D5.3

Deep Learning Master model and Radiomic Signatures

Related Work Package	WP5.3 – Deep Learning Master model and Radiomic Signatures
Related Tasks	Task 5.3 — Development and performance evaluation of the master model using retrospective data; Task 6.2 — Deep learning methods for semi-automatic segmentation
Lead Beneficiary	FCHAMPALIMAUD
Contributing Beneficiaries	FCHAMPALIMAUD; CNR; FORTH; FPO; ADVANTIS; QUIBIM; HULAFE
Document version	vf
Deliverable Type	R
Distribution level	PU
Contractual Date of Delivery	28 Feb 2023
Actual Date of Delivery	13 Nov. 2023

Authors	José Guilherme de Almeida; Ana Carolina Rodrigues; Nuno Rodrigues; Raquel Moreno; Nickolas Papanikolaou; Rossana Buongiorno; Claudia Caudai; Giulio Del Corso; Danila Germanese; Eva Pachetti; Maria Antonietta Pascali; Grigorios Kalliatakis; Dimitrios Zaridis; Eugenia Mylona; Avtantil Dimitriadis; Dimitris Agraniotis; Zoi Giavri; Valentina Giannini; Giovanni Maimone; Simone Mazzetti; Daniele Regge; Manuel Marfil Trujillo; David Vallmanya Poch; Leonor Cerdá Alberich; Jose Munuera Mora; Ana Jímenez Pastor
Contributors	FCHAMPALIMAUD; CNR; FORTH; FPO; ADVANTIS; QUIBIM; HULAFE
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

Chapter 1

Executive summary

Deliverable 5.3, led by partner FCHAMPALIMAUD, contains the work performed by the ProCAncer-I consortium on master models using radiomics and deep learning techniques. ’Master models’ — models which can act as a foundation for other models — were developed for radiomics for all relevant use cases (UC2, UC3, UC5, UC6, UC7a, UC7b and UC8) through the development of consistent and robust pipelines, while deep learning was used only for UC1, UC2 and UC5 due to its more demanding data requirements. Radiomics master models were developed by three partners (FCHAMPALIMAUD, FORTH and CNR), while deep learning master models were developed and investigated by six different partners (FCHAMPALIMAUD, CNR, FORTH, ADVANTIS, FPO, QUIBIM). Through this approach, several aspects of deep learning models were investigated and consistent approaches and trends were identified. We finally note that the concept of a ’master model’ is similar to that of a foundation model; in that light, this deliverable reflects that insight. We describe the work in terms of foundation models and provide an overview of all experiments performed to arrive at foundation models.

Chapter 2

Introduction

The development of master models — general models which act as foundations for other models — entails different prerequisites depending on the task at hand and the methodology. Here, we focus on classification tasks to develop radiomics master models (chapter 4) and deep learning (DL) master models (chapter 5), as well as on a segmentation task to develop whole prostate gland, prostate zone and lesion segmentation master models (also a part of chapter 5). We also show how orphan data — data which has no clinically useful annotations — can be used to pre-train two-dimensional models which can be deployed in 3D classification (chapter 6). We also note particular challenges that we came across in terms of data curation due to problems in the data uploading stage (chapter 3).

Two major differences in terms of methodology are key to understand the difference in approaches between radiomics and deep-learning models — feature extraction and data requirements. While radiomics feature extraction pipelines are relatively well defined through several popular packages, the central paradigm surrounding machine-learning states that the feature extraction process is itself an essential part of the learning process. This leads to differences in terms of data requirements — given that DL models have to learn how to characterize images, they typically require much larger volumes of data, making their application to small use cases unreasonable. The fact that radiomics workflow is to some extent more standardised motivates a more uniform approach among partners involved in these tasks. in contrast to the multiple issues investigated in the case of DL models. It is worth noting that various choices might have an impact on the development of these methods. However, it should be noted that DL models have an advantage over radiomics models — while the latter requires anatomical or lesion segmentations to be available (predicted or manually annotated), DL models are able to learn patterns without having access to this information. The partners worked together to coordinate the activities through regular calls to discuss and agree on approaches and solutions to common problems. Based on the concept of creating basic models and evaluating their performance regardless of the data acquisition vendor, no data harmonisation techniques were applied. Robustness was comprehensively evaluated through precise testing of model performance, including learning curves and fairness analyses, when possible.

Radiomics models (chapter 4). **FCHAMPALIMAUD** focused on using a consistent feature extraction and machine-learning pipeline based on automatic whole prostate gland segmentations to assess its predictive performance across several different use cases (UC2, UC3, UC5, UC6, UC7b and UC8). To gain a better understanding of the requirements of different models, fairness, learning curve, and feature importance analyses were performed. This also helped us understand how different features impact the performance levels exhibited across tasks and use cases. **FORTH** contributed with a performance analysis using manually annotated lesion or whole prostate gland segmentation masks, drawing an important comparison between what each model requires. **CNR** inspected the performance of radiomics features on UC7a using predicted whole prostate gland segmentations, identical to those used by FCHAMPALIMAUD (the models used to develop these masks are detailed in section 5.7).

Deep-learning master models (chapter 5). Given that the amount of data is of paramount importance when developing models, and that DL models are relatively less well-established when compared with radiomics approaches and take longer to train/develop, the development of these models was decentralized, with more partners participating in the parallel and independent development of models. This results in more diverse approaches and findings. **FCHAMPALIMAUD (section 5.1)** performed a comprehensive study on how the definition of the UC2 target could have an impact on classification performance, while also studying how different models, the inclusion of clinical features, crop sizes and amounts of training data can have an impact on performance. Additionally, the impact of conformal prediction on performance was investigated. **CNR (section 5.2)** investigated how a generic central crop compares with an adaptive image cropping technique around the prostate using masks (identical to those used in chapter 4 and developed in section 5.7) impacted the performance in both UC2 and UC5 classification, while also performing a standard learning curve analysis. **FORTH (section 5.3)** compared unsupervised with supervised approaches to determine how each compares in terms of performance in UC1; additionally a learning curve analysis is also presented. **ADVANTIS (section 5.4)** study compare how 2D and 3D data lead to differences in performance in prostate and lesion segmentation models. **FPO (section 5.5)** study how using architectures which process the input sequences differently impacts lesion segmentation models using mpMRI data. **QUIBIM (section 5.6)** validate their automatic prostate segmentation model and present novel developments that make it more robust. **FCHAMPALIMAUD (section 5.7)** show how ProstateNet (a larger and more diverse dataset) can have a positive impact on performance when compared with smaller and less diverse datasets. Finally, **HULAFE (section 5.8)** designed a deep learning-based lesion segmentation model in ProstateNet's T2-weighted axial images, covering data processing, 2D and 3D model specifics, and strategies to combat over-fitting.

Self-supervised learning chapter (chapter 6). Finally, FCHAMPALIMAUD trained self-supervised learning (SSL) models on 2D data and applied them to 3D classification to understand whether this constituted an improvement over 3D models. The purpose of this analysis was to understand if 2D orphan data (usually stored in DICOM format) could be used to improve the performance of DL models on deep-learning tasks by comparing the performance of SSL models with models trained using what was learned during the development of chapter 5.

Together, we believe these analyses and results constitute an important and considerable volume of work which not only represents the value of ProstateNet when compared with other data sources, but also constitutes important building blocks in terms of future DL model developments for prostate cancer MRI data.

Chapter 3

Data curation for prostate cancer mpMRI

Chapter summary

Here, the fundamental aspects of an automatic data curation method for prostate cancer mpMRI is detailed. Particularly, this protocol provides adequate filtering steps to convert a set of disorganized DICOM studies into a organized dataset of volumes appropriate for training deep-learning models with correctly identified mpMRI data for each study (T2-weighted sequences (T2W), apparent diffusion coefficients (ADC), high b-value sequences in diffusion weighted imaging (DWI)).

We note that this was necessary as the uploaded studies initially in the "staging area" of ProstateNet exhibited some data quality and organizational issues (e.g. series types were hard to identify), making model development impossible. At a latter stage, clinical partners provided practically all of the necessary annotations, but these automatic data curation methods (implemented and significant time had been allocated developing them) were tools for the data quality assessment of ProstateNet. It is worth to note that the latter represented a significant delay in the development of the AI models, which is why we detail here the issues which were necessary to overcome prior to the clinical experts reevaluation of the series types.

3.1 Problem statement

While comprehensive, ProstateNet, the largest worldwide PCa-related mpMRI dataset, showed many data quality issues. Specifically, while three series were expected in each study (”T2W”, ”ADC” and ”DWI”, all of which should be axial), this was not the case for most studies (Figure 3.1) — only 1,790 studies had the correct number of series. This made an automatic curation strategy necessary to avoid including irrelevant information into the following analyses.

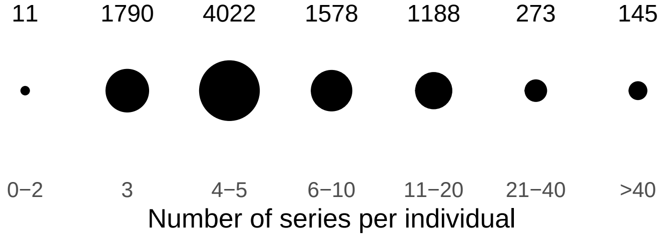


Figure 3.1: Number of studies stratified by the number of series on each study.

Upon an early inspection of ProstateNet data, it was determined that cases where excess series were identified could be attributed to:

- T1-weighted series
- Synthetic DWI (as opposed to regular DWI)
- Exponential ADC (as opposed to regular ADC)
- T2W-star (as opposed to regular T2W)
- Non-axial orientations in T2W series
- Several b-values in the same DWI series (as opposed to a single, $b = 1,400$ DWI series as recommended in PI-RADS [7])

Finally, in very few instances, non-prostate (i.e. abdominal) studies were also incorrectly included.

3.2 Problems and adopted heuristics

The uploaded data demonstrated a number of quality related issues, although specific and detailed instructions have been delivered to the clinical centers and repeated training sessions were organised. For simplicity, we will provide the observed issues as subsection titles and the identified solutions below. In any case they were consistently based on DICOM tags. We note that most solutions are based on heuristics derived upon the observation of multiple series in ProstateNet and are, as such, specific to ProstateNet as it currently is.

3.2.1 Problem: sequences were not annotated

While simple heuristics can be used to obtain a relatively satisfactory result, the variability in metadata information is considerable and can cause specific heuristics to fail. For this reason, we annotated 38,942 series with their corresponding sequence type (i.e. T2W, DWI, DCE, ADC and Others) and trained a CatBoost model on these data, using only a subset of metadata tags (Table 3.1). Since the DICOM series metadata is highly redundant (most instances in a given series will have the same metadata value), we keep only the unique values for each metadata tag for each series. As such, we consider each series to be a sample and classify it accordingly given its unique metadata values. Finally, we use an additional feature to aid in classification - the number of instances in each series.

Metadata value preprocessing.

Before concatenating all values for a given metadata key in a series, some minor preprocessing was applied. Particularly, we replace all recurring non-alphanumeric separators (`|`, `-`, `;`, `,`, `.`, `(`, `)`, `_`, `:`) with an empty space. All unique values are then concatenated as a single string using the space character as a separator.

Training and validation of CatBoost model.

We split the training and testing data into two separate training and testing sets (with 80% and 20% of the data, respectively). Using the training data, CatBoost models were optimised using 5-fold cross-validation with 250 iterations and the gradient leaf estimation method. Average weighted F1-scores of 98.9% and 98.9% were observed was for CV (Table 3.2) and testing sets (Table 3.3), respectively. To obtain an average for the testing set, a consensus prediction (majority vote) using the individual predictions of the 5 models from cross-validation was obtained given that they all performed remarkably well.

Series-classification heuristics.

While this CatBoost model performs well, we determined that some minimal and reasonable heuristics could be applied to the final results in order to make them more robust. Particularly, these heuristics are as follows (here ”word” is used as a string of alphanumeric characters delimited by spaces):

- If a non-zero or non-missing b-value tag is present and the ”adc” substring is not present in either the lowercase series description or in the lowercase image type we consider this series to be **”DWI”**;
- If the ”adc” substring is present in the lower case image type we consider this series to be **”ADC”**;

Tag name	Hex code
Image type	(0008—0008)
Modality	(0008—0060)
Manufacturer	(0008—0070)
Manufacturer model name	(0008—1090)
Scanning sequence	(0018—0020)
Sequence variant	(0018—0021)
Scan options	(0018—0022)
MR acquisition type	(0018—0023)
Slice thickness	(0018—0050)
Repetition time	(0018—0080)
Echo time	(0018—0081)
Inversion time	(0018—0082)
Number of echos	(0018—0086)
Magnetic field strength	(0018—0087)
Number of phase encoding steps	(0018—0089)
Echo train length	(0018—0091)
Pixel bandwidth	(0018—0095)
Software versions	(0018—1020)
Reconstruction matrix	(0018—1100)
Receive coil name	(0018—1250)
Transmit coil name	(0018—1251)
Acquisition matrix	(0018—1310)
In place phase encoding direction	(0018—1312)
Flip angle	(0018—1314)
Patient position	(0018—5100)
Diffusion b-value	(0018—9087)
Image orientation (patient)	(0020—0037)
Number of temporal positions	(0020—0105)
Temporal resolution	(0020—0110)
Photometric interpretation	(0028—0004)
Pixel spacing	(0028—0030)

Table 3.1: List of metadata tags used in CatBoost model.

Label	Mean	Minimum	Maximum
T2W	98.8%	98.5%	99.2%
DWI	99.3%	98.9%	99.6%
ADC	99.4%	99.2%	99.9%
DCE	98.4%	97.0%	99.2%
Others	98.6%	98.1%	99.0%
Average	**98.9%**	**98.4%**	**99.3%**
Weighted average	**98.9%**	**98.5%**	**99.3%**

Table 3.2: 5-fold cross-validation F1-score for the CatBoost model.

Label	Mean	Minimum	Maximum
T2W	98.5%	98.5%	98.5%
DWI	99.2%	99.2%	99.2%
ADC	99.4%	99.4%	99.4%
DCE	99.2%	99.2%	99.2%
Others	98.6%	98.6%	98.6%
Average	**99.0%**	**99.0%**	**99.0%**
Weighted average	**98.9%**	**98.9%**	**98.9%**

Table 3.3: Test set F1-score for the CatBoost model.

3.2.2 Problem: not all T2W sequences were axial

Several strategies were identified to identify axial T2W sequences as opposed to other sequence types:

- Using the Series Description attribute of the DICOM header (0008,103E), we can exclude series where the series description contains “cor”, “cor”, “sag” or “sag” and include series containing “ax”, “tra” and “ptra”;

- Using the Image Orientation attribute of the DICOM header (0020,0037), we can classify images as being axial. This parameter specifies the cosines of the first row (x) and first column (y); however, to classify an image as being axial, we need the cosines of the slice (z) direction. This can be calculated as the cross-product of x and y (i.e. `z = abs(cross(x,y))`). Taking the index for the maximum argument (`argmax`) of the absolute of z we get a simple way of inferring direction:
  - If `argmax(abs(x)) is 0`: “sagital”;
  - If `argmax(abs(x)) is 1`: “coronal”;
  - If `argmax(abs(x)) is 2`: “axial”;
- Another method, also using the Image Orientation attribute, is by rounding the absolute of the matrix described above and comparing it to the identity matrix - these should be quite similar if the plane is axial/oblique axial;

3.2.3 Problem: some T2W sequences were fat-suppressed

An additional data quality related issue was the fact that some uploaded T2W sequences were fat-suppressed **??**. To ensure that no fat-suppressed (fs) sequences were included, the following steps were taken:

- Exclusion of series with the Series Description (0008,103E; assuming that only characters and spaces were present) containing: ”fs dixon”, ”spair”, ”spir”;
- Exclusion of series with the following Scan Options (0018,0022): ”FS”;
- It was determined that all FS sequences had low interquartile ranges, but only the first two heuristics were adopted as they were simpler and faster to implement.

