferred, while for UCs 2 and 5 tuning was performed with an exhaustive grid search. The overall pipeline can be found in Fig. 9.1 and the hyperparameter space used can be found in Table 9.12.

Model Post-processing

For all models developed, the ROC curve was analyzed and the probability decision threshold that resulted in the highest youden index was selected for the remaining analysis.

All final models were analyzed in two main areas: explainability and fairness.

Regarding model explainability, a SHapley Additive exPlanations (SHAP) analysis (Python package shap version 0.41.0) [17] was used to identify the most relevant variables for the prediction in the hold-out test set. The 20 most relevant variables for the output of each model were displayed. Each dot in the graph represents a feature's SHAP value for one observation in the hold- out test set. The SHAP value's position on the x-axis expresses whether it is associated with a positive or negative prediction. The red color indicates higher values of a feature and the blue color means lower value.

GridSearch (UCs 2, 3, 5, 6 and 8)

```
pipe = CatBoostClassifier(loss_function='Logloss',
                          eval_metric='AUC',
                          cat_features=cat,
                          random_seed=42,
                          logging_level='Silent')
param_grid = {'n_estimators': [100, 500, 1000], # estimators
              'learning_rate': [0.01, 0.03, 0.1], # Learning rate for gradient boosting
              'max_depth': [4, 6, 10]}
```

```
pipe = Pipeline([('classifier', CalibratedClassifierCV(LGBMClassifier(), method='isotonic'))])
param_grid = dict(classifier__base_estimator__n_estimators = [100, 500],
                  classifier__base_estimator__num_leaves = [5, 10, 30],
                  classifier__base_estimator__learning_rate = [0.01, 0.1],
                  classifier__base_estimator__subsample = [0.1, 0.3, 0.5, 0.75],
                  classifier__base_estimator__colsample_bytree = [0.1, 0.3, 0.5, 0.75])
```

RandomSearch (UC 7b)

```
pipe = Pipeline([('classifier', CatBoostClassifier(loss_function='MultiClass',
                                                   eval_metric='AUC',
                                                   cat_features=cat,
                                                   logging_level='Silent',
                                                   random_seed=seed))])
param_distributions = {'classifier__n_estimators': np.array([1000]),
                       'classifier__bootstrap_type': np.array(['Bayesian']),
                       'classifier__learning_rate': np.linspace(0.1, 0.9, num=10),
                       'classifier__learning_rate': np.linspace(0.001, 0.01, num=100),
                       'classifier__max_depth': np.array([4, 6, 8, 10]),
                       'classifier__l2_leaf_reg': np.array([1, 3, 5, 7, 9]),
                       'classifier__border_count': np.array([32, 64, 128]),
                       'classifier__bagging_temperature': np.linspace(0.5, 2, num=10),
                       'classifier__random_strength': np.linspace(0.5, 2, num=10)}
```

```
pipe = Pipeline([('classifier', LGBMClassifier(random_state=seed, metric='auc_mu'))])
param_distributions = {'classifier__n_estimators': np.array([1000]),
                       'classifier__boosting_type': np.array(['goss']),
                       'classifier__num_leaves': np.linspace(10, 100, num=10, dtype=int),
                       'classifier__learning_rate': np.linspace(0.001, 0.01, num=100),
                       'classifier__max_depth': np.array([4, 6, 8, 10]),
                       'classifier__min_child_samples': np.linspace(10, 50, num=5, dtype=int),
                       'classifier__subsample': np.linspace(0.5, 1.0, num=10),
                       'classifier__colsample_bytree': np.linspace(0.5, 1.0, num=10),
                       'classifier__reg_alpha': np.logspace(-3, 3, num=10),
                       'classifier__reg_lambda': np.logspace(-3, 3, num=10),
                       'classifier__min_split_gain': np.random.uniform(low=0, high=1, size=10),
                       'classifier__num_boost_round': np.linspace(100, 500, num=5, dtype=int),
                       'classifier__scale_pos_weight': np.linspace(1, 5, num=5, dtype=int)}
```

```
pipe = Pipeline([('classifier', SVC(random_state=seed, probability=True))])
param_distributions = {'classifier__C': np.logspace(-3, 3, num=10),
                       'classifier__kernel': np.array(['linear', 'poly', 'rbf', 'sigmoid']),
                       'classifier__degree': np.array([2, 3, 4]),
                       'classifier__gamma': np.logspace(-3, 3, num=10),
                       'classifier__coef0': np.linspace(0, 1, num=10),
                       'classifier__shrinking': [True, False],
                       'classifier__tol': np.logspace(-6, -2, num=10)}
```

Table 9.12: Hyperparameter space used for optimization.

In terms of fairness, model performance was tested for different subgroups of the data with the fairlearn python package. ROC-AUC, f2-score, precision and recall are reported for each subgroup, as well as subgroup size on the train and test sets and test set label distribution. For subgroups where only one target label is present the ROC-AUC metric is replaced with Accuracy.

9.3 Results

Use Case 2 - ISUP 1 vs 2,3,4,5

Model Performance

Fig. 9.2 shows the cross-validation and hold-out test set ROC-AUC model performance for the 64 models trained for UC2, in four spider plots colored according to MRI volumes included. At first glance, we can see that training with PHILIPS exams leads to the highest cross-validation and hold-out test set performances (Fig. 9.2). Furthermore, the inclusion of the three sequences, T2W, DWI and ADC, is beneficial for the model performance, both in the cross-validation and hold-out test set. The highest performing models for each manufacturer were the following: radclin_uc2_T2&DWI&ADC_SIEMENS_CatBoost (0.5 on cross-validation, 0.6453 on hold-out test set), radiomics_uc2_T2&DWI&ADC_PHILIPS_LGBM (0.6756 on cross-validation, 0.6894 on hold-out test set) and hybrid_uc2_DWI_GE_CatBoost (0.5019 on cross-validation, 0.6355 on hold-out test set).

Figure 9.2: Cross-validation ROC-AUC model performance of 64 models trained to predict disease aggressiveness in UC 2 (ISUP 1 vs 2345). The observations are color-coded according to MRI volumes used for training.

Sub-cohort Analysis of the Best Models

Figure 9.3 shows the subcohort analysis performed on the models selected for each scanner manufacturer in the previous section. We would expect that the model would achieve the highest performance on data from the same vendor that was used for training, however, this was not verified. The highest performances overall were achieved by the PHILIPS model on SIEMENS data (23 negative cases and 85 positive cases) and by the SIEMENS model on GE data (5 negative cases and 30 positive cases).

Figure 9.3: Performance of the models selected for each scanner on different subsets of the hold-out test set, divided by scanner manufacturer.

Use Case 2 - ISUP 1,2 vs 3,4,5

Model Performance

Fig. 9.4 shows the cross-validation and hold-out test set ROC-AUC model performance for the 64 models trained for UC2, in four spider plots colored according to MRI volumes included. At first glance, we can see that on all subplots, the outer rings represent the hold-out test set performance of models trained with DWI-only or all sequences. The highest performing models for each manufacturer were the following: radiomics_uc2_T2_SIEMENS_LGBM (0.5526 on cross-validation, 0.6716 on hold-out test set), hybrid_uc2_DWI_PHILIPS_CatBoost (0.6265 on cross-validation, 0.7215 on hold-out test set) and hybrid_uc2_T2_GE_noERC_CatBoost (0.6492 on cross-validation, 0.6661 on hold-out test set).

Sub-cohort analysis of the best models

Figure 9.5 shows the subcohort analysis performed on the models selected for each scanner manufacturer in the previous section. The SIEMENS model consistently showed the highest recall and F2 performance, across the different sub-cohorts of the test set, with the exception of the SIEMENS subcohort, where it, surprisingly, ranked in second place, after the model trained with all scanner manufacturers.

Use Case 2 - ISUP 1,2,3 vs 4,5

Model Performance

Fig. 9.6 shows the cross-validation and hold-out test set ROC-AUC model performance for the 64 models trained for UC2, in four spider plots colored according to the MRI volumes included in the training. At

Figure 9.4: Cross-validation ROC-AUC model performance of 64 models trained to predict disease aggressiveness in UC 2 (ISUP 12 vs 345). The observations are color-coded according to MRI volumes used for training.

first glance, we can see that on all subplots, the outer rings represent the hold-out test set performance of models trained with DWI-only or all sequences. The highest performing models for each manufacturer were the following: hybrid_uc2_DWI_SIEMENS_CatBoost (0.5945 on cross-validation, 0.7827 on hold-out test set), radiomics_uc2_T2&DWI&ADC_PHILIPS_LGBM (0.5619 on cross-validation, 0.7815 on hold-out test set) and hybrid_uc2_DWI_GE_CatBoost (0.5437 on cross-validation, 0.7427 on hold-out test set).

Sub-cohort analysis of the best models

Figure 9.7 shows the subcohort analysis performed on the models selected for each scanner manufacturer in the previous section.

Figure 9.5: Performance of the models selected for each scanner on different subsets of the hold-out test set, divided by scanner manufacturer.

Use Case 5 - Post-surgery

Model Performance

Fig. 9.8 shows the cross-validation and hold-out test set ROC-AUC model performance for the 64 models trained for UC5, in four spider plots colored according to the MRI volumes included in the training. The highest performing models for each manufacturer were the following: radclin_uc5_ADC_SIEMENS_CatBoost (0.5383 on cross-validation, 0.8188 on hold-out test set), hybrid_uc5_T2_PHILIPS_CatBoost (0.5170 on cross-validation, 0.6655 on hold-out test set) and radclin_uc5_DWI_GE_CatBoost (0.7976 on cross-validation, 0.6899 on hold-out test set).

Use Case 6

Model Performance

Fig. 9.9 shows the cross-validation and hold-out test set ROC-AUC model performance for the 64 models trained for UC5, in four spider plots colored according to the MRI volumes included in the training. The highest performing models for each manufacturer were the following: radclin_uc6_T2_PHILIPS_CatBoost (??? on cross-validation, 0.6275 on hold-out test set) and hybrid_uc6_T2_GE_CatBoost (0.8 on cross-validation, 0.8824 on hold-out test set).

Use Case 8

Model Performance

Fig. 9.10 shows the cross-validation and hold-out test set ROC-AUC model performance for the 64 models trained for UC5, in four spider plots colored according to the MRI volumes included in the training. The highest performing models for each manufacturer were the following: hybrid_uc8_ADC_SIEMENS_CatBoost (0.5 on cross-validation, 0.9474 on hold-out test set) and hybrid_uc8_T2_GE_CatBoost (??? on cross-validation, 0.5263 on hold-out test set).

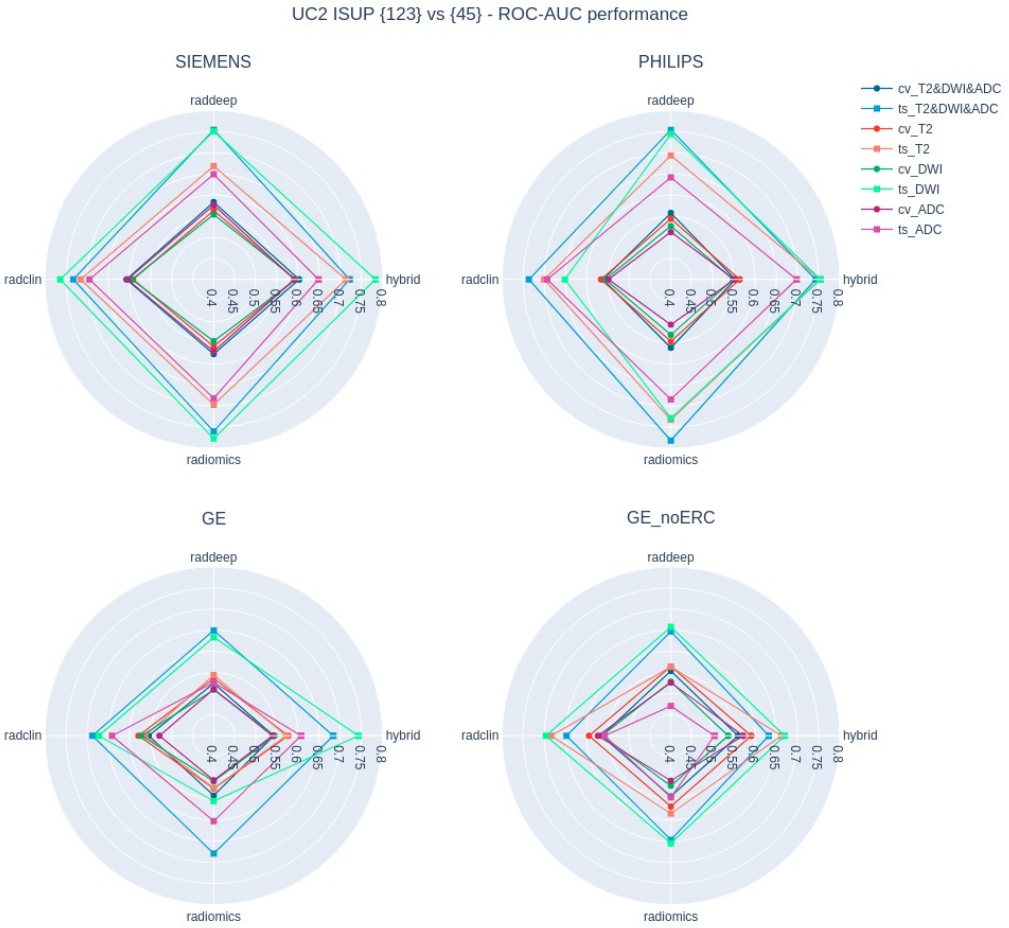


Figure 9.6: Cross-validation ROC-AUC model performance of 64 models trained to predict disease aggressiveness in UC 2 (ISUP 123 vs 45). The observations are color-coded according to MRI volumes used for training.

Figure 9.7: Performance of the models selected for each scanner on different subsets of the hold-out test set, divided by scanner manufacturer.

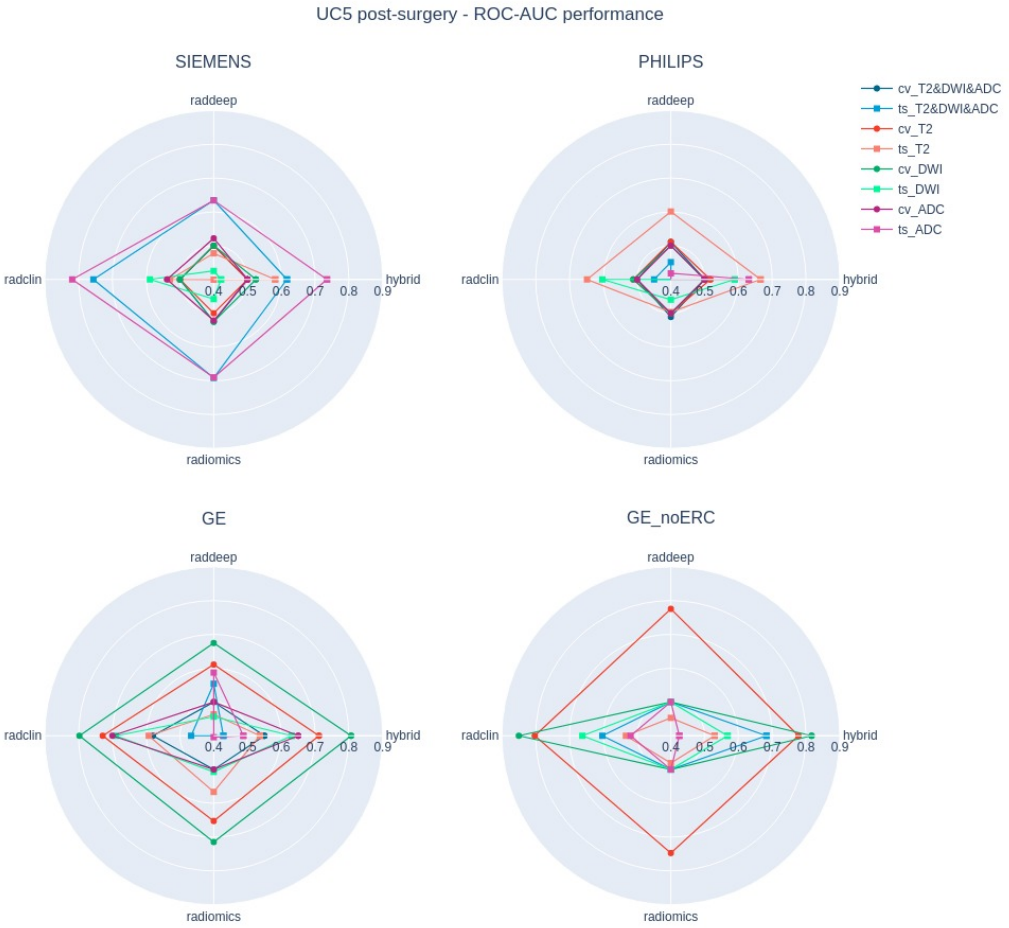


Figure 9.8: Cross-validation ROC-AUC model performance of 64 models trained to predict disease aggressiveness in UC 5 (post-surgery scenario). The observations are color-coded according to MRI volumes used for training.

Figure 9.9: Cross-validation ROC-AUC model performance of 64 models trained to predict disease aggressiveness in UC 6. The observations are color-coded according to MRI volumes used for training.

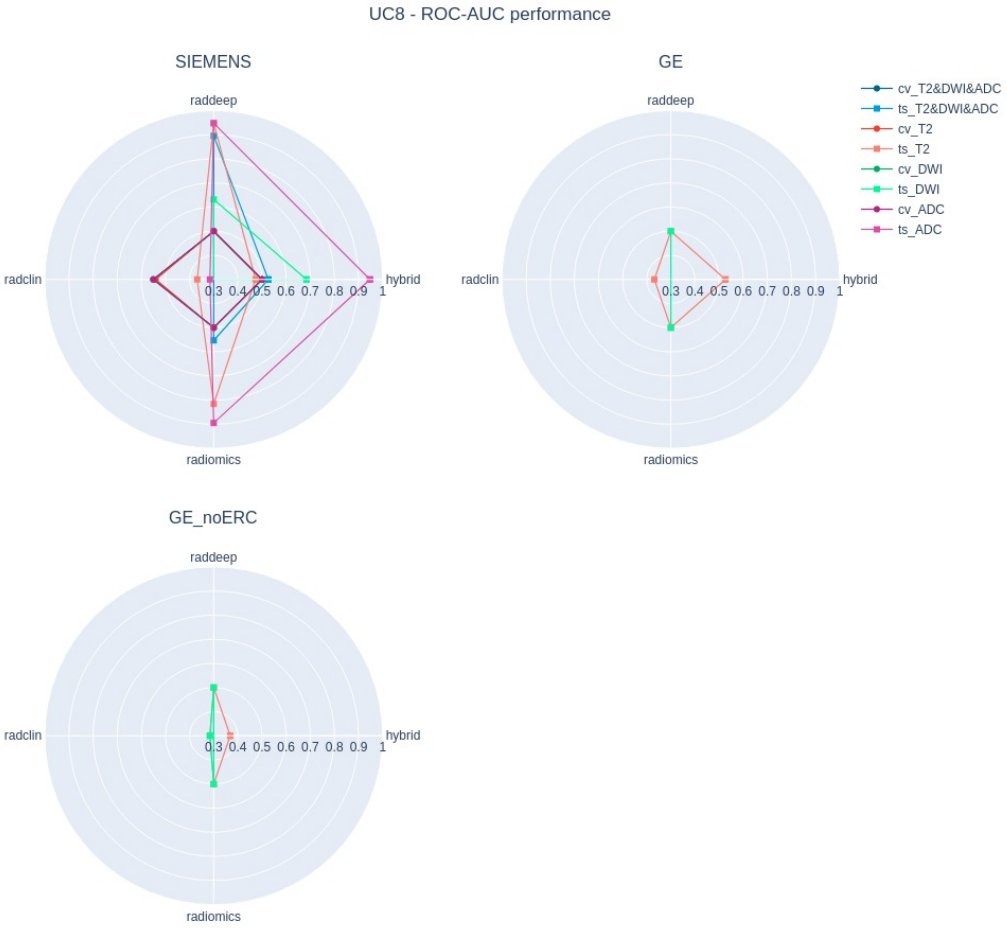


Figure 9.10: Cross-validation ROC-AUC model performance of 64 models trained to predict disease aggressiveness in UC 5 (post-surgery scenario). The observations are color-coded according to MRI volumes used for training.

9.4 Discussion

The development of vendor specific master models constitutes the second stage of model development, following the development of master models (first stage), within the ProCAncer-I project's body of work. The initial idea was to use the master models developed and fine-tune them for each scanner manufacturer (SIEMENS, PHILIPS and GE). Although this is possible in the deep learning arm of the project, in radiomics and classical machine learning, it is not as straightforward. For this reason, the vendor specific radiomics models were trained from scratch utilizing only exams from each vendor. This lead to the first issue encountered: data scarcity. As we divided the dataset by scanner manufaturer, we found that the smaller use cases did not have enough data for model development in every subgroup. This was especially evident in use case 3, where no cohort had 30 cases or more (Table 9.7). Similarly, for use case 6, only PHILIPS and GE models were trained (Table 9.9), and for use case 8, only SIEMENS and GE (Table 9.11).

Regarding UC 2, the different target definitions showed similar patterns in what type of data was the most useful to model it. Considering ISUP 1 vs 2345 and ISUP 12 vs 345, the PHILIPS models achieved the highest performance, both in cross-validation and hold-out test set. For the third label definition (ISUP 123 vs 45), the highest performances were achieved with both SIEMENS and PHILIPS models, showing equivalent hold-out test set performances and only slightly different cross-validation performances. Overall, despite not being the most prevalent in the overall dataset, PHILIPS data shows the highest overall generalization power both internally (in PHILIPS cases) and externally, to cases from other vendors. Furthermore, it is interesting to find that, for the first and third label distribution, the data type that achieved the highest performance was the same as in the master models (Table 3.1), radiomics data extracted from all three MRI volumes. Despite the high performance achieved by the PHILIPS models, it still did not surpass the performance of the corresponding master models on the same hold-out test set.

In use cases 5, 6 and 8, we see a much higher weight put on clinical variables than we did in use case 2, given that, for all scanner vendors, the selected models were trained with radclin or hybrid data.

Chapter 10

Vendor Specific Deep Learning Models (Experiments Set 1)

10.1 Chapter Summary

For this deliverable (D6.1), FCHAMPALIMAUD expanded the analysis presented in deliverable 5.3 (D5.3) and trained the same models on subsets of data stratified by manufacturer (vendor) — GE, Philips and Siemens — and further stratify by endorectal coil (ERC) use. More concretely, we study how training and testing on different vendors/manufacturers leads to differences in performance which may be problematic for model performance upon deployment in distinct medical settings. To better understand this, we use different target definitions (as in D5.3 — ISUP=1,2 vs. ISUP=3,4,5; ISUP=1 vs. ISUP=2,3,4,5; ISUP=2,3 vs. ISUP=4,5), contemplate different ways of including clinical variables, and offer additional visualizations and analyses in terms of domain shifts associated with manufacturer and endorectal coil use. Finally, we evaluate these models on prospective data.

10.2 Methods

For the sake of completeness and despite its similarities with D5.3, we detail the training and testing protocol used for this deliverable. Given that only GE scans from UNIPI and FPO used ERC, we finally define 5 relevant subsets — GE (ERC), GE (no ERC), Philips, Siemens and Full (which uses all data). We further test all models on the same subsets in the hold-out test set and prospective testing stages. We used the models trained on all scanners (Full) that were presented in D5.3.

Data Description

We used the retrospective cases available through ProstateNet until March 13th, 2023 (8,891 cases), of which 5,478 were specific for use case 2. Using an automated DICOM-to-NIFTI conversion pipeline, we obtained a total of 5,352 PCa studies with any relevant sequence. Of these, 4,975 had T2-weighted sequences (T2w), whereas 4,574 had all three sequences for multiparametric MRI (mpMRI) – T2w, diffusion weighted imaging sequences (DWI), and apparent diffusion coefficient sequences (ADC). Given that we are interested in assessing the impact of clinical data – prostate specific antigen (PSA) and age at baseline – we further calculate the amount of sequences with clinical data (4,764 studies with T2w and 4,380 complete mpMRI studies). Using the set of studies with all 4,380 mpMRI studies and clinical data, we constructed 5 non-overlapping validation folds using 85% of the data ($n = [741, 744, 747, 746, 745]$) and use the remaining 15% as a hold-out test set (n=657). Validation folds and the hold-out test set were obtained by considering ISUP scores (1, 2, 3, 4, 5), scanner manufacturer (GE Medical Systems (GE), Philips, Siemens) and endorectal coil usage (yes/no) as stratifying variables.

As noted, we consider five relevant subsets in our data (endorectal coils were only used with studies acquired with GE scanners and as such subsets with/without endorectal coils were not considered for neither

Siemens or Philips):

- Studies acquired with GE scanners and with endorectal coil (GE ERC)
- Studies acquired with GE scanners and without endorectal coil (GE no ERC)
- Studies acquired with Philips scanners (Philips)
- Studies acquired with Siemens scanners (Siemens)
- Studies acquired with any scanner (Full)

Three different ISUP-based target variables were considered (as noted previously in D5.3):

- **Low vs. possibly high** — ISUP 1 vs. ISUP 2-5 – a clinical application of this would enable the stratification of patients considering a low-risk class (ISUP=1) and a possibly high risk class (ISUP=2-5)
- **Possibly low vs. high** — ISUP 1-2 vs. ISUP 3-5 – a clinical application of this would enable the stratification of patients considering a possibly low-risk class (ISUP=1,2) and a high risk class (ISUP=3-5)
- **Intermediate vs. high** — ISUP 2-3 vs. ISUP 4-5 – a clinical application of this would enable the stratification of patients considering an intermediate risk class (ISUP=2,3) and a high risk class (ISUP=4,5)

The complete training set and hold-out test set composition is provided in Table 10.1. We note once again here that the data was split in such a way that an approximately equal proportion of all ISUP-manufacturer intersections is present across training and hold-out test sets.

Manufacturer	ISUP=1	ISUP=2	ISUP=3	ISUP=4	ISUP=5	Total
Training set (cross-validation)						
GE (ERC)	143 (28.9%)	191 (38.6%)	88 (17.8%)	51 (10.3%)	22 (4.4%)	495
GE (no ERC)	216 (22.7%)	417 (43.9%)	170 (17.9%)	55 (5.8%)	92 (9.7%)	950
Philips	550 (37.0%)	525 (35.3%)	251 (16.9%)	87 (5.8%)	75 (5.0%)	1488
Siemens	515 (24.5%)	804 (38.2%)	342 (16.3%)	185 (8.8%)	256 (12.2%)	2102
Hold-out test set						
GE (ERC)	14 (31.8%)	16 (36.4%)	7 (15.9%)	5 (11.4%)	2 (4.5%)	44
GE (no ERC)	17 (19.8%)	37 (43.0%)	17 (19.8%)	6 (7.0%)	9 (10.5%)	86
Philips	84 (37.5%)	81 (36.2%)	36 (16.1%)	13 (5.8%)	10 (4.5%)	224
Siemens	69 (21.8%)	124 (39.2%)	55 (17.4%)	25 (7.9%)	43 (13.6%)	316
Total						
GE (ERC)	157 (29.1%)	207 (38.4%)	95 (17.6%)	56 (10.4%)	24 (4.5%)	539
GE (no ERC)	233 (22.5%	454 (43.8%)	187 (18.1%)	61 (5.9%)	101 (9.7%)	1036
Philips	634 (37.0%)	606 (35.4%)	287 (16.8%)	100 (5.8%)	85 (5.0%)	1712
Siemens	584 (24.2%)	928 (38.4%)	397 (16.4%)	210 (8.7%)	299 (12.4%)	2418

Table 10.1: Data distribution across different ISUP scores and manufacturers.

Data Preparation

All sequences were resampled to 0.5x0.5x3.0mm spacing and a 128x128x24 voxel central crop was extracted, similar to previous studies on PCa aggressiveness prediction using multiparametric MRI data 18. T2w and DWI were individually normalized to values between 0 and 1, while ADC were first converted to mm2/s (if necessary) and multiplied by $\frac{1}{3}$. This enables us to keep the dynamic value range for ADC while ensuring that values are approximately between 0 and 1. In models using more than one sequence all three sequences are concatenated in the 0-th dimension (the input for a three sequence model is 3x128x128x24 voxels). As in D5.3, models were trained/tested using a 192x192x24 voxel-size crop to inspect the effect of different crop sizes on performance.

Model	Batch size (per GPU)	Warmup epochs	Number of epochs	Learning rate	Weight decay	Dropout rate
VGG	64 (16)	10	100	5 * 10-4	0.005	0.1
ConvNeXt	128 (32)					
ViT	64 (32)			5 * 10-5	0.1	
F. ViT	64 (32)					

Table 10.2: Training hyperparameters for deep learning networks (F. Vit is Factorized ViT).