3.2.4 Problem: T2W sequences for other anatomical regions were uploaded

An additional data quality related issue was the fact that in a few cases T2W sequences for other anatomical regions were uploaded. In order to ensure that only pelvic T2W sequences were included the following heuristics were adopted, series whose Series Description (0008,103E) contains ”whole pelvis”, ”bh”, ”star” or ”kidneys” were excluded.

3.2.5 Problem: some ADC sequences were exponential

To exclude exponential, rather than regular, ADC cases, series whose Image Type (0008,0008) contains “EADC” were excluded.

3.2.6 Problem: some HBV/DWI sequences contained more than one b-value

In some instances, the same series will have multiple acquisitions at multiple b-values. The correct identification of high b-value diffusion images is predicated on the existence of b-value information. However, when this is absent, it is possible to infer which series has the highest b-value:

- If the Diffusion b-value attribute of the DICOM header (0018,9087 for the public attribute, 0043,1039 for the private GE attribute and 0019,100c for the private Siemens attribute) is available:
  - The PI-RADS recommends that sequences with b-values closer to 1400 should be kept [7];

  If Diffusion b-value is not available (most common in Siemen’s data):
  - Each slice has two useful attributes here: the Instance Number (0020,0013) and the Image Position (Patient) (0020,0032). These correspond to the instance number (IN; for a given sequence, slices are acquired with a given order) and to the position of the image (PI);
  - IN was used to sort the last value PI, which will give us a sequence of slice positions in the z-plane (i.e. -3, -2, -1, 0). If more than one sequence is present, we will get instead -3, -2, -1, 0, -3, -2, -1, 0, -3, -2, -1, 0, etc.;

- We can then identify breakpoints in this vector of positions and keep only the first contiguous sequence of images. Alternatively, it would be possible to compute the median intensity of a region of interest out of the prostate (e.g., the upper/lower corner) and the different volumes according to their intensity values. This is made possible by the fact that lower intensities correspond to higher b-values;
- For information on the exact b-value, it may be possible to determine this from the series description tag as, in some instances, this makes explicit the b-values used during acquisition;

3.2.7 Problem: some HBV/DWI sequences were synthetic

There are few (very few) diffusion weighted sequences which are synthetic, i.e. they have been obtained artificially from (typically) two b-values. These are quite distinct as the scales will appear to be inverted - the gland will have a lower intensity than its surroundings. These are clinically valuable but are, in effect, artifacts for the models presented in this deliverable. To remove them, HBV sequences where the Series Description (0008,103E) contains "DW_Synthetic" were removed.

3.2.8 Problem: DICOM Segmentation to Nifti Conversion Limitations

The conversion of DICOM files to a more manageable format like Nifti, or others such as nrrd and mhd, is a crucial step that facilitates any subsequent processing. The DICOM series files in the ProstateNet database can be converted successfully with common tools like "dcm2niix". However, the DICOM segmentation files in ProstateNet pose a challenge as they cannot be converted successfully with standard tools such as "dcm2niix", "segimage2itkimage" (from the "dcmqi" library), or "dicomtonifti" (from the "VTK" library). Some of these tools appear to rely on the DICOM tag "Spacing Between Slices (0018,0088)", which is often incorrect in many DICOM segmentation files (e.g., in more than 70% of the index lesion segmentations). However, not all tools rely on this tag, and the root cause of the failure in other tools remains unidentified.

A solution to this issue seems to be the extraction of segmentation data from the DICOM files using a library like "pydicom", followed by the definition of the metadata (e.g., the "affine" matrix) through the associated series on which the segmentation was based. This solution, however, implies in general a limitation: the segmentation DICOM files likely cannot be converted correctly to a more manageable format independently of the source series.

Chapter 4

Radiomics Master Models

Chapter summary

In this Chapter, we report on the development of radiomics-based models in seven separate use cases, within the prostate cancer disease continuum (Table ??): the prediction of disease aggressiveness (UC2), development of metastasis within 6 months (UC3), biochemical recurrence risk after prostatectomy or radiation therapy (UC5 and UC6, respectively), toxicity after radiation therapy (UC7a), quality of life after prostatectomy (UC7b) and early withdrawal from the active surveillance program (UC8). With this goal in mind, whole gland radiomic features were extracted from three MRI volumes (T2W, DWI and ADC) and combined with clinical information and/or deep features. All developed models were evaluated according to FutureAI guidelines for medical AI technology.

Section	Partner
Segmentation quality assessment	FCHAMPALIMAUD
Use Case 2 - ISUP 1 vs 2345	FCHAMPALIMAUD
Use Case 2 - ISUP 12 vs 345	FCHAMPALIMAUD
Use Case 2 - ISUP 123 vs 45	FCHAMPALIMAUD
Use Case 2 - ISUP 1 vs 23 vs 45	FCHAMPALIMAUD
Use Case 2 - Binary Classification of Clinical Significant Prostate Cancer (Radiomics Analysis)	FORTH
Use Case 3	FCHAMPALIMAUD
Use Case 5 - post-surgery	FCHAMPALIMAUD
Use Case 5 - pre-surgery	FCHAMPALIMAUD
Use Case 6	FCHAMPALIMAUD
Use Case 7a	CNR
Use Case 7b	FCHAMPALIMAUD
Use Case 8	FCHAMPALIMAUD

Table 4.1: List of sections in this chapter and the responsible partners.

4.1 Methodology

4.1.1 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.

4.1.2 Segmentation

Automatic segmentation of the whole prostate gland was performed on T2W sequences using an in-house developed segmentation model. The full details of this model are shown in section 5.7.

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.

4.1.3 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 4.2. 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 4.2: Co-registration parameter file.

4.1.4 Segmentation quality assessment

A radiologist was asked to assess the segmentation quality of 125 T2W volumes and respective DWI. A three level grade was assigned in the following manner: 1 was given to good/decent segmentations; 2 to masks where minor corrections were needed; and 3 where major corrections were needed (less than 50% of the gland covered). The radiologist also provided notes/observations about the images and masks that allowed us to construct a dataframe with 17 variables: small_FOV_T2, small_FOV_DWI, big_FOV_DWI, poor_quality_T2, poor_quality_DWI, erc, previous_resection, missed_apex, missed_posterior, mis-sed_anterior, missed_base/superior, missed_PZ, included_seminal_vesicle, lesion_ difficulties, mismatch_apex/base, mismatch_antero/posterior and mismatch_la-tero/lateral, where FOV corresponds to field of view, erc corresponds to presence of endorectal coil, lesion_difficulties indicates whenever a lesion made it more difficult to segment the gland and the several mismatches correspond to masks dislocated from the actual gland due to co-registration errors. This information was used to train logistic regression models to predict mask quality, and the trained coefficients were analyzed for insights into the causes of low quality.

4.1.5 Radiomic features extraction

Bias field correction was performed on T2W sequences using the N4 Bias Field Correction algorithm [87] and the Python package Simple ITK (version 2.0.0) [96]. First, each image's x-, y- and z-spacing were assessed 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 also 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 4.3.

Radiomic features were extracted from the whole gland segmentation using the Pyradiomics package (version 3.0) [89] in Python (version 3.7.9) [90]. 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/.

4.1.6 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.

4.1.7 Clinical features

The clinical variables included for each use case can be found in Table 4.4. 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 UC5, two contexts were considered: presurgery and postsurgery. For the latter, the clinical variables included are the ones listed in Table 4.4. 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.

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 4.3: Radiomics Extraction parameters.

4.1.8 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) [25]. 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 4.5 and 4.6, for binary and multiclass tasks, respectively. For the smaller use cases, namely 3, 6 and 8, only the cross-validation performance is reported.

Different data subsets were tested for their training ability. Pure radiomics datasets were appended clinical and/or deep features and their performance was compared. 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. This resulted in 32 training combinations. The discriminated data sizes of the training set are shown in Tables 4.7 - 4.11 for each use case. In UC3 (Table 4.8), the settings ADC_All_noERC and T2W&DWI&ADC_All_noERC were discarded, since there were no ERC patients to remove.

For UC7b, 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 4.11.

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 4.4: Clinical variables included for each use case. “num” indicates a numerical variable; “bool” indicates a binary variable; “cat” indicates a categorical variable.

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

4.1.9 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.

4.1.10 Training

For models using radiomics and using radiomics together with deep features, a light gradient boosting machine (LGBM) [43] was trained, while for radclin or hybrid models, which may include categorical data, the CatBoost [68] algorithm was used. Regarding the smaller UCs, a support vector machine (SVM) classifier was selected for UC7b 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. 4.1 and the hyperparameter space used can be found in Table 4.13.

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

UC2 - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
Cases	All	4983	4605	4107	4107
	All_noERC	4485	4119	3899	3899

Table 4.7: Discriminated data sizes of the training sets for UC2.

UC3 - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
Cases	All	62	50	47	47
	All_noERC	59	47	47	47

Table 4.8: Discriminated data sizes of the training sets for UC3.

UC5 - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
Cases	All	716	683	657	657
	All_noERC	676	643	623	623

Table 4.9: Discriminated data sizes of the training sets for UC5.

UC6 - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
Cases	All	116	114	77	77
	All_noERC	56	54	50	50

Table 4.10: Discriminated data sizes of the training sets for UC6.

UC7b - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
Cases	All	234	232	228	228
	All_noERC	-	-	-	-

Table 4.11: Discriminated data sizes of the training sets for UC7.

UC8 - Train Sets		sequence			
		T2W	DWI	ADC	T2W&DWI&ADC
Cases	All	118	112	75	75
	All_noERC	113	108	72	72

Table 4.12: Discriminated data sizes of the training sets for UC8.

4.1.11 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) [53] 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.

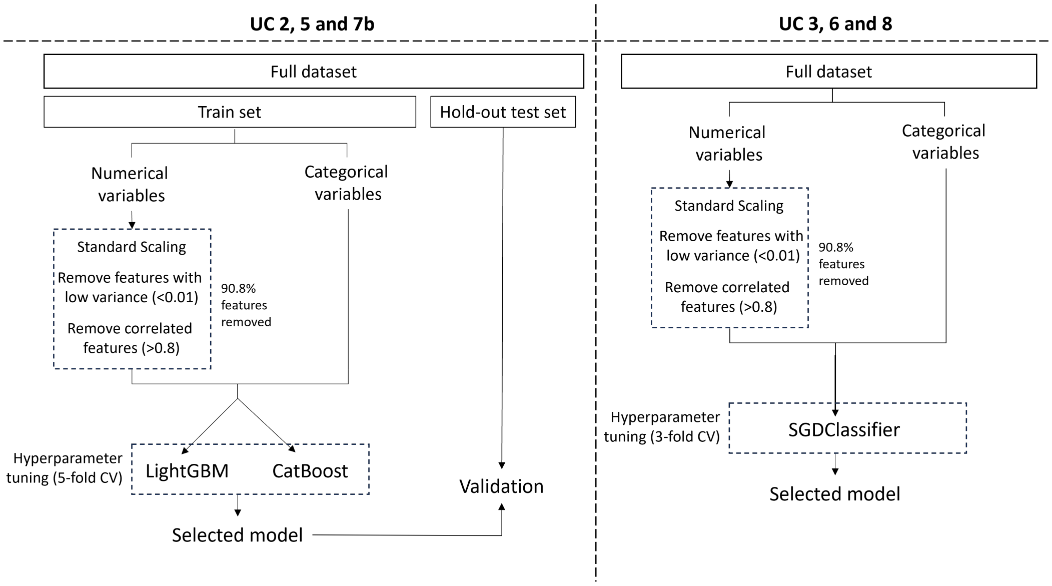


Figure 4.1: Radiomics model development pipeline.

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.

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 (UC7b)

```
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 4.13: Hyperparameter space used for optimization.

4.2 Results

4.2.1 Data description

The total dataset is composed of 5474 patients that meet the requirements for use case (UC2), the disease aggressiveness use case. Of these, 814 patients are also suitable for UC5, the biochemical recurrence use case, and 272 patients for UC7b, regarding quality of life after prostatectomy. The dataset size changes during the workflow are described in Table 4.14 for each use case.

	UC2	**UC3**	**UC5**	**UC6**	**UC7a**	**UC7b**	**UC8**
Initial number of patients	5474	113	814	137	137	272	156
T2 available	5297	107	771	136	136	264	151
T2 segmentation	5295	107	771	136	136	264	151
T2 extraction	5294	107	771	136	136	264	151
Available ground truth	5183	78	771	136	136	260	138
DWI exists	5025	105	760	135	-	260	150
DWI exists and T2 mask available	5025	105	760	135	-	260	150
DWI extraction	4858	95	737	134	-	260	145
Available ground truth	4805	66	737	134	-	256	132
ADC exists	4719	102	744	97	-	260	111
ADC exists and T2 mask available	4718	102	744	97	-	260	111
ADC extraction	4360	90	711	97	-	256	107
Available ground truth	4307	63	711	97	-	252	95

Table 4.14: Data workflow, specifying the number of patients in each use case after each step.

4.2.2 Segmentation quality assessment

The whole gland segmentations generated for T2W volumes proved to be of high quality (93% with quality level 1), with only 7% (9/125) of masks needing small corrections (quality level 2) and no masks requiring major corrections (quality level 3). Table 4.15 shows that the major contributors to the low quality were the presence of lesions that made the gland margins more difficult to identify or the presence of endorectal coil. This is followed by the absence of parts of the gland (previous resection) and the low quality of the DWI sequence. Additionally, the evidence of the slight drop in quality seems to be in missed apexes and wrongful inclusion of seminal vesicles.

Features	**Coefficient**	**exp(coefficient)**
lesion_difficulties	1.487972	4.428107
erc	0.753808	2.125077
missed_apex	0.703056	2.019917
included_seminal_vesicle	0.589769	1.803572
previous_resection	0.478329	1.613376
poor_quality_DWI	0.374318	1.453999
poor_quality_T2	0	1
mismatch_apex/base	-0.0748	0.927929
missed_PZ	-0.08214	0.921141
small_FOV_T2	-0.09743	0.907166
mismatch_antero/posterior	-0.22899	0.795338
mismatch_latero/lateral	-0.27934	0.756281
missed_anterior	-0.33524	0.715168
big_FOV_DWI	-0.41192	0.662381
small_FOV_DWI	-0.46918	0.625514
missed_base/superior	-0.54922	0.577401
missed_posterior	-0.71678	0.488323

Table 4.15: Coefficients of logistic regression trained to predict low quality of the whole-gland segmentation of T2W volumes.

In contrast, the whole gland segmentations generated for DWI volumes did not show as high quality. Here, 37% (46/125) of masks had quality level 2 and 9% (11/125) showed quality level 3, requiring major corrections. Table 4.16 shows that the major contributors to the low quality were the absence of parts of the gland (previous resection), the low quality of the DWI sequence and the small field of view of the T2 sequence. Regarding the middle quality (exp(coef_2)), it seems to be mostly due to several mismatches, which indicate imperfect coregistrations.

Features	coef_1	coef_2	coef_3	exp(coef_1)	exp(coef_2)	exp(coef_3)
previous_resection	-0.64316	-0.79734	1.440495	0.525631	0.450527	4.222786
poor_quality_DWI	-0.12803	-0.62484	0.752878	0.879824	0.535345	2.123101
small_FOV_T2	0.151386	-0.39528	0.243895	1.163446	0.673491	1.27621
small_FOV_DWI	0.377822	-0.40817	0.030348	1.459103	0.664866	1.030813
big_FOV_DWI	0.240165	-0.26686	0.026693	1.271459	0.765781	1.027053
poor_quality_T2	0	0	0	1	1	1
lesion_difficulties	-0.17899	0.18172	-0.00273	0.836116	1.199278	0.997272
missed_apex	0.394028	-0.37801	-0.01602	1.482942	0.685226	0.984106
missed_PZ	-0.23069	0.257765	-0.02708	0.793987	1.294034	0.973286
missed_base/superior	-0.31075	0.369413	-0.05866	0.732897	1.446886	0.943024
mismatch_apex/base	-0.13709	0.232008	-0.09492	0.871896	1.26113	0.909443
included_seminal_vesicle	-0.09962	0.206007	-0.10638	0.905177	1.228762	0.89908
missed_anterior	0.968572	-0.78994	-0.17863	2.634179	0.453871	0.836416
mismatch_antero/posterior	-0.91905	1.099245	-0.18019	0.398896	3.001898	0.835111
mismatch_latero/lateral	-0.44828	0.705045	-0.25676	0.638725	2.023937	0.773552
erc	-0.3212	0.809423	-0.48823	0.725281	2.246612	0.613713
missed_posterior	-0.01052	0.517407	-0.50689	0.989534	1.677673	0.602368

Table 4.16: Coefficients of logistic regression trained to predict low quality of whole-gland segmentation of DWI volumes.

4.2.3 Use Case 2 - ISUP 1 vs 2345

Exploratory Data Analysis of Clinical variables

Fig. 4.2 shows the distribution of each clinical variable used in UC2. There is no variable where the distinction between the two classes is clear, however we can see that the proportion of clinically significant cases increases as the PIRADS increases and there is higher proportion of clinically significant cases when the index lesion is located in the peripheral zone (PZ) or the transitional zone (TZ).

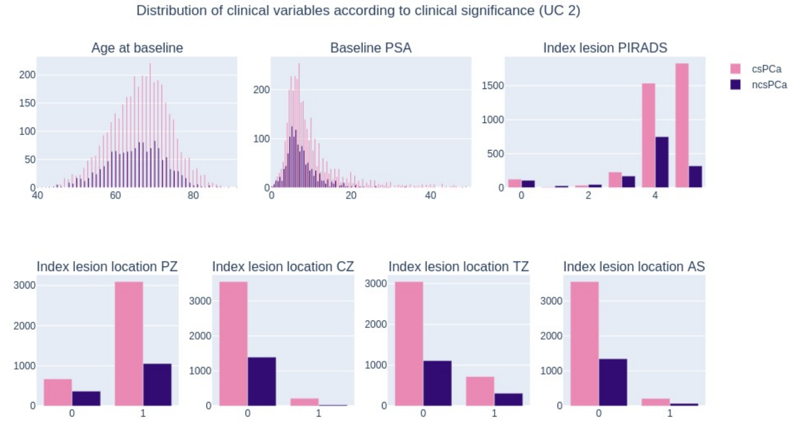


Figure 4.2: Distribution of clinical variables according to clinical significance defined as ISUP 1 vs 2,3,4 and 5.

Model Performance

Fig. 4.3 shows the cross-validation and hold-out test set ROC-AUC model performance for the 32 models trained for UC2, in four spider plots colored according to inclusion/exclusion of endorectal coil patients. At first glance, we can see that 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 exclusion of endorectal coil patients doesn't seem to have much impact on the models' generalizability. Overall, the highest hold-out test set performance reached was 0.765 AUC with a model trained with radiomics data.

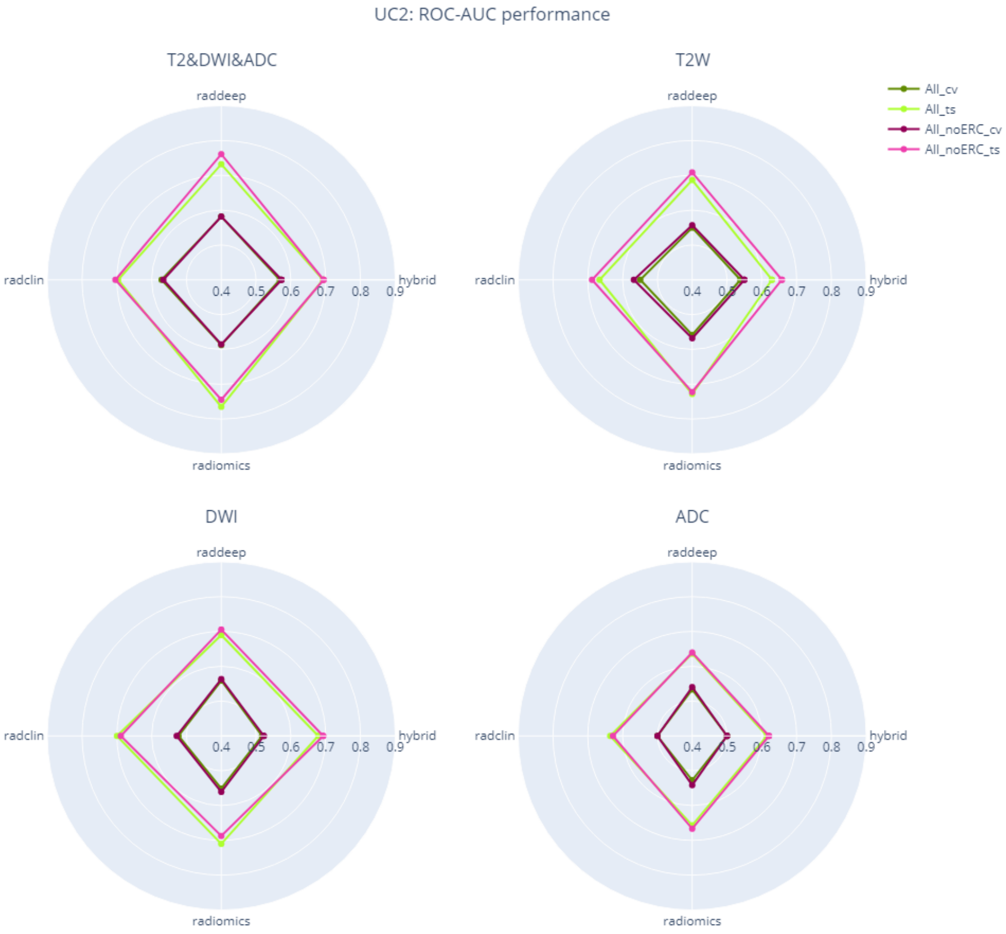


Figure 4.3: Cross-validation and hold-out test set ROC-AUC model performance of 32 models trained to predict disease aggressiveness in UC2 (ISUP 1 vs 2, 3, 4 & 5). The observations are color coded according to the inclusion/exclusion of endorectal coil patients: all patients in green; excluding ERC cases in pink.

Model Selection

Given the results of the previous section, we decided to move forward analyzing a radiomics model using all sequences and all patients. Upon inspection of the ROC curve for a subset of 50 patients from the hold-out test set (Fig. 4.4), we decided the appropriate probability decision threshold would be 0.7477, above which the model gives an output of clinical significance. The model's performance at this probability decision threshold is shown in Table 4.17 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Fig. 4.5.

Explainability analysis

Fig. 4.6 shows the SHAP analysis performed on the radiomics_uc2_T2&DWI& ADC_LGBM model. The plot shows the features ordered by their impact on model output. It is worth pointing out the shape feature

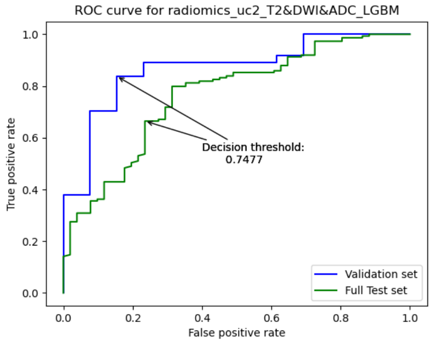


Figure 4.4: Hold-out test set receiver operator characteristics curve of the radiomics_uc2_T2&DWI&ADC_LGBM model.

radiomics_uc2_T2&DWI&ADC_LGBM	
AUC	0.7648
Sensitivity/Recall/TPR	0.6510
Specificity/TNR	0.7647
Precision/PPV	0.8899
F1	0.7519
F2	0.6879
Cohen's Kappa	0.3305

Table 4.17: Multi-metric performance of the radiomics_uc2_T2&DWI&ADC_LGBM model at 0.7477 probability threshold.

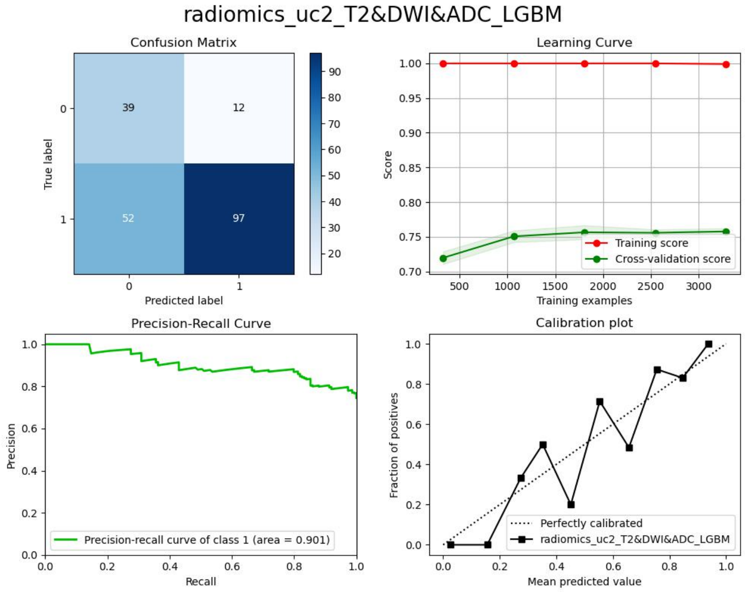


Figure 4.5: Analysis of the radiomics_uc2_T2&DWI&ADC_LGBM model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

Maximum2DDiameter Slice, which is negatively associated with the aggressive output.

Fairness and sub-cohort analysis

Tables 4.18 through Table 4.25 show the UC2 model's performance on different subsets of the hold-out test set. All tables are sorted by the number of cases on the train set, "Train counts". Regarding scanner manufacturer (Table 4.18) or usage of endorectal coil (Table 4.19), the model seems to perform relatively fairly.

Regarding lesion location, the model seems fair when it comes to peripheral zone (Table 4.20), however, for the transitional (Table 4.21) or central zones (Table 4.22), the model performs better when there is no lesion in these areas. The opposite relationship is true for anterior stroma (Table 4.23), where the model performs higher when a lesion is located here.

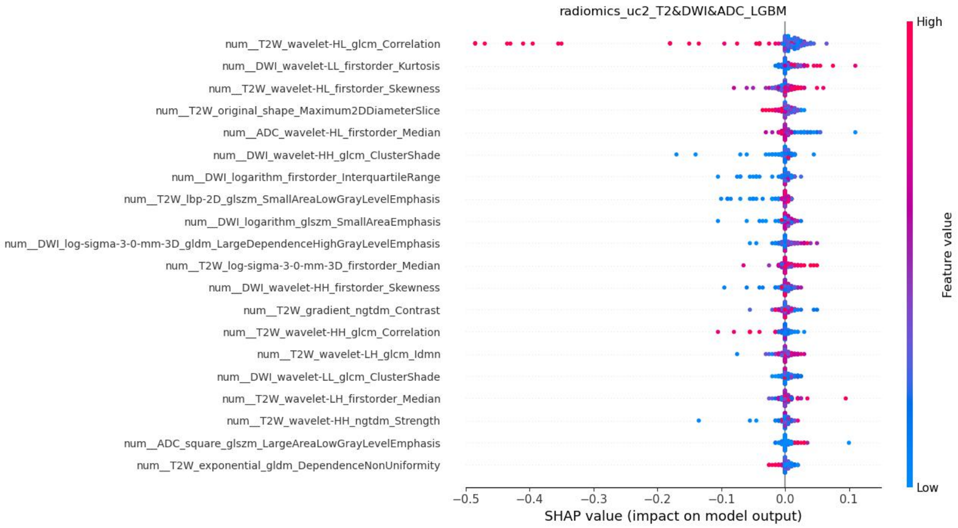


Figure 4.6: SHAP values contribution of features for radiomics_uc2_T2&DWI&ADC_LGBM model.

manufacturer	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
SIEMENS	0.639642	0.70197	0.863636	0.670588	108	23	85	2077
PHILIPS	0.759552	0.649351	0.909091	0.606061	56	23	33	1365
GE MEDICAL SYSTEMS	0.73871	0.719178	0.954545	0.677419	36	5	31	665

Table 4.18: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by scanner manufacturer.

Regarding country of origin (Table 4.24), we would expect the countries with the most data in the train set to achieve the best performance, but this is not the case. For Portugal and Spain the model performs the highest, while it shows the lowest performance for the UK and Lithuania.

Regarding ISUP grade (Table 4.25), we can see that as the ISUP grade increases so does the model's performance. With the lowest performance residing in ISUP 2 cases.

4.2.4 Use Case 2 - ISUP 12 vs 345

Model Performance

Fig. 4.7 shows the cross-validation and hold-out test set ROC-AUC model performance for the 32 models trained for UC2 (ISUP 1 and 2 vs 3, 4 and 5), in four spider plots colored according to the inclusion/exclusion of endorectal coil patients. At first glance, we can see that, the inclusion of clinical variables (radclin and hybrid) leads to a higher cross-validation performance, though that is no setting that particularly stands out. However, in terms of generalizability, we can see that training with all sequences leads to the highest test set performance. Overall, the highest hold-out test set performance reached was 0.7347 AUC with a model trained with all sequences and radclin data.

Model Selection

Given the results of the previous section, we decided to move forward analyzing a radclin model using all sequences and excluding patients with endorectal coil from the training. Upon inspection of the ROC curve for a subset of 50 patients from the hold-out test set (Fig. 4.8), we decided the appropriate probability decision threshold would be 0.2158, above which the model gives an output of clinical significance. The model's performance at this probability decision threshold is shown in Table 4.26 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Fig. 4.9.

Endorectal coil	ROC-AUC	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Patients without ERC	0.716352	0.698027	0.893204	0.661871	187	48	139	3899
Patients with ERC	0.633333	0.638298	0.857143	0.6	13	3	10	208

Table 4.19: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by presence/absence of endorectal coil.

index_lesion _location_PZ	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
1	0.714774	0.696203	0.905882	0.65812	152	35	117	3258
0	0.703125	0.686275	0.84	0.65625	48	16	32	849

Table 4.20: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the peripheral zone.

index_lesion _location_TZ	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.736066	0.709343	0.911111	0.67213	162	40	122	3284
1	0.614478	0.625	0.8	0.59259	38	11	27	823

Table 4.21: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out 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	0.688776	0.697674	0.888889	0.66207	196	51	145	3919
1	0.5	0.555556	1	0.5	4	0	4	188

Table 4.22: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the central zone.

index_lesion _location_AS	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.7	0.686275	0.883495	0.65	188	48	140	3859
1	0.888889	0.813953	1	0.77778	12	3	9	248