Deep Learning Model Specification

We trained 4 distinct 3D deep learning architectures – a VGG-based model (consisting only of convolutions, Gaussian error linear unit activations, batch normalizations and max-pooling operations) [32], a ConvNeXt model [15], a 3D vision transformer (ViT) model [6] and a variation of the 3D ViT that separates within and between slice processing (factorized ViT). General training details are provided in Table 10.2. All models output a probability value between 0 and 1 – 0 if it belongs to the lower risk class, 1 if it belongs to the higher risk class. Particular details about each architecture are provided below:

- **VGG**. The VGG model was composed of 3 blocks with depth d following a conv(d)-gelu-batchnorm-conv(d*2)-gelu-batchnorm structure. In other words, for a given depth d, each element is passed through a convolution (conv), a Gaussian error linear unit (gelu), a batch normalization (batchnorm) and this process is repeated with the double of the depth. This is followed by a 2x2x2 max-pooling operation and repeated three times with depths [64,128,256]. After the last pooling operation, a global max-pooling operation is applied to the image, yielding a 512-dimension vector. A multilayer perceptron (with structure [512,512,512,1] and gelu activations and batchnorm) is then applied to this feature vector, yielding a uni-dimensional prediction.
- **ConvNeXt**. For the ConvNeXt model, we used the block architecture specified in the original paper [15] with no modifications. This block is repeated 4 times with depths $[32, 64, 128, 256]$ and the output vector with size 512 is then used as the input to a multilayer perceptron (with structure $[512, 512, 512, 1]$ and gelu activations and batchnorm).
- **ViT and factorized ViT**. For the ViT, we rely on replicating the original implementation [6] with no modifications. We use an 8 ViT block structure with a convolutional embedding size of 768 and 12 heads. For the multilayer perceptron structure of each block we used a $[768, 2048, 768]$ structure.

Data augmentation. During training, images are randomly augmented in real-time. For this, we used a wide array of augmentations from MONAI [20], namely:

- Identity (no transform)
- Random contrast adjustment (gamma = $[0.5, 1.5]$)
- Random standard shift in intensity (range = $[-0.1, 0.1]$)
- Random shift in intensity (range = $[-0.1, 0.1]$)
- Random Rician noise (std = 0.02)
- Random bias field (degree = 3; T2W-only)
- Affine transforms (translation range = $[4, 4, 1]$, rotation range = $\frac{\pi}{16}, \frac{\pi}{16}, \frac{\pi}{16}$])
- Horizontal flip

Each study is augmented with one of the above-mentioned transforms, which is picked at random with uniform probability (as per [22]).

Optimization. The AdamW optimizer [16], a modification to the Adam optimizer that corrects the application of weight decay with a standard cross-entropy loss, was used to train all models. Class weights were used to account for class imbalance (each positive instance is multiplied by pos/neg, where pos and neg are the number of positive and negative cases, respectively).

Additional model specification and enumeration. As mentioned above, we trained models considering 4 distinct deep-learning architectures and using only T2w and using T2w, DWI and ADC. Finally, we also assessed how clinical/demographic features – PSA, age at baseline – could have an effect on prediction. This assessment was performed in two different manners:

- Retraining the models and concatenating each normalized feature (minus the mean and divided by the standard deviation, each calculated for each fold) – we call this approach the "hybrid model" approach
- Extracting the probability scores from each sequence-only deep-learning model and calculating a binomial linear model which combines this with PSA and age at baseline. We call this approach the binomial linear model approach. Given that this warrants additional flexibility and reduced computational costs, we also train models which make use of PI-RADS

In total, we train 4 architectures with 2 distinct sequence inputs and with the inclusion/exclusion of clinical/demographic features. Each of these 16 combinations is trained using 5 different data subsets, yielding a total of 80 models, each of which is trained using 5-fold cross validation for a total of 400 training runs, each with 100 epochs.

Model evaluation. Each model is evaluated with its AUC using 5-fold cross-validation according to the best observed AUC during training and its generalizability is assessed using the hold-out test set. To assess how models perform on different subsets, we use the hold-out test set with different data subsets.

Sensitivity analysis and learning curves. As noted earlier, to understand the effect of crop size on model performance, we train the best performing model using a larger crop size ($192 \times 192 \times 24$). Additionally, to understand how the amount of data impacts model performance we train the best performing model using different fractions of the total amount of data – 0.1, 0.3, 0.5 and 0.7.

Multi-dimensional data visualization and dataset distances. To understand how the multi-dimensional features of the best performing model are distributed, we use t-SNE [37] on the last convolutional layer of our models for the complete hold-out test set. This technique allows us to have a two-dimensional representation of a multi-dimensional space. However, t-SNE is not a quantitative approach to this – indeed, it only provides a 2D visualization of our data by preserving the local neighborhood of each point. For this reason, we also calculate optimal transport dataset distances (OTDD) as suggested by Alvarez-Melis and Fusi [1]. In essence, OTDD assumes that features are distributed according to a multivariate normal distribution and uses the mean and covariance to calculate a generalized Wasserstein distance between each data description.

10.3 Results

We have organized results according to the ISUP-based target, particularly:

- Possibly low vs. high — classifying ISUP scores 1 and 2 vs. ISUP scores 3-5
- Low vs. possibly high — classifying ISUP score 1 vs. ISUP scores 2-5
- Possibly low vs. high — classifying ISUP scores 2 vs. ISUP scores 4 and 5

Possibly Low vs. High (ISUP 1,2 vs. ISUP 3,4,5)

Cross-validation results. We observe wide variability in performance across different manufacturers and architectures (Figure 10.1). Particularly, we observe that:

- **mpMRI has greater diagnostic value than T2-weighted sequences alone.** While T2-weighted sequences provide greater anatomical resolution, using these alone in prediction leads to lacklustre results; indeed, if ADC/DWI is available, there is no reason not to use them in prediction as they consistently lead to improved results ($6 * 10^{-13}$ for a paired rank sum test comparing T2W vs. T2W+DWI+ADC)
- **VGG outperforms more complex models.** Across different data subsets, VGG models have a tendency to outperform more complicated and modern models such as ViT or ConvNext. Indeed, in all cases, the average AUC for VGG models is higher than that of other models. While there is no statistical significance when performing direct comparisons between VGG models and the second best performing model (Table 10.3), this may be associated with a lack of power due to the relatively small number of folds (k=5). Taking this into consideration, the relationship between AUC and deep-learning models while controlling for manufacturer and sequence type is analyzed with a multivariate linear model. VGG models are associated with higher AUC (at least 5% AUC when compared to other models; Table 10.4). Finally, the use of ViT architectures appears to be detrimental - in most instances, ViT-based models perform worse than CNN-based models (Table 10.5); this can be explained by their lack of inductive biases, making them more dependent on high volumes of data. While these models are trained, to the best knowledge of the authors, on the largest mpMRI dataset available, the original ViT models were trained on at least 1 million images [6], 200 times more images than those in ProstateNet. As such, their apparent success in other image-based tasks cannot be easily transferred to most tasks in biological and biomedical image analysis
- **Within-manufacturer performance is distinct across manufacturers.** One of the key expected insights was manufacturer-specific effects on predictive performance. Firstly, the performance transference from models trained and tested on the same domain (same manufacturer) is analyzed:
  - Generally, models trained/tested on GE data with endorectal coils are worse than models trained/tested on other types of data, whereas models trained/tested on other manufacturers appear to perform better. This can be due to one of two things: either the relevant signal is weaker in GE scanners, or the amount of data in GE scanners (GE scanner data is the most underrepresented of the three in ProstateNet) is not sufficient to create good models
  - Models trained on data from Philips scanners with VGG models - nearly half as prevalent as Siemens studies - seem to perform the best out of all three scanners. This finding is common to both T2W and T2W+DWI+ADC.

Manufacturer	Model (other)	Mean VGG AUC	Mean other AUC	Sequences	p-value
GE (ERC)	Regular ViT	0.6643	0.6793	T2W	0.8125
GE (no ERC)	Regular ViT	0.6725	0.6309	T2W	0.3125
Philips	ConvNeXt	0.7255	0.6709	T2W	0.0625
Siemens	ConvNeXt	0.662	0.6026	T2W	0.0625
Full	ConvNeXt	0.6377	0.6032	T2W	0.0625
GE (ERC)	Regular ViT	0.6094	0.6836	T2W+DWI+ADC	0.625
GE (no ERC)	Regular ViT	0.7295	0.6409	T2W+DWI+ADC	0.0625
Philips	ConvNeXt	0.793	0.7258	T2W+DWI+ADC	0.0625
Siemens	Regular ViT	0.7047	0.6723	T2W+DWI+ADC	0.125
Full	ConvNeXt	0.6873	0.6589	T2W+DWI+ADC	0.1875

Table 10.3: p-values for paired Wilcoxon rank sum tests comparing VGG models with the second best model for each manufacturer for the possibly low vs. high target definition.

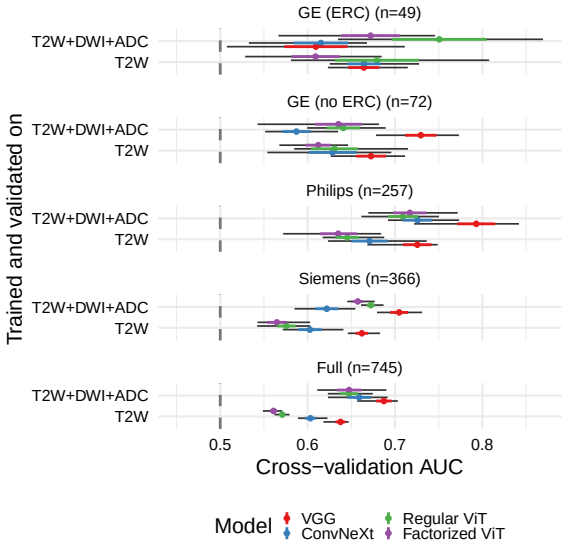


Figure 10.1: Cross validation area under the curve (AUC) of different models on different manufacturer datasets.

Variable	Estimate	Std. Error	t-value	p-value
Intercept	0.5934	0.0111	53.293	2.10E-131
Sequences used (vs. T2)				
T2W+DWI+ADC	0.0363	0.007	5.1606	5.30E-07
Manufacturer (vs. all manufacturers)				
GE (ERC)	0.0183	0.0122	1.4987	1.40E-01
GE (no ERC)	0.0154	0.0122	1.2628	2.10E-01
Philips	0.076	0.0122	6.2276	2.20E-09
Siemens	0.006	0.0122	0.488	6.30E-01
Deep-learning models (vs. ConvNext)				
Factorized ViT	-0.002	0.01	-0.1984	8.40E-01
Regular ViT	0.0107	0.01	1.079	2.80E-01
VGG	0.0517	0.01	5.1873	4.70E-07

Table 10.4: Coefficients for a linear model where AUC is the dependent variable and sequence type, manufacturer and deep-learning model are independent variables for the possibly low vs. high target definition.

Manufacturer	Mean CNN AUC	Mean ViT AUC	Sequences	p-value
GE (ERC)	0.6644	0.6442	T2W	5.57E-01
GE (no ERC)	0.6505	0.6215	T2W	0.2324
Philips	0.6982	0.6402	T2W	9.80E-03
Siemens	0.6323	0.5702	T2W	0.002
Full	0.6204	0.566	T2W	2.00E-03
GE (ERC)	0.5935	0.6779	T2W+DWI+ADC	8.40E-02
GE (no ERC)	0.6584	0.6381	T2W+DWI+ADC	4.32E-01
Philips	0.7594	0.713	T2W+DWI+ADC	2.73E-02
Siemens	0.6633	0.6648	T2W+DWI+ADC	0.8457
Full	0.6731	0.6473	T2W+DWI+ADC	3.71E-02

Table 10.5: Paired Wilcoxon rank sum test comparing convolutional models with transformer-based (ViT) models for the possibly low vs. high target definition.

Hold-out test results. After analyzing the performance of our models across different training and validation subsets, we now seek to assess how transferable these models are to a hold-out test set. We arrive at some key conclusions:

- Different models show distinct generalizability. There is a clear, albeit inconsistent, drop in perfor-

mance, as shown in Figure 10.2 and Figure 10.3. This drop in performance, however, appears to be more evident in the VGG (especially for Philips) and regular ViT models. For other models (ConvNeXt and factorized ViT) this drop in performance is specific to only a few of the manufacturer models and follows a relatively consistent linear trend

- Diverse training data leads to better generalization. Having such a dataset with plentiful cases of different manufacturers enables a more substantial analysis - particularly, it is possible to assess how models trained on data from specific manufacturers are affected when applied to data obtained using other manufacturers. As shown in Figure 10.4 and Figure 10.5 and as expected, models trained on specific scanners generally perform better when tested on data obtained using the same scanner. However, a few interesting aspects should be highlighted:
  - Models trained on all manufacturers (Full) perform similarly across different scanners. This is less so the case for data obtained using GE scanners
  - There appears to be some evidence showing that training (and possibly fine-tuning) scanner-specific models is highly useful, but it is evident that training models on large collections of images from different scanners leads to similar performance
  - Hold-out test performance is not independent of DL architecture - once again, VGG outperforms other models ($p = 0.0002$ considering performance across all scanners). Looking at model performance on individual scanners, however, shows a different picture - while the average VGG performance is consistently better in T2W+DWI+ADC models, this is not statistically significant except for models trained on Siemens data.

On the inclusion of clinical data. Age at baseline and PSA are two useful clinical variables which can be helpful in early detection programs. Here, "hybrid models" - deep-learning models combining both sequence and clinical information - are compared with "sequence-only models" - models using only sequence information - to determine whether these clinical covariates can be beneficial for prediction.

Analyzing the CV performance for transformer-based models reveals that there is little to be gained from including these specific clinical variables when all three mpMRI sequences are used (Figure 10.6 and Figure 10.7). Further considering the results for the hold-out test set shows that including clinical variables is unlikely to be useful in both T2W and T2W+DWI+ADC models. Indeed, as visible in Figure 10.8 and Figure 10.9, it is evident that the inclusion of clinical variables in T2W models does not lead to improvements in most instances. Finally, we note that training linear classification models and even including PI-RADS scores does not lead to any changes in these conclusions — by combining sequence-only probabilities, age and PSA, we show that performance does not improve by combining sequence only probabilities (or deep probabilities) with age, PSA or PI-RADS (Figure 10.10). This would suggest that our sequence-only models are already capturing the relevant information that would be provided by PI-RADS.

Learning curve analysis. To better understand the relationship between the amount of data and performance, sequence-only VGG models with all sequences were trained with different amounts of training data. In terms of cross-validated performance, there is an expected relation between the amount of training data (the fraction of available training data) and performance for all manufacturers (Figure 10.11). This trend, however, is not as clear for the hold-out test set - while for most cases one sees an upwards trend when training and testing on the same data, this is unpredictable when testing on data from other manufacturers (Figure 10.12). For instance, increasing the amount of Siemens data when training VGG models leads to improved performance on GE data but is detrimental when testing on Philips data; on the other hand, when training with data from all manufacturers, performance plateaus at approximately half (0.5) except for GE ERC, where it generally remains poor regardless of the amount of training data used.

Sensitivity analysis to crop size. While the crop size used in this work ($128 \times 128 \times 24$) is not uncommon [23], a better understanding on whether this could lead to a loss of signal was needed. Hence, the effect of a larger crop size ($192 \times 192 \times 24$) on performance was tested, showing that the performance in both CV and hold-out test set does not improve (Figure 10.13 and Figure 10.14), with models trained on GE data with

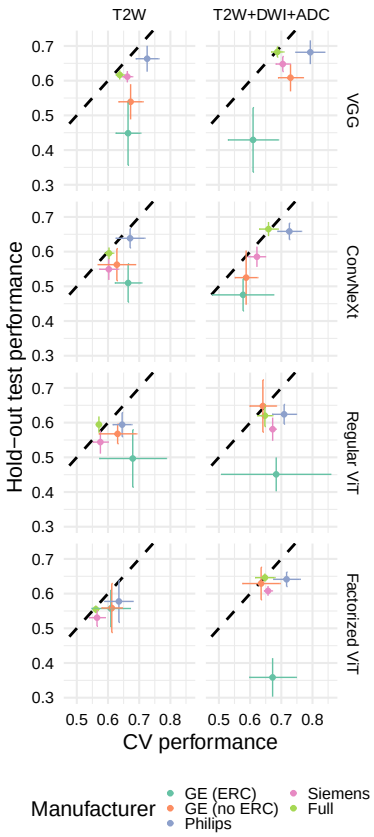


Figure 10.2: Comparison of cross-validated (CV) and test area under the curve (AUC). Each point represents the average AUC, whereas the vertical and horizontal error bars represent the mean with the addition and subtraction of the standard error, respectively. The diagonal dashed line represents equality between both axes.

larger crops appearing to perform worse than models trained on smaller crops (this difference is, however, not statistically significant; Table 10.6).

Manufacturer	Mean 128x128 AUC	Mean 192x192 AUC	p-value
GE (ERC)	0.4547	0.4486	0.8119
GE (no ERC)	0.5987	0.6042	0.4908
Philips	0.637	0.6328	0.8949
Siemens	0.5959	0.5949	0.7915
Full	0.653	0.6635	0.7112

Table 10.6: p-values for paired Wilcoxon rank sum tests comparing performance on different crop sizes ($128 \times 128 \times 24$ and $192 \times 192 \times 24$) for the possibly low vs. high target definition.

Multi-dimensional data visualization and dataset distances. While DL methods can perform relatively well, the associations between deep features, learned by these models, and aggressiveness or manufacturer can further illuminate which features are being learned. For this, deep features are first visually inspected using t-SNE on the features obtained in the best performing sequence-only T2W+DWI+ADC VGG fold, showing how the local neighborhood structure in this representation shows how samples from the same manufacturer cluster together. Particularly, samples from GE scanners — mostly stemming from dataset provider pseudonym B — share very few neighbors with samples from other manufacturers (Figure 10.15), hinting that performance may be tied not only with scanner manufacturers but also with center-specific

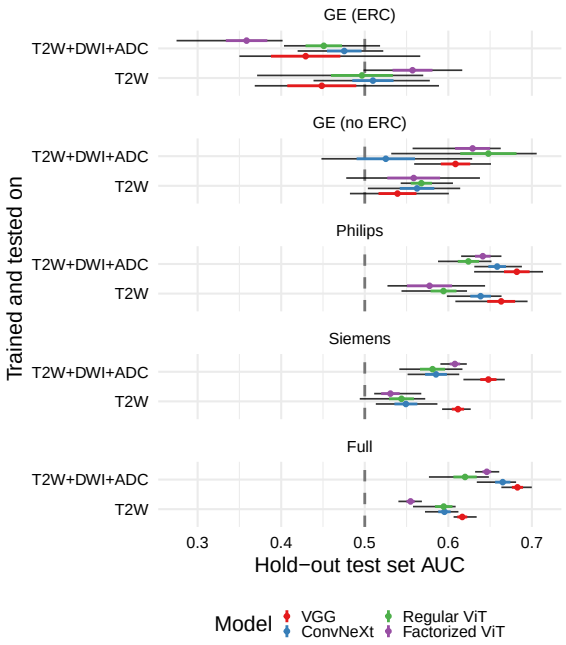


Figure 10.3: Test area under the curve (AUC) of different models on different manufacturer testing datasets.

protocols. Indeed, as notable in the same figure, the application of an endorectal coil during examination is present in 96.8% of the studies (31/32 studies), possibly explaining the predictive underperformance in GE ERC examinations — features appear to be different when comparing studies with and without endorectal coils.

Since t-SNE is qualitative, quantitative dataset comparisons were performed using the dataset distance proposed in 26 and stratifying by aggressiveness and by manufacturer. Using this, the qualitative findings in Figure 10.15 are recapitulated (Figure 10.16): data hailing from the same manufacturer (and protocol) is, in general, more similar than data with the same classification (aggressiveness).

Low vs. Possibly High (ISUP 1 vs. ISUP 2,3,4,5)

In general, we observe that the same conclusions regarding manufacturer performance hold for this target definition, with some notable exceptions; however, possibly due to reasons which will be discussed ahead, performance is in general better than that observed for the previous target definition. In the following discussion we offer brief discussions using the same sections to facilitate comparison between both targets.

We note here that comparing the performance between targets is not sensible as we are effectively comparing two distinct tasks. Instead, we focus on observing changes in differences between manufacturer, clinical preparation, model and other conditions.

Cross-validation results. Compared with the previous target definition, we note here that the same conclusions can be drawn. Particularly, mpMRI models outperform T2w-only models, VGG outperforming other, more complex models and within-scanner performance shows wide variability Figure 10.17. For this task, however, Siemens underperforms when compared with other models, whereas for the previous task it was comparable with Philips models. On the other hand, Philips and Full models show improved performance, with the Philips T2W-only model showing good performance (73% average AUC). Given that data was stratified based on ISUP grades and scanners, we do not believe these discrepancies — Siemens underperforming for this target — is due to shifts in manufacturer distribution. Other trends — performance differences between VGG and the best performing model (Table 10.7; Table 10.8) and between CNN-based and ViT-based models (Table 10.9) — are also recapitulated here.

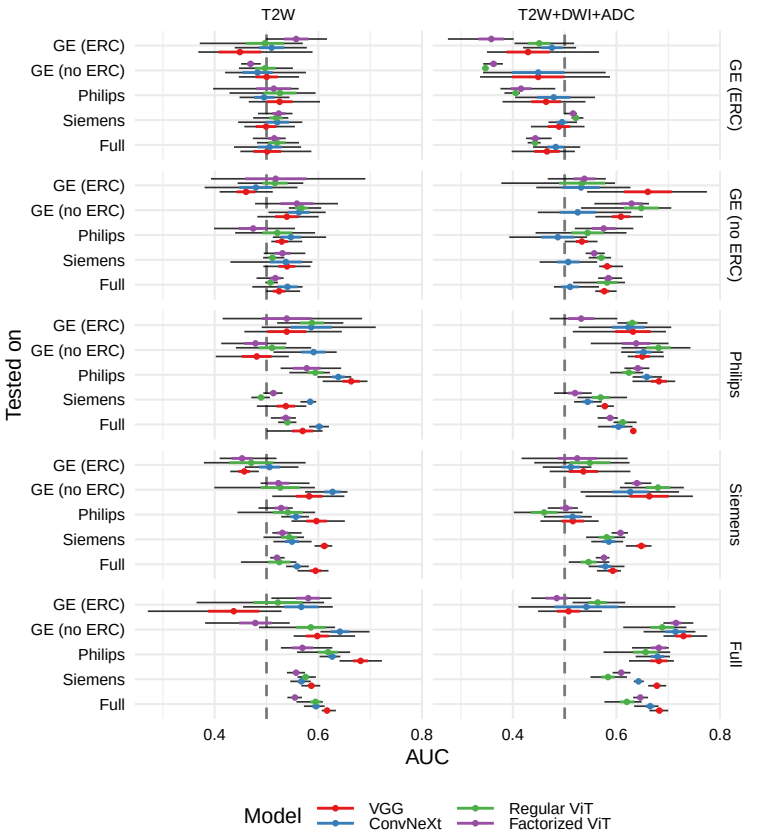


Figure 10.4: Test AUC of models trained and tested on different scanners. The y-axis refers to the data used to train each model and the y-facet (text on the right side of the image) refers to the data used to test each model.

Manufacturer	Model (**other**)	**Mean VGG AUC**	**Mean other AUC**	**Sequences**	**p-value**
GE (ERC)	ConvNeXt	0.6808	0.68	T2W	1
GE (no ERC)	ConvNeXt	0.7387	0.732	T2W	1
Philips	ConvNeXt	0.7295	0.6495	T2W	0.0625
Siemens	Factorized ViT	0.6118	0.5668	T2W	0.0625
Full	ConvNeXt	0.6532	0.5922	T2W	0.0625
GE (ERC)	Regular ViT	0.607	0.6339	T2W+DWI+ADC	0.625
GE (no ERC)	ConvNeXt	0.7632	0.7142	T2W+DWI+ADC	0.3125
Philips	Regular ViT	0.7786	0.72	T2W+DWI+ADC	0.0625
Siemens	Regular ViT	0.6515	0.617	T2W+DWI+ADC	0.0625
Full	ConvNeXt	0.7109	0.6813	T2W+DWI+ADC	0.125

Table 10.7: p-values for paired Wilcoxon rank sum tests comparing VGG models with the second best model for each manufacturer for the low vs. possibly high target definition.

Hold-out test results. The main difference is that, excluding GE data, drops in performance are not as striking — this holds for both Siemens and Philips models, with GE (no ERC) models suffering a considerable drop in performance (for mpMRI VGG models, the performance drops from 76% on CV to 67% on the hold-out test set; Figure 10.18 and Figure 10.19). Full models are still the ones showing the smallest variability and drop in performance. An interesting finding is that Full models outperform Siemens models on Siemens data, whereas Philips VGG and Regular ViT models outperform GE (no ERC) models on GE (no ERC) data (this effect was already somewhat visible in the possibly low vs. high target but here it becomes even more apparent; Figure 10.20 and Figure 10.21). Together with the cross-validation results for this target and the

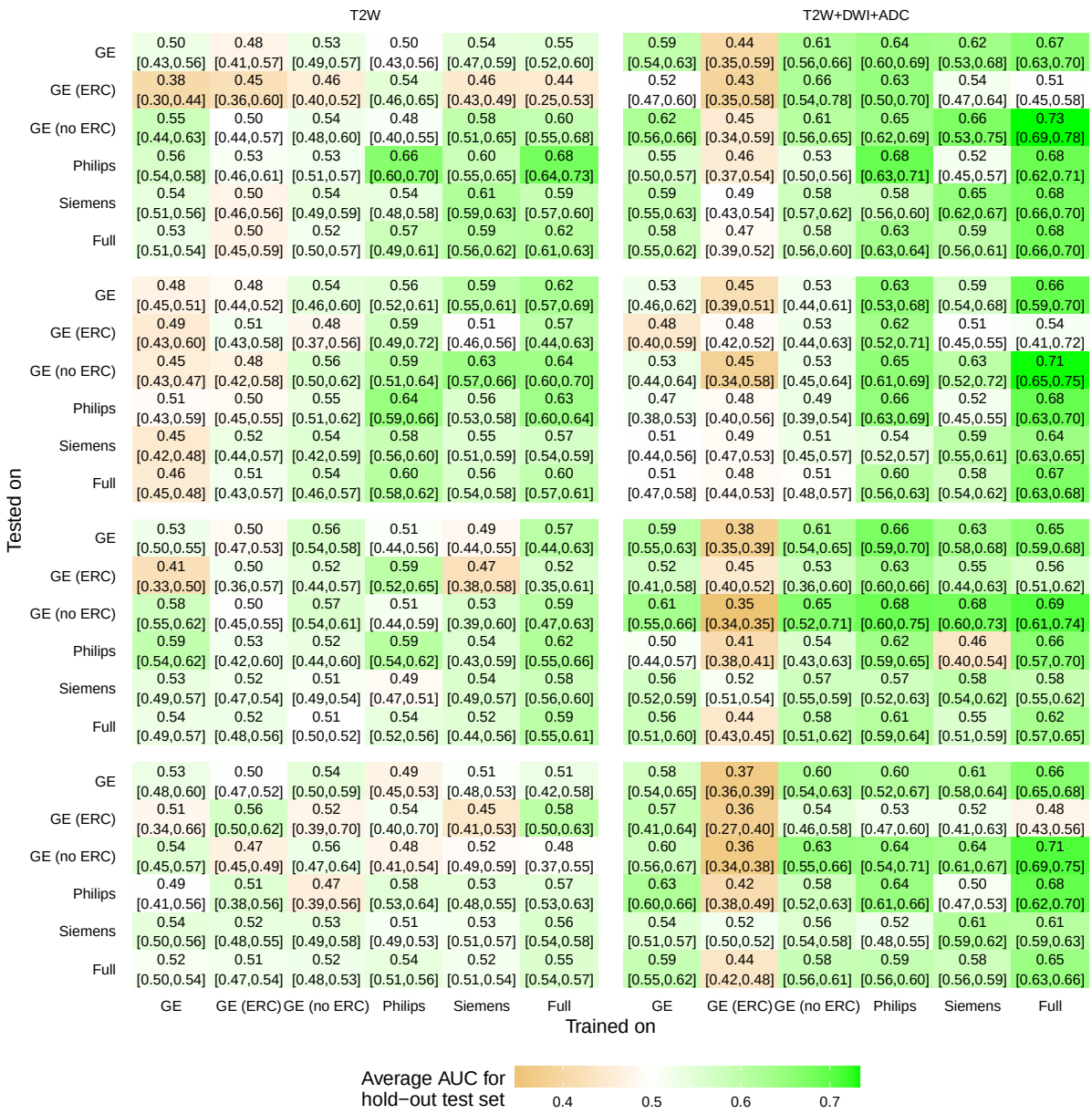


Figure 10.5: Test AUC of models trained and tested on different scanners. The text corresponds to the average, minimum and maximum AUC values (minimum and maximum values are between brackets) and the colour corresponds to the average AUC value.

results for the previous target, we suggest that there may be some transferability with Philips models to GE studies obtained without endorectal coils. However, it still holds that training models on data from multiple different scanners is the best approach in terms of transferability of performance to different models.

On the inclusion of clinical data. We obtain results similar to those presented using the possibly low vs. high target — indeed, no noteworthy gains are observed when including clinical or demographic data in our DL models (Figure 10.22; Figure 10.23; Figure 10.24; Figure 10.25).