Table 4.23: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the anterior stroma.

country	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Netherlands	0.641176	0.667702	0.86	0.632353	88	20	68	1622
Portugal	0.928571	0.882353	1	0.857143	22	1	21	576
Lithuania	0.5	0	0	0	23	14	9	466
UK	0.546667	0.785124	0.904762	0.76	28	3	25	447
Turkey	0.733333	0.689655	0.8	0.666667	11	5	6	359
Italy	0.654762	0.681818	0.9	0.642857	17	3	14	281
Spain	0.816667	0.833333	0.833333	0.833333	11	5	6	243
France	0	0	0	0	0	0	0	95
Greece	0	0	0	0	0	0	0	18

Table 4.24: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by country of origin of the data.

Explainability analysis

Fig. 4.10 shows the SHAP analysis performed on the radclin_uc2_T2&DWI&ADC_noERC_CatBoost model. The plot shows the features ordered by their impact on model output. The four most relevant features for

ISUP grade	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
2	0.615385	0.666667	1	0.615385	91	0	91	1593
1	0.764706	0	0	0	51	51	0	1113
3	0.653846	0.702479	1	0.653846	26	0	26	712
5	0.875	0.897436	1	0.875	16	0	16	391
4	0.6875	0.733333	1	0.6875	16	0	16	298

Table 4.25: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by ISUP grade.

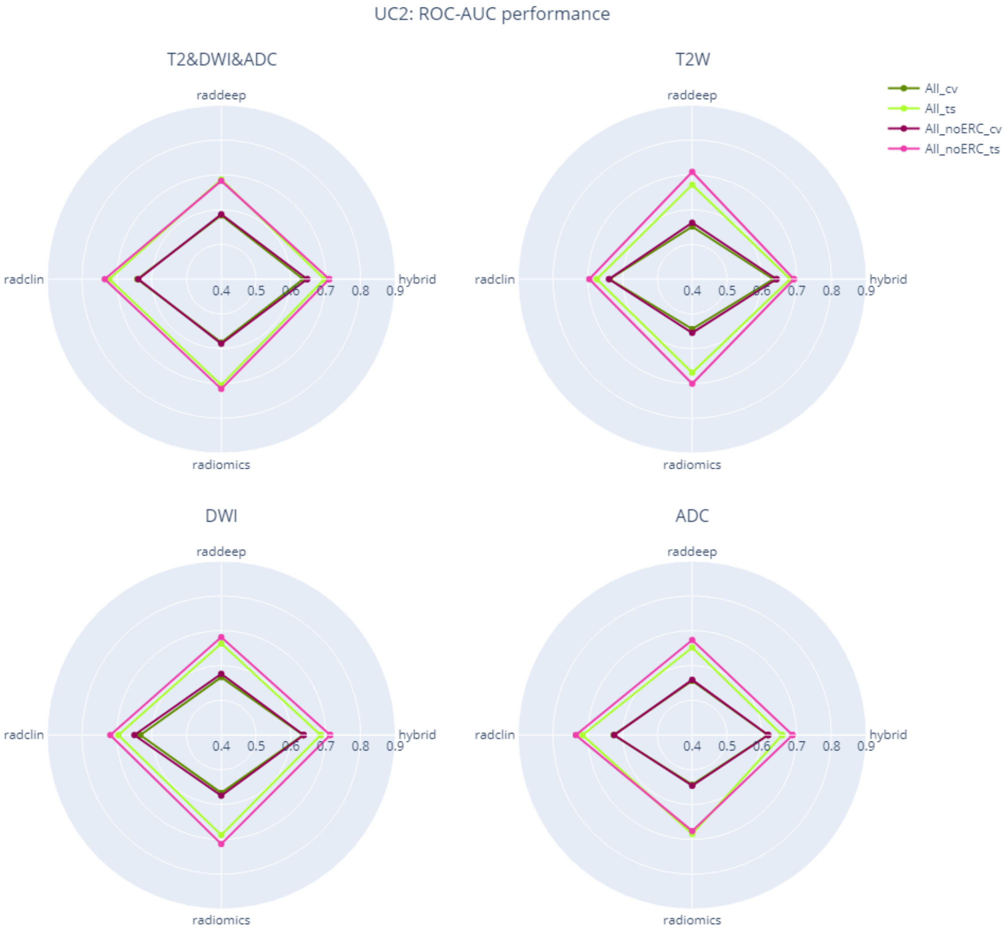


Figure 4.7: Cross-validation and hold-out test set ROC-AUC model performance of 32 models trained to predict disease aggressiveness in UC2 (ISUP 1 & 2 vs 3, 4 & 5). The observations are color coded according to the inclusion/exclusion of endorectal coil patients: all patients in green; excluding ERC cases in pink.

the model output are clinical variables, namely baseline PSA, patient age, PZ lesion location and PIRADS (Fig. 4.11), which are all positively associated with an aggressive output.

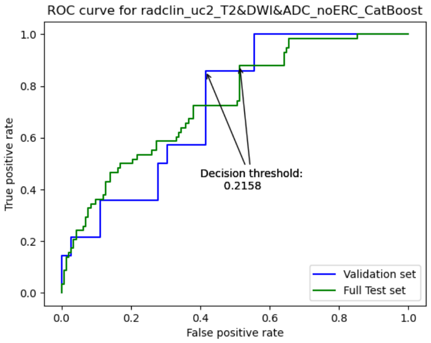


Figure 4.8: Hold-out test set receiver operator characteristics curve of the radclin_uc2_T2&DWI&ADC_noERC_Cat-Boost model.

radclin_uc2_T2&DWI&ADC_noERC_CatBoost	
AUC	0.7347
Sensitivity/Recall/TPR	0.7931
Specificity/TNR	0.4859
Precision/PPV	0.3866
F1	0.5198
F2	0.6553
Cohen's Kappa	0.2128

Table 4.26: Multi-metric performance of the radclin_uc2_T2&DWI&ADC_noERC_Cat-Boost model at 0.2158 probability threshold.

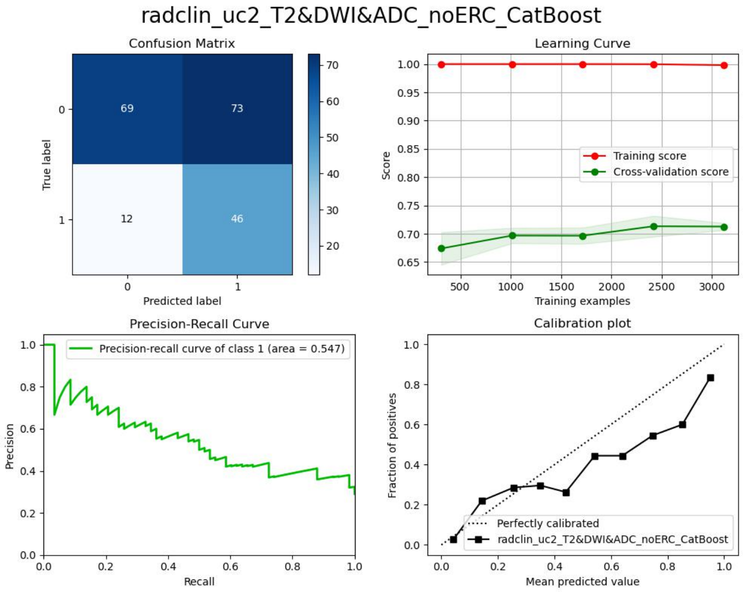


Figure 4.9: Analysis of the radclin_uc2_T2&DWI&ADC_noERC_CatBoost model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

Fairness and sub-cohort analysis

Tables 16 through Table 23 show the radclin_uc2_T2&DWI&ADC_noERC_CatBoost model's performance on different subsets of the hold-out test set. All tables are sorted by the number of cases on the train set, "Train counts". In terms of scanner manufacturer (Table 16), the model performs higher on SIEMENS data (F-score of 0.6982).

Regarding lesion location, the model seems to perform better when there is no lesion in the peripheral zone (Table 4.29), however, for the remaining anatomical zones, the opposite relationship is observed, though for the CZ (Table 4.31) and AS (Table 4.32) there are not enough samples in the minority sub-cohort to make fair conclusions.

Regarding country of origin (Table 4.33), most sub-cohorts do not have enough samples for fair conclusions to be made about the performance of the model in these countries.

Regarding ISUP grade (Table 4.34), we can see that as the ISUP grade increases so does the model's performance. With the lowest performance residing in ISUP 2 cases.

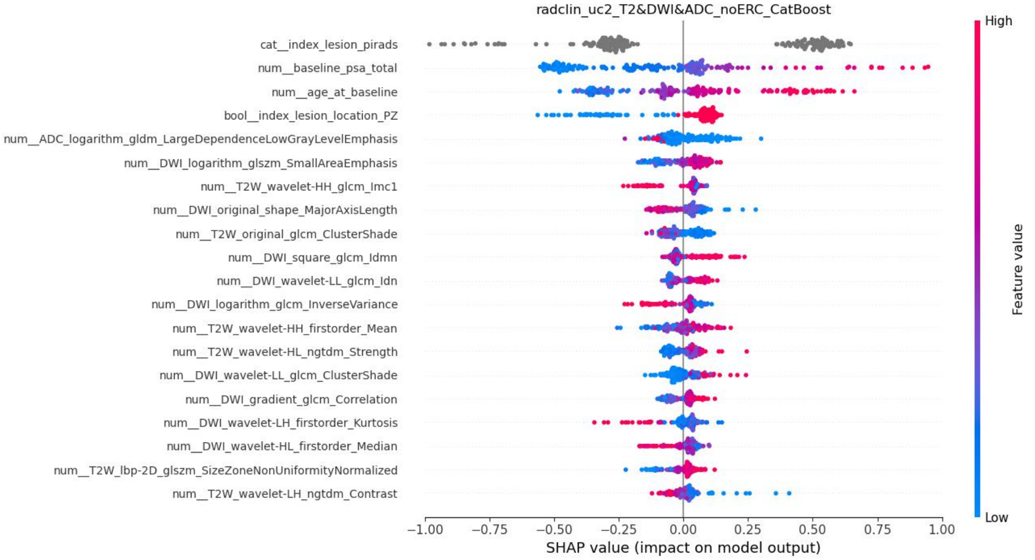


Figure 4.10: SHAP values contribution of features for radclin_uc2_T2&DWI&ADC_noERC_CatBoost model.

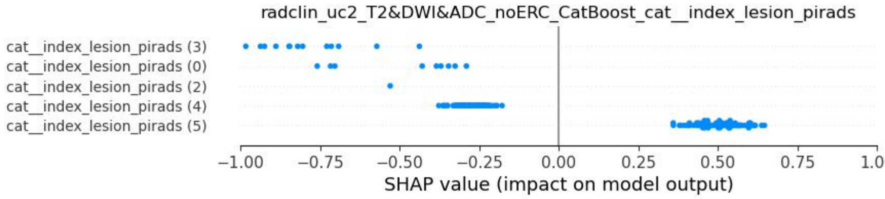


Figure 4.11: SHAP summary plot of the categorical variable index_lesion_pirads within model radclin_uc2_T2&DWI&ADC_noERC_CatBoost

4.2.5 Use Case 2 - ISUP 123 vs 45

Model Performance

Fig. 4.12 shows the cross-validation and hold-out test set ROC-AUC model performance for the 32 models trained for UC2 (ISUP 1, 2 and 3 vs 4 and 5), in four spider plots colored according to the inclusion/exclusion of endorectal coil patients. At first glance, we can see that, overall, the inclusion of clinical variables (radclin and hybrid) leads to a higher cross-validation performance. However, in terms of generalizability, we can see that training with radiomics data and all sequences leads to the highest hold-out test set performance of 0.8427 AUC.

Model Selection

Given the results of the previous section, we decided to move forward analyzing a radiomics model using all sequences and all patients. Upon inspection of the ROC curve for a subset of 50 patients from the hold-out test set (Fig. 4.13), we decided the appropriate probability decision threshold would be 0.2552, above which the model gives an output of clinical significance. The model's performance at this probability decision threshold is shown in Table 4.35 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Fig. 4.14.

Explainability analysis

Fig. 4.15 shows the SHAP analysis performed on the radiomics_uc2_T2&DWI&ADC_LGBM model. The plot shows the features ordered by their impact on model output.

manufacturer	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
SIEMENS	0.6042	0.6982	0.3974	0.8611	108	72	36	2077
PHILIPS	0.7041	0.5208	0.25	0.7143	56	49	7	1365
GE MEDICAL SYSTEMS	0.5714	0.6173	0.4762	0.6667	36	21	15	457

Table 4.27: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the hold-out test set, divided by scanner manufacturer.

Endorectal coil	ROC-AUC	Fbeta_2	Precision	Recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Patients without ERC	0.6360	0.6290	0.3679	0.7647	187	136	51	3899
Patients with ERC	0.5	0.8537	0.5385	1	13	6	7	0

Table 4.28: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the hold-out test set, divided by presence/absence of endorectal coil.

index_lesion _location_PZ	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
1	0.6154	0.6446	0.4066	0.7551	152	103	49	3077
0	0.7564	0.7031	0.3214	1	48	39	9	822

Table 4.29: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the peripheral zone.

Index_lesion _location_TZ	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.6118	0.6494	0.4	0.7692	162	110	52	3113
1	0.7969	0.6977	0.3158	1	38	32	6	786

Table 4.30: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the hold-out 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	0.5663	0.6399	0.3707	0.7818	196	141	55	3716
1	1	1	1	1	4	1	3	183

Table 4.31: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the central zone.

Index_lesion _location_AS	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.6353	0.6495	0.3874	0.7818	188	133	55	3653
1	0.7222	0.75	0.375	1	12	9	3	246

Table 4.32: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the anterior stroma.

Fairness and sub-cohort analysis

Tables 4.36 - 4.43 show the radiomics_uc2_T2&DWI&ADC_LGBM model's performance on different subsets of the hold-out test set. All tables are sorted by the number of cases on the train set, "Train counts". In terms of scanner manufacturer (Table 4.36), there is a drop in performance for GE cases, which can be explained by the drop in performance for endorectal coil cases 4.37, since most GE exams were taken with endorectal coil.

Regarding lesion location, the model seems to be relatively robust to the different index lesion locations

country	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Netherlands	0.5455	0.725	0.4531	0.8529	88	54	34	1622
Portugal	0.5	0.6452	0.2667	1	22	18	4	576
Lithuania	0.9130	0	0	0	23	23	0	466
UK	0.5	0.1471	0.1	0.1667	28	22	6	447
Turkey	0.6364	0.7143	0.3333	1	11	9	2	359
Spain	0.4545	0.5263	0.2857	0.6667	11	8	3	243
France	0	0	0	0	0	0	0	95
Italy	0.5294	0.7843	0.5333	0.8889	17	8	9	73
Greece	0	0	0	0	0	0	0	18

Table 4.33: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the hold-out test set, divided by country of origin of the data.

ISUP grade	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
2	0.4505	0	0	0	91	91	0	1501
1	0.5490	0	0	0	51	51	0	1061
3	0.6923	0.7377	1	0.6923	26	0	26	671
5	1	1	1	1	16	0	16	383
4	0.75	0.7895	1	0.75	16	0	16	283

Table 4.34: radclin_uc2_T2&DWI&ADC_noERC_CatBoost model performance on sub cohorts of the hold-out test set, divided by ISUP grade.

(Tables 4.38 - 4.40), with the exception of anterior stroma (Table 4.41), where the performance drops when the index lesion is located here.

Regarding country of origin (Table 4.42), the performance is relatively high across the table. However, for example in the case of UK and Turkey, the F-score of 0 indicates that the model classified everything into the non-aggressive class.

Regarding ISUP grade (Table 4.43), the values closest to the binary cutoff (ISUP 3 and 4) are expected to be the hardest for the model to classify, however, even though the model was very successful classifying ISUP 3, this was not the case for ISUP 4.

4.2.6 Use Case 2 - ISUP 1 vs 23 vs 45

Model Performance

Two of the previously described models (ISUP 1 vs 2345 and ISUP 123 vs 45) were combined to produce a multiclass classifier: ISUP 1 vs 2&3 vs 4&5. It's performance and confusion matrix are in Table 4.44 and Figure 4.16, respectively.

4.2.7 Use Case 2 - Binary Classification of Clinical Significance Prostate Cancer presence (Radiomics Analysis)

Population Dataset

For the specific UC2 model development and internal validation (FORTH contributor), the ProstateNET dataset was used consisting of 465 patients without endorectal coil with mpMRI sequences from Siemens, Philips and GE vendors. Manual lesion segmentations were available while the prostate gland was segmented automatically using the nn-Unet algorithm [40]. Additionally, for external model validation the Prostatex-2 dataset[20] was used which consists of 204 patient mpMRI acquired on two types of Siemens 3T scanners. The corresponding prostate gland and lesion segmentations were generated manually[50].

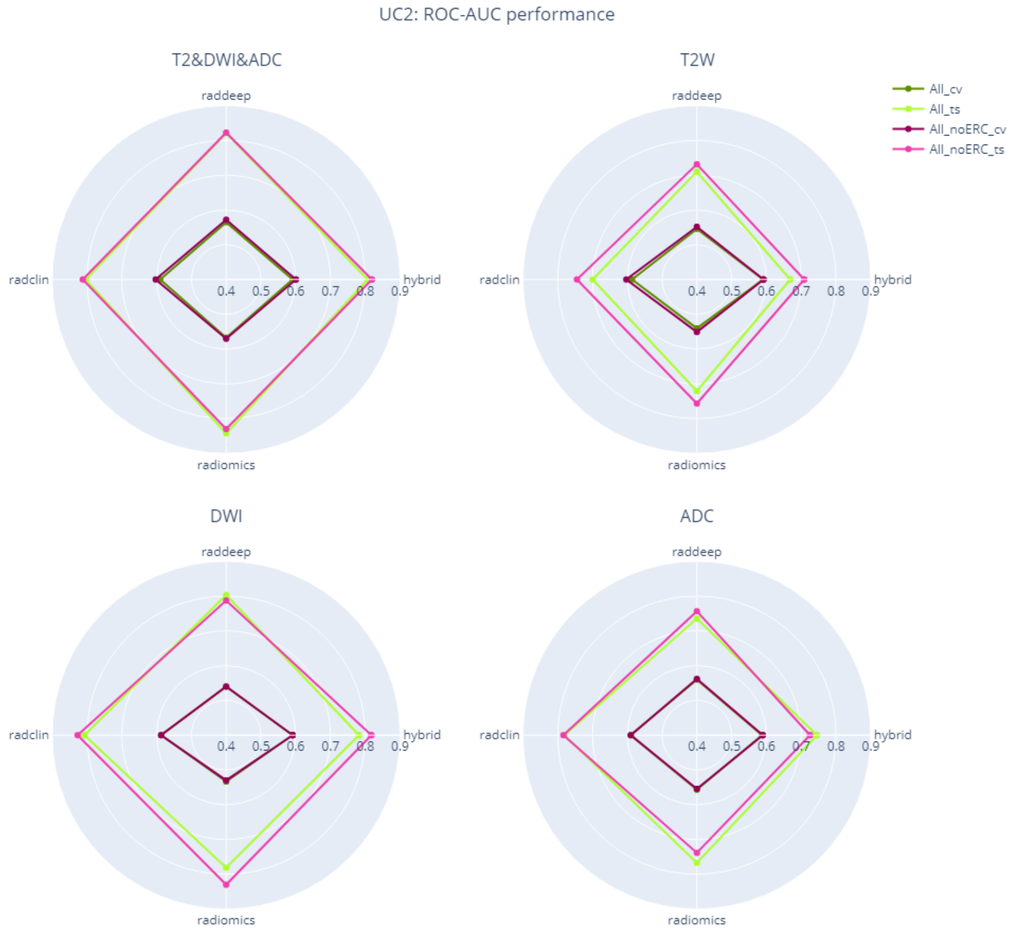


Figure 4.12: Cross-validation and hold-out test set ROC-AUC model performance of 32 models trained to predict disease aggressiveness in UC2 (ISUP 1, 2 & 3 vs 4 & 5). The observations are color coded according to the inclusion/exclusion of endorectal coil patients: all patients in green; excluding ERC cases in pink.

Image processing

Image cropping was performed in order to bring all the sequences at the same starting and ending points, in all three directions. Same process was performed also on lesion and whole gland segmentations. The reason cropping was selected instead of resampling, was to keep the original intensities intact. Bias field correction was performed using the N4 Bias Field Correction algorithm and the Python package Simple ITK (version 2.2.12.0.0) . Normalization was performed using the Pyradiomics (version 2.2.0). All images were resampled by pyradiomics at [1, 1, 1] mm pixel spacing. For histogram discretization, the absolute discretization (fixed bin size) approach was adopted as it has been found to preserve a higher number of reproducible features for MRI compared to relative discretization (fixed bin number)[22]. The optimal bin width was defined so that the number of bins in each image histogram would range from 16 to 128 bins. Detailed description of the extraction parameters is provided in Table 4.45.

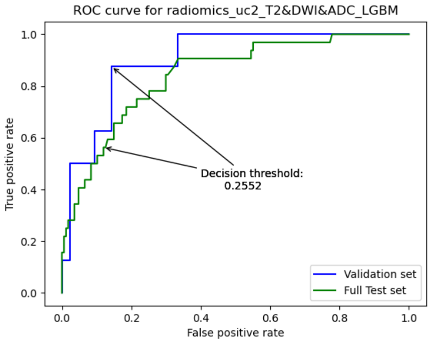


Figure 4.13: Hold-out test set receiver operator characteristics curve of the radiomics_uc2_T2&DWI&ADC_LGBM model.

radiomics_uc2_T2&DWI&ADC_LGBM	
AUC	0.8427
Sensitivity/Recall/TPR	0.5313
Specificity/TNR	0.8802
Precision/PPV	0.4595
F1	0.4928
F2	0.5152
Cohen's Kappa	0.3870

Table 4.35: Multi-metric performance of the radiomics_uc2_T2&DWI&ADC_LGBM model at 0.2552 probability threshold.

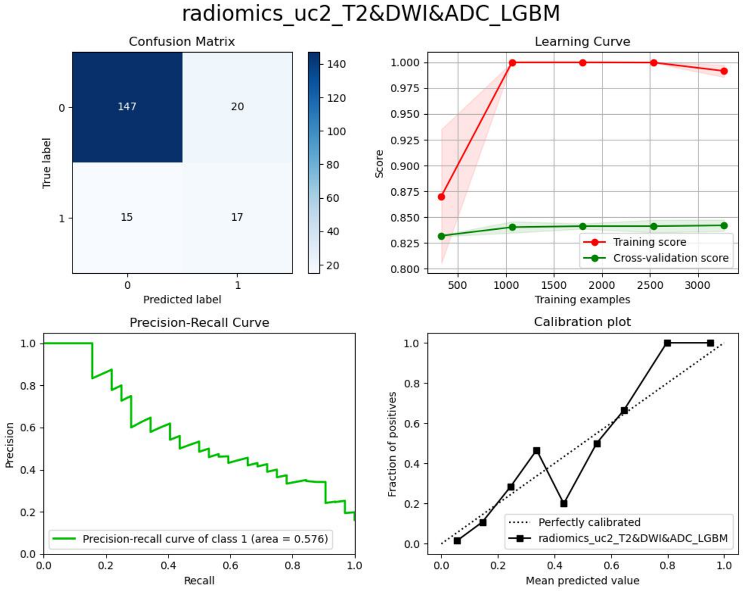


Figure 4.14: Analysis of the radiomics_uc2_T2&DWI&ADC_LGBM model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

Feature Reduction and Feature Selection

In total 1246 radiomic features were extracted from each MRI sequence, namely the T2-weighed (T2w) images and the Apparent Diffusion Coefficient (ADC) maps and each 3D region of interest (ROI), namely the whole gland and the tumor. Given the vast number of features extracted, feature reduction was required prior to modelling to improve algorithms' performance. This process aims to find the minimally sized feature subset that is necessary and sufficient to describe the target concept. First, highly correlated features and variables irrelevant to the outcome of interest were removed. Features with low variance were excluded using a variance threshold of 0.01 and multicollinear features were eliminated using a threshold of 0.85. If two variables had a correlation higher than the threshold, the one with the largest mean absolute correlation was removed. Then, supervised feature selection was applied only to the training data to avoid data leakage to the test set. To retain only the features correlated with the outcome variable, a Wilcoxon rank sum test was performed with a significance threshold of 0.1 as the main goal is to prioritize a subset of features for further feature

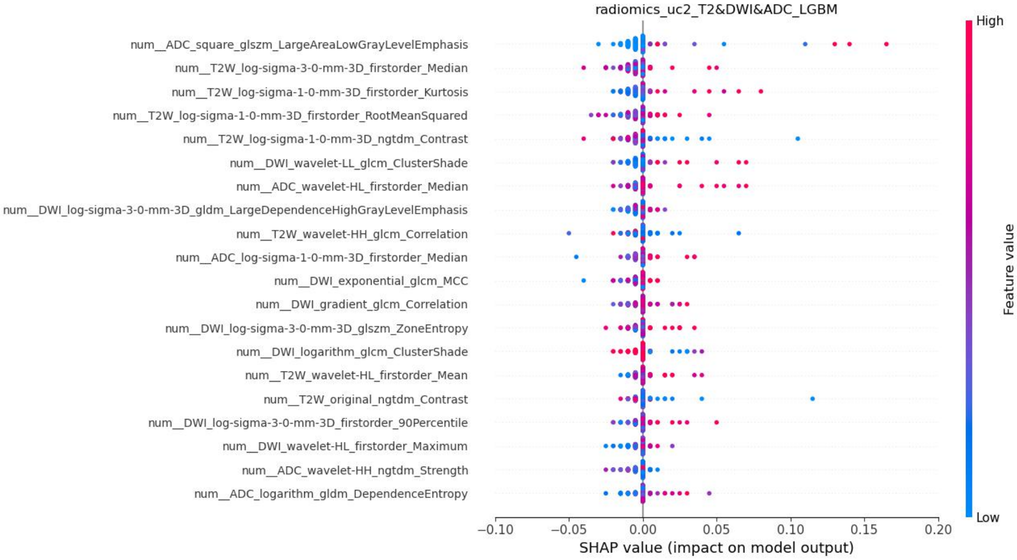


Figure 4.15: SHAP values contribution of features for radiomics_uc2_T2&DWI&ADC_LGBM model.

manufacturer	ROC-AUC	fbeta_2	precision	recall	test counts	test_counts target_0	test_counts target_1	train counts
SIEMENS	0.7113	0.5859	0.4688	0.6250	108	84	24	2077
PHILIPS	0.7407	0.5000	0.5000	0.5000	56	54	2	1364
GE MEDICAL SYSTEMS	0.6323	0.3571	0.5000	0.3333	35	29	6	650

Table 4.36: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by scanner manufacturer.

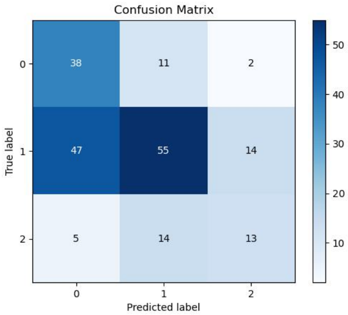


Figure 4.16: Confusion matrix.

	low	medium	high
Balanced Accuracy	0.5327	0.5327	0.5327
CohensKappa	0.2538	0.2538	0.2538
Precision/PPV	0.4222	0.6875	0.4483
Sensitivity/ Recall/TPR	0.7451	0.4741	0.4062
F1	0.5390	0.5612	0.4262
F2	0.6463	0.5055	0.4140

Table 4.44: Multi-metric performance.

selection. This process led to a significant reduction to the number of features in each dataset (¡150 variables) allowing a more exhaustive feature selection to be performed. For comparison, 10 state-of-the-art feature selection techniques were implemented: Minimum Redundancy Maximum Relevance (mRMR)[70] , Boruta algorithm[46] , Relief algorithm[44] , Recursive Feature Elimination (RFE)[28] , Statistically equivalent multiple feature subsets (SES)[48] , Fast Correlation-Based Filter (FCBF)[65] , Correlation-based Feature Selection (CFS) with forward selection strategy [92], Random Forest (RF) variable importance, LASSO[64] , and AUC-based feature selection.

ERC	ROC-AUC	fbeta_2	precision	recall	test	test_counts	test_counts	train
					counts	target_0	target_1	counts
0	0.7329	0.5592	0.4722	0.5862	187	158	29	3897
1	0.6111	0.3571	0.5000	0.3333	12	9	3	194

Table 4.37: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by presence/absence of endorectal coil.

index_lesion_location_PZ	ROC-AUC	fbeta_2	precision	recall	test	test_counts	test_counts	train
					counts	target_0	target_1	counts
1	0.7124	0.5200	0.5200	0.5200	151	126	25	3245
0	0.7596	0.6098	0.3846	0.7143	48	41	7	846

Table 4.38: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the peripheral zone.

index_lesion_location_TZ	ROC-AUC	fbeta_2	precision	recall	test	test_counts	test_counts	train
					counts	target_0	target_1	counts
0	0.7266	0.5517	0.5517	0.5517	161	132	29	3271
1	0.7333	0.4762	0.2222	0.6667	38	35	3	820

Table 4.39: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the transitional zone.

index_lesion_location_CZ	ROC-AUC	fbeta_2	precision	recall	test	test_counts	test_counts	train
					counts	target_0	target_1	counts
0	0.7227	0.5414	0.4595	0.5667	195	165	30	3903
1	0.75	0.5556	1.0	0.5	4	2	2	188

Table 4.40: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the central zone.

index_lesion_location_AS	ROC-AUC	fbeta_2	precision	recall	test	test_counts	test_counts	train
					counts	target_0	target_1	counts
0	0.7329	0.5592	0.4722	0.5862	187	158	29	3844
1	0.6111	0.3571	0.5	0.3333	12	9	3	247

Table 4.41: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the anterior stroma.

country	accuracy	fbeta_2	precision	recall	test	test_counts	test_counts	train
					counts	target_0	target_1	counts
Netherlands	0.7273	0.5952	0.5	0.625	88	64.0	24.0	1622
Portugal	1.0	0.0	0.0	0.0	22	22.0	0.0	575
Lithuania	1.0	0.0	0.0	0.0	23	23.0	0.0	466
UK	0.8571	0.0	0.0	0.0	28	27.0	1.0	447
Turkey	0.8182	0.0	0.0	0.0	11	9.0	2.0	362
Italy	0.8125	0.5263	0.6667	0.5	16	12.0	4.0	266
Spain	0.9091	0.8333	0.5	1.0	11	10.0	1.0	243
France	0.0	0.0	0.0	0.0	0	0.0	0.0	95
Greece	0.0	0.0	0.0	0.0	0	0.0	0.0	15

Table 4.42: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by country of origin of the data.

Model Training

Two training schemes were considered: 1) The ProstateNET dataset was split into training (80%) and test (20%) sets so that both datasets have the same label distribution, and 2) The entire ProstateNET was

ISUP grade	accuracy	fbeta_2	precision	recall	test counts	test_counts target_0	test_counts target_1	train counts
2.0	0.8111	0.0	0.0	0.0	90.0	90	0	1591
1.0	0.9412	0.0	0.0	0.0	51.0	51	0	1105
3.0	1.0	0.0	0.0	0.0	26.0	26	0	707
5.0	0.6875	0.7333	1.0	0.6875	16.0	0	16	391
4.0	0.4375	0.4929	1.0	0.4375	16.0	0	16	297

Table 4.43: radiomics_uc2_T2&DWI&ADC_LGBM model performance on sub cohorts of the hold-out test set, divided by ISUP grade.

Image Type	Feature Class	Setting
Original: {}	first order	Normalize: True
Wavelet: {}	glcm	normalizeScale: 1
Gradient: {}	glrlm	interpolator: sitkBSpline

LoG: {
sigma: [1.0, 2.0, 3.0, 4.0]
}
	glszm	resampledPixelSpacing:
		[1, 1, 1]

-	gldm	padDistance:5
-	shape	-

Table 4.45: Parameters for Radiomic Feature Extraction.

Dataset	T2 radiomics	ADC Radiomics	T2+ADC Radiomics
ProstateNET	465 (74% ISUP≥2)	419 (75% ISUP≥2)	419 (75% ISUP≥2)
ProstateX-2	Gland : 186 (37% ISUP≥2) ; Tumor : 299 (25% ISUP≥2)		

Table 4.46: Number of patients and proportion of the target class (ISUP score ≥ 2).

used to train the models and for external validation the Prostate X dataset was used. The total number of patients available and the proportion belonging to the target class (ISUP score ≤ 2) for each classification scenario are shown in Table 4.46. In total 8 classification algorithms were implemented, namely, Random Forest (RF), Support Vector Machines (SVM) with kernel function, Extreme Gradient Boosting (XGB), Adaptive Boosting (AdaBoost), Boosted Generalized Linear Models (GLM), Bayesian GLM, and LASSO Regularization. All the models were implemented in R using the 'caret' function. Considering that radiomic features were extracted from the different ROIs (2) and MR sequences (3) and several feature selection methods (10) and ML algorithms (8) were used, the analysis resulted to a total of 480 models.

Statistical Analysis

A receiver operating characteristic (ROC) curve was generated to determine model accuracy and discriminative performance. ROC AUC, Precision, Recall, Precision-Recall (PR) AUC, and the F1 score were used to estimate model performance. Calibration curves based on the Hosmer-Lemeshow test. All statistical analyses were performed using R (version 4.1.0; R Foundation for Statistical Computing, Vienna, Austria).

Results

Table 4.47 shows optimal performance obtained for each classification task, encompassing the tumor and whole gland ROIs and three sets of input radiomic features (T2 only, ADC only, T2 and ADC combined). For each classification task, the combination of the feature selection method and the ML algorithm that resulted to the higher ROC AUC is shown along with the other performance metrics. The highest performance, both in internal (AUC=0.86) and external (AUC=0.75) validation, was noticed for the radiomic features extracted from both the T2 and ADC sequenced of the tumor, although marginally better than the ADC-only tumor model. Conversely, all the radiomic models of the whole gland exhibited inferior performarce, especially in the external validation (AUC≤0.61).

Figure 4.17 (a) highlights the calibration plots for WG models for internal and external validation while Figure 4.17 (b) highlights the calibration plots for Tumor models for internal and external validation. The

Model Characteristics					Performance					
Roi	Sequence	Best Feature Selection Method	N# of features selected	Best Model	Validation	ROC AUC	PR AUC	Precision	Recall	F1
Tumor	T2	Boruta	24	SVM	Internal	0.75	0.9	0.85	0.68	0.75
					External	0.64	0.41	0.27	0.78	0.4
	ADC	RF vh	8	SVM	Internal	0.84	0.94	0.94	0.78	0.85
					External	0.75	0.5	0.36	0.87	0.51
	T2+ADC	Relief	9	LASSO	Internal	0.86	0.95	0.92	0.77	0.84
					External	0.75	0.52	0.33	0.86	0.48
WG	T2	Relief	7	AdaBoost	Internal	0.77	0.91	0.84	0.75	0.79
					External	0.59	0.41	0.46	0.49	0.47
	ADC	AUC	4	SVM	Internal	0.69	0.87	0.83	0.67	0.74
					External	0.54	0.42	0.37	0.59	0.45
	T2+ADC	Relief	11	LASSO	Internal	0.78	0.91	0.83	0.79	0.81
					External	0.61	0.45	0.41	0.72	0.42

Table 4.47: The characteristics and the performance of the best models for each classification task.

selected radiomic features and the corresponding variable importance for the best models of the tumor and WG radiomics are shown in Figure 4.17.

4.2.8 Use Case 3

Exploratory Data Analysis of Clinical variables

Fig. 4.18 shows the distribution of each clinical variable used in UC3. We can see that the PSA histogram of the non-metastatic cases is skewed to the lower PSA values. Additionally, the proportion of metastatic cases is much higher for PIRADS 5 when compared to PIRADS 4.

Model Performance

Fig. 4.19 shows the cross-validation ROC-AUC model performance for the 32 models trained for UC3. Even though the differences are not extensive, the highest performance is achieved with a hybrid DWI dataset.

Model Selection

Given the results of the last section, we decided to move forward analyzing a hybrid model using DWI sequences and excluding ERC patients. The model's performance at a 0.5 probability decision threshold is shown in Table 4.48 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Fig. 4.20.

hybrid_uc3_DWI_noERC_SGD	
AUC	0.8077
Sensitivity/Recall/TPR	0.6155
Specificity/TNR	1.0
Precision/PPV	1.0
F1	0.7595
F2	0.6657
Cohen's Kappa	0.3466

Table 4.48: Multi-metric cross-validation performance of the hybrid_uc3_DWI_noERC_SGD model at 0.5 probability threshold.

Explainability analysis

Fig. 4.21 shows the SHAP analysis performed on the hybrid_uc3_DWI_noERC_SGD. There are no clinical variables among the 20 with the most impact on model output.

Fairness and sub-cohort analysis

Tables 4.49 through Table 4.54 show the fairness analysis for model hybrid_uc3_DWI_noERC_SGD. Regarding scanner manufacturer (Table 4.49), there is a significant drop in performance for SIEMENS cases (dropped around 15% in accuracy and F2-score), despite being well represented in the training set. Regarding index lesion location (Tables 4.88 to 4.90), we see significant drops in performance (up to 25% accuracy and F2-score) from the largest cohorts seen in training to the least represented ones, this is specially significant for the PZ, however the small cohort size (6 cases) doesn't allow fair conclusions to be drawn. Regarding country of origin (Table 4.91), the lowest performances are found for cases from Turkey and the Netherlands, despite being the second and third largest cohort seen in training.

manufacturer	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
PHILIPS	0.7537	0.7573	1	0.7175	31	5	26	31
SIEMENS	0.5899	0.5892	1	0.5381	22	2	20	22
GE MEDICAL SYSTEMS	0.7222	0.6467	1	0.6111	9	3	6	9

Table 4.49: hybrid_uc3_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train 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	0.6963	0.6909	1	0.6443	56	9	47	56
0	0.4444	0.4615	0.6667	0.4444	6	1	5	6