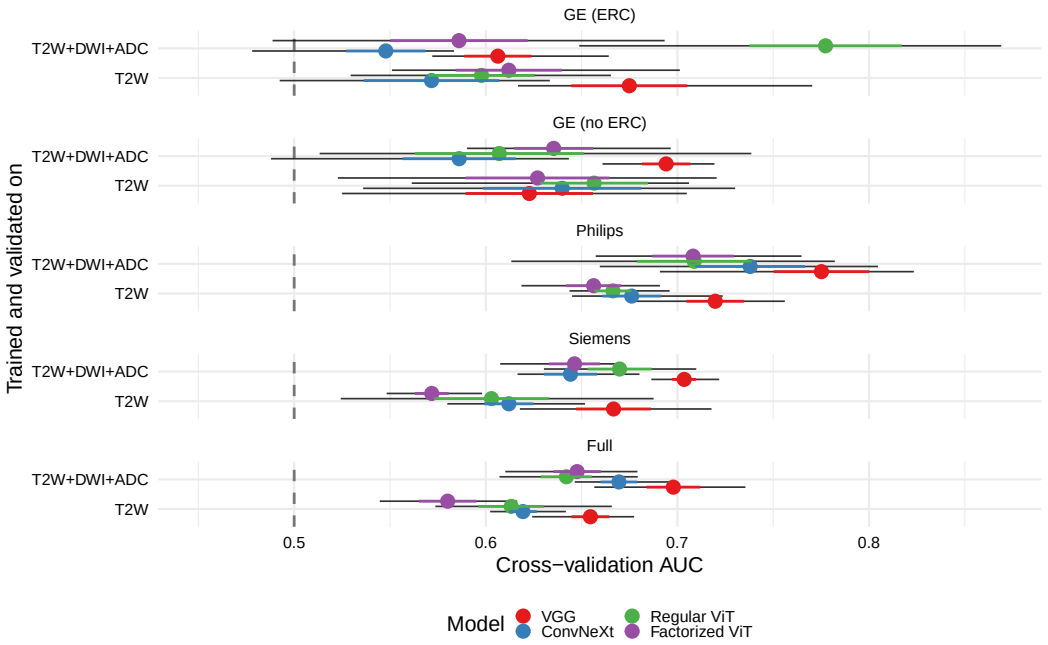


Figure 10.6: Cross validation area under the curve (AUC) of different hybrid models (mpMRI + clinical) on different manufacturer datasets.

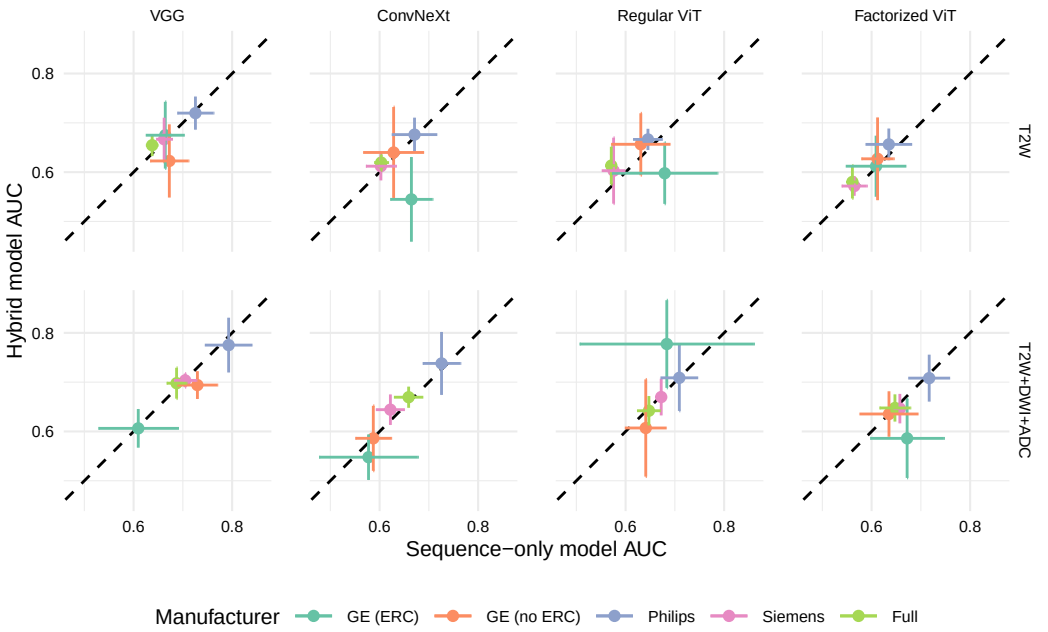


Figure 10.7: Comparison of AUC for sequence-only and hybrid models. Each point represents the average AUC, whereas the vertical and horizontal error bars represent the mean with the addition and subtraction of the standard error, respectively. The diagonal dashed line represents equality between both axes.

Learning curve analysis. Similarly to what was observed for the previous target, we note that there is no clear saturation effect (despite the clear observation of a "deceleration"; Figure 10.27). This suggests that

Variable	Estimate	Std. Error	t-value	p-value
Intercept	0.6213	0.0113	55.1213	1.50E-134
Sequences used (vs. T2)				
T2W+DWI+ADC	0.034	0.0071	4.7715	3.30E-06
Manufacturer (vs. all manufacturers)				
GE (ERC)	-0.0018	0.0123	-0.1475	8.80E-01
GE (no ERC)	0.0685	0.0123	5.5493	7.90E-08
Philips	0.0605	0.0123	4.8961	1.80E-06
Siemens	-0.0444	0.0123	-3.5949	4.00E-04
Deep-learning models (vs. ConvNext)				
Factorized ViT	-0.0224	0.0101	-2.218	2.80E-02
Regular ViT	-0.014	0.0101	-1.3903	1.70E-01
VGG	0.0361	0.0101	3.5777	4.20E-04

Table 10.8: Coefficients for a linear model where AUC is the dependent variable and sequence type, manufacturer and deep-learning model are independent variables for the low vs. possibly high target definition.

Manufacturer	Mean CNN AUC	Mean ViT AUC	Sequences	p-value
GE (ERC)	0.6804	0.6384	T2W	2.75E-01
GE (no ERC)	0.7353	0.6506	T2W	0.0098
Philips	0.6895	0.6389	T2W	9.80E-03
Siemens	0.5859	0.5594	T2W	0.1309
Full	0.6227	0.5743	T2W	2.00E-03
GE (ERC)	0.6074	0.6196	T2W+DWI+ADC	6.95E-01
GE (no ERC)	0.7387	0.7025	T2W+DWI+ADC	1.31E-01
Philips	0.7478	0.7187	T2W+DWI+ADC	4.88E-02
Siemens	0.6179	0.6123	T2W+DWI+ADC	0.6953
Full	0.6961	0.6598	T2W+DWI+ADC	2.00E-03

Table 10.9: Paired Wilcoxon rank sum test comparing convolutional models with transformer-based (ViT) models for the low vs. possible high target definition.

the inclusion of more data can add further benefit to our models. As before, the performance of GE (ERC) models remains poor, further suggesting that endorectal coil use during scan acquisition, while beneficial for contrast, is detrimental for deep learning models. When observing the hold-out test set learning curves this non-saturating effect is only evident for GE (no ERC), Philips and Full models (Figure 10.27). Indeed, for Siemens models, there appears to be no relevant improvement in hold-out test set performance that is associated with increasing the volume of training data.

Sensitivity analysis to crop size. As shown earlier, crop size does not appear to have a predominant effect on the performance of our models (Figure 10.29 and Figure 10.30), suggesting that the relevant signal is contained within the $128 \times 128 \times 24$ central crop. It should be noted, however, two exceptions from this otherwise clear trend in the hold-out test set (Figure 10.30; Table 10.10) — when training on all manufacturers (Full models), using a larger crop size ($192 \times 192 \times 24$) leads to better performance with GE (no ERC) data. However, and interestingly this does not hold when analyzing GE (no ERC) models. Secondly, a drop in performance happens in Full models when considering Philips data. Considering that, in general, performance does not drop when considering the performance for all manufacturers, we suggest here that there may be an added benefit to using a larger crop size for specific manufacturers (i.e. on GE (no ERC) data), and that this added benefit should be carefully weighed against potential detrimental effects (i.e. on Philips data).

Multi-dimensional data visualization and dataset distances. Finally, we note that observing the multi-dimensional distribution of features in two dimensions leads to similar results as before (Figure 10.31 and Figure 10.32). However, we note here that there appears to be a clearer separation of both classes (ISUP=1 vs. ISUP=2,3,4,5). Nonetheless, neighbors in feature space appear to be mostly determined by manufacturer/protocol rather than by classification.

Manufacturer	Mean 128x128 AUC	Mean 192x192 AUC	p-value
GE (ERC)	0.5060	0.5055	0.8949
GE (no ERC)	0.5878	0.5785	0.8119
Philips	0.6382	0.6281	0.7112
Siemens	0.5263	0.5304	0.9158
Full	0.6609	0.6665	0.6721

Table 10.10: p-values for paired Wilcoxon rank sum tests comparing performance on different crop sizes (128x128x24 and 192x192x24) for the low vs. possibly high target definition.

Intermediate vs. High (ISUP 2,3 vs. ISUP 4,5)

In general, we know that the conclusions regarding manufacturer performance are follow from those observed in the previous section. However, we note that models demonstrate worse generalizability, possibly due to the relatively smaller amounts of available data for training (no ISUP=1 cases were used).

Cross-validation results As before, mpMRI models outperform T2W-only models (Figure 10.33), and other similar trends, particularly performance differences between VGG and the best performing model (Table 10.11; Table 10.12) and between CNN-based and ViT-based models (Table 10.13) are also observable here.

Manufacturer	Model (**other**)	Mean VGG AUC	Mean other AUC	Sequences	p-value
GE (ERC)	ConvNeXt	0.6808	0.6800	T2W	1.0000
GE (no ERC)	ConvNeXt	0.7387	0.7320	T2W	1.0000
Philips	ConvNeXt	0.7295	0.6495	T2W	0.0625
Siemens	Factorized ViT	0.6118	0.5668	T2W	0.0625
Full	ConvNeXt	0.6532	0.5922	T2W	0.0625
GE (ERC)	Regular ViT	0.6070	0.6339	T2W+DWI+ADC	0.6250
GE (no ERC)	ConvNeXt	0.7632	0.7142	T2W+DWI+ADC	0.3125
Philips	Regular ViT	0.7786	0.7200	T2W+DWI+ADC	0.0625
Siemens	Regular ViT	0.6515	0.6170	T2W+DWI+ADC	0.0625
Full	ConvNeXt	0.7109	0.6813	T2W+DWI+ADC	0.1250

Table 10.11: p-values for paired Wilcoxon rank sum tests comparing VGG models with the second best model for each manufacturer for the low vs. possibly high target definition.

Variable	Estimate	Std. Error	t-value	p-value
Intercept	0.6213	0.0113	55.1213	0.0000
Sequences used (vs. T2)				
T2W+DWI+ADC	0.0340	0.0071	4.7715	0.0000033
Manufacturer (vs. all manufacturers)				
GE (ERC)	-0.0018	0.0123	-0.1475	0.8829
GE (no ERC)	0.0685	0.0123	5.5493	0.0000001
Philips	0.0605	0.0123	4.8961	0.0000018
Siemens	-0.0444	0.0123	-3.5949	0.0003970
Deep-learning models (vs. ConvNext)				
Factorized ViT	-0.0224	0.0101	-2.2180	0.0275
Regular ViT	-0.0140	0.0101	-1.3903	0.1658
VGG	0.0361	0.0101	3.5777	0.0004

Table 10.12: Coefficients for a linear model where AUC is the dependent variable and sequence type, manufacturer and deep-learning model are independent variables for the low vs. possibly high target definition.

Hold-out test results A considerably larger drop in performance is observed for this target when considering the hold-out test set. This is more evident for mpMRI models, whose performance deteriorates considerably (in particular for GE (ERC) and Philips models; Figure 10.34 and Figure 10.35). As before,

Manufacturer	Mean CNN AUC	Mean ViT AUC	Sequences	p-value
GE (ERC)	0.6804	0.6384	T2W	0.2754
GE (no ERC)	0.7353	0.6506	T2W	0.0098
Philips	0.6895	0.6389	T2W	0.0098
Siemens	0.5859	0.5594	T2W	0.1309
Full	0.6227	0.5743	T2W	0.0020
GE (ERC)	0.6074	0.6196	T2W+DWI+ADC	0.6953
GE (no ERC)	0.7387	0.7025	T2W+DWI+ADC	0.1309
Philips	0.7478	0.7187	T2W+DWI+ADC	0.0488
Siemens	0.6179	0.6123	T2W+DWI+ADC	0.6953
Full	0.6961	0.6598	T2W+DWI+ADC	0.0020

Table 10.13: Paired Wilcoxon rank sum test comparing convolutional models with transformer-based (ViT) models for the low vs. possible high target definition.

Full models show a smaller amount of variability, but even for this case there is a larger drop in mpMRI models when compared with T2w models (Figure 10.36; Figure 10.37).

On the inclusion of clinical data We observe little improvements for DL models incorporating a hybrid setup (Figure 10.38; Figure 10.39; Figure 10.40; Figure 10.41); Figure 10.42).

Learning curve analysis Considering the CV learning curves, there is no clear evidence of a saturation effect (Figure 10.43). However, it is fairly evident that in the test-set this is not as clear, particularly when considering tests from GE machines (Figure 10.44).

Sensitivity analysis to crop size The small crop size used does not appear to have a deleterious effect on performance, with most instances of changes in performance being a decrease in performance when larger crop sizes are used. This holds for both CV and hold-out test set performance (Figure 10.45 and Figure 10.46; Table 10.14).

Manufacturer	Mean 128x128 AUC	Mean 192x192 AUC	p-value
GE (ERC)	0.5060	0.5055	0.8949
GE (no ERC)	0.5878	0.5785	0.8119
Philips	0.6382	0.6281	0.7112
Siemens	0.5263	0.5304	0.9158
Full	0.6609	0.6665	0.6721

Table 10.14: p-values for paired Wilcoxon rank sum tests comparing performance on different crop sizes (128x128x24 and 192x192x24) for the low vs. possibly high target definition.

Multi-dimensional data visualization and dataset distances Finally, we note that observing the multi-dimensional distribution of features in two dimensions leads to similar results as before (Figure 10.47 and Figure 10.48). However, the separation between GE (ERC) and other manufacturers is not as evident as was observed for the previous targets.

Prospective Validation

The aforementioned models were also tested on different subsets. When analysing this (Figure 10.49), it is observable that, while this failure is relatively widespread, some meaningful trends can be observed:

- While generalization is poor, models trained on data from all manufacturers are still capable of generalising as well as models trained on specific subsets of data, suggesting that this is still a good strategy to maximise the transferability of models;
- Siemens models appear to offer relatively good generalisation with the exception of T2W low vs. possibly high models;

- GE models with and without ERC are relatively poor at generalising even when being tested in the same data.

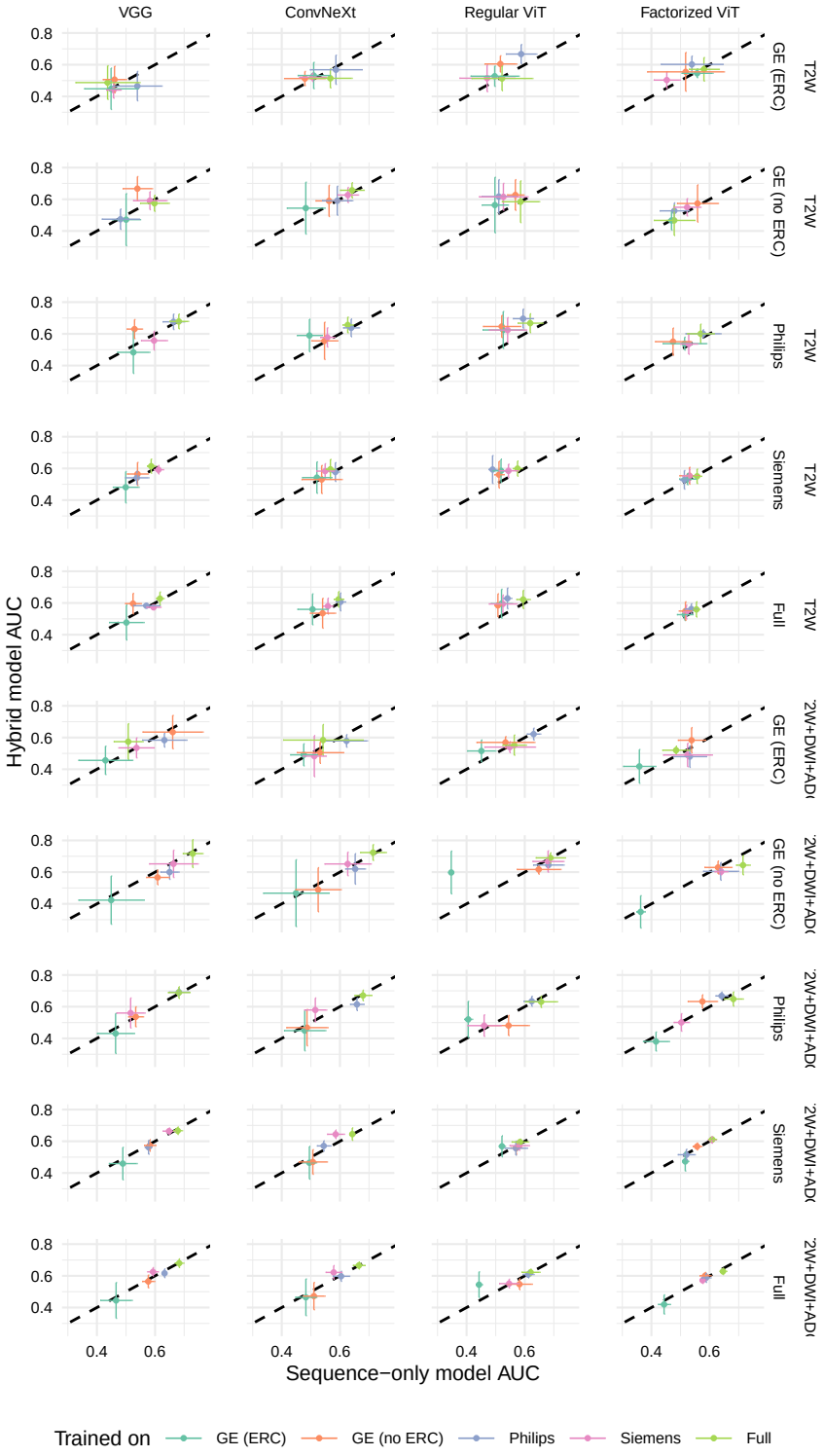


Figure 10.8: Comparison of hold-out test set AUC for sequence-only and hybrid models. Each point represents the average AUC, whereas the vertical and horizontal error bars represent the mean with the addition and subtraction of the standard error, respectively. The diagonal dashed line represents equality between both axes.

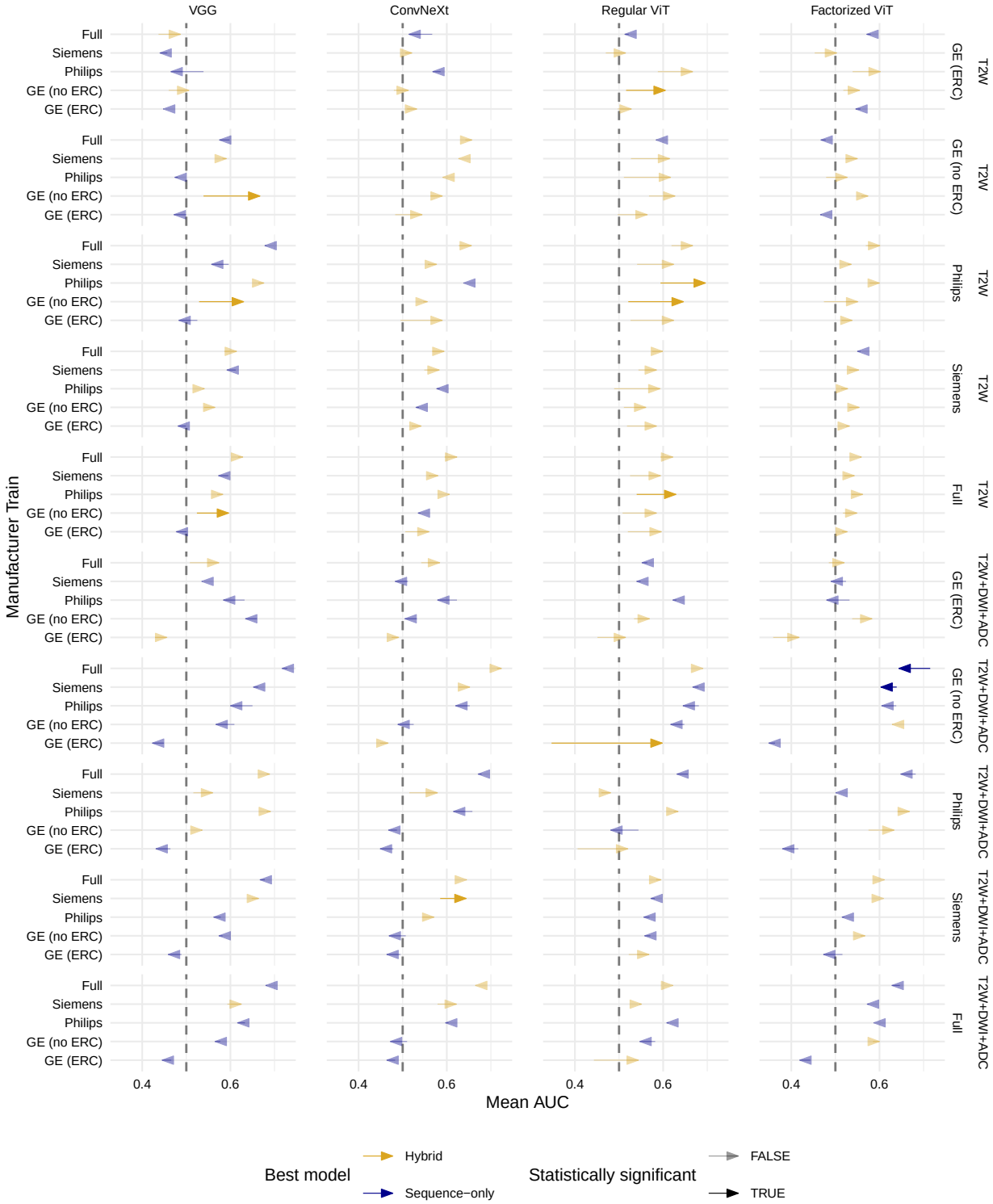


Figure 10.9: Test AUC of models trained and tested on different scanners. The arrow represents sequence-only and hybrid model performance (base and tip, respectively) and the colour of the arrow represents which model performed best. The y-axis refers to the data used to train each model and the y-facet (text on the right side of the image) refers to the data used to test each model.

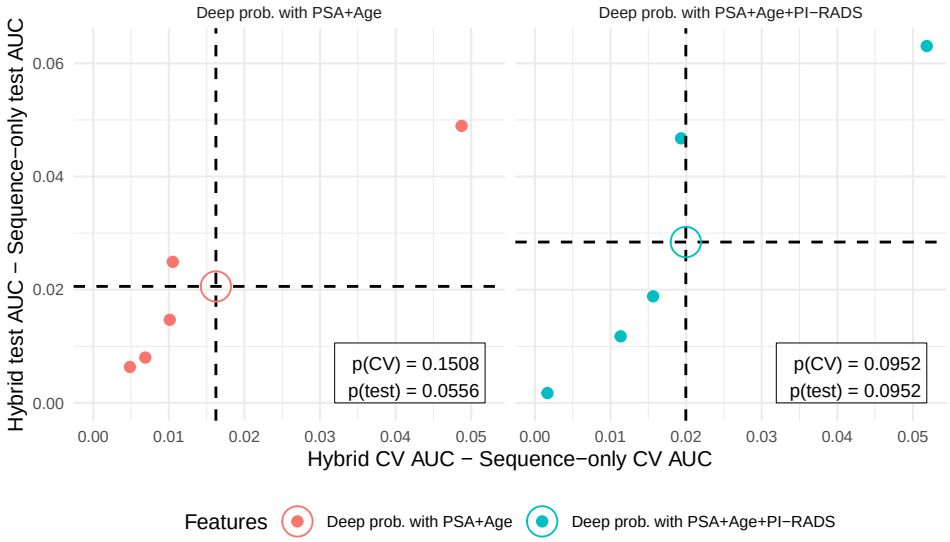


Figure 10.10: Difference between sequence-only and elastic net-regularized linear classification model AUC. Both CV (x axis) and test (y axis) AUC are represented, with the average value noted as a circle at the intersection of the dashed lines. The p-values shown in the figure were obtained using a one-sample Wilcoxon rank sum test.

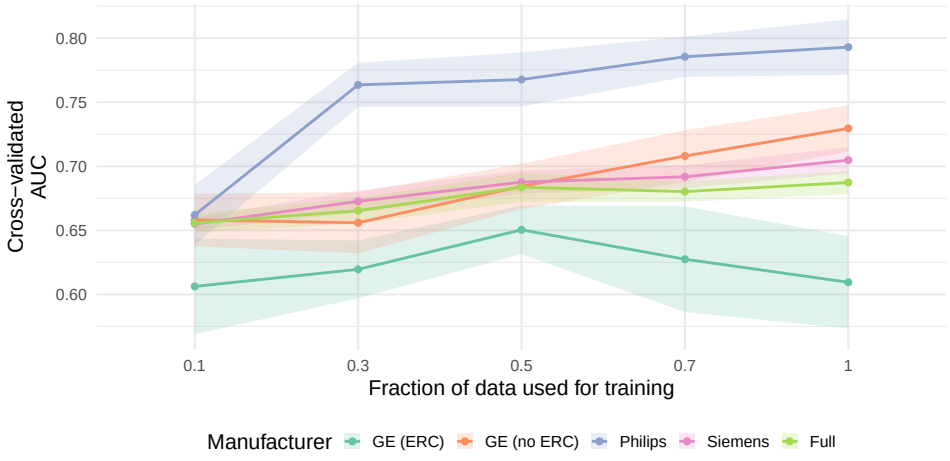


Figure 10.11: Learning curve for cross-validation AUC.

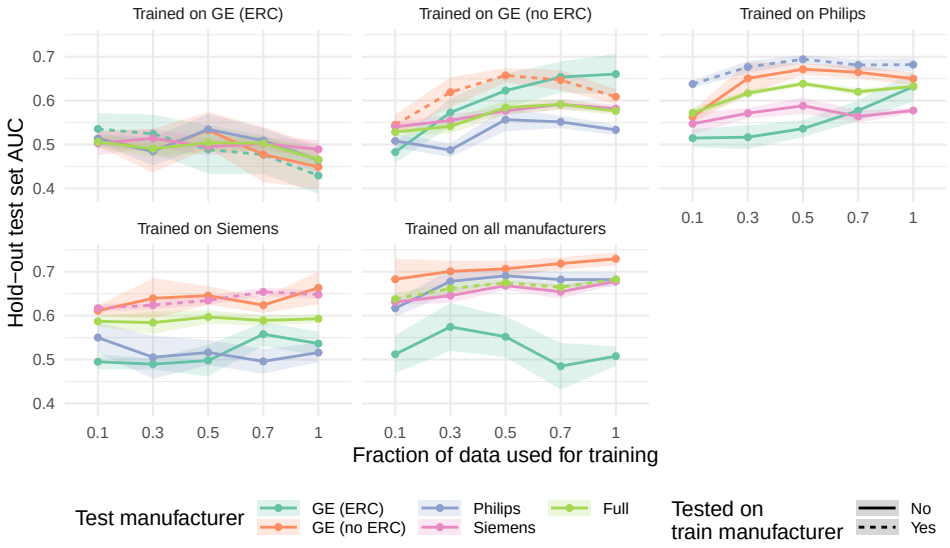


Figure 10.12: Learning curve for hold-out test set AUC.

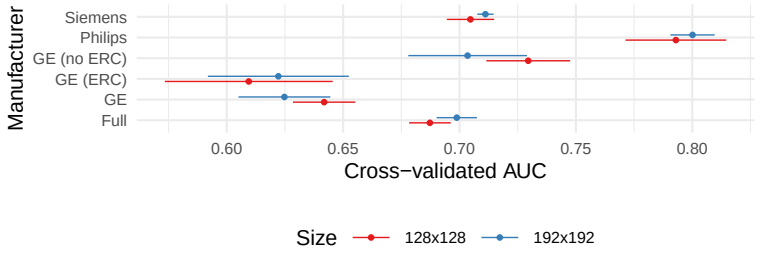


Figure 10.13: Impact of crop size on cross-validation AUC.

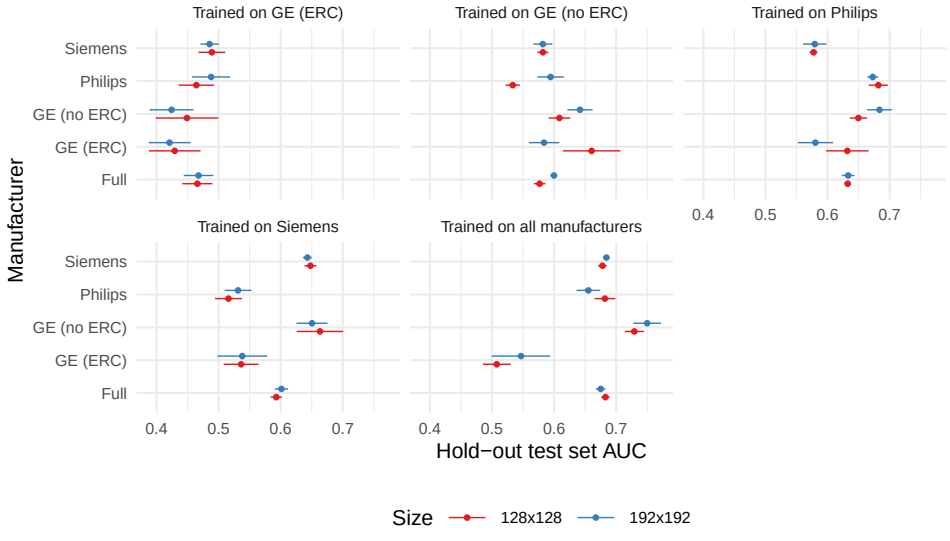


Figure 10.14: Impact of crop size on hold-out test set AUC.