Table 4.50: hybrid_uc3_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train 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	0.7312	0.6952	1	0.6473	42	10	32	42
1	0.5333	0.5837	1	0.5333	20	0	20	20

Table 4.51: hybrid_uc3_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train 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	0.7037	0.6768	1	0.6269	48	10	38	48
1	0.5778	0.6277	1	0.5778	14	0	14	14

Table 4.52: hybrid_uc3_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train 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	0.6839	0.6665	1	0.6167	49	9	40	49
1	0.5278	0.5659	1	0.5159	13	1	12	13

Table 4.53: hybrid_uc3_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train set, divided by presence/absence of lesion in the anterior stroma.

4.2.9 Use Case 5 - post-surgery

Exploratory Data Analysis of Clinical variables

Fig. 4.22 shows the distribution of each clinical variable used in UC5. There is no variable where the distinction between the two classes is clear, however we can see that the proportion of biochemical recurrence

country	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Portugal	0.7955	0.8252	1	0.7955	21	0	21	21
Netherlands	0.5381	0.5020	1	0.4548	18	2	16	18
Turkey	0.6984	0.3456	0.6667	0.3111	17	8	9	17
Spain	0.8750	0.8947	1	0.8750	5	0	5	5
Italy	1	1	1	1	1	0	1	1

Table 4.54: hybrid_uc3_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train set, divided by country of origin of the data.

cases increases as the PIRADS increases. This proportion is also higher for the central ISUP grades, namely ISUP 2 and 3, and, as expected, it is also higher when there is perineural invasion, seminal vesicle invasion, extraprostatic extension or when the resection margins are positive after prostatectomy. A much higher proportion of recurrence was also found after laparoscopic prostatectomies compared with robotic-assisted or retro-pubic.

Model Performance

Fig. 4.23 shows the cross-validation and hold-out test set ROC-AUC model performance for the 32 models trained for UC5 (post-surgery context). At first glance, we can see that training with clinical data (radclin and hybrid models) significantly improves generalizability.

Model Selection

Given the results of the last section, we decided to move forward to analyzing a hybrid model using T2W sequences and excluding ERC patients. Due to the smaller size of the hold-out test set for this use case, we chose the decision threshold using the whole test set. Upon inspection of the respective ROC curve (Fig. 4.24), we decided the appropriate probability decision threshold would be 0.0528, above which the model outputs a high risk of biochemical recurrence. The model's performance at this probability decision threshold is shown in Table 4.55 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Fig. 4.25.

Explainability analysis

Fig. 4.26 shows the SHAP analysis performed on the hybrid_uc5_T2_noERC_CatBoost model. Several clinical variables are displayed among the top most influential to the model output: gleason1 (the most prominent histological pattern), baseline PSA and ISUP grade, which are positively associated with biochemical recurrence. The impact of the 2 categorical variables, extraprostatic extension and resection margin status, is shown in figures 4.27 and 4.28, respectively. Both the presence of extraprostatic extension and positive resection margins are associated with the biochemical recurrence output.

Fairness and sub-cohort analysis

Tables 4.56 through 4.60 show the fairness analysis for model hybrid_uc5_T2_noERC_CatBoost. Since there were no ERC cases or patients with index lesion located in the CZ, these two settings were removed from the subcohort analysis. In terms of scanner manufacturer (Table 4.56), the model generalizes well to philips data, but there are not enough samples on the hold-out test set to extract conclusions for GE.

Similarly, regarding lesion location (Table 4.57-4.59) and country of origin (Table 4.60), it is very difficult to conclude about the model's overall performance on the minority sub-cohorts, due to the low representation of these cases in the hold-out test set.

manufacturer	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
PHILIPS	0.6111	0.4688	0.1875	0.75	36	32	4	429
SIEMENS	0.8750	0.9375	0.7500	1	8	5	3	182
GE MEDICAL SYSTEMS	0.3333	0	0	0	3	3	0	65

Table 4.56: hybrid_uc5_T2_noERC_CatBoost model performance on sub cohorts of the hold-out 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	0.6667	0.625	0.3	0.8571	45	38	7	602
0	0	0	0	0	2	2	0	74

Table 4.57: hybrid_uc5_T2_noERC_CatBoost model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the peripheral zone.

index_lesion _location_TZ	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.7361	0.5952	0.2778	0.8333	42	36	6	561
1	0.6250	0.6250	0.2500	1	5	4	1	115

Table 4.58: hybrid_uc5_T2_noERC_CatBoost model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the transitional zone.

index_lesion _location_AS	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.7193	0.5682	0.25	0.8333	44	38	6	616
1	0.7500	0.8333	0.50	1	3	2	1	60

Table 4.59: hybrid_uc5_T2_noERC_CatBoost model performance on sub cohorts of the hold-out 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
Portugal	0.6216	0.5952	0.2778	0.8333	37	31	6	340
Spain	1	1	1	1	4	3	1	96
Lithuania	0	0	0	0	0	0	0	88
Turkey	0.6	0	0	0	5	5	0	82
Netherlands	0	0	0	0	0	0	0	68
Italy	0	0	0	0	1	1	0	2

Table 4.60: hybrid_uc5_T2_noERC_CatBoost model performance on sub cohorts of the hold-out test set, divided by country of origin of the data.

4.2.10 Use Case 5 - pre-surgery

Model Performance

Fig. 4.29 shows the cross-validation and hold-out test set ROC-AUC model performance for the 32 models trained for UC5 (pre-surgery context). Overall, in terms of cross-validation, it seems that the inclusion of clinical variables is detrimental to the performance. However, the highest generalizability is consistently found with hybrid models.

Model Selection

Given the results of the last section, we decided to move forward to analyzing a hybrid model using T2W sequences and all patients. Due to the smaller size of the hold-out test set for this use case, we chose the decision threshold using the whole test set. Upon inspection of the respective ROC curve (Fig. 4.30), we decided the appropriate probability decision threshold would be 0.0038, above which the model outputs a high risk of biochemical recurrence. The model's performance at this probability decision threshold is shown in Table 4.61 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Fig. 4.31.

Explainability analysis

Fig. 4.32 shows the SHAP analysis performed on the hybrid_uc5_T2_CatBoost model. Three clinical variables are displayed among the top 20 most influential to the model output: gleason1 (the most prominent histological pattern), baseline PSA and ISUP grade, which are positively associated with biochemical recurrence. Additionally, several radiomic features extracted from the log-sigma transformation of the T2 volume show predictive power.

Fairness and sub-cohort analysis

Tables 4.62 through 4.66 show the fairness analysis for model hybrid_uc5_T2_CatBoost. Since there was only one ERC case and no patients with index lesion located in the CZ, these two settings were removed from the subcohort analysis. In concordance with the post-surgery context, the model generalizes well to philips data (Table 4.62), but there are not enough samples on the hold-out test set to extract conclusions for GE.

manufacturer	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
PHILIPS	0.6111	0.4688	0.1875	0.75	36	32	4	429
SIEMENS	0.7500	0.8824	0.6000	1	8	5	3	182
GE MEDICAL SYSTEMS	0	0	0	0	3	3	0	98

Table 4.62: hybrid_uc5_T2_CatBoost model performance on sub cohorts of the hold-out test set, divided by scanner manufacturer.

Similarly, regarding lesion location (Table 4.63-4.65) and country of origin (Table 4.66), it is very difficult to conclude about the model's overall performance on the minority sub-cohorts, due to the low representation of these cases in the hold-out test set.

index_lesion _location_PZ	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
1	0.6222	0.6000	0.2727	0.8571	45	38	7	633
0	0	0	0	0	2	2	0	76

Table 4.63: hybrid_uc5_T2_CatBoost model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the peripheral zone.

index_lesion _location_TZ	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.7222	0.5814	0.2632	0.8333	42	36	6	592
1	0.5000	0.5556	0.2000	1	5	4	1	117

Table 4.64: hybrid_uc5_T2_CatBoost model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the transitional zone.

4.2.11 Use Case 6

Exploratory Data Analysis of Clinical variables

Fig. 4.33 shows the distribution of each clinical variable used in UC6. There is no variable where the distinction between the two classes is clear, however we can see that the proportion of biochemical recurrence

index_lesion _location_AS	ROC-AUC	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.6930	0.5435	0.2273	0.8333	44	38	6	649
1	0.7500	0.8333	0.50	1	3	2	1	60

Table 4.65: hybrid_uc5_T2_CatBoost model performance on sub cohorts of the hold-out 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
Portugal	0.5946	0.5814	0.2632	0.8333	37	31	6	340
Spain	1	1	1	1	4	3	1	96
Lithuania	0	0	0	0	0	0	0	88
Turkey	0.4	0	0	0	5	5	0	82
Netherlands	0	0	0	0	0	0	0	68
Italy	0	0	0	0	1	1	0	35

Table 4.66: hybrid_uc5_T2_CatBoost model performance on sub cohorts of the hold-out test set, divided by country of origin of the data.

cases increases as the PIRADS increases. This proportion is also higher when a patient presents a lesion in the PZ.

Model Performance

Fig. 4.34 shows the cross-validation ROC-AUC model performance for the 32 models trained for UC6. Even though the differences are not extensive, the highest performance is achieved with a raddeep DWI dataset.

Model Selection

Given the results of the last section, we decided to move forward analyzing a raddeep model using DWI sequences and excluding ERC patients. The model's cross-validation performance at a 0.5 probability decision threshold is shown in Table 4.67 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Fig. 4.35.

raddeep_uc6_DWI_noERC_SGD	
AUC	0.8393
Sensitivity/Recall/TPR	0.8056
Specificity/TNR	0.8730
Precision/PPV	0.5
F1	0.6111
F2	0.7119
Cohen's Kappa	0.5340

Table 4.67: Multi-metric cross-validation performance of the raddeep_uc6_DWI_noERC_SGD model at 0.5 probability threshold.

Given the model's perfect performance on the train set, the fairness analysis was not carried out, since the subcohort performance would also be perfect.

Explainability analysis

Fig. 4.36 shows the SHAP analysis performed on the raddeep_uc6_DWI_noERC_SGD model. Here, shape information seems to be highly impactful to the model output.

Fairness and sub-cohort analysis

Tables 4.68 through Table 4.73 show the fairness analysis for model raddeep_uc6_DWI_noERC_SGD. Though it is difficult to conclude about the performance on the smaller subcohort (SIEMENS), we can say the model generalized relatively well to GE data, with a drop in accuracy of around 17.6% but a rise in F2-score of 25% (Table 4.68). Regarding, index lesion location (Tables 4.70 to 4.72), the model shows minimal drops in performance on TZ, CZ and AS (< 8% reduction in accuracy). On PZ we see a larger drop in performance (Table 4.69), however the small cohort size (4 cases) doesn't allow fair conclusions to be drawn. Regarding country of origin (Table 4.73), the model performs relatively well on cases coming from Portugal or Spain (around 90% accuracy), but the performance drops almost 20% for cases from Lithuania and Italy.

manufacturer	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
PHILIPS	0.8982	0.6111	0.5	0.6667	59	53	6	59
GE MEDICAL SYSTEMS	0.7222	0.8586	0.5556	1	11	7	4	11
SIEMENS	0.6667	0	0	0	3	3	0	3

Table 4.68: raddeep_uc6_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train 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	0.8839	0.7405	0.5667	0.8056	69	59	10	69