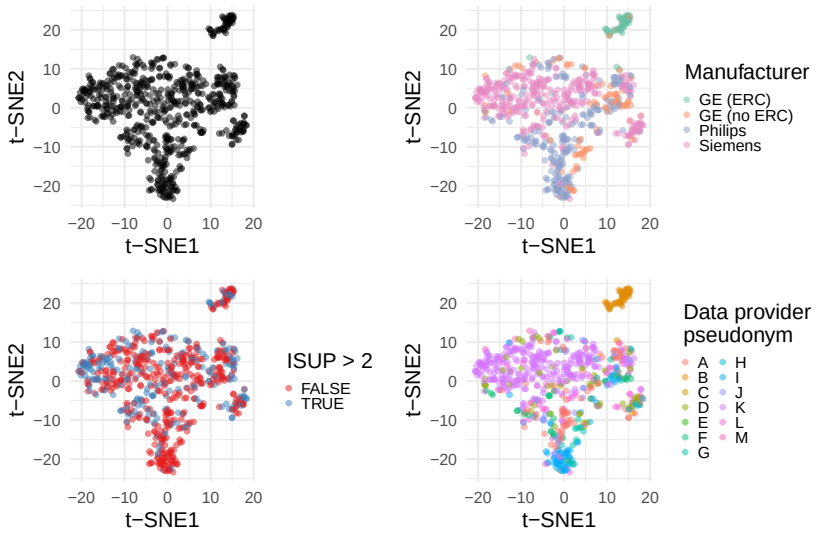


Figure 10.15: t-distributed stochastic neighbor embedding (t-SNE) visualization of all data (n=560 studies; first panel) and stratified by manufacturer (second panel) and by aggressiveness. The embedding is the same across panels and t-SNE1 and t-SNE2 represent the t-SNE dimensions.

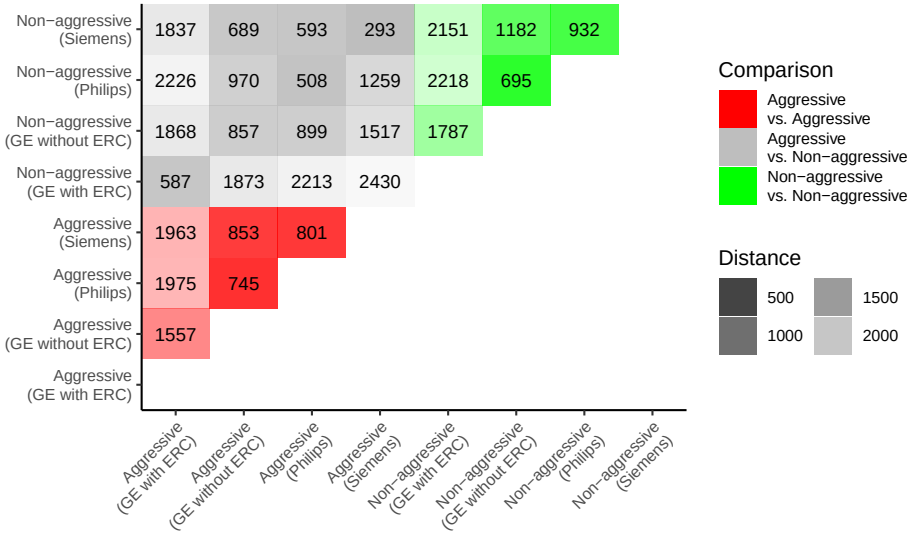


Figure 10.16: Optimal transport dataset distance between different data subsets. The colours correspond to different aggressiveness comparisons and the transparency of each grid cell corresponds to the distance between data subsets (higher values imply greater dissimilarity).

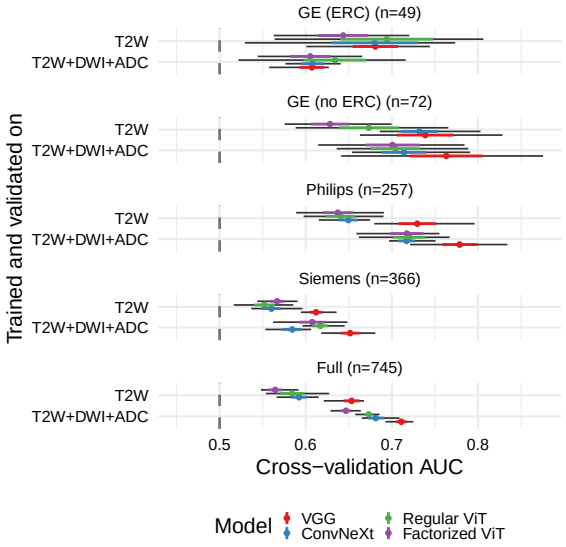


Figure 10.17: Cross validation area under the curve (AUC) of different models on different manufacturer datasets.

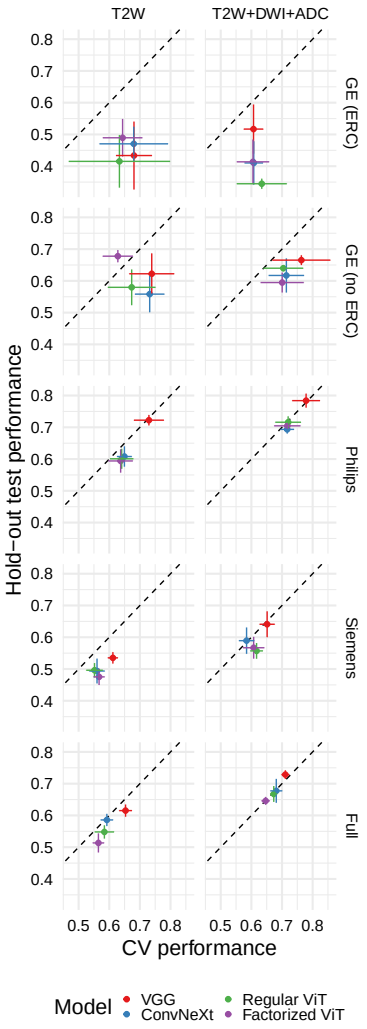


Figure 10.18: Comparison of cross-validated (CV) and test area under the curve (AUC). Each point represents the average AUC, whereas the vertical and horizontal error bars represent the mean with the addition and subtraction of the standard error, respectively. The diagonal dashed line represents equality between both axes.

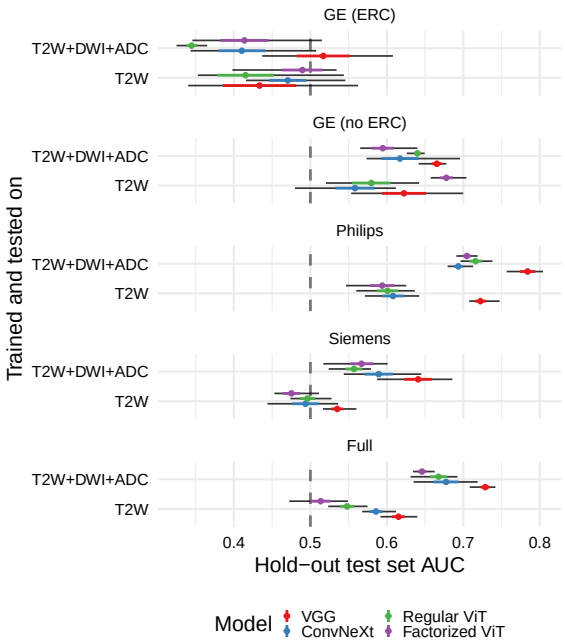


Figure 10.19: Test area under the curve (AUC) of different models on different manufacturer testing datasets.

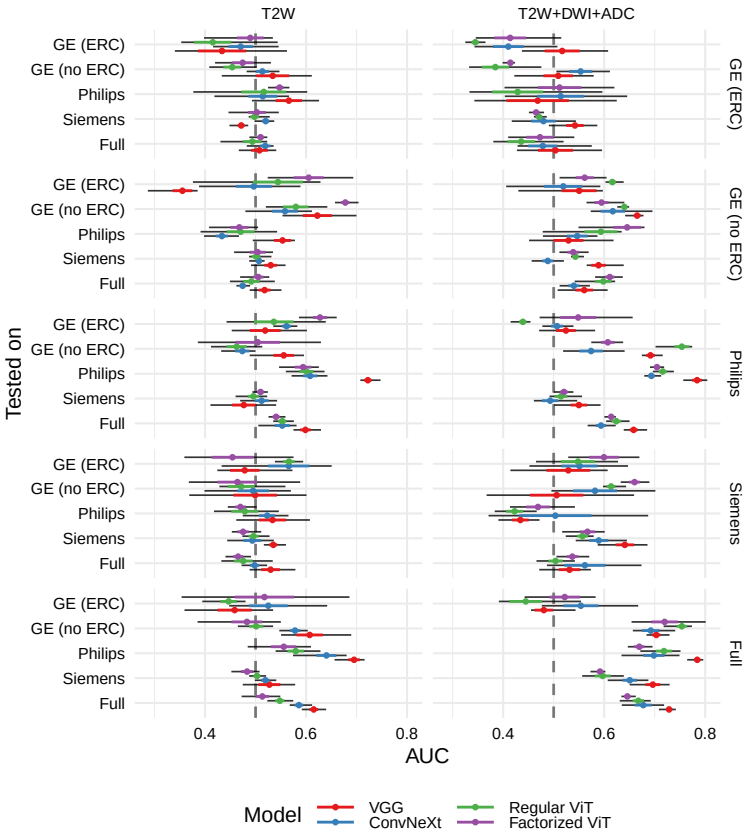


Figure 10.20: Test AUC of models trained and tested on different scanners. The y-axis refers to the data used to train each model and the y-facet (text on the right side of the image) refers to the data used to test each model.

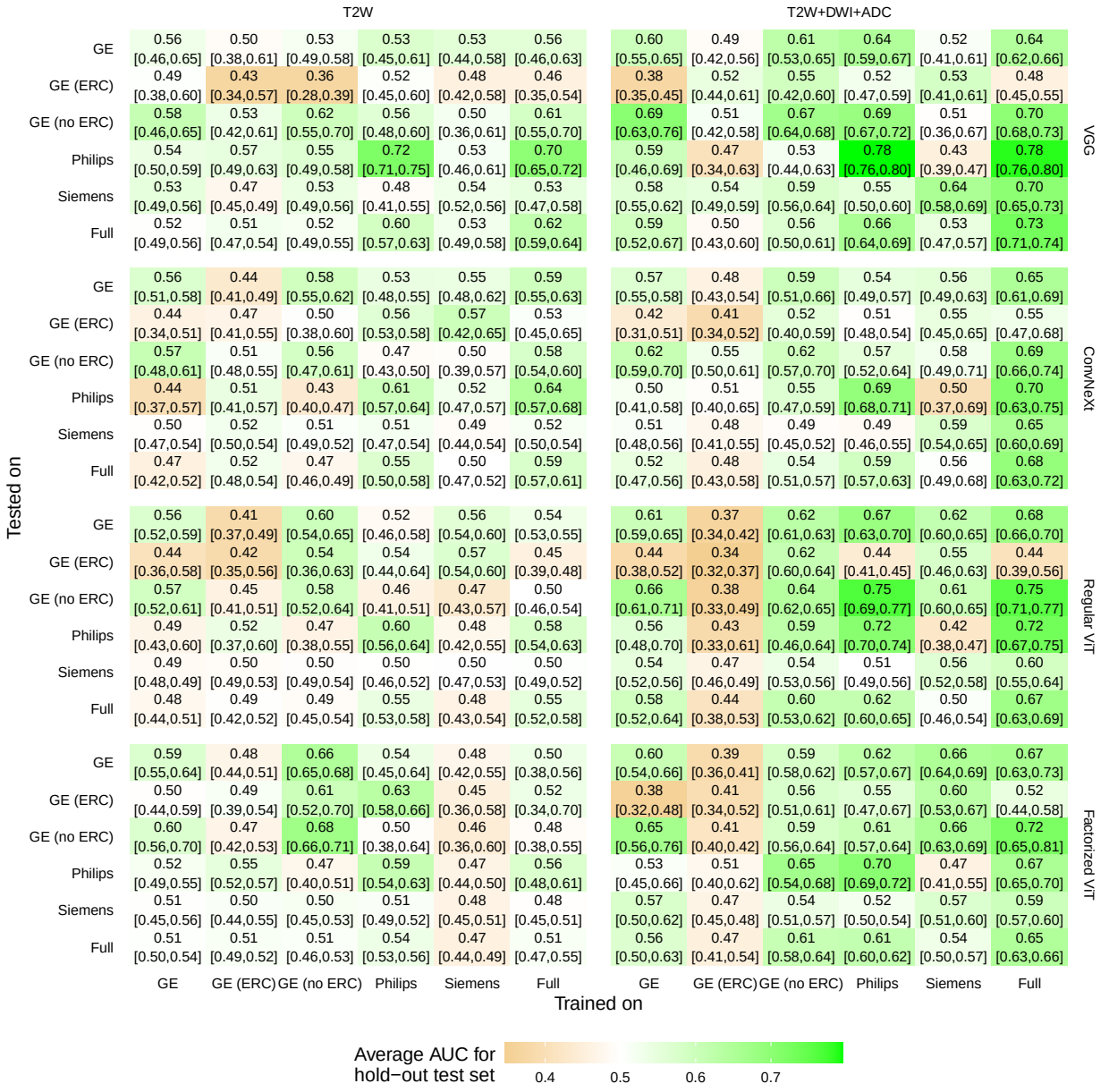


Figure 10.21: Test AUC of models trained and tested on different scanners. The text corresponds to the average, minimum and maximum AUC values (minimum and maximum values are between brackets) and the colour corresponds to the average AUC value.

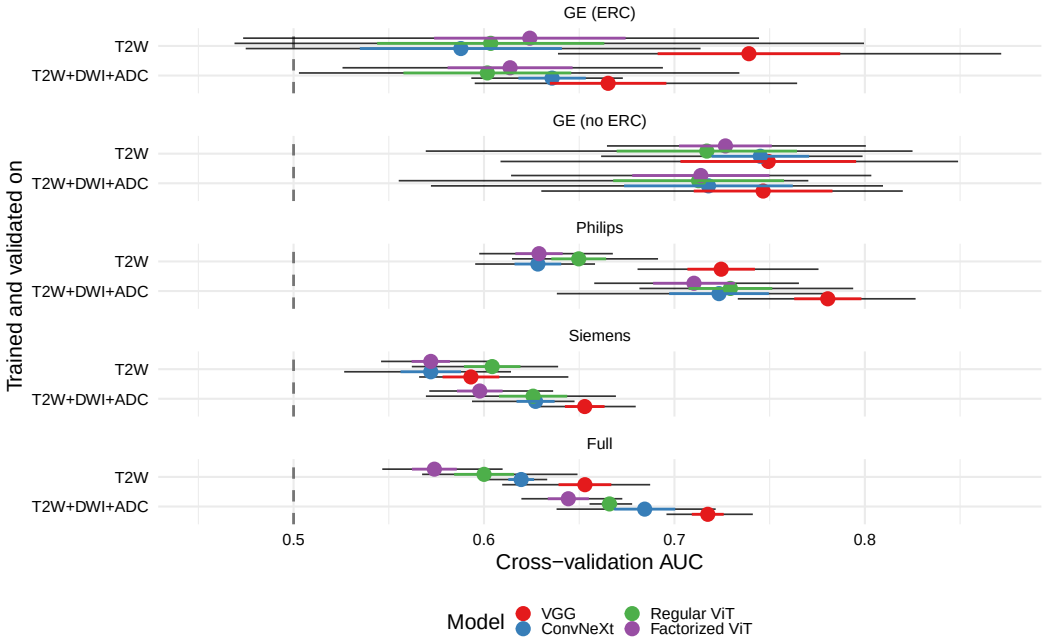


Figure 10.22: Cross validation area under the curve (AUC) of different hybrid models (mpMRI + clinical) on different manufacturer datasets.

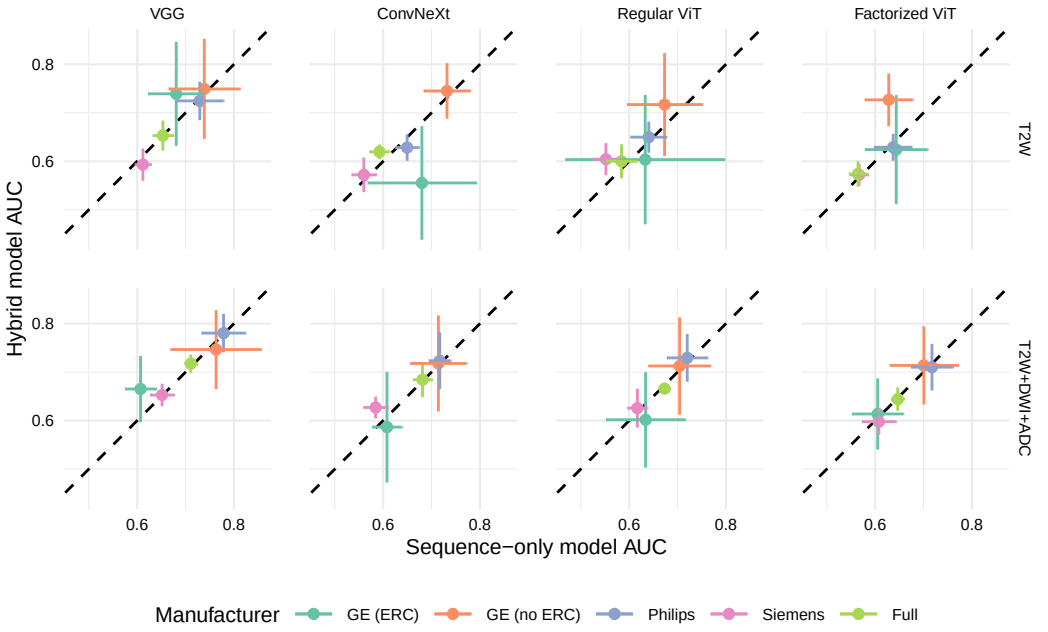


Figure 10.23: Comparison of AUC for sequence-only and hybrid models. Each point represents the average AUC, whereas the vertical and horizontal error bars represent the mean with the addition and subtraction of the standard error, respectively. The diagonal dashed line represents equality between both axes.

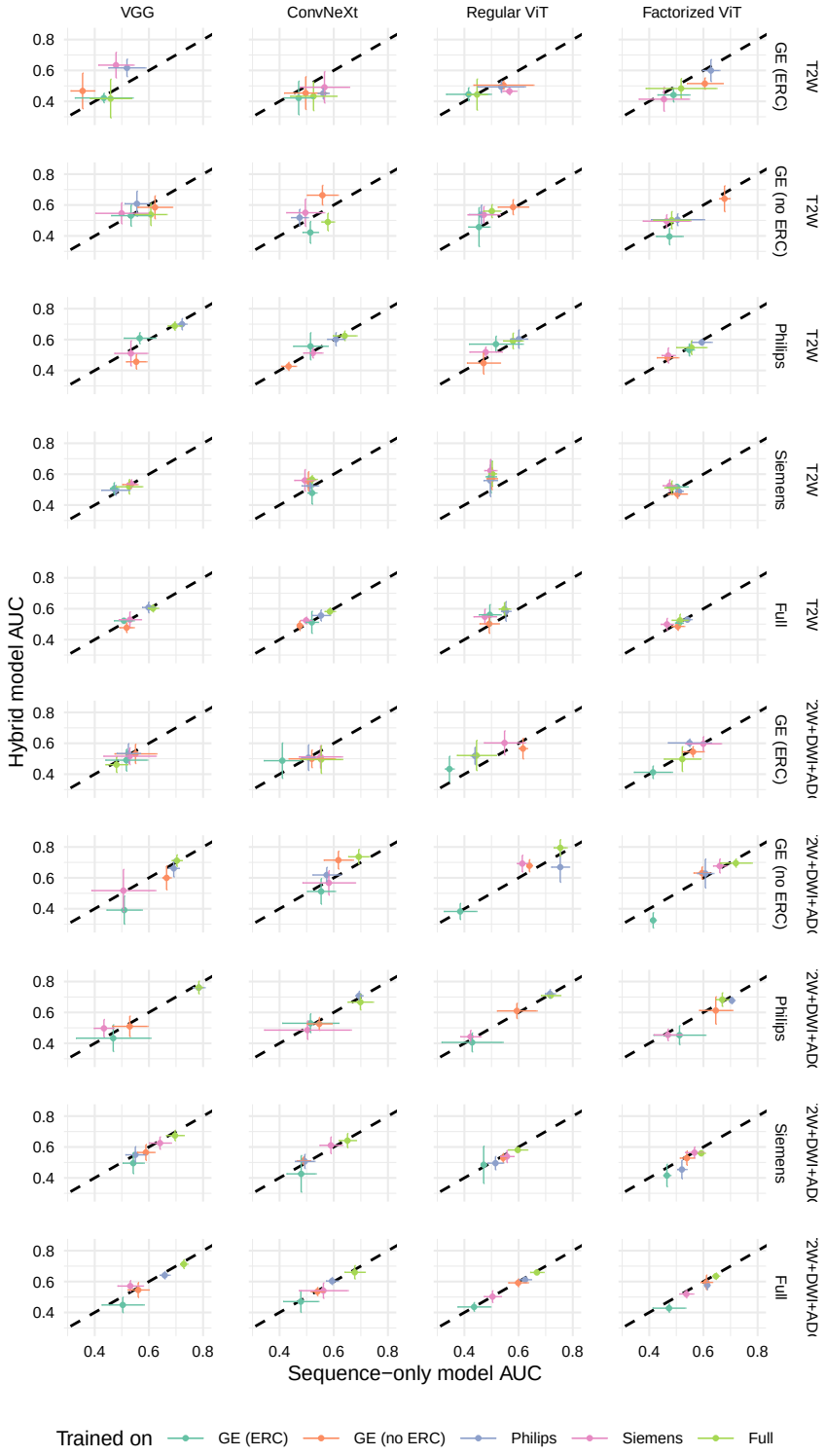


Figure 10.24: Comparison of hold-out test set AUC for sequence-only and hybrid models. Each point represents the average AUC, whereas the vertical and horizontal error bars represent the mean with the addition and subtraction of the standard error, respectively. The diagonal dashed line represents equality between both axes.

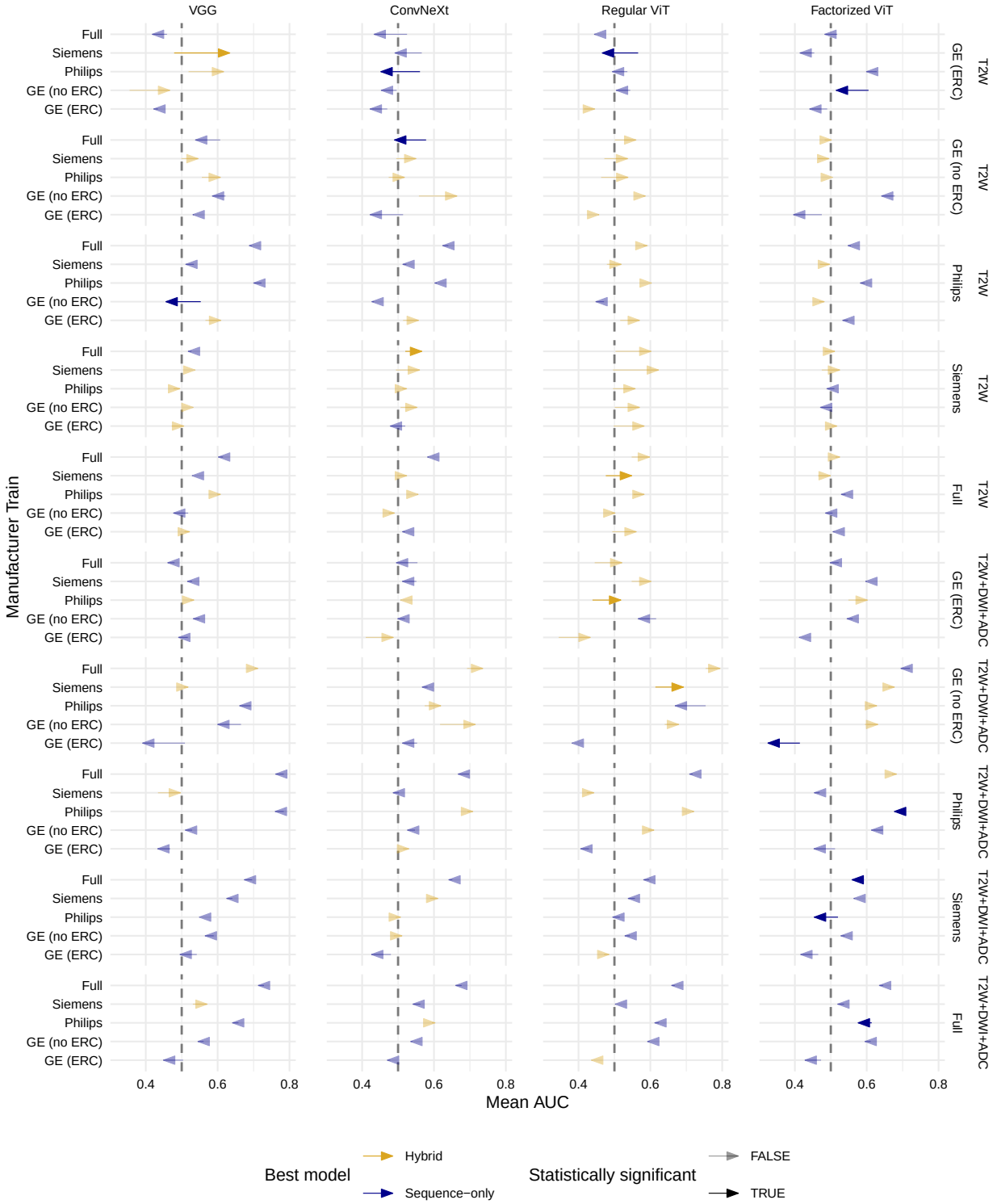


Figure 10.25: Test AUC of models trained and tested on different scanners. The arrow represents sequence-only and hybrid model performance (base and tip, respectively) and the colour of the arrow represents which model performed best. The y-axis refers to the data used to train each model and the y-facet (text on the right side of the image) refers to the data used to test each model.

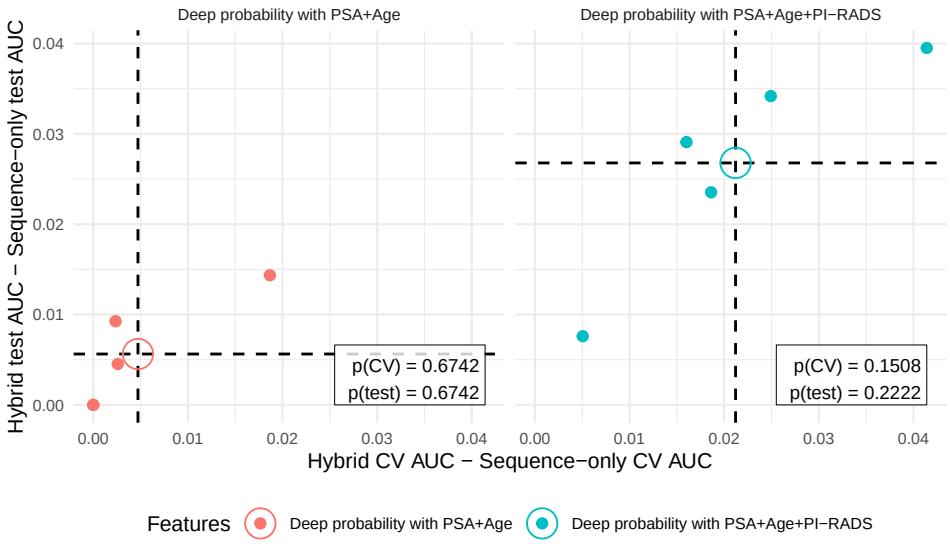


Figure 10.26: Difference between sequence-only and elastic net-regularized linear classification model AUC. Both CV (x axis) and test (y axis) AUC are represented, with the average value noted as a circle at the intersection of the dashed lines. The p-values shown in the figure were obtained using a one-sample Wilcoxon rank sum test.

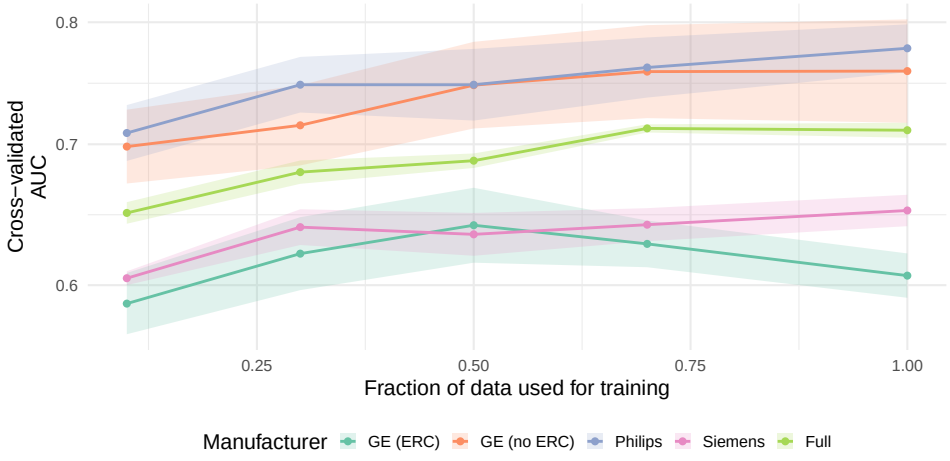


Figure 10.27: Learning curve for cross-validation AUC.