0	0.5	0	0	0	4	4	0	4

Table 4.69: raddeep_uc6_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train 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	0.8789	0.7639	0.5	0.8889	58	51	7	58
1	0.8000	0.4629	0.5	0.5	15	12	3	15

Table 4.70: raddeep_uc6_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train 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	0.8472	0.6608	0.4333	0.6269	66	58	8	66
1	1	0.6667	0.6667	0.6667	7	5	2	7

Table 4.71: raddeep_uc6_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train 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	0.8656	0.7242	0.5333	0.8056	67	57	10	67
1	0.8333	0	0	0	6	6	0	6

Table 4.72: raddeep_uc6_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train set, divided by presence/absence of lesion in the anterior stroma.

4.2.12 Use Case 7a

The variables of the ecrf indicating the presence of side effects are four: Rectal Toxicity Acute, Rectal Toxicity Chronic, Urinary Toxicity Acute, and Urinary Toxicity Chronic. Each of them has been assessed

country	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
Portugal	0.9203	0.6111	0.5	0.6667	46	43	3	46
Lithuania	0.7389	0	0	0	12	10	2	12
Spain	0.8889	0.9444	0.8333	1	9	4	5	9
Italy	0.7222	0	0	0	6	6	0	6

Table 4.73: raddeep_uc6_DWI_noERC_SGD model cross-validation performance on sub cohorts of the train set, divided by country of origin of the data.

by clinicians with a score in $\{1, \ldots, 5\}$. Hence, the **UC7a** could be tackled in different ways, depending on which output we are interested in: to predict the generic presence of side effects, or to predict independently both urinary toxicity and rectal toxicity. As the sample size is not so big (less than 150 cases), we decided to train binary classification models addressing different classification tasks, namely the prediction of (i) rectal toxicity and urinary toxicity independently; (ii) any side effect, but weighting more the chronic side effect with respect to the acute one.

Independent prediction of Rectal Toxicity and Urinary Toxicity

In **UC7a** with rectal toxicity, a total of 136 patients (ProstateNet) were studied using the T2W axial images. The whole gland masks for these patients were derived utilizing the pre trained `nnU-net` model. These masks were automatically generated for all 136 patients, with the distribution between Grade 1 and Grade 2 being [119, 17] respectively.

Moreover, in **UC7a** with urinary toxicity, a total of 136 patients were studied using T2W axial images. The whole gland masks for these patients were derived utilizing the pre-trained `nnU-net` model. These masks were automatically generated for all 136 patients, with the distribution between Grade 1 and Grade 2 being [94, 40] respectively

Methodology

For each patient, 642 radiomics were derived using bin widths of 20, 50, and 80. To measure the performance of the models, a 3-fold cross-validation was conducted over 100 iterations. Several feature selection methods, including Variance threshold, ANOVA, and L1-based methods, were employed. To determine the optimal classifier, a range of seven classifiers was assessed, encompassing `GPC_RBF`, `KNN`, `Decision tree`, `polynomial SVM`, `Linear SVM`, `rbf_SVM`, and `sigmoid_SVM`. By experimenting with different values, the optimal values for the hyperparameters k (number of features to be selected) and `c_value` (regularization) were identified as k being 40, 80, or 120 and `c_value` being 0.4, 0.8, or 1. Finally, learning curves were generated for the best performing classifier.

Results

In analyzing Rectal and Urinary Toxicity under Use Case 7a, a series of classifiers and parameter combinations were examined to identify the optimal model for each case. The results, summarized in Table 4.74, reveal that for Rectal Toxicity, the Sigmoid SVM classifier, with a bin width of 80, k best of 80, and C value of 1, yielded the highest performance across all metrics. Similarly, for Urinary Toxicity the Sigmoid SVM classifier, with a bin width of 80, k best of 80, and C value of 0.4, demonstrated optimal performance.

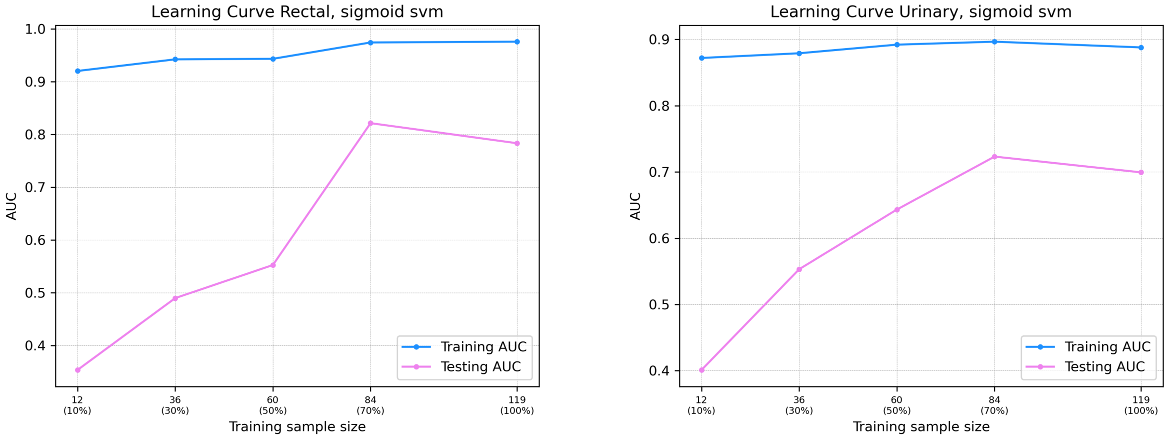
Discussion

The confusion matrices, Figure 4.37, for Rectal and Urinary Toxicity illustrate both the strengths and weaknesses of our models. For Rectal Toxicity, the model correctly identified most of the Grade 1 samples but struggled with several False Positives, indicating a tendency to overestimate the presence of Grade 2 toxicity. In contrast, the Urinary Toxicity model demonstrated a fair balance in identifying both grades but still had misclassifications. Finally, the learning curves Figure 4.38 reveal that the models are inclined to overfit and exhibit poor performance on the test set, especially when the amount of training data is reduced. We also observe that the model yields the best results on the test set when 70% of the total amount of data

	Rectal Toxicity	Urinary Toxicity
Classifier	Sigmoid SVM	Sigmoid SVM
Parameters	Bin width 80 k best 80 C value 1	Bin width 80 k best 80 C value 0.4
ACC_MEAN	0.7214 [± 0.046]	0.6694 [± 0.03]
AUC_MEAN	0.7835 [± 0.03]	0.6988 [± 0.037]
SN_MEAN	0.7654 [± 0.097]	0.7323 [± 0.05]
SP_MEAN	0.6774 [± 0.03]	0.606 [± 0.04]
F1_MEAN	0.7179 [± 0.06]	0.6836 [± 0.03]

Table 4.74: Comparison of Results for Rectal and Urinary Toxicity.

is used for training. We suspect this is attributable to the removal of data with an endorectal coil during the data reduction process.



(a) Rectal Toxicity

(b) Urinary Toxicity

Figure 4.38: Learning curves from the optimal models for Rectal and Urinary Toxicity.

Prediction of any side effect

In **UC7a** a total of 134 patients (ProstateNet) were studied using the T2W axial images. The whole gland masks for these patients were provided by the Champalimoud team. These masks have been visually inspected and some correction have been applied with respect to the presence of multiple connected components (occurring in 29/134 cases), and to any other issues (e.g. one mirrored mask), Fig. 4.39.

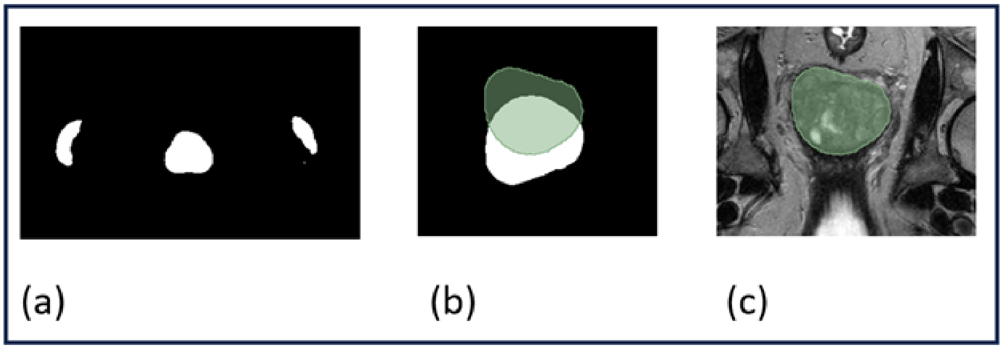


Figure 4.39: Segmentation issues. In (a) multiple connected components (PatientID: PCa-73840158220672581509385408987452733166); in (b) the original mask (white) and the mirrored one (green), which perfectly fits the prostate shape in (c) (PCa-309443526081327790079195286136548121193).

Among the 134 T2W axial series, about the 50% of cases (66 patients) were acquired using an endorectal coil. Hence, we decided to perform three different experimentation for predicting side effects after radiotherapy: the first using all the subjects (All_subjs), the second and the third one splitting the dataset with respect to the coil used (ERC, and noERC).

Looking at the sum of the grading over the 4 variables for each patient, we obtained a population with the distribution of side effects reported in Table 4.75

Side effect sum	4	5	6	7	8	9	10	11	12
# of patients	65	16	32	10	9	0	1	0	1

Table 4.75: Distribution of the cumulative grades of side effects in the UC7a population, not regarding the type of side effect.

This distribution could be used to split the dataset into two classes: subjects with no side effects at all (all grading are ones), and subjects with at least one side effect assessed as grade 2. This splitting would result in a binary classification: 65 subjects with a sum equal to 4 and 69 subjects with a sum greater than 4.

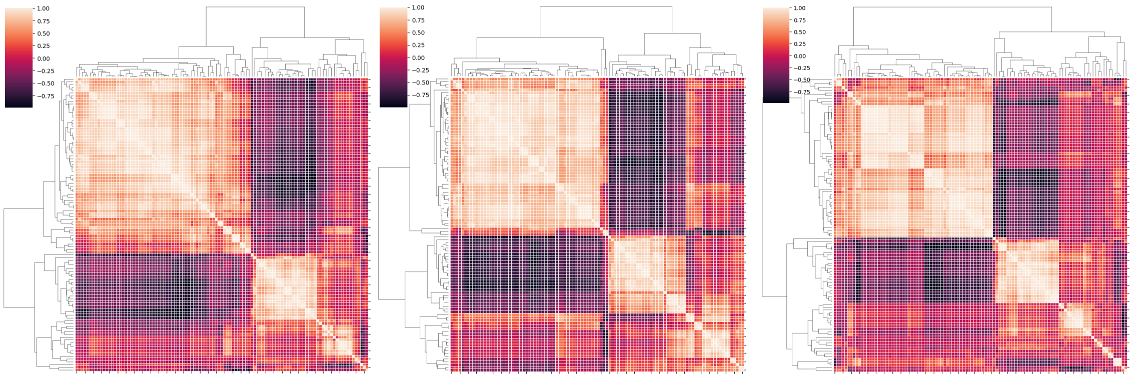
Conversely, both type of side effects (Rectal and Urinary Toxicity) has been evaluated in terms of their nature, distinguishing between Acute and Chronic effects. Naturally, the chronic side effects hold greater significance compared to the acute ones. Therefore, a division that better respects the clinical context was chosen: in the class with no side effects were included also the subjects who showed a grade 2 in only one among rectal toxicity acute and urinary toxicity acute. Such a ground truth produced a [79, 55] splitting in the whole dataset: 79 subjects with no or light side effects and 55 subjects with side effects.

In the end, also considering the coil used for the MRI examination, we have the following splittings for the three dataset:

- All subjects: 79 subjects without side effects, hence [79, 55];
- ERC: 29 subjects without side effects, hence [29, 37];
- noERC: 50 subjects without side effects, hence [50, 18].

Methodology

The radiomic features were derived from the original T2W images labeled with the prostate mask, using Simple-ITK and Pyradiomics libraries. In the end, 107 radiomic features (belonging to the classes: firstorder, shape, glszm, ngtdm, glcm, glrlm, gldm) were extracted for each patient using the default bin width of 25. The clustering plot in Figure 4.40 shows the redundancy among the radiomic features extracted.



(a) All subjects (b) ERC (c) noERC

Figure 4.40: Radiomic features: clustering analysis.

The PCA analysis showed that in each dataset the 85% of features variance is retained in the first 4 principal components. For each dataset, the pairs of features with Pearson correlation > 0.9 have been identified in order to reduce the number of radiomic features to train the classification models: (i) for the whole dataset we got 36 features, (ii) for the ERC dataset we reduced to 25 features, and (iii) for the noERC dataset to 33 features.

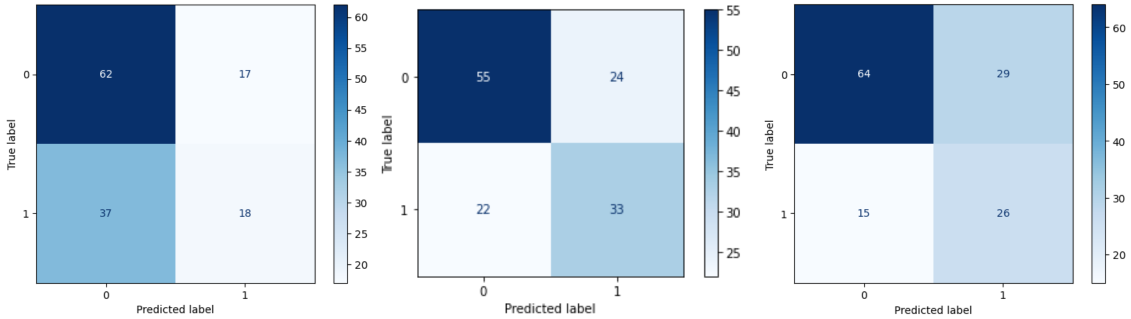
To study **UC7a**, we trained 3 different types of models: a classical k-nearest neighbor (kNN) with $k = 2$, a Support Vector Machine (SVM) with $kernel$='sigmoid', $gamma$= 10 and C=1 and a Random Decision Forest (RDF), trained on 500 trees.
The performance of the trained models were compared using a 5-fold cross-validation scheme (i.e., each model was trained 5 times, using 4 folds as the training set and evaluating performance on the fifth unseen set). Patients were stratified by both ground truth and acquisition modality (ERC/noERC) in order to ensure that each fold was homogeneous compared to the others.
Metrics were calculated on the test fold for each of the 5-folds. The area under the receiver operating characteristic (AUROC) of each of the test fold was reported [*Median AUROC*]. The score was given as median and [0%-100%] percentiles due to the small number of samples and the non-normality of the values. Considering the size of each fold (median value 27 [25-28]), an aggregated score can be more stable to the splitting procedure. Therefore, we defined the *Combined AUROC* as the AUROC calculated by aggregating the 5 test folds. The DeLong confidence interval was added to this score. The *Confusion Matrix*, obtained as the sum of the confusion matrices of each fold, where the selection thresholds are calibrated using the 4 training folds, was also reported. Using this matrix we derived the following scores: *Sensitivity*, *Specificity*, and *Balanced Accuracy* (to handle data imbalances).

Results

Tables 4.76, 4.77, and 4.78 report all the performance metrics for the three models trained on each dataset, while confusion matrices are displayed in Figures 4.41, 4.42, and 4.43. Considering only the balanced accuracy, classification based on kNN is always the worst; while SVM and RDF achieved similar performances: both of them showed near 65% of balanced accuracy in the whole dataset experimentation. On the other hand, in the other two dataset, SVM models showed a stronger stability, while, probably due to the small number of subjects, the performances of RDF dropped down. Also, the SVM showed the best balance between sensitivity and specificity both in the whole (0.6 sens. and 0.7 spec.) and in the noERC dataset (0.61 sens. and 0.68 spec.), despite the limited sample size.

All subjects	**Median AUROC**	**Combined AUROC**	**Sensitivity**	**Specificity**	**Balanced Accuracy**