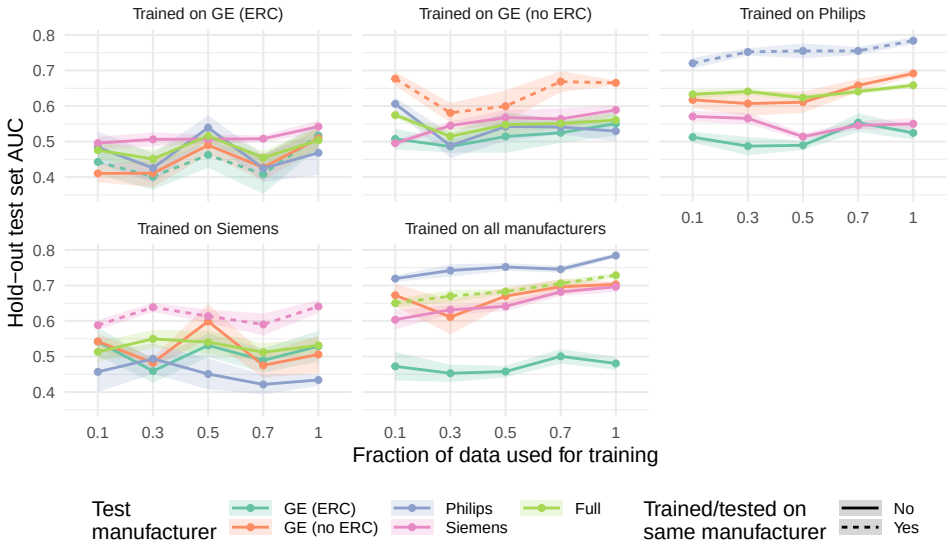


Figure 10.28: Learning curve for hold-out test set AUC.

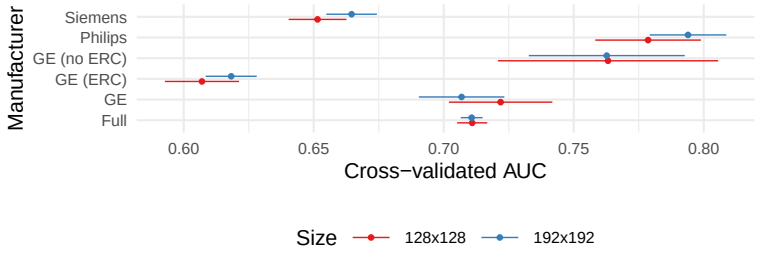


Figure 10.29: Impact of crop size on cross-validation AUC.

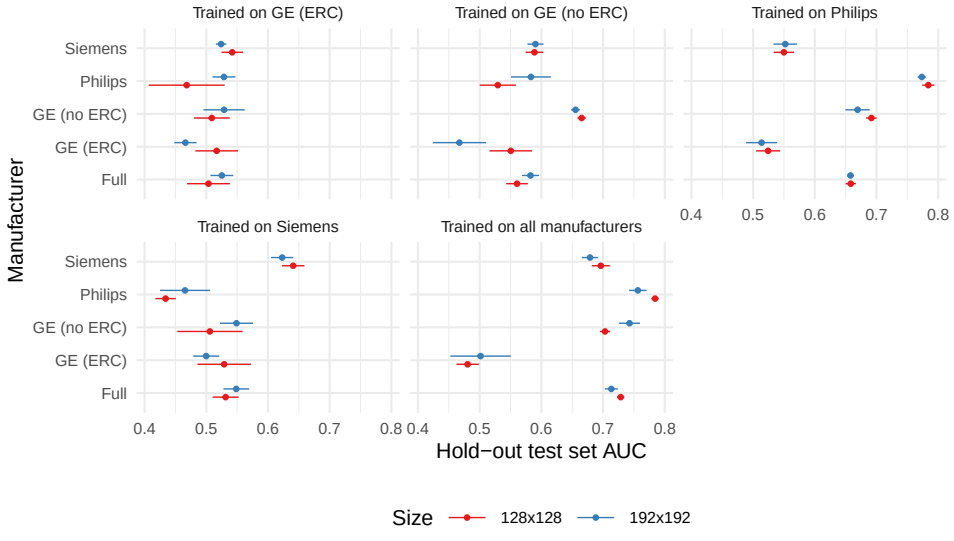


Figure 10.30: Impact of crop size on hold-out test set AUC.

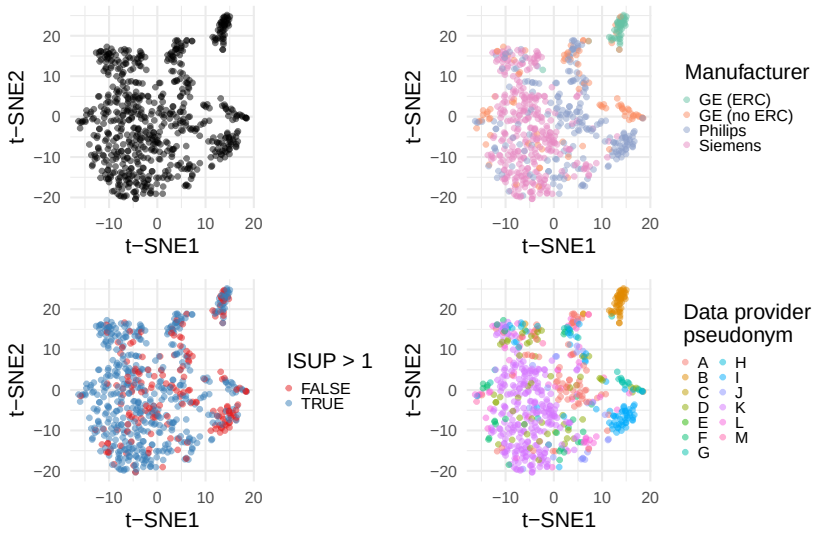


Figure 10.31: t-distributed stochastic neighbor embedding (t-SNE) visualization of all data (n=560 studies; first panel) and stratified by manufacturer (second panel) and by aggressiveness. The embedding is the same across panels and t-SNE1 and t-SNE2 represent the t-SNE dimensions.

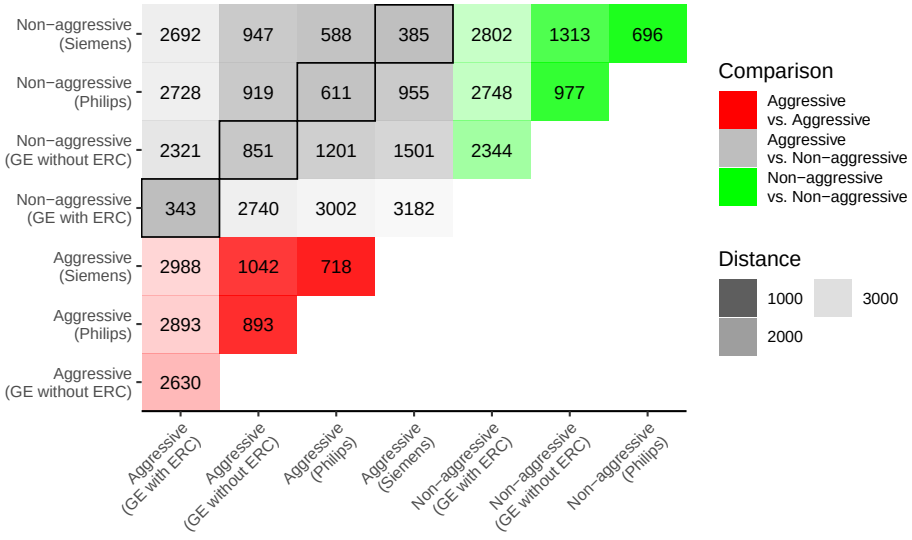


Figure 10.32: Optimal transport dataset distance between different data subsets. The colours correspond to different aggressiveness comparisons and the transparency of each grid cell corresponds to the distance between data subsets (higher values imply greater dissimilarity).

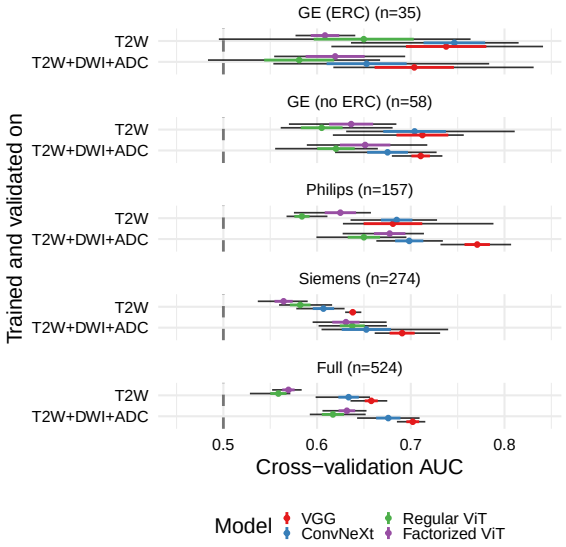


Figure 10.33: Cross validation area under the curve (AUC) of different models on different manufacturer datasets.

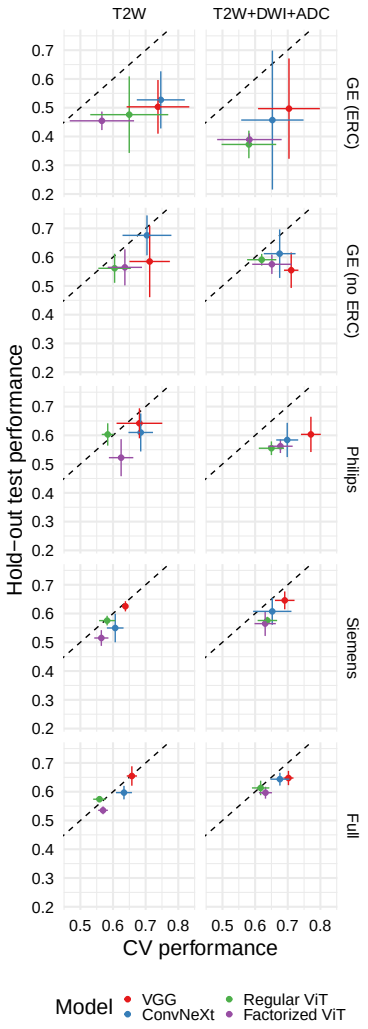


Figure 10.34: Comparison of cross-validated (CV) and test area under the curve (AUC). Each point represents the average AUC, whereas the vertical and horizontal error bars represent the mean with the addition and subtraction of the standard error, respectively. The diagonal dashed line represents equality between both axes.

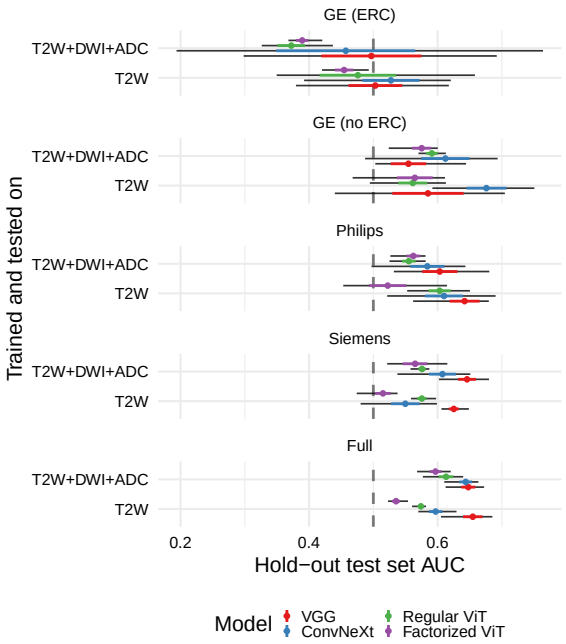


Figure 10.35: Test area under the curve (AUC) of different models on different manufacturer testing datasets.

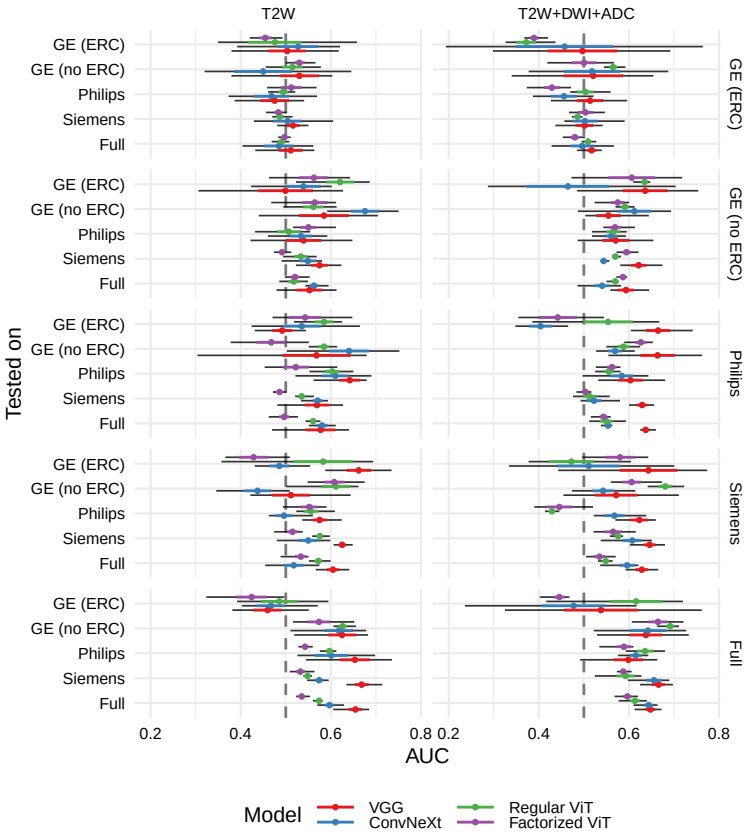


Figure 10.36: Test AUC of models trained and tested on different scanners. The y-axis refers to the data used to train each model and the y-facet (text on the right side of the image) refers to the data used to test each model.

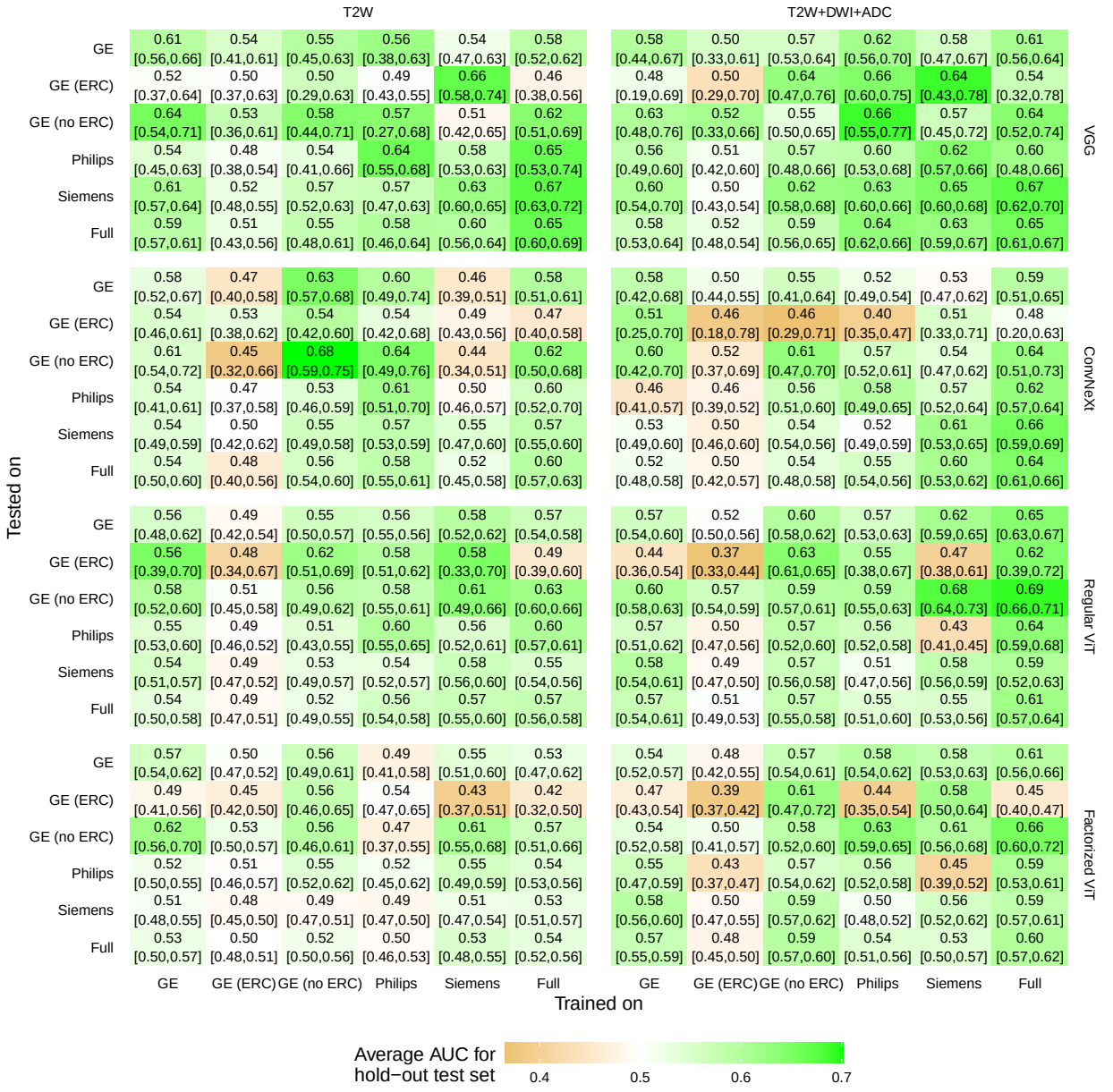


Figure 10.37: Test AUC of models trained and tested on different scanners. The text corresponds to the average, minimum and maximum AUC values (minimum and maximum values are between brackets) and the colour corresponds to the average AUC value.

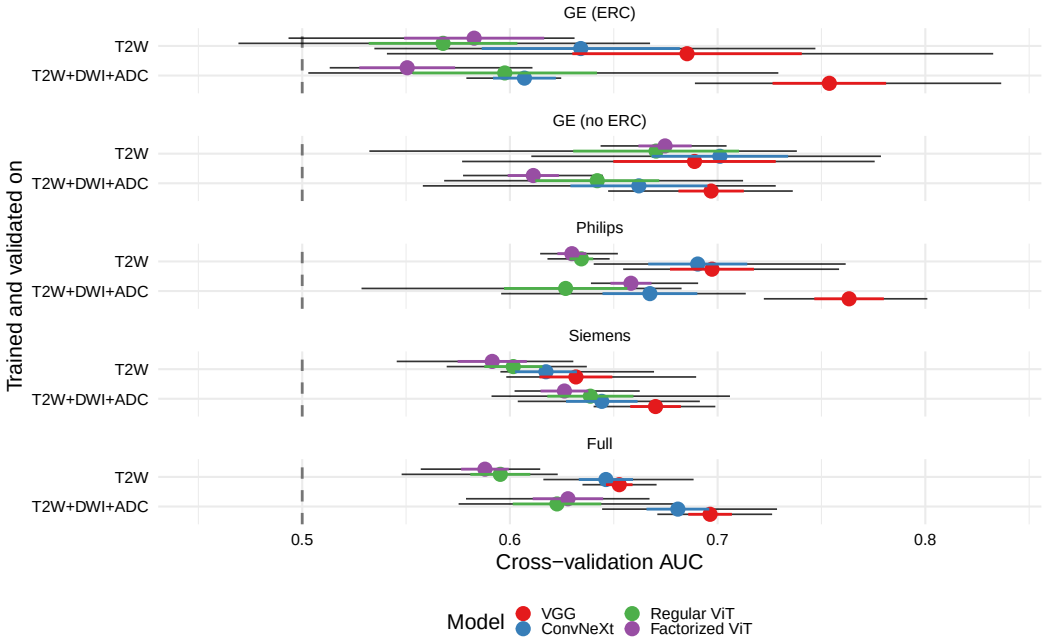


Figure 10.38: Cross validation area under the curve (AUC) of different hybrid models (mpMRI + clinical) on different manufacturer datasets.

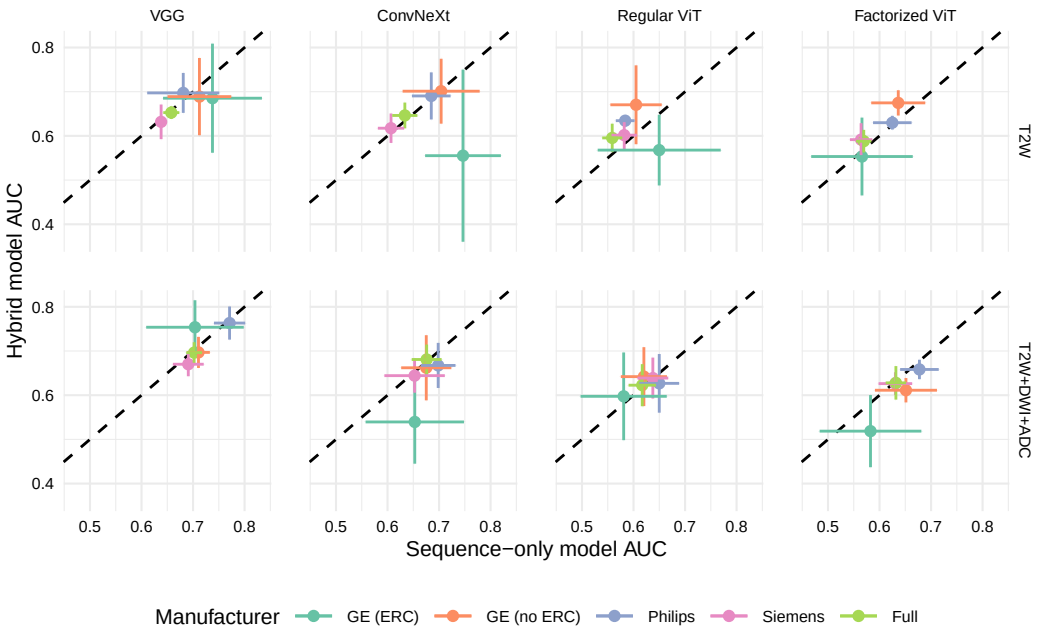


Figure 10.39: Comparison of AUC for sequence-only and hybrid models. Each point represents the average AUC, whereas the vertical and horizontal error bars represent the mean with the addition and subtraction of the standard error, respectively. The diagonal dashed line represents equality between both axes.

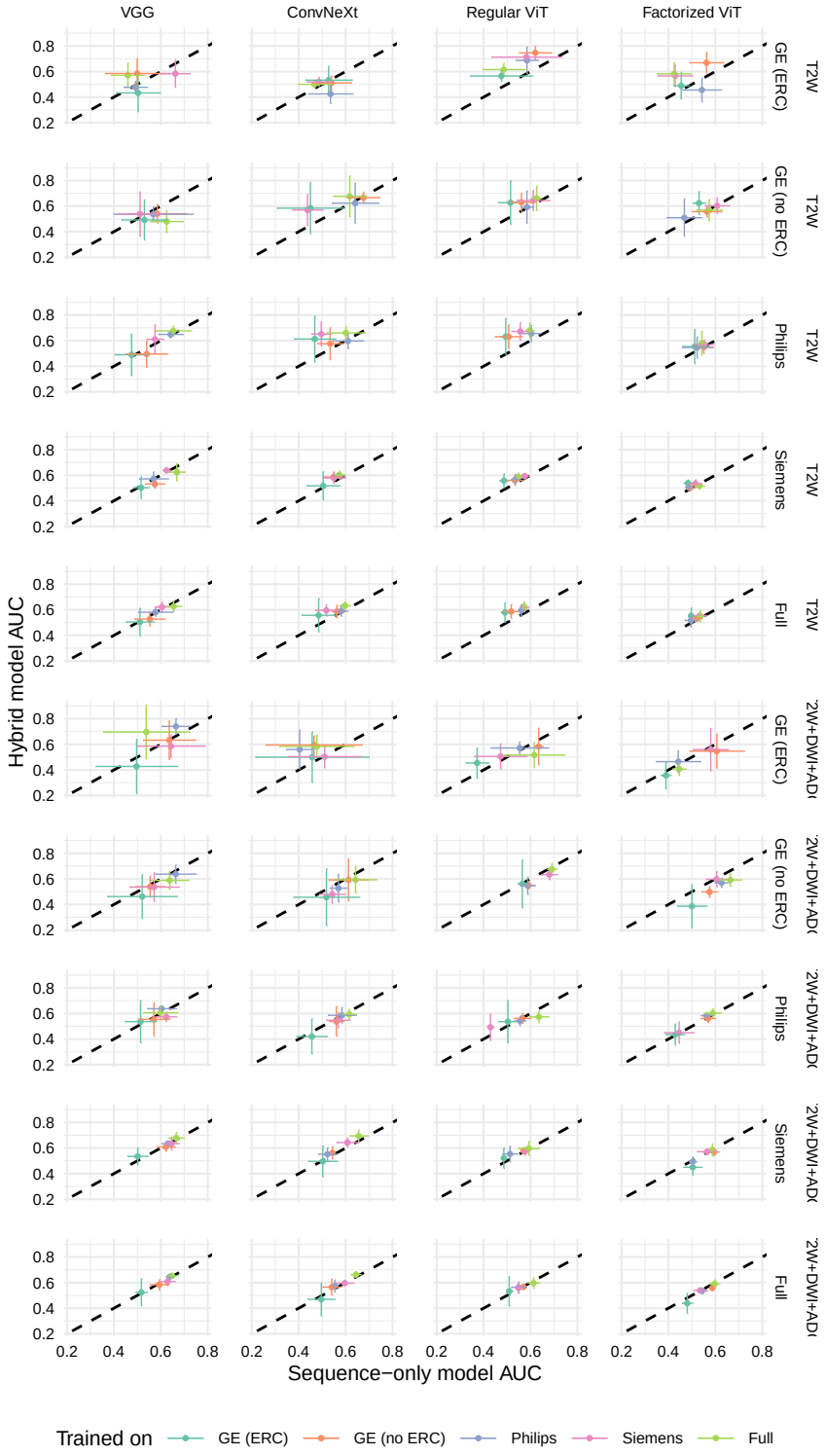


Figure 10.40: Comparison of hold-out test set AUC for sequence-only and hybrid models. Each point represents the average AUC, whereas the vertical and horizontal error bars represent the mean with the addition and subtraction of the standard error, respectively. The diagonal dashed line represents equality between both axes.

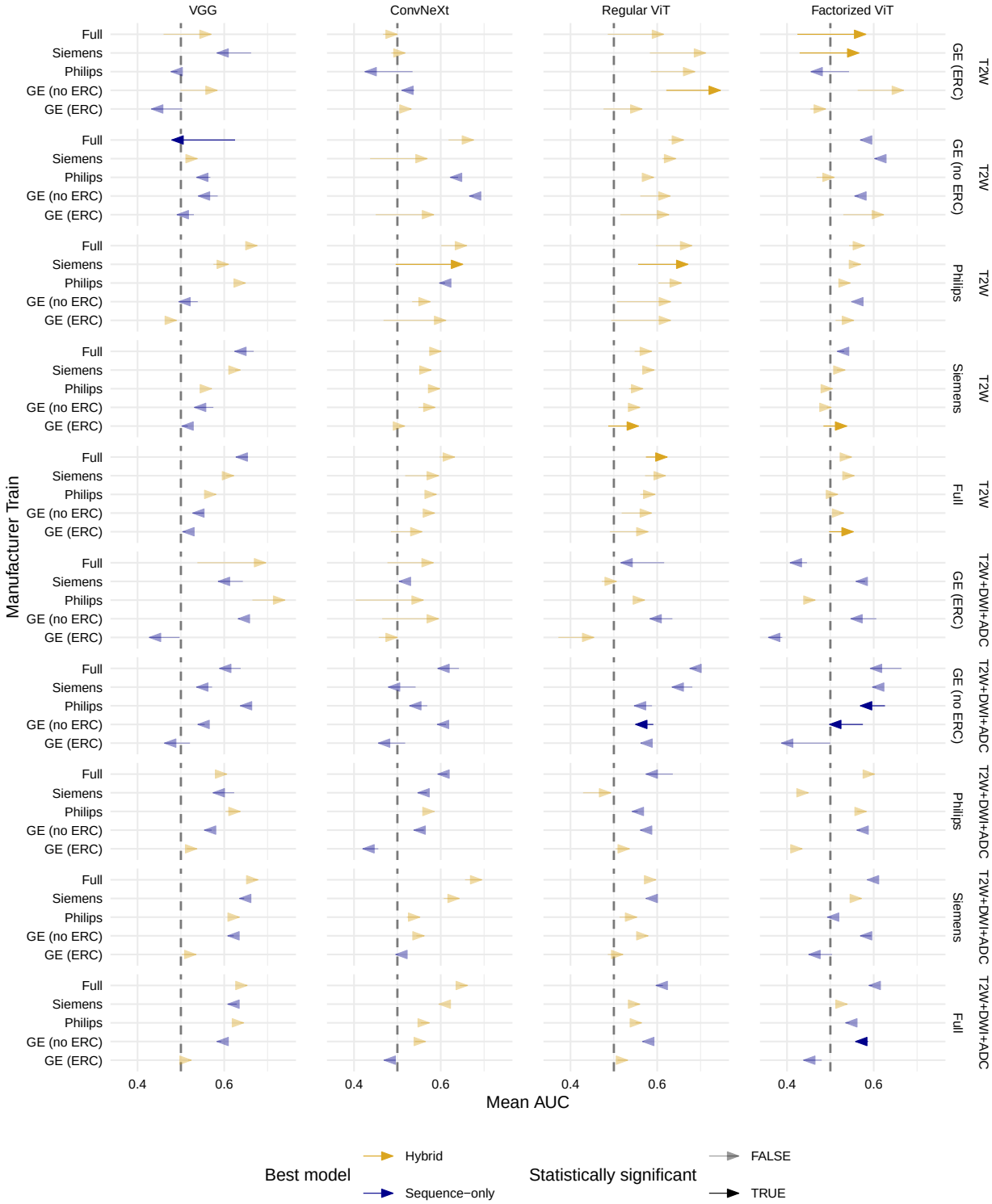


Figure 10.41: Test AUC of models trained and tested on different scanners. The arrow represents sequence-only and hybrid model performance (base and tip, respectively) and the colour of the arrow represents which model performed best. The y-axis refers to the data used to train each model and the y-facet (text on the right side of the image) refers to the data used to test each model.

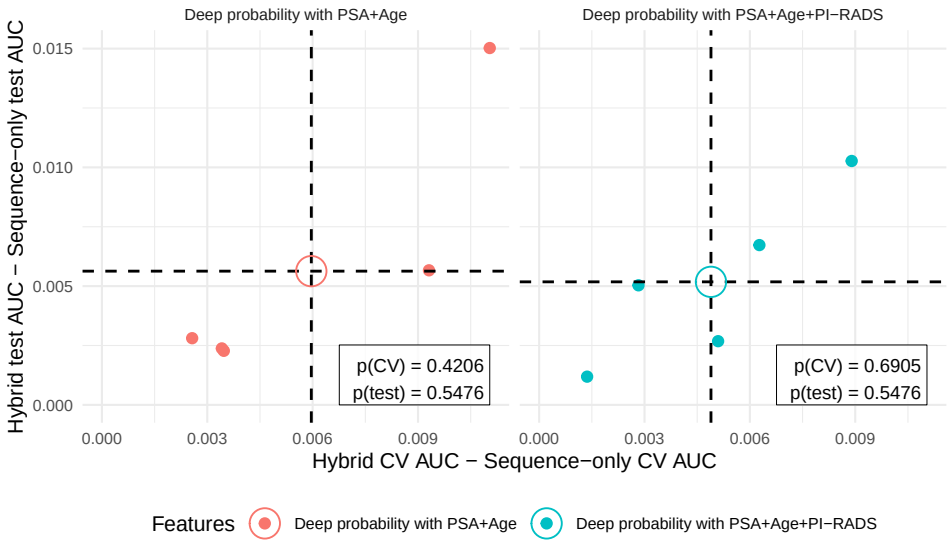


Figure 10.42: Difference between sequence-only and elastic net-regularized linear classification model AUC. Both CV (x axis) and test (y axis) AUC are represented, with the average value noted as a circle at the intersection of the dashed lines. The p-values shown in the figure were obtained using a one-sample Wilcoxon rank sum test.

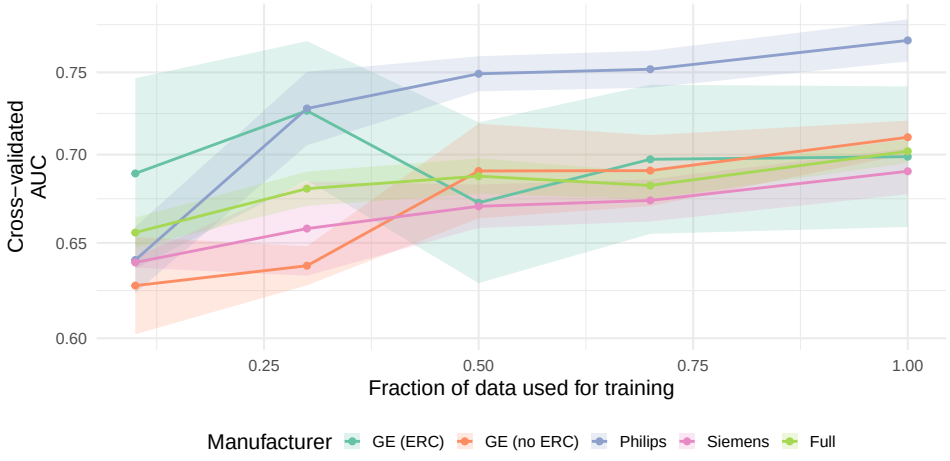


Figure 10.43: Learning curve for cross-validation AUC.

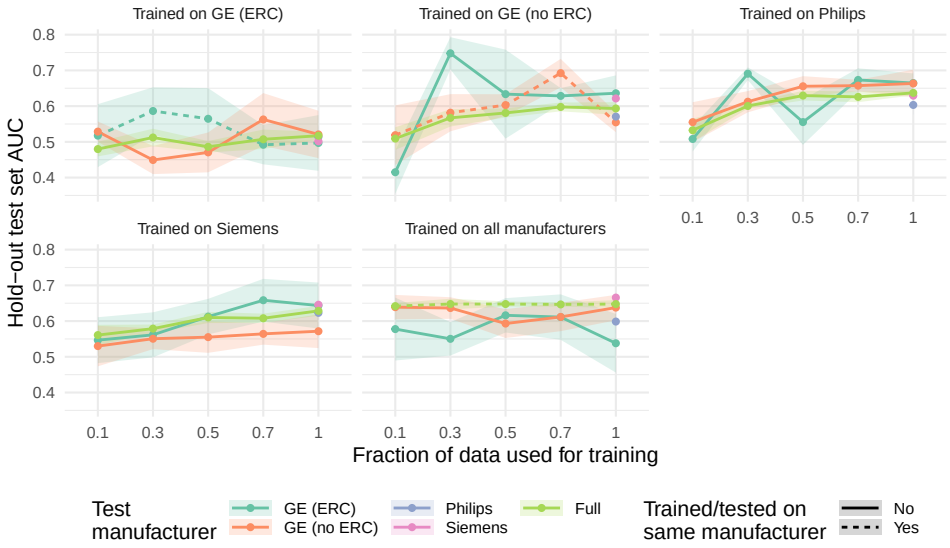


Figure 10.44: Learning curve for hold-out test set AUC.

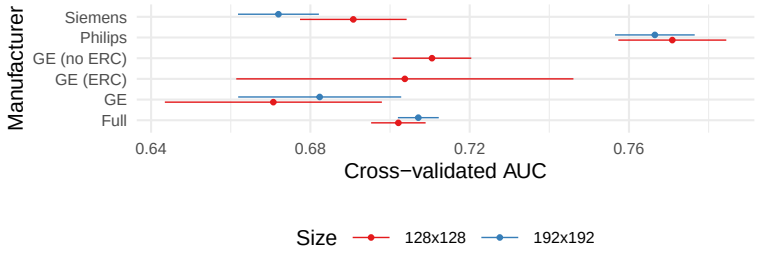


Figure 10.45: Impact of crop size on cross-validation AUC.

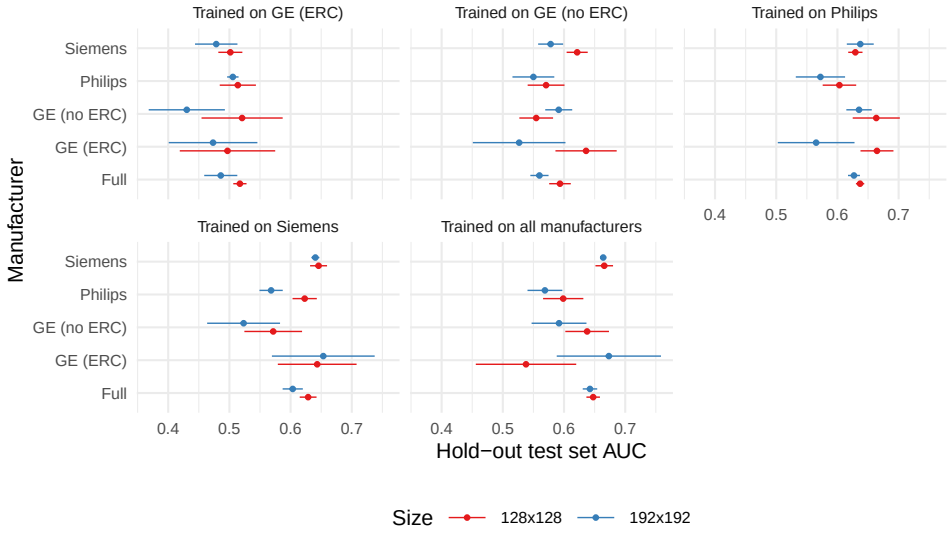


Figure 10.46: Impact of crop size on hold-out test set AUC.

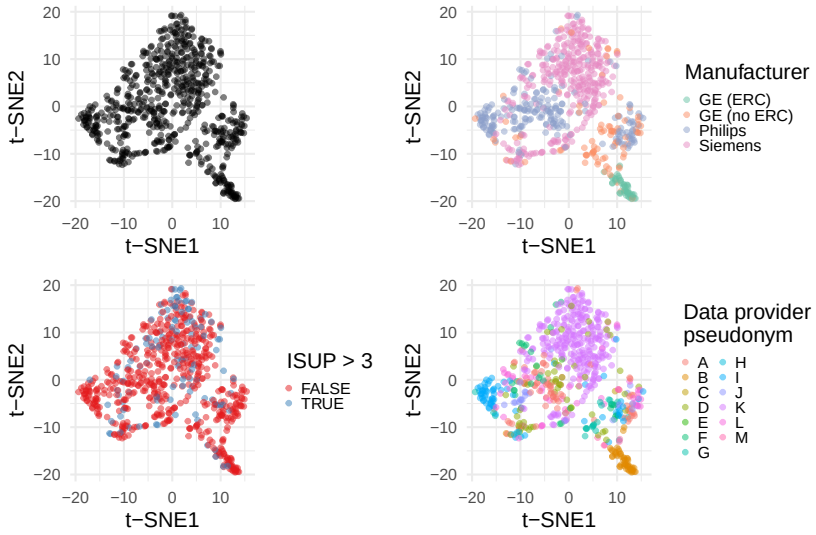


Figure 10.47: t-distributed stochastic neighbor embedding (t-SNE) visualization of all data (n=560 studies; first panel) and stratified by manufacturer (second panel) and by aggressiveness. The embedding is the same across panels and t-SNE1 and t-SNE2 represent the t-SNE dimensions.

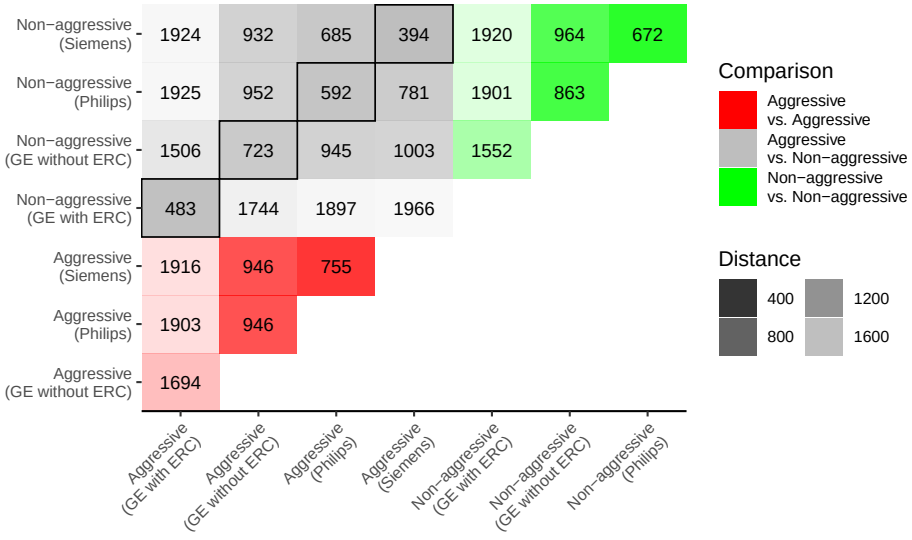


Figure 10.48: Optimal transport dataset distance between different data subsets. The colours correspond to different aggressiveness comparisons and the transparency of each grid cell corresponds to the distance between data subsets (higher values imply greater dissimilarity).

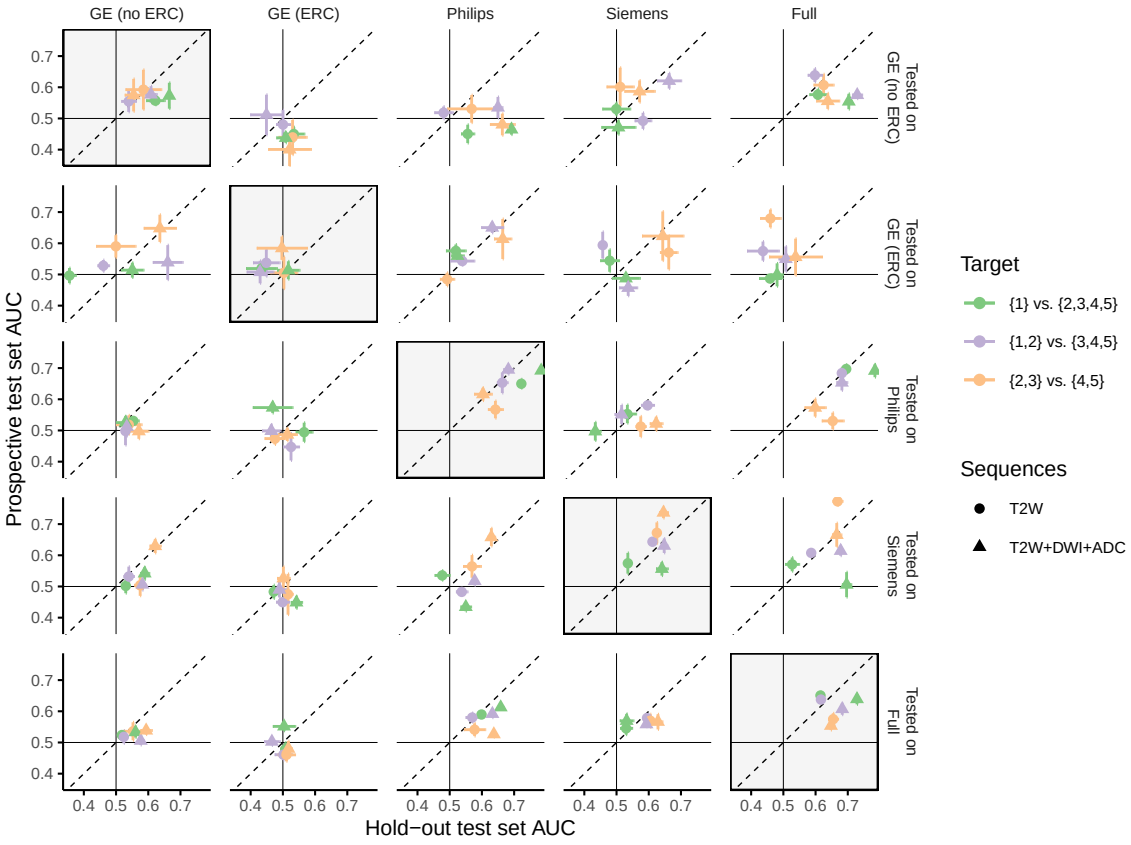


Figure 10.49: Hold-out and prospective test set performance of models trained and tested in different data subsets. Circles represent the mean of the 5-folds and the horizontal/vertical lines represent the standard error around the mean. Panels enclosed in a square represent data trained and tested on data from the same manufacturer.

10.4 Discussion

As we noted, direct comparisons of performance between targets is unwise and should be avoided. Nonetheless, we attempt here to summarize trends which are common to both target definitions.

Relevant Commonalities and Differences

In general, we observe the following to be applicable to all target types (we note exceptions, particularly for the intermediate vs. high risk target, whenever relevant):

1. **VGG models outperform other, more recent models.** This may be a consequence of more recent models requiring larger amounts of data as they have been developed with modern natural image datasets, typically comprising of hundreds of thousands or millions of images. This is particularly the case for ViT models

2. **mpMRI outperforms T2w-only models.** In general this is sensible — PI-RADS, the protocol used to evaluate prostate MRI studies recommends the use of both high b-value DWI and ADC to ensure the best possible results [2]. Interestingly, mpMRI models in the intermediate vs. high risk target suffered a considerable drop in performance when tested on a hold-out test set, oftentimes making them perform comparably to T2W-only models. This entails that there may be little information to be gained in functional sequences (ADC, DWI) when classifying between intermediate and high risk cases, or that this information is more complicated to learn for DL models

3. **Performance can drop when using a hold-out test set but training with multiple scanners overcomes this.** In general, we observe that generalizability — the ability of models to perform as well on a hold-out test set — is not perfect for models trained on data hailing from specific manufacturers (as noted earlier). The extent of this is quite variable and may be target dependent — low vs. possibly high Philips models suffer a minimal drop in performance when tested on Philips data, but this drop in performance is approximately 10% when evaluating possibly low vs. high Philips models on the hold-out test set. Nonetheless, by using data from multiple scanners during training, this can be overcome — indeed, no drop in performance is observed for Full models in the possibly low vs. high and low vs. possibly high targets. However and as noted, mpMRI models suffer a considerable drop in performance in intermediate vs. high risk cases. We posit that this may be due to the relatively smaller amounts of data which can lead to more dramatic cases of overfitting [21]

4. **Clinical data (age, PSA, PI-RADS) fails to improve the performance of DL models.** While multiple different models were assessed, we failed to see any improvement gains by using PSA, age or PI-RADS as additional predictors in a model. While this complicates future additions to this model as other clinical data is relatively more complicated to obtain, we note that this also shows that our DL models are learning the information that otherwise would require additional mpMRI interpretation to derive a PI-RADS score

5. **More data is likely to be beneficial.** Our learning curve analyses show that there is a general association between data volume and performance. This, however, is not always the case — for Siemens data using the low vs. possibly high target definition this dependence is not strict, particularly for the hold-out test set. The upstream causes of this difference is, to the best of our knowledge, hard to assess — indeed this is not the case when evaluating these models with the Full models, perhaps further highlighting the necessity of data variability when training these models

6. **A central crop is sufficient to contain the relevant signal.** One of the main concerns for this project was the definition of a crop that would not require additional input from clinical practitioners regarding the location of the prostate. Here, we show that using a central and relatively small crop is sufficient, validating an approach used in earlier studies with smaller datasets [23]. This indicates that, generally, we can expect the prediction-relevant signal to be centered around the prostate and in the middle of the image

7. **Manufacturer and protocol overpowers the feature landscape.** Finally, we note an important aspect of this analysis — while training on the complete set of manufacturers leads to models which are relatively performant, generalisable and transferable, the effect that manufacturer and protocol (or even data provider) have on the distribution of features at high dimensions is predominant when compared with classification. This highlights an important aspect that is likely to be crucial to the applicability of these models in new clinical settings and centers — a minimal amount of finetuning is a likely necessity for the incorporation of possible deviations in the feature space.

Considerations on the Utility of Models with Different Targets

Different objectives can be accomplished with each models trained on different targets using the first two targets:

1. **Low vs. possibly high.** Having a model capable of separating low risk cases from those that can potentially be high risk reduces patient discomfort and clinician burden — considering that ISUP=1 accounts for approximately one third of all cases, a well-performing model could potentially reduce the number of biopsies required and allow clinicians to focus on cases representing a higher risk for the patient

2. **Possibly low vs. high.** This case represents a different application — indeed, while it would be irresponsible to skip biopsy for patients with ISUP=1,2, it would be safer to make a clear statement about the necessity of more aggressive treatments. More concretely, in a scenario where the predicted target is possibly low, clinicians could advise patients to undergo a less aggressive form of treatment.

As such, **low vs. possibly high** is a target to reduce the necessary biopsies, whereas **possibly low vs. high** is a target to reduce the chance of overtreatment. Considering the relative performance of each model, we suggest that **low vs. possibly high** is likely to be the most impactful as it still allows for relevant patient stratification.

Considerations on Target Definition through ISUP Grading

In general, we tendentially observe better performance for the Low vs. possibly high (ISUP=1 vs. ISUP=2,3,4,5) when compared with the Possibly low vs. high (ISUP=1,2 vs. ISUP=3,4,5). This is reasonable — from a histopathological perspective, ISUP=1 is characterized as having no clear indications of pathogenicity, whereas ISUP¿2 should have some clear signs. On the other hand, ISUP=2 is characterized by some indicative signs that the lesion is growing, whereas ISUP¿3 has clear indications of abnormal prostate cells.

However, it should be noted that from a prognostic point of view this relationship is not as clear cut — while ISUP=1 and ISUP=2 are generally considered to stratify patients in terms of overall survival [34], the evidence for stratification in recurrence-free survival is mixed [34, 33, 19]. Additionally, there may be missing information in ISUP scores and relevant differences in grading between experts — a 2015 study has shown that ISUP=2 without cribriform structures may be similar to ISUP=1 [14], whereas another showed that reevaluation of Gleason scores leads to a different grading in approximately 20% of instances [40]. Indeed, ISUP is a useful, albeit noisy, grading and we believe this is consequential in terms of defining a target variable for prediction.

Considerations on Prospective Validation of Vendor-Specific Models

The prospective validation of vendor-specific models shows that the generalisation of some models can be improved by training models on specific subsets — this is particularly the case for Siemens models (excluding low vs. possibly high T2W+DWI+ADC models) and in some cases for Philips. In any case no particular trend is detectable, making the generic deployment of these models complicated and careful evaluation should be afforded to deployed models.

Chapter 11

Vendor Specific Deep Learning Models (Experiments Set 2)

11.1 Chapter Summary

For this task of deliverable 6.1, FPO expanded the analysis presented in deliverable 5.3 and trained the model that reached the highest performances in terms of detection rate (DR) and true negative rate (TNR) on subsets of data stratified by vendors, i.e., GE Medical System (GE), Philips, and Siemens. The selected model was the one that used an Unet in which the encoder was replaced with a Resnet50 and that received as input a 3-channel 2D image in which T2w, ADC map, and DWI images were concatenated. We tested both detection and segmentation on a set of retrospective data used as a test set, while we only evaluated detection on prospective data since no manual masks were provided for these patients. Training and test sets are detailed in the following section.

11.2 Methods

Data Description

We used the same subset of patients used for the master model to fine tune and test the networks. The dataset was composed of 371 retrospective patients who were splitted into a training set (n=312, 85% of cases of each vendor) and a test set (n=59, 15% of cases of each vendor). All patients had the manual segmentation of the tumor. The construction set was further divided into training set (80%) and validation set (20%), resulting in the following numbers:

- for GE: 100 for training (80 train + 20 validation) and 14 for testing
- for Philips: 125 for training (100 train + 25 validation) and 27 for testing
- for Siemens: 87 for training (70 train + 17 validation) and 18 for testing

Figure 11.1 shows the distribution of patients in the training and test set across centers and vendors.

Data Preprocessing

Before feeding the networks, some pre-processing steps were applied. First, in case T2w and hbDWI/ADC didn't have the same slice thickness, they were co-registered with the T2w image, using an elastic transformation and the mutual information as metric. Then, all sequences were cropped and resampled in order to have the same resolution and field of view (FOV), and the N4 bias correction filter was applied to the T2w image to correct inhomogeneities due to the coil. Finally, a *in-house* developed algorithm to automatically segment the prostate was applied and each sequence was cropped around the automatically segmented prostate area using a bounding box of 224x224 pixels to ease the network training and reduce the computational cost.

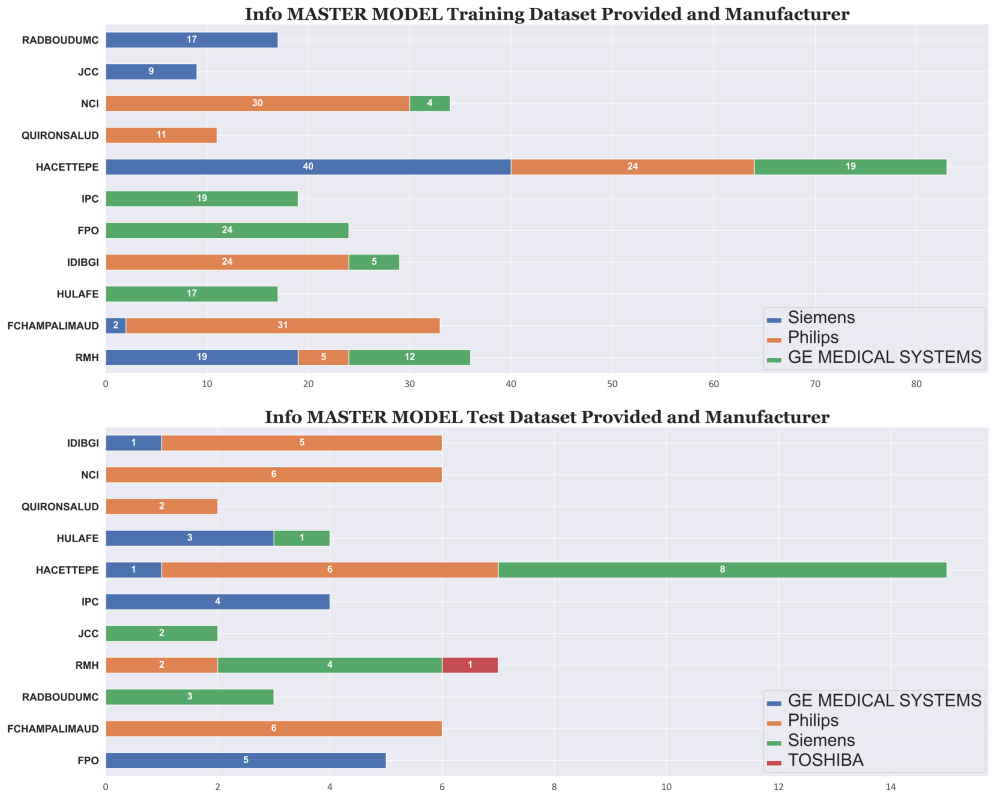


Figure 11.1: Distribution of patients in the training and test set across centers and vendors

Then, a pixel standardization using the z-score technique was applied at the patient level. Pixel intensities values were rescaled between 0 and 1, and all voxels outside the prostate area were set to 0. Finally, 2D slices were transformed into RGB images in which each RGB channel is represented by a different sequence (T2w, ADC, and hbDWI). Once the output images were generated, a binary threshold filter was applied to the probability maps returned by the networks to obtain the automatic masks of the tumors. Then, connected areas smaller than 50 voxels were discarded.

11.3 Results

Table 11.1 shows results obtained by the three models on the validation set.

Manufacturer	per-patient DR (%[rate])	Per-lesion DR (%[rate])	DSC	N_FP/N_Voxel
GE	80[16/20]	75[15/20]	0.72	0/1010
Philips	76[19/25]	76[19/25]	0.73	0/580
Siemens	88[15/17]	76[13/17]	0.64	0/473

DR = detection rate; DSC = Dice Similarity Score (DSC); N_FP/N_Voxel = Average Number of false positive lesions per patient/Median Number of false positive Voxel

Table 11.1: Results of vendor specific models on the validation set

Table 11.2 shows the performance obtained by the three Vendor Specific models (GE, Philips, and Siemens) on the retrospective test set compared to the results obtained by the master model stratified according to vendors.

Table 11.3 and 11.4 show the performances of the vendor specific models on prospective cases stratified per vendors, respectively on positive and negative patients. For each vendor specific model, we also reported

Manufacturer	per-patient DR (%[rate])	Per-lesion DR (%[rate])	DSC	N_FP/N_Voxel
Master Model on GE	79[11/14]	71[10/14]	0.62	1/1250
GE	79[11/14]	71[10/14]	0.60	1/472
Master Model on Philips	81[22/27]	70[19/27]	0.56	1/921
Philips	81[22/27]	70[19/27]	0.46	0/773
Master Model on Siemens	61[11/18]	50[9/18]	0.61	0/70
Siemens	67[12/18]	56[10/18]	0.48	0/168

DR = detection rate; DSC = Dice Similarity Score (DSC); N_FP/N_Voxel = Average Number of false positive lesions per patient/Median Number of false positive Voxel

Table 11.2: Results of vendor specific models on the test set

the performances obtained by the master model stratified per vendor, on the same subset of patients.

Vendor	Detection rate (%)	Detection rate (rate)
Master model on GE	86	79/91
GE	89	80/91
Master model on Philips	60	27/45
Philips	53	24/45
Master model on Siemens	70	23/33
Siemens	64	21/33

Table 11.3: Results obtained using vendor specific and master models on positive prospective patients

Vendor	True Negative Rate (%)	True Negative Rate (rate)
Master model on GE	29	9/31
GE	32	10/31
Master model on Philips	40	19/48
Philips	44	21/48
Master model on Siemens	57	12/21
Siemens	38	8/21

Table 11.4: Results obtained using vendor specific and master models on negative prospective patients

11.4 Discussion

Results obtained on GE model were statistically higher than those obtained on both Philips and Siemens ($p \leq 0.001$). This might be due to two different reasons: a) Siemens dataset is very small, and could not be sufficient to train a robust network, and b) most of Philips cases did not have the same slice thickness between T2w and ADC/hbDWI images, therefore image registration might have introduced some biases. Despite the challenges faced, i.e., prospective dataset size and heterogeneity in the different use cases, performances were similar between validation and test set, meaning that the networks were able to generalize to a different cohort. However, further research is needed to increase the sample size and optimize the registration between T2w and ADC/hbDWI images.

Chapter 12

Vendor Specific Deep Learning Models (Experiments Set 3)

12.1 Chapter Summary

Within the scope of deliverable 6.1, and T6.3, Radboudumc investigated the viability and efficacy of vendor-specific deep learning models for UC1 and UC2 in multiple parts. This section presents our study on harmonizing several state-of-the-art techniques from recent literature to develop a novel vendor-specific end-to-end 3D deep learning system. We trained, tuned, and tested this system at autonomously generating voxel-level detections of clinically significant ISUP $\geq$ 2 prostate cancers (i.e., UC1, UC2) in 2.7K patient examinations from clinical routine at two tertiary care centers based in the Netherlands. All images across both centers were acquired using MRI scanners from Siemens Healthineers. We investigated whether vendor-specific deep learning models trained using vendor-specific data from one center can generalize to data obtained using the same vendor at another center. We also determined the minimum number of training cases required to achieve expert radiologist performance in such a setting. Next, we expanded our dataset to 7.7K patient examinations by including non-annotated cases from the same distribution (same centers, same vendor). We investigated semi-supervised learning strategies and rule-based systems (using automated sparse information from diagnostic reports) to automatically create pseudo-labels for non-annotated cases. Such a method allows us to employ vastly larger training datasets for supervised models (within ProstateNET and in general) while reducing the manual workload for radiologists and the overall annotation budget. Our findings from these studies have been published as peer-reviewed papers at the `Medical Imaging Meets NeurIPS Workshops of the 34th and 35th Conference on Neural Information Processing Systems` [25, 27], and as full journal articles in `Medical Image Analysis` [28], `European Radiology` [9] and `Radiology: Artificial Intelligence` [3].

12.2 Methods

Single-Vendor, Multi-Center ISUP$\geq$2 Detection

Dataset The primary dataset was a cohort of 2436 consecutive prostate MRI studies (2317 patients) from Radboud University Medical Center (RUMC), acquired over the period January 2016–January 2018. All cases were paired with radiologically-estimated annotations of csPCa derived via PI-RADS v2. From here, 1584 (65%), 366 (15%), and 486 (20%) scans were split into training, validation, and testing (TS1) sets, respectively, via double-stratified sampling –preserving the same class balance (*benign* or *malignant*) while ensuring non-overlapping patients, between each subset of data. Additionally, 296 prostate MRI scans (296 patients) from Ziekenhuisgroep Twente (ZGT), acquired over the period March 2015–January 2017, were used to curate an external testing set (TS2). TS2 annotations included biopsy-confirmed histological ISUP grades. Patients were biopsy-naive men (RUMC: {median age: 66 yrs, IQR: 61–70}, ZGT: {median age: 65 yrs, IQR: 59–68}) with elevated levels of PSA (RUMC: {median level: 8 ng/mL, IQR: 5–11}, ZGT: {median level:

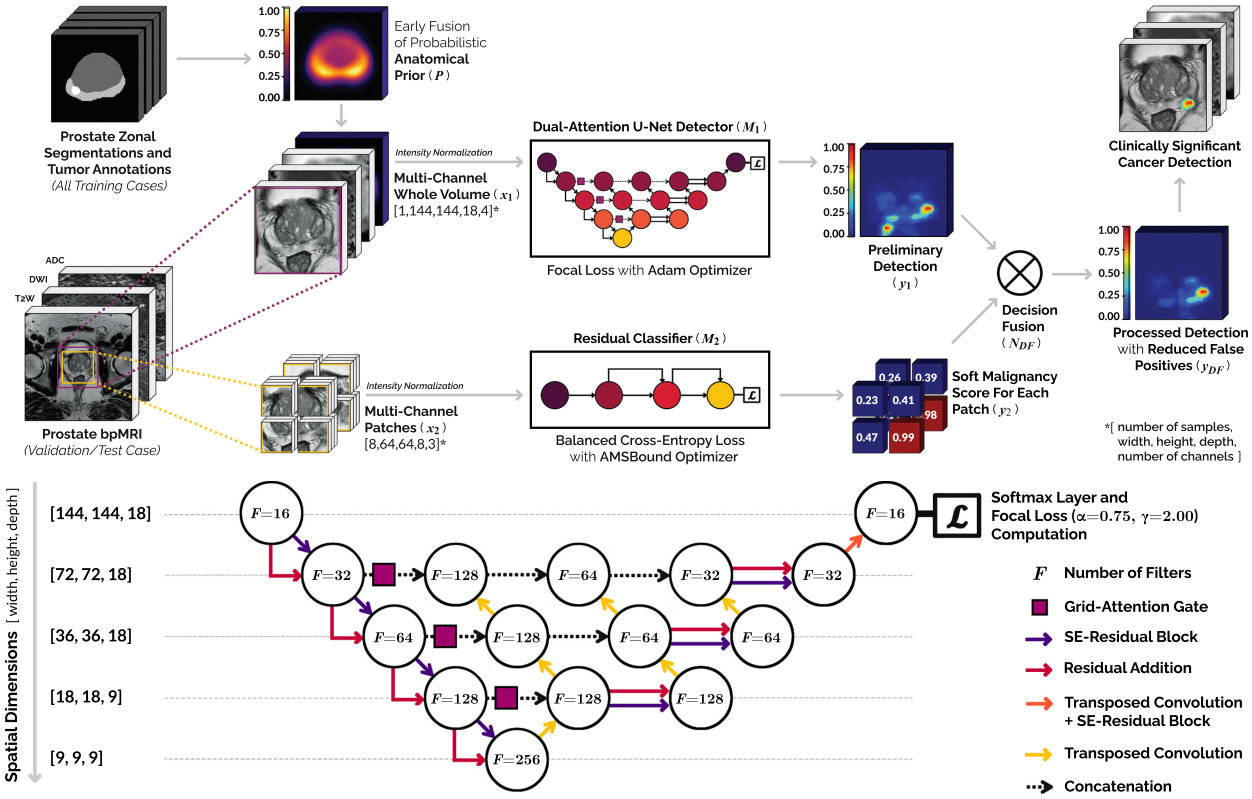


Figure 12.1: ***(top)*** Proposed end-to-end framework for computing voxel-level detections of csPCa in validation/test samples of prostate bpMRI. The model center crops two ROIs from the multi-channel concatenation of the patient's T2W, DWI, and ADC scans for inputting its detection and classification 3D CNN sub-models (M_1, M_2). M_1 leverages an anatomical prior P in its input x_1 to synthesize spatial priori and generate a preliminary detection y_1. M_2 infers on a set of overlapping patches x_2 and maps them to a set of probabilistic malignancy scores y_2. Decision fusion node N_{DF} aggregates y_1, y_2 to produce the model output y_{DF} in the form of a post-processed csPCa detection map with high sensitivity and reduced false positives. ***(bottom)*** Architecture schematic for the Dual-Attention U-Net (M_1). M_1 is a modified adaptation of the UNet++ architecture [44], utilizing a pre-activation residual backbone [8] with *Squeeze-and-Excitation* (SE) channel-wise attention mechanism [10] and grid-attention gates [30]. All convolutional layers in the encoder and decoder stages are activated by ReLU and LeakyReLU, respectively, and use kernels of size $3 \times 3 \times 3$ with L_2 regularization ($\beta = 0.001$). Both downsampling and upsampling operations throughout the network are performed via anisotropic strides. Dropout nodes ($rate = 0.50$) are connected at each scale of the decoder to alleviate train-time overfitting.

6.6 ng/mL, IQR: 5.1–8.7}). Imaging was performed on 3T MR scanners with surface coils (RUMC: {89.9% on Magnetom Trio/Skyra, 10.1% on Prisma}, ZGT: {100% on Skyra}; Siemens Healthineers, Erlangen). In this study, we used bpMRI sequences only, which included T2-weighted (T2W) and diffusion-weighted imaging (DWI). Apparent diffusion coefficient (ADC) maps, and high b-value DWI (b≥1400 s/mm^2) were computed from the raw DWI scans. All patient cases were read during regular clinical routine via PI-RADS v2. At RUMC, all cases were read by at least one of six radiologists (4- 25 years of experience), and difficult cases were jointly examined with an expert radiologist (25 years of experience with prostate MRI). At ZGT, all cases were read by two radiologists (6, 24 years of experience) and independently reviewed by two expert radiologists (5, 25 years of experience with prostate MRI) in consensus. In this study, we flagged any detected lesions marked PI-RADS 4 or 5 as csPCa(PR). All patients at ZGT underwent TRUS-guided biopsies performed by a urologist, blinded to the imaging results. In the presence of any suspicious lesions (PI-RADS> 2), patients underwent in-bore MRI-guided biopsies. All tissue samples were graded by general

pathologists and independently reviewed by an experienced uropathologist (25 years of experience), where cores containing cancer were assigned ISUP grades. Any lesion graded ISUP ≥ 2 was marked as csPCa$^{(GS)}$. All instances of csPCa$^{(PR)}$ and csPCa$^{(GS)}$ were carefully delineated on a voxel-level basis by trained students (6–18 months of expertise), under the supervision of an experienced radiologist (7 years of experience). Upon complete annotation, the RUMC and ZGT datasets contained 1527 and 210 *benign* cases, along with 909 and 86 *malignant* cases (≥ 1 csPCa lesion), respectively. Moreover, on a lesion-level basis, the RUMC dataset contained 1092 csPCa$^{(PR)}$ lesions (mean frequency: 1.21 lesions per *malignant* scan; median size: 1.05 cm^3, range: 0.01–61.49 cm^3), while the ZGT dataset contained 97 csPCa$^{(GS)}$ lesions (mean frequency: 1.05 lesions per *malignant* scan; median size: 1.69 cm^3, range: 0.23–22.61 cm^3).

For a secondary experiment on investigating semi-supervised learning strategies to leverage non-annotated training cases, the dataset above was expanded to 7756 examinations (6380 patients) from 9275 consecutive examinations (7430 patients) performed between January 2014 and December 2020 at Radboud University Medical Center. Within this overall set, the manually labeled development dataset ($D_{dev,labeled}$) comprised 3050 examinations performed between January 2016 and August 2018.

Model Architecture for Evaluating State-of-the-Art Deep Learning Methods Multi-class segmentation of prostatic transitional zones (TZ) and peripheral zones (PZ) were generated for each scan using a multi-planar, anisotropic 3D U-Net. Trained using a subset of 40 patient scans from the RUMC training cohort, this network achieved an average Dice Similarity Coefficient (DSC) of 0.90±0.01, 0.85±0.02 and 0.63±0.03 for whole-gland, TZ and PZ segmentation, respectively, over 5×5 nested cross-validation. We used these zonal segmentations to construct and apply an anatomical prior, as detailed in [27]. For this study, the goal of the zonal segmentations was to establish object-level, prior-to-image correspondence rather than voxel-level matching. Thus, high-quality segmentations with precise contour definitions were not mandatory. The architecture of our proposed end-to-end 3D deep learning system comprised of two parallel 3D CNNs (detection network, M_1; classification network for false positive reduction, M_2) followed by a decision fusion node N_{DF}, as shown in Fig. 12.1. We opted for anisotropically-strided 3D convolutions in both M_1 and M_2 to process the bpMRI data, which resembled multi-channel stacks of 2D images rather than full 3D volumes. Prior to usage, all acquisitions were spatially resampled to a common axial in-plane resolution of 0.5 mm^2 and slice thickness of 3.6 mm via B-spline interpolation. T2W and DWI channels were normalized to zero mean and unit standard deviation, while ADC channels were linearly normalized from [0,3000] to [0,1] in order to retain their clinically relevant numerical significance. Anatomical prior P, constructed using the prostate zonal segmentations and csPCa$^{(PR)}$ annotations in the training dataset, was encoded in M_1 to infuse spatial priori (as detailed in [27]). At train-time, M_1 and M_2 were independently optimized using different loss functions and target labels. At test-time, N_{DF} was used to aggregate their predictions (y_1, y_2) into a single output detection map y_{DF}. For more details on the model architecture, please refer to [28].

Model Architecture for Evaluating Pseudo Labels Our novel semi-supervised learning method leverages diagnostic reports to guide the generation of pseudo labels for semi-supervised prostate cancer detection. At a high level, our method consists of the four steps listed below (and as shown in Fig. 12.2). For more details, please refer to [3].

1. Train a supervised model with manually labeled cases, the *Teacher Model*.
2. Automatically parse the diagnostic reports using a natural language processing (NLP) algorithm to assess the number of clinically significant findings in unlabeled cases, n_{sig}.
3. Predict the cancer likelihood heatmap for unlabeled cases with the *Teacher Model*. Generate pseudo labels by iteratively extracting the n_{sig} most likely lesion candidate from the heatmap.
4. Train a semi-supervised model on the full dataset with manually and automatically labeled cases, the *Student Model*.

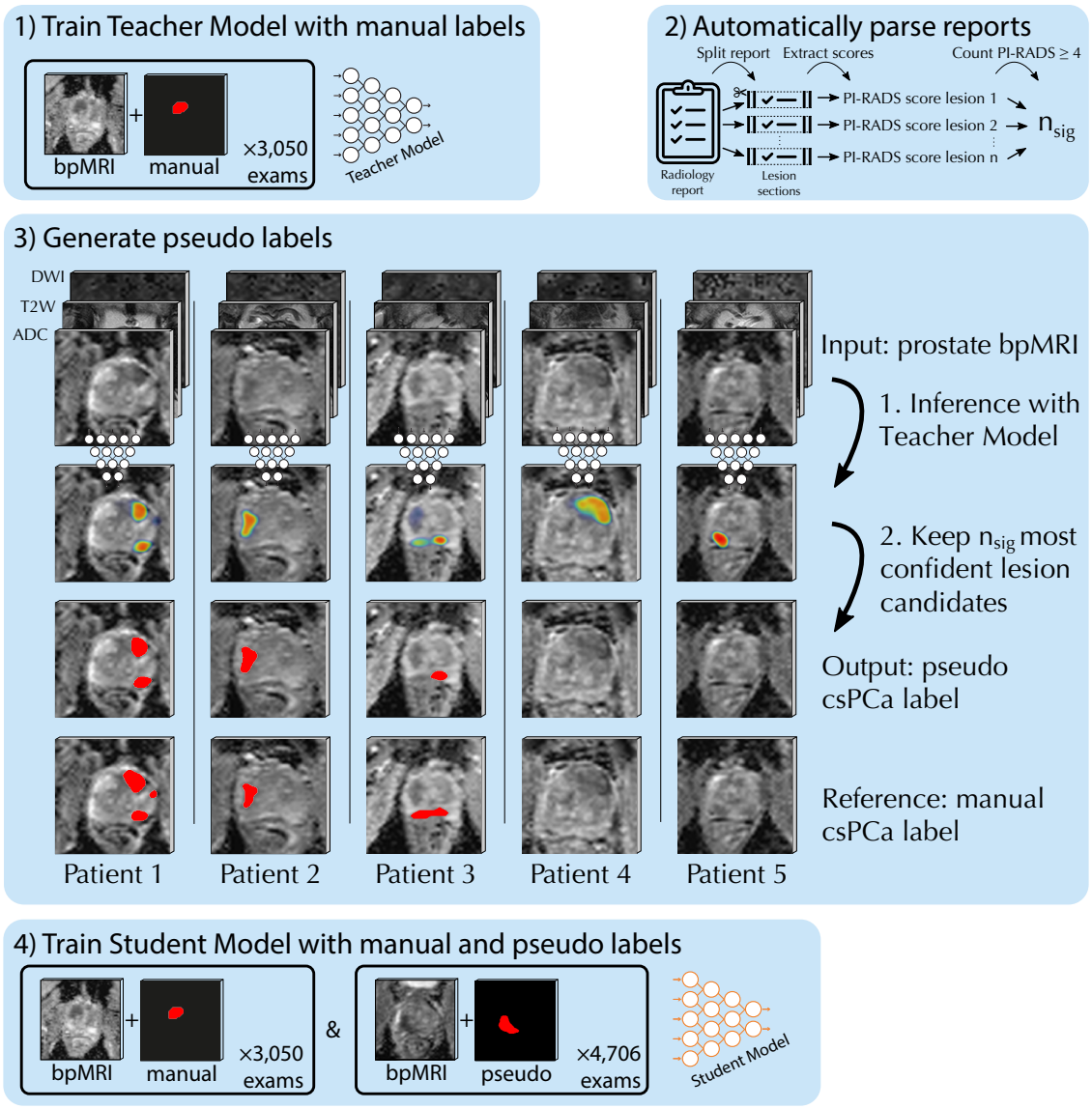


Figure 12.2: Overview of our novel semi-supervised learning method for malignancy detection. 1) Train the *Teacher Model* with manual labels. 2) Count the number of clinically significant lesions described in the report, . 3) Localize and segment the lesions, by keeping the most confident lesion candidates of the *Teacher Model*. 4) Train the *Student Model* with manual and pseudo labels.

12.3 Results

Single-Vendor, Multi-Center ISUP≥2 Detection

Effects of Architecture and Label Noise on Classification To determine the effect of selecting a specific classification architecture for M_2, five different 3D CNNs (ResNet-v2, Inception-ResNet-v2, Residual Attention Network, SEResNet, SEResNeXt) were implemented and tuned across their respective hyper-parameters to maximize patient-based AUROC over 5-fold cross-validation. Furthermore, each candidate CNN was trained using whole-images and patches, in separate turns, to draw out a comparative analysis surrounding the merits of spatial context versus localized labels. In the latter case, we studied the effect of τ (i.e. an ad-hoc hyperparameter to regular the decision fusion of predictions from M_1 and M_2) on patch-wise label assignment. We investigated four different values of τ: 0.0%, 0.1%, 0.5%, 1.0%; which correspond to

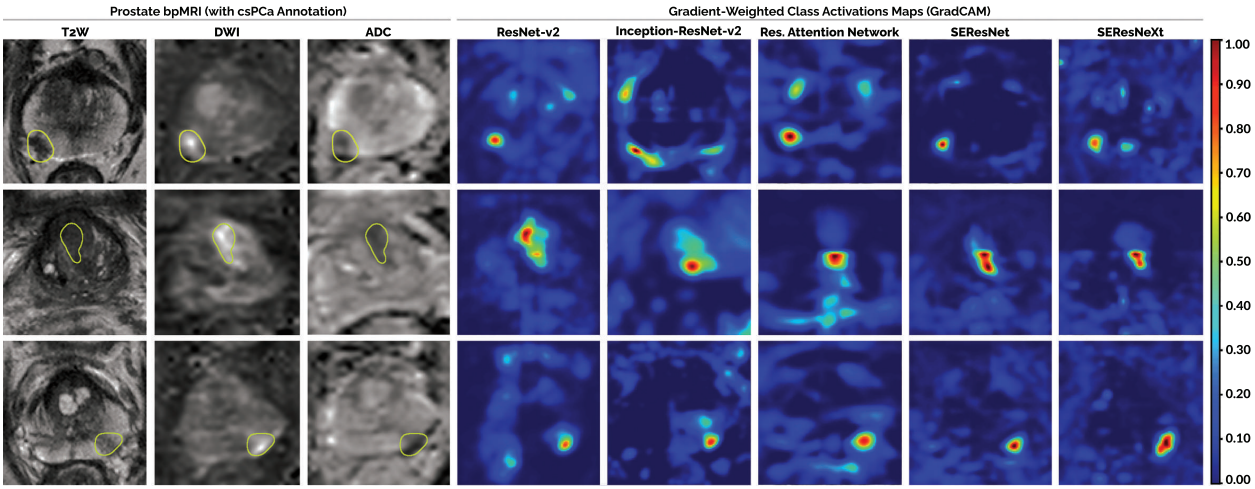


Figure 12.3: Model interpretability of the candidate CNN architectures for classifier M_2 at τ =0.1%. Gradient-weighted class activation maps (GradCAM) and their corresponding T2W, DWI and ADC scans for three patient cases from the validation set are shown above. Each case included a single instance of csPCa$^{(PR)}$ located in the prostatic TZ (***center row***) or PZ (***top, bottom rows***), as indicated by the yellow contours. Whole-image GradCAMs were generated by restitching and normalizing (*min-max*) the eight patch-level GradCAMs generated per case. Maximum voxel-level activation was observed in close proximity of csPCa$^{(PR)}$, despite training each network using patch-level binary labels only.

minimum csPCa volumes of 9, 297, 594 and 1188 mm^3, respectively. Each classifier was assessed qualitatively via 3D GradCAMs [31] to ensure adequate interpretability for clinical usage (as shown in Fig. 12.3). From the results noted in Table 12.1, we observed that the SEResNet architecture consistently scored the highest AUROC across every training scheme. However, in each case, its performance remained statistically similar ($p \geq 0.01$) to the other candidate models. We observed that a higher degree of supervision from patch-wise training proved more useful than the near 8× additional spatial context provided per sample during whole-image training. Increasing the value of τ consistently improved performance for all candidate classifiers (up to 10% in patch-level AUROC). While we attribute this improvement to lower label noise at train-time, it is important to note that the total csPCa volume per patient is typically small. If τ is set too large, not only are patch labels regulated, as intended, but multiple patch-level label swaps can compound to the point where entire patient cases can swap labels –resulting in an inaccurate evaluation.

Model	Params	AUROC (Whole-Image)	AUROC (Patches)			
			$\tau = 0.0\%$	$\tau = 0.1\%$	$\tau = 0.5\%$	$\tau = 1.0\%$
ResNet-v2 [8]	0.089 M	0.819±0.018	0.830±0.010	0.844±0.011	0.868±0.013	0.897±0.008
Inception-ResNet-v2 [35]	6.121 M	0.823±0.017	0.822±0.014	0.860±0.015	0.883±0.009	0.905±0.008
Res. Attention Network [41]	1.233 M	0.826±0.024	0.837±0.012	0.850±0.007	0.876±0.008	0.901±0.008
SEResNet [10]	0.095 M	0.836±0.014	0.842±0.019	0.861±0.005	0.886±0.008	0.912±0.008
SEResNeXt [10]	0.128 M	0.820±0.022	0.833±0.013	0.843±0.005	0.875±0.009	0.896±0.012

Table 12.1: Patient-based diagnosis performance of the candidate CNN architectures and training schemes (whole-image versus patch-wise training with four different values of τ to regulate label noise) for classifier M_2. Performance scores indicate mean of 5-fold cross-validation, followed by 95% confidence intervals estimated as twice the standard deviation.

Effects of Architecture and Clinical Priori on Detection We analyzed the effect of the M_1 architecture, in comparison to the four baseline 3D CNNs (U-SEResNet, UNet++, nnU-Net, Attention U-Net) that inspired its design. We evaluated the end-to-end 3D deep learning system, along with the individual

Model	Params	VRAM	Inference	Maximum Sensitivity {FP/Patient}	
				TS1 – csPCa$^{(PR)}$	TS2 – csPCa$^{(GS)}$
U-SEResNet [10]	1.615 M	0.94 GB	1.77±0.20 s	85.63%±4.70 {2.44}	84.42%±7.36 {2.26}
UNet++ [44]	14.933 M	2.97 GB	1.79±0.19 s	86.41%±4.54 {1.74}	82.28%±7.62 {2.25}
nnU-Net [11]	30.599 M	4.69 GB	2.09±0.03 s	84.34%±4.40 {1.44}	77.23%±8.14 {1.12}
Attention U-Net [30]	2.235 M	1.96 GB	1.77±0.19 s	90.46%±3.63 {2.07}	82.43%±7.79 {2.32}
Dual-Attention U-Net – M_1	15.250 M	3.01 GB	1.79±0.19 s	92.29%±3.24 {1.94}	84.60%±7.45 {2.31}
$M_1 \otimes M_2$	15.335 M	3.75 GB	1.89±0.23 s	92.29%±3.24 {1.69}	84.60%±7.45 {2.22}
$M_1 \otimes M_2$ with *Prior*	15.335 M	3.98 GB	1.90±0.23 s	93.19%±2.96 {1.46}	90.03%±5.80 {1.67}
CAD∗	40.069 M	9.85 GB	2.41±0.42 s	93.69%±3.13 {2.36}	91.05%±5.24 {1.29}

Table 12.2: Computational requirements (in terms of the number of trainable parameters, VRAM usage and the average time taken per patient scan during inference on a single NVIDIA RTX 2080 Ti) against the localization performance (in terms of the maximum csPCa detection sensitivity achieved and its corresponding false positive (FP) rate across both testing datasets) for each candidate detection system.

contributions of its constituent components (M_1, M_2, P), to examine the effects of false positive reduction and clinical priori. Additionally, we applied the ensembling heuristic of the nnU-Net framework [11] to create CAD∗, i.e. an ensemble model comprising of multiple CAD instances, and we studied its impact on overall performance. Each candidate setup was tuned over 5-fold cross-validation and benchmarked on the testing datasets (TS1, TS2). Fig. 12.4 and Table 12.2 show the patient-level performance, lesion-level performance and the computation requirements of the models. Fig. 12.5 shows the spatial congruence analysis of the predictions from our proposed deep learning system. Fig. 12.6 highlights six example successful and failure cases encountered by our proposed deep learning system in the external testing dataset TS2.

Effect of Training Dataset Size We observe that models trained with larger training datasets performed better than models trained using smaller training datasets (ranging from 50 to 1586 cases), as shown in Fig. 12.7. Notably, we observed that exponentially larger datasets are required to continue improving diagnostic performance –indicating the need for international consortium initiatives to curate datasets at scale and reach expert-level performance.

Viability of Pseudo Labels Our NLP score extraction algorithm identified the correct number of PI-RADS ≥ 4 lesions for 3024 out of the 3044 (99.3%) radiology reports in $D_{dev,labeled}$ (as shown in Fig. 12.8). Negative cases (PI-RADS ≤ 3) were identified with 99.7% accuracy. Report-guided semi-supervised learning (iteration 2) with 300 manual labels exceeded case-based AUROC performance of supervised learning with 2440 manually labeled exams (as shown in Fig. 12.9). Performance with 100 manual labels came close to supervised learning. Interpolation suggests that supervised performance is matched with 169 manual labels (annotation burden reduction of 14×). Report-guided semi-supervised learning (iteration 2) with 1000 manual labels exams exceeded lesion-based pAUC performance of supervised learning with 2440 manually labeled exams. Performance with 300 manual labels came close to supervised learning. Interpolation suggests the supervised performance is matched with 431 manual labels (annotation burden reduction of 6×).

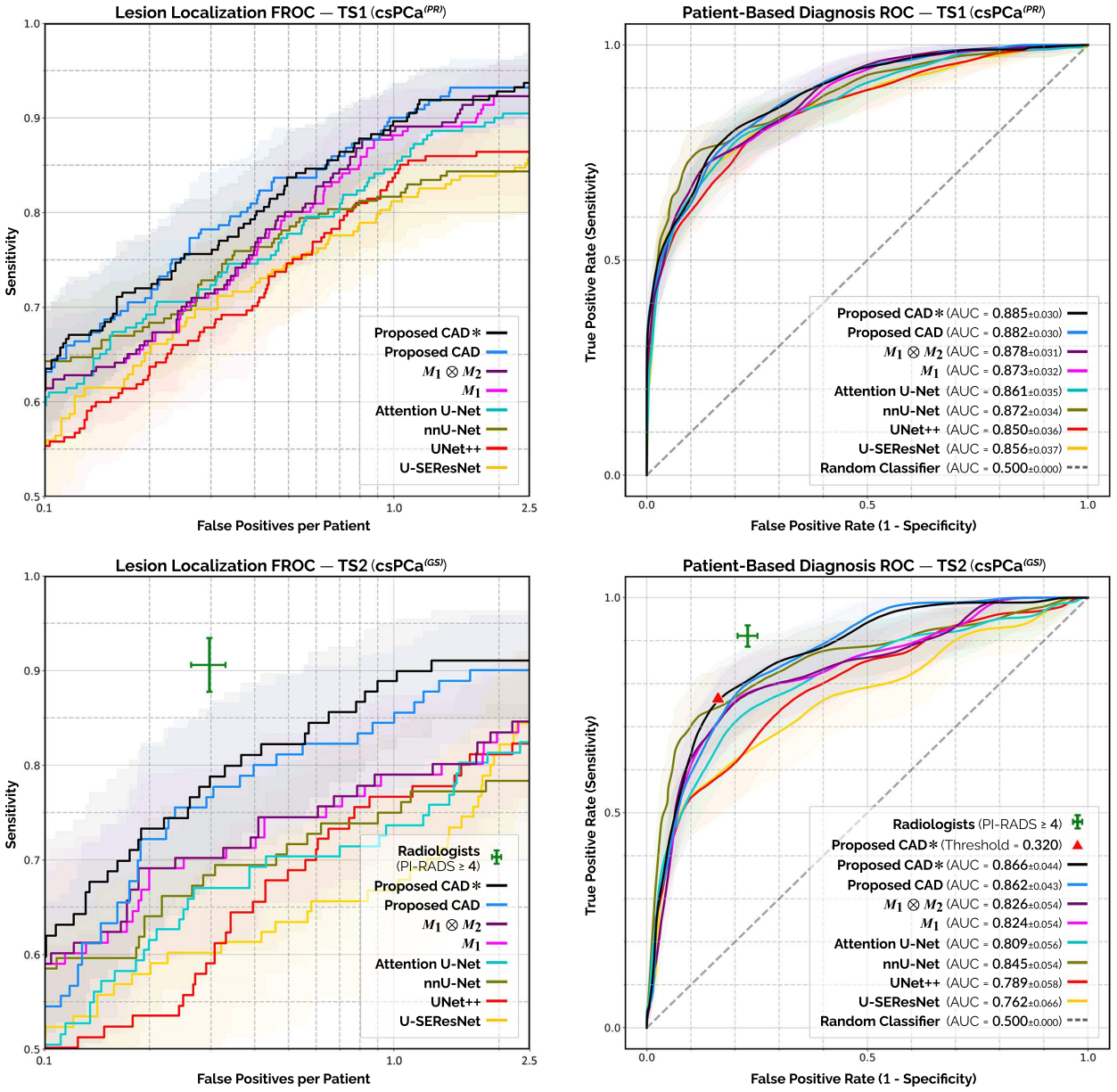


Figure 12.4: Lesion-level FROC (***left***) and patient-based ROC (***right***) analyses of csPCa$^{(PR)}$ (***top row***) / csPCa$^{(GS)}$ (***bottom row***) detection sensitivity against the number of false positives generated per patient scan using the baseline, ablated and proposed detection models on the institutional testing set TS1 (***top row***) and the external testing set TS2 (***bottom row***). Transparent areas indicate the 95% confidence intervals. Mean performance for the consensus of expert radiologists and their 95% confidence intervals are indicated by the centerpoint and length of the green markers, respectively, where all observations marked PI-RADS 4 or 5 are considered positive detections.

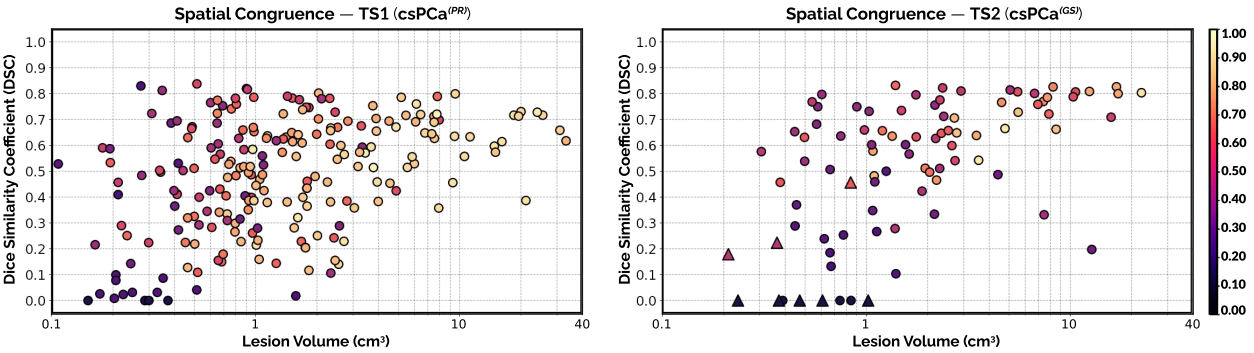


Figure 12.5: Distribution of per-lesion Dice Similarity Coefficient (DSC) (relative to csPCa lesion volume) for CAD∗ detections against the ground-truth annotations of csPCa$^{(PR)}$ in the institutional testing TS1 ***(left)*** and csPCa$^{(GS)}$ in the external testing set TS2 ***(right)***. All DSC values were computed in 3D for the model-specific operating point with maximum detection sensitivity (91.05±5.24%). Encoded color for each marker indicates its corresponding likelihood of malignancy, as predicted by CAD∗. Triangular markers for TS2 ***(right)*** indicate csPCa$^{(GS)}$ lesions missed by the radiologists.

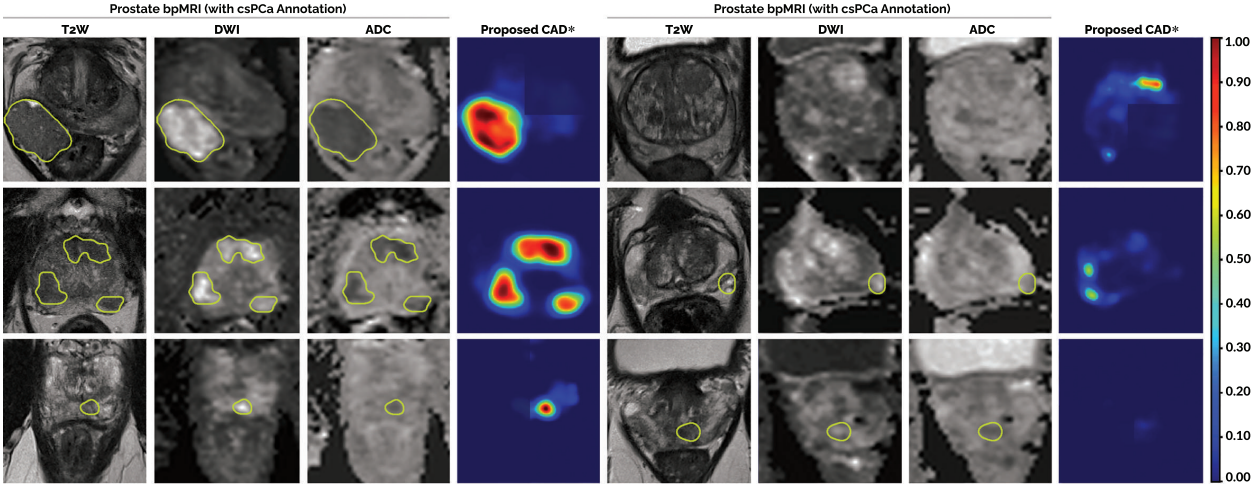


Figure 12.6: Six patient cases from the external testing set TS2 and their corresponding csPCa detection maps, as predicted by the proposed CAD∗ system. Yellow contours indicate csPCa$^{(GS)}$ lesions, if present. While CAD∗ was able to successfully localize large, multifocal and apical/basal instances of csPCa$^{(GS)}$ ***(left)***, in the presence of severe inflammation/fibrosis induced by other non-malignant conditions (eg. BPH, prostatitis), CAD∗ misidentified smaller lesions, resulting in false positive/negative predictions ***(right)***.

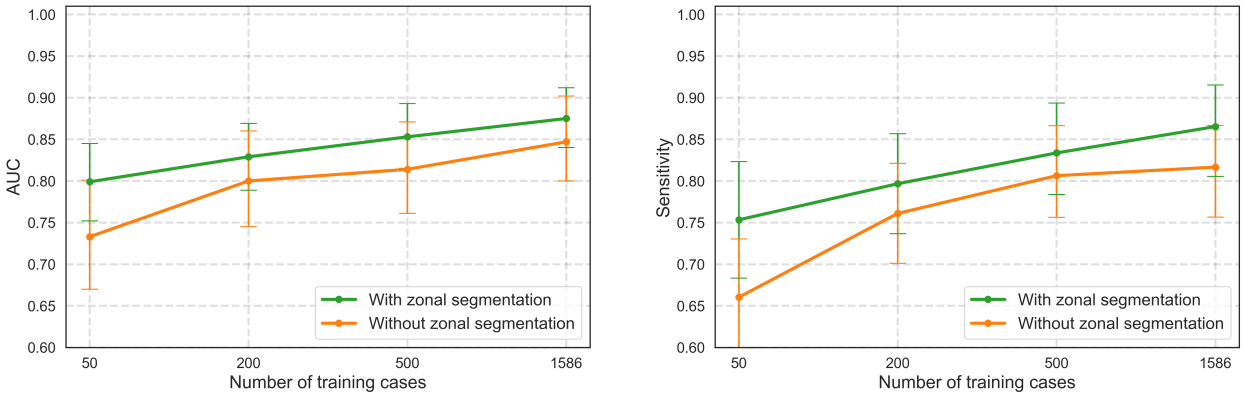


Figure 12.7: ***(left)*** Effect of training set size on patient-based performance on the institutional test set TS1. ***(right)*** Effect of training set size on lesion-based performance on the institutional testing set TS1. Sensitivities are at on average 1 FP lesion prediction per patient.

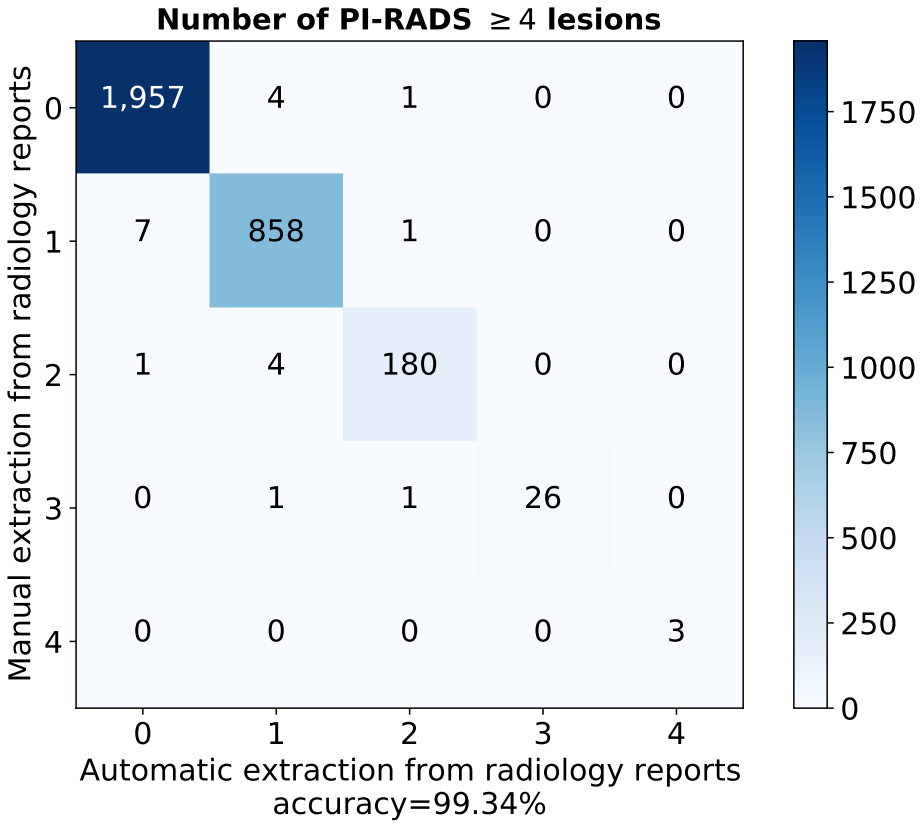


Figure 12.8: Accuracy of our NLP score extraction algorithm, as depicted by the confusion matrix for number of clinically significant findings in a radiology report. Evaluated on $D_{dev,labeled}$.

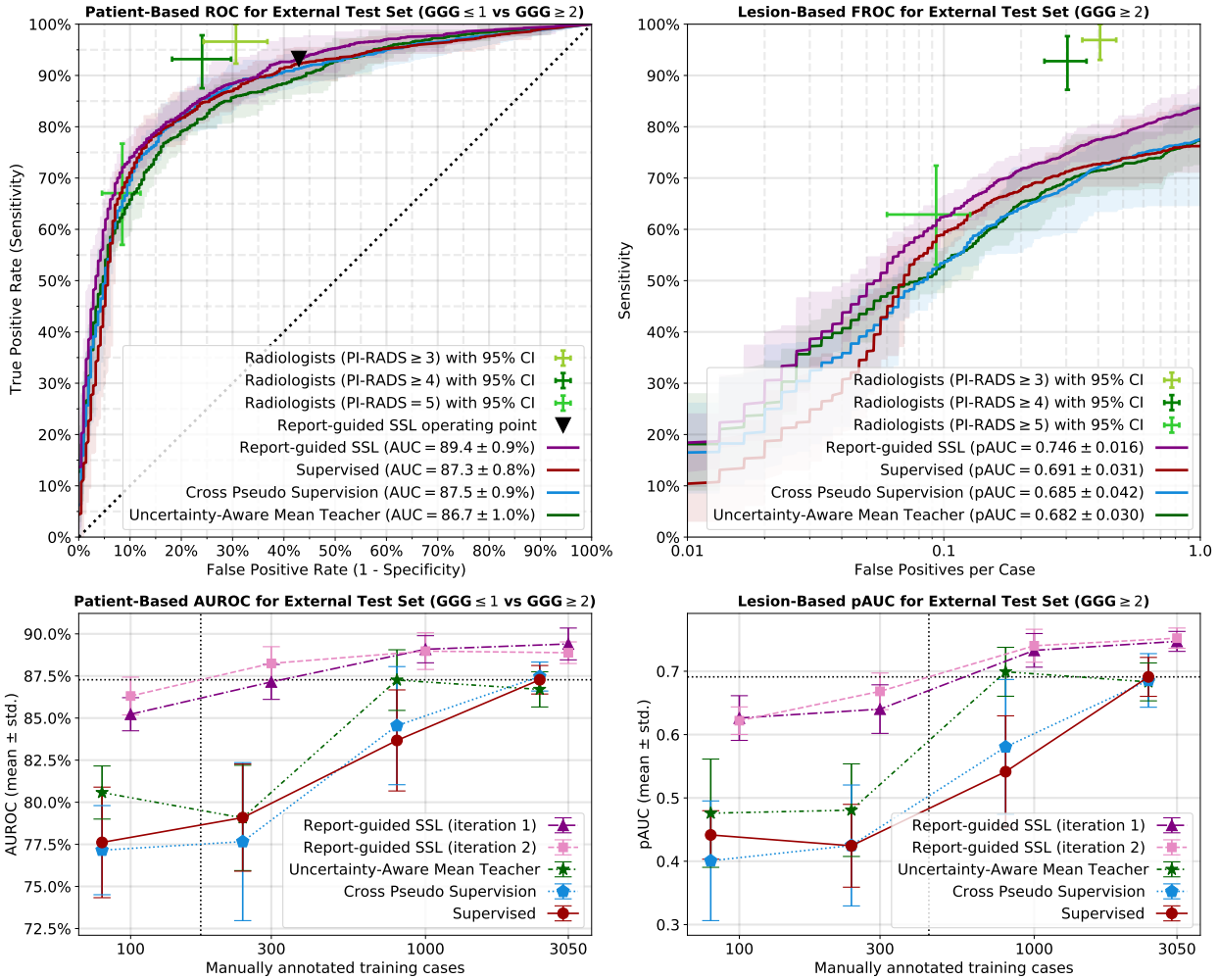


Figure 12.9: Model performance for (semi-)supervised learning. ***(top row)*** Supervised models are trained with 5-fold cross-validation on 3050 manually labeled exams, and semi-supervised learning (SSL) also includes 4706 unlabeled exams. Report-guided SSL significantly outperforms supervised learning as well as the baseline SSL methods. ***(bottom row)*** Model performance for 100, 300, 1000 or 3050 manually labeled exams, combined with 7656, 7456, 6756 or 4706 unlabeled exams, respectively. Report-guided SSL significantly outperforms the baseline SSL methods as well as supervised learning at each annotation budget, except for case-based AUROC of Uncertainty-Aware Mean Teacher trained with 1000 labeled exams. The ***(left)*** panels show ROC performance for case-based diagnosis of exams with at least one ISUP ≥ 2 lesion, and the ***(right)*** panels show FROC performance for lesion-based diagnosis of ISUP ≥ 2 lesions. All models are trained with radiology-based PI-RADS ≥ 4 labels and evaluated on the external test set with histopathology-confirmed ground truth. Shaded areas indicate the 95% confidence intervals from 15 or 5 independent training runs. Error bars indicate standard deviation across 15 or 5 independent training runs.

12.4 Discussion

From our experiments, we observe that within the context of UC1 and UC2, where ample data is available with an adequate reference standard, vendor-specific deep learning models incorporating state-of-the-art methods approach human expert-level diagnostic performance (as shown by the moderate agreement observed between the AI system and both expert radiologists and pathologists [28]). We also observe that deep learning models continue to scale up in performance, but only with exponentially larger datasets. Finally, we demonstrate that AI-generated pseudo labels are of sufficient quality to train fully supervised diagnostic models, thereby vastly reducing the annotation budget and manual workload required for training such systems. We plan to use and investigate the viability of such AI-generated annotations across the ProstateNET dataset, for training different models in future studies.

Chapter 13

Vendor Specific Deep Learning Models (Experiments Set 4)

13.1 Chapter Summary

In this chapter, FORTH expands upon the analysis presented in Deliverable 5.3, focusing specifically on the task of lesion segmentation. It details the implementation and training of two deep learning models: nnU-Net, as described by Isensee et al. [12], and our custom-developed ProLesA-Net. Each model was independently trained on datasets from Siemens, GE, and Philips vendors and subsequently validated on data from the non-training vendors to ensure robust cross-vendor generalization but also on a validation set for the same vendor to assess whether the models learn vendor-specific features.

13.2 Methods

Data Description

Table 13.1 presents the distribution of cases across each vendor. Philips constitutes the majority with 182 cases, followed by Siemens with 127, and GE with 110. Toshiba contributes the least, with only 1 case. Prior to analysis, the dataset underwent preprocessing, during which cases deemed as extreme outliers were identified and subsequently excluded to ensure the quality and consistency of the data.

Cases	Siemens	Philips	GE	Toshiba
419	127	181	110	1

Table 13.1: Number of cases per vendor

Deep learning model specification

This analysis employs two deep learning models: nnU-Net [12] and ProLesA-Net. Each model was trained over 1000 epochs, utilizing checkpoint strategies to facilitate effective learning progression. We adopted a hybrid loss function that integrates binary cross-entropy loss with Dice Loss to optimize segmentation performance. Additionally, in order to maintain the integrity of vendor-specific results, validation splits for each vendor were stratified by lesion size, ensuring a balanced representation and minimizing size-induced bias.

Model evaluation

In order to assess the performance of the deep learning models, Dice Score (DS), Hausdorff Distance(HD), Recall, Precision, and Average Surface Distance (ADC) were employed. DS is a well-known segmentation

metric that measures the proportion of overlap between the prediction and the GT image, while HD and ASD are distance-based metrics measuring the furthest and average distance between predicted and GT voxels.

13.3 Results

The ProLesA-Net model's vendor-specific results, outlined in Table 13.2, depict its performance across different vendors. When trained and tested with the same vendor, the model reaches its peak DS 52.80% for Siemens, significantly higher than GE and Philips. It indicates a strong vendor bias, particularly for Siemens, where the model also attains its lowest HD and ASD at 7.64 mm and 1.44 mm respectively. Notably, the model's Recall is exceptionally high at 60.47% for Siemens, indicating a high true positive rate. For Philips, the highest DS achieved is 42.77% when trained on Philips data, reflecting a better generalization on Philips than on GE, where the highest DS is evidently lower at 24.65%. These outcomes demonstrate that the ProLesA-Net model exhibits substantial variability in generalization performance across vendors, with the best results obtained for vendor-specific settings, particularly for Siemens, and a significant drop when applied to data from other vendors.

			Tested		
			GE	Siemens	Philips
Trained	**DS (%)**	GE	24.65	21.89	**25.34**
	HD (mm)		**20.18**	24.45	20.59
	ASD (mm)		11.02	8.71	**5.22**
	Recall (%)		28.05	28.22	**35.79**
	Precision (%)		**35.04**	30.00	29.21
		Siemens	16.18	**52.80**	25.83
			27.01	**7.64**	23.15
			11.34	**1.44**	6.21
			16.21	**60.47**	30.18
			28.67	**55.59**	33.17
		Philips	20.69	23.66	**42.77**
			25.98	23.77	**13.00**
			12.83	7.54	**2.17**
			30.16	29.4	**51.89**
			22.82	32.22	**41.79**

Table 13.2: Vendor-specific results for ProLesA-Net model

The nnU-Net model's performance, shown in Table 13.3, reveals distinct variations in generalization across different vendors. The model demonstrates the highest DS, 33.28%, when both trained and tested on GE data. Notably, cross-vendor evaluation typically results in decreased performance, with the model trained on GE and tested on Siemens and Philips, showcasing a significant drop in DS. The lowest HD is observed when the model is trained on GE and it is tested on Philips, suggesting better boundary delineation in that particular setup. Precision and Recall are highest when the model is trained and tested on Siemens, scoring 53.53% and 35.27%, respectively.

			Tested		
			GE	Siemens	Philips
Trained	**DS (%)**	GE	**33.28**	31.90	29.62
	HD (mm)		18.74	17.02	**15.73**
	ASD (mm)		**3.72**	3.73	4.49
	Recall (%)		30.48	**31.43**	27.79
	Precision (%)		**46.39**	42.34	40.77
		Siemens	22.39	**35.51**	29.21
			26.07	**12.04**	17.55
			9.77	**2.83**	5.09
			20.66	**35.27**	27.90
			36.65	**53.53**	40.16
		Philips	24.75	**31.31**	30.91
			21.65	15.97	**15.06**
			9.92	**3.91**	4.20
			22.00	**29.60**	28.19
			39.47	33.70	38.69

Table 13.3: Vendor-specific results for nnU-Net model

13.4 Discussion

The nnU-Net model seems to have a more stable performance across vendors. It shows the least variability in DS when trained and tested on different vendors, which suggests a higher level of robustness and generalization in its segmentation capability. The best DS for nnU-Net comes when trained and tested on GE, but the performance does not significantly drop when tested on other vendors' data.

On the contrary, ProLesA-Net shows a much higher DS when trained and tested on data from the same vendor, especially for Siemens, with the DS reaching 52.80%, which is a significant improvement over the best DS of nnU-Net at 33.28%. That said, ProLesA-Net has a high degree of specialization and performs exceptionally well when the training and testing distributions are consistent. However, ProLesA-Net's performance appears to decline significantly than nnU-Net's when tested on data from vendors not seen during training, suggesting that ProLesA-Net is more sensitive to vendor-specific features.

Concluding, the nnU-Net model seems to be more domain-agnostic, maintaining more consistent performance across different vendors, but with a generally lower performance for vendor-specific tasks. In contrast, ProLesA-Net can achieve higher scores but seems to be more domain-specific, showing a decline in performance when dealing with cases from vendors not included in its training data.

Bibliography

[1] David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Adv. Neural Inf. Process. Syst.*, 33:21428–21439, 2020.

[2] T Barrett, B Turkbey, and P L Choyke. PI-RADS version 2: what you need to know. *Clin. Radiol.*, 70(11):1165–1176, November 2015.

[3] Joeran S. Bosma, Anindo Saha, Matin Hosseinzadeh, Ivan Slootweg, Maarten de Rooij, and Henkjan Huisman. Semisupervised learning with report-guided pseudo labels for deep learning–based prostate cancer detection using biparametric mri. *Radiology: Artificial Intelligence*, 5(5):e230031, 2023.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *arXiv*, June 2016.

[5] Maarten de Rooij, Bas Israël, Marcia Tummers, Hashim U. Ahmed, Jochen Walz, and Jelle O. Barentsz et al. ESUR/ESUI consensus statements on multi-parametric MRI for the detection of clinically significant prostate cancer: quality requirements for image acquisition, interpretation and radiologists' training. *European Radiology*, 30(10):5404–5416, Oct 2020.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, October 2020.

[7] Raul Garreta and Guillermo Moncecchi. *Learning scikit-learn: Machine Learning in Python*. Packt Publishing Ltd, November 2013.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian" Sun. Identity Mappings in Deep Residual Networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 630–645. Springer International Publishing, 2016.

[9] Matin Hosseinzadeh, Anindo Saha, Patrick Brand, Ilse Slootweg, Maarten de Rooij, and Henkjan Huisman. Deep learning–assisted prostate cancer detection on bi-parametric mri: minimum training data size requirements and effect of prior knowledge. *European Radiology*, pages 1–11, 2021.

[10] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 7132–7141, 2019.

[11] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, Dec 2020.

[12] Fabian Isensee, Paul F Jaeger, Simon A A Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods*, 18(2):203–211, February 2021.

[13] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.*, 30, 2017.

[14] Charlotte F Kweldam, Mark F Wildhagen, Chris H Bangma, and Geert J L H van Leenders. Disease-specific death and metastasis do not occur in patients with gleason score ≤6 at radical prostatectomy. *BJU Int.*, 116(2):230–235, August 2015.

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, June 2022.

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. November 2017.

[17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*, 30, 2017.

[18] Oskar Maier, Alex Rothberg, Pradeep Reddy Raamana, Rémi Bèges, Fabian Isensee, Michael Ahern, mamrehn, VincentXWD, and Jay Joshi. loli/medpy: Medpy 0.4.0, February 2019.

[19] Daimantas Milonas, Žilvinas Venclovas, Inga Gudinaviciene, Stasys Auskalnis, Kristina Zviniene, Nemira Jurkiene, Algidas Basevicius, Ausvydas Patasius, Mindaugas Jievaltas, and Steven Joniau. Impact of the 2014 international society of urological pathology grading system on concept of High-Risk prostate cancer: Comparison of Long-Term oncological outcomes in patients undergoing radical prostatectomy. *Front. Oncol.*, 9:1272, November 2019.

[20] MONAI Consortium. MONAI: Medical open network for AI, June 2023.

[21] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Overfitting, model tuning, and evaluation of prediction performance. In Osval Antonio Montesinos López, Abelardo Montesinos López, and José Crossa, editors, *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, pages 109–139. Springer International Publishing, Cham, 2022.

[22] Samuel G Muller and Frank Hutter. TrivialAugment: Tuning-free yet state-of-the-art data augmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021.

[23] Eva Pachetti, Sara Colantonio, and Maria Antonietta Pascali. On the effectiveness of 3D vision transformers for the prediction of prostate cancer aggressiveness. In *Image Analysis and Processing. ICIAP 2022 Workshops*, pages 317–328. Springer International Publishing, 2022.

[24] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.*, 31, 2018.

[25] A. Saha, J. Bosma, J. Linmans, M. Hosseinzadeh, and H. Huisman. Anatomical and Diagnostic Bayesian Segmentation in Prostate MRI Should Different Clinical Objectives Mandate Different Loss Functions? In *Medical Imaging Meets NeurIPS Workshop–35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.

[26] A. Saha, J. Bosma, J. Twilt, B. van Ginneken, D. Yakar, M. Elschot, J. Veltman, J. Fütterer, M. de Rooij, and H. Huisman. Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI - The PI-CAI Challenge. In *Medical Imaging in Deep Learning (MIDL 2023)*, 2023.

[27] A. Saha, M. Hosseinzadeh, and H. Huisman. Encoding Clinical Priori in 3D Convolutional Neural Networks for Prostate Cancer Detection in bpMRI. In *Medical Imaging Meets NeurIPS Workshop–34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[28] Anindo Saha, Matin Hosseinzadeh, and Henkjan Huisman. End-to-end prostate cancer detection in bpmri via 3d cnns: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Medical Image Analysis*, 73:102155, 2021.

[29] Anindo Saha, Jasper Jonathan Twilt, Joeran Sander Bosma, Bram van Ginneken, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen Fütterer, Maarten de Rooij, and Henkjan Huisman. Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge (Study Protocol), June 2022.

[30] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images. *Medical Image Analysis*, 53:197 – 207, 2019.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for Large-Scale image recognition. *arXiv*, September 2014.

[33] D E Spratt, W C Jackson, A Abugharib, S A Tomlins, R T Dess, P D Soni, J Y Lee, S G Zhao, A I Cole, Z S Zumsteg, H Sandler, D Hamstra, J W Hearn, G Palapattu, R Mehra, T M Morgan, and F Y Feng. Independent validation of the prognostic capacity of the ISUP prostate cancer grade grouping system for radiation treated patients with long-term follow-up. *Prostate Cancer Prostatic Dis.*, 19(3):292–297, September 2016.

[34] Guang-Xi Sun, Peng-Fei Shen, Xing-Ming Zhang, Jing Gong, Hao-Jun Gui, Kun-Peng Shu, Jiang-Dong Liu, Jinge Zhao, Yao-Jing Yang, Xue-Qin Chen, Ni Chen, and Hao Zeng. Predictive efficacy of the 2014 international society of urological pathology gleason grading system in initially diagnosed metastatic prostate cancer. *Asian J. Androl.*, 19(5):573–578, 2017.

[35] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4278–4284. AAAI Press, 2017.

[36] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging*, 29(6):1310–1320, June 2010.

[37] Lvdmaaten van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. `https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl`, 2008. Accessed: 2023-7-13.

[38] Joost J M van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G H Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J W L Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.*, 77(21):e104–e107, November 2017.

[39] Guido Van Rossum and Fred Drake. *The python language reference manual*. Network Theory, Bristol, England, March 2011.

[40] B W H van Santvoort, G J L H van Leenders, L A Kiemeney, I M van Oort, S E Wieringa, H Jansen, R W M Vernooij, C A Hulsbergen-van de Kaa, and K K H Aben. Histopathological re-evaluations of biopsies in prostate cancer: a nationwide observational study. *Scand. J. Urol.*, 54(6):463–469, December 2020.

[41] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual Attention Network for Image Classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6458, 2017.

[42] Ziv Yaniv, Bradley C Lowekamp, Hans J Johnson, and Richard Beare. SimpleITK Image-Analysis notebooks: a collaborative environment for education and reproducible research. *J. Digit. Imaging*, 31(3):290–303, June 2018.

[43] V. Yeghiazaryan and I. Voiculescu. Family of boundary overlap metrics for the evaluation of medical image segmentation. *J Med Imaging (Bellingham)*, 5(1):015006, Jan 2018.

[44] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2020.