kNN	-	-	0.785	0.327	0.556
SVM	0.729 [0.578-0.787]	0.692 CI [0.603-0.786]	0.6	0.696	0.648
RDF	0.684 [0.677-0.787]	0.698 CI [0.605-0.792]	0.473	0.81	0.641

Table 4.76: Performance metrics of the three models (kNN, SVM, RDF) trained on the whole dataset.

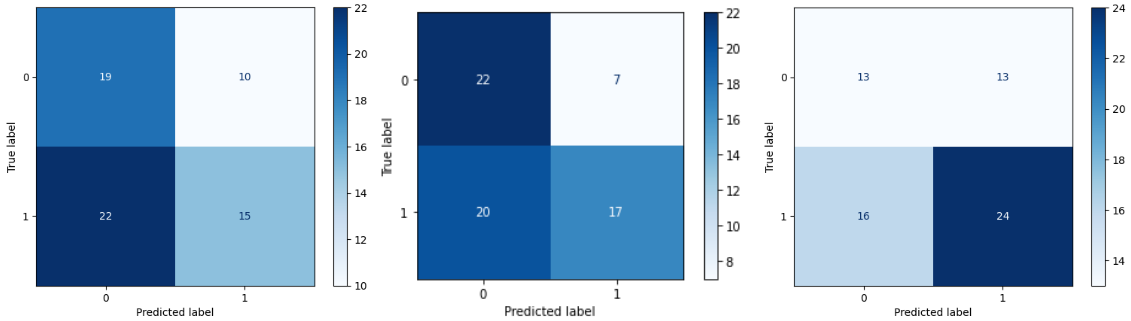


(a) kNN (b) SVM (c) RDF

Figure 4.41: Confusion Matrices from the optimal models, for the whole dataset.

ERC	**Median AUROC**	**Combined AUROC**	**Sensitivity**	**Specificity**	**Balanced Accuracy**
kNN	-	-	0.655	0.405	0.53
SVM	0.628 [0.604-0.833]	0.640 CI [0.503-0.771]	0.459	0.758	0.609
RDF	0.586 [0.52-0.75]	0.571 CI [0.43-0.712]	0.649	0.448	0.549

Table 4.77: Performance metrics of the three models (kNN, SVM, RDF) trained on the *ERC* dataset.

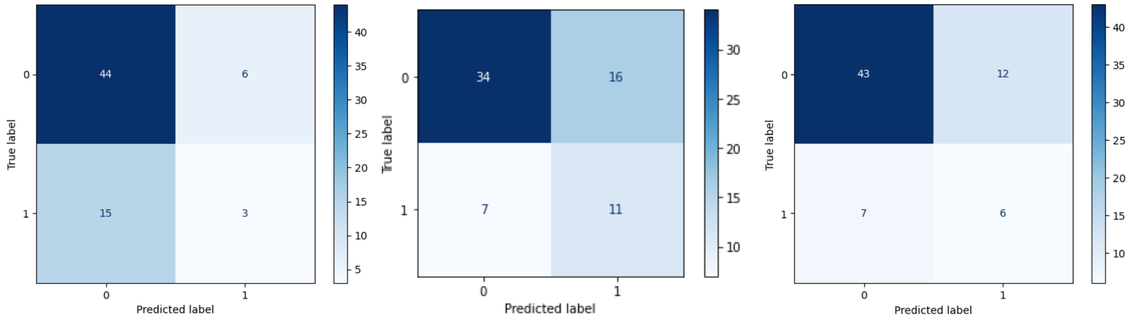


(a) kNN (b) SVM (c) RDF

Figure 4.42: Confusion Matrices from the optimal models, for the *ERC* dataset.

noERC	**Median AUROC**	**Combined AUROC**	**Sensitivity**	**Specificity**	**Balanced Accuracy**
kNN	-	-	0.88	0.167	0.523
SVM	0.775 [0.46-0.866]	0.706 CI [0.564-0.848]	0.611	0.68	0.645
RDF	0.60 [0.533-0.90]	0.685 CI [0.535-0.835]	0.333	0.86	0.597

Table 4.78: Performance metrics of the three models (kNN, SVM, RDF) trained on the *noERC* dataset.



(a) kNN (b) SVM (c) RDF

Figure 4.43: Confusion Matrices from the optimal models, for the *noERC* dataset.

Discussion

Both SVM and RDF performed quite well in predicting the presence of side effects after radiotherapy, even if SVM showed a stronger ability of exploiting the homogeneity of the training dataset, despite the reduction of the sample size (clear if considering performance metrics achieved in the whole dataset and in the ERC dataset).

On the other hand, the RDF models are more explainable ML tools, as they make it possible to perform an importance analysis of the radiomic features. In Table 4.79 the 5 features scoring highest for each dataset are reported: it's worth to remark that most of them are texture features, as expected, even if some features derived from the prostate shape (e.g. elongation, Max 2D diameter) also occur.

All subjects	***ERC***	***noERC***
GLCM Correlation 5.7	Large Area Emphasis 4.0	Elongation 4.8
Elongation 4.5	GLSZM Size Zone Non Uniformity 2.7	Max 2D diameter 3.9
GLSZM Zone Entropy 4.2	Minimum 2.6	Least Axis Length 3.3
GLRLM Run Length Non Uniformity 4.1	GLCM Correlation 0.7	Flatness 2.5
Maximum 3.8	GLCM Inverse Variance 0.6	GLCM LDMN 2.0

Table 4.79: Mean decrease accuracy for the RDF models. The 5 highest medians of the folds are reported.

4.2.13 Use Case 7b

Model Performance (multiclass)

Figures 4.45 and 4.46 show, respectively, the cross-validation and hold-out test set ROC-AUC and Brier model performance for the 48 models trained for UC7b.

Observing Fig. 4.44, we can see that during cross-validation the support vector classifier (SVC) produces the best results when evaluating ROC-AUC, however, these do not hold when testing on the holdout data. There we can observe a very high degree of overfitting from the SVC models, while the remaining ones produce more consistent results. Overall, the stochastic gradient descent classifier (SGD) is the one that generalizes the best, with its best version using Raddeep data with ADC features, obtaining a ROC-AUC score of 0.60. However, when observing the Brier score, which provides an estimate of the model calibration, we can see that despite providing the best ROC-AUC scores, the SGD models are highly unreliable. On the other hand, a model that provides both a high ROC-AUC (0.55) and low Brier (0.68) is the CatBoost trained on Radclin data with DWI features, which will be our chosen model.

In order to assess the impact of the different modalities, we compare the both ROC-AUC and Brier score obtained on each modality, stratified by both dataset type and predictive model type (Figs. 4.45 and 4.46). During cross-validation, it can be observed that, regarding ROC-AUC, both Raddeep and SVC tend to provide, generally, the best results, despite generating more miscalibrated classifiers in the case of Raddeep. Looking at the holdout results, it can be seen that the results do not hold, with Radclin and SGD providing the best models, although the SGD results are severely miscalibrated. Interestingly, neither in cross-validation nor in holdout is there a clear best modality, with all of them producing very similar results.

Model Selection

Given the results of the previous section, we decided to move forward analyzing the CatBoost Radclin model using DWI sequences. Observing the confusion matrix (Fig. 4.47) we can see that the model is severely underperforming, throwing the vast majority of predictions to the “sufficient” class (class 2). Inspecting the ROC (Fig. 4.48) curve confirms this as we can observe that all thresholds have very similar values.

Explainability analysis

From Fig. 4.49 to Fig. 4.54, the SHAP analysis performed on the CatBoost model using Radclind data and DWI features is shown. For each class two plots are shown, the standard shap summary plot and an additional one describing the impact of each category of a categorical variable on model output.

Regarding class 1, or low quality of life (Fig. 4.49), the model is relying heavily on three clinical variables, namely extra prostatic extension, PIRADS and resection margin status, and on one shape feature, Sphericity, inversely associated with the low quality of life output. Lower in the plot we can also find the age, positively associated with low quality of life, and shape Flatness with an inverse relationship to the target. Analyzing Fig. 4.50 we find the presence of extraprostatic extension associated with the low quality of life output. Surprisingly, all PIRADS values are positively associated with a lower quality of life, with the exception of PIRADS 5 which shows an inverse relationship to the target. It was also found that negative resection margins lead to a higher risk of low quality of life.

Fairness and sub-cohort analysis

Tables 4.80 to 4.85 show the fairness analysis for the CatBoost model using Radclind data and DWI features is shown. This analysis was, for the most part, inconclusive due to the low representation of minority subgroups. Regarding scanner manufacturers (Table 4.80), the model seems robust to the two scanners represented in the hold-out test set (PHILIPS and SIEMENS). Regarding PZ lesion location (Table 4.81), CZ lesion location (Table 4.83) and country (Table 4.85), no conclusions can be drawn due to the low representation of minority subgroups, with only 3, 2, and 1 patients respectively in the minority sub-cohorts. Regarding TZ (Table 4.82) and AS (Table 4.84), we can cautiously conclude that the model performs better when there is a lesion in these areas.

manufacturer	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_1	Test counts target_2	Test counts target_3	Train counts
PHILIPS	0.4	0.379132	0.391534	0.4	40	12	16	12	157
SIEMENS	0.555556	0.498575	0.555556	0.555556	9	2	4	3	47

Table 4.80: CatBoost model using Radclind data and DWI features performance on sub cohorts of the hold-out test set, divided by scanner manufacturer.

Index_lesion _location_PZ	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_1	Test counts target_2	Test counts target_3	Train counts
1	0.434783	0.407855	0.427323	0.434783	46	13	20	13	193
0	0.333333	0.37037	0.666667	0.333333	3	1	0	2	13

Table 4.81: CatBoost model using Radclind data and DWI features performance on sub cohorts of the hold-out 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_1	Test counts target_2	Test counts target_3	Train counts
0	0.409091	0.382223	0.410349	0.409091	44	12	18	14	179
1	0.6	0.530303	0.366667	0.6	5	2	2	1	27

Table 4.82: CatBoost model using Radclind data and DWI features performance on sub cohorts of the hold-out 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_1	Test counts target_2	Test counts target_3	Train counts
0	0.425532	0.401386	0.426821	0.425532	47	13	19	15	190
1	0.5	0.416667	0.25	0.5	2	1	1	0	16

Table 4.83: CatBoost model using Radclind data and DWI features performance on sub cohorts of the hold-out 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_1	Test counts target_2	Test counts target_3	Train counts
0	0.418605	0.392187	0.430523	0.418605	43	12	18	13	187
1	0.5	0.469697	0.388889	0.5	6	2	2	2	19

Table 4.84: CatBoost model using Radclind data and DWI features model performance on sub cohorts of the hold-out test set, divided by presence/absence of lesion in the anterior stroma.

country	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_1	Test counts target_2	Test counts target_3	Train counts
Portugal	0.4375	0.412941	0.433824	0.4375	48	13	20	15	197
Spain	0	0	0	0	1	1	0	0	9

Table 4.85: CatBoost model using Radclind data and DWI features performance on sub cohorts of the hold-out test set, divided by country of origin of the data.

4.2.14 Use Case 8

Exploratory Data Analysis of Clinical variables

Fig. 4.55 shows the distribution of each clinical variable used in UC8. No patient appears to have the index lesion located in the CZ, so this variable is not informative. Regarding the remaining variables, there is none where the distinction between the two classes is clear.

Model Performance

Fig. 4.56 shows the cross-validation ROC-AUC model performance for the 32 models trained for UC8. Here, we see the performance is enhanced in the presence of deep features (hybrid and raddeep datasets). Even though the differences are not extensive, the highest performance is achieved with a raddeep ADC dataset.

Model Selection

Given the results of the last section, we decided to move forward analyzing a raddeep model using ADC sequences and excluding ERC patients. The model's cross-validation performance at a 0.5 probability decision threshold is shown in Table 4.86 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Fig. 4.57.

raddeep_uc8_ADC_noERC_SGD	
AUC	0.6905
Sensitivity/Recall/TPR	0.6667
Specificity/TNR	0.7143
Precision/PPV	0.15
F1	0.2444
F2	0.3935
Cohen's Kappa	0.1361

Table 4.86: Multi-metric cross-validation performance of the raddeep_uc8_ADC_noERC_SGD model at 0.5 probability threshold.

Given the model's perfect performance on the train set, the fairness analysis was not carried out, since the subcohort performance would also be perfect.

Explainability analysis

Fig. 4.58 shows the SHAP analysis performed on the raddeep_uc8_ADC_noERC_SGD model.

Fairness and sub-cohort analysis

Tables 4.87 through Table 4.91 show the fairness analysis for model raddeep_uc8_ADC_noERC_SGD. Though it is difficult to conclude about the performance on the smaller subcohorts, we can say the model generalized relatively well to PHILIPS data, with a drop in accuracy of around 6% (Table 4.87). Regarding, index lesion location (Tables 4.88 to 4.90) and country of origin (Table 4.91), the model shows minimal drops in performance (1.5% - 6% reduction in accuracy) from the largest cohorts seen in training to the least represented ones. We also observe a rise in accuracy, but only on subcohorts composed uniquely of the majority class label. This is the case for GE MEDICAL SYSTEMS (Table 4.87), index lesions located in the anterior stroma (Table 4.90) and cases from Turkey (Table 4.91).

manufacturer	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
SIEMENS	0.7273	0.4273	0.1769	0.6667	81	73	8	81
PHILIPS	0.6667	0	0	0	10	10	0	10
GE MEDICAL SYSTEMS	1	0	0	0	1	1	0	1

Table 4.87: raddeep_uc8_ADC_noERC_SGD model cross-validation performance on sub cohorts of the train 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	0.7237	0.4074	0.1619	0.6667	58	54	4	58
0	0.7071	0.3704	0.1333	0.6667	34	30	4	34

Table 4.88: raddeep_uc8_ADC_noERC_SGD model cross-validation performance on sub cohorts of the train 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	0.7339	0.4257	0.1742	0.6667	74	68	6	74
1	0.6714	0	0	0	18	16	2	18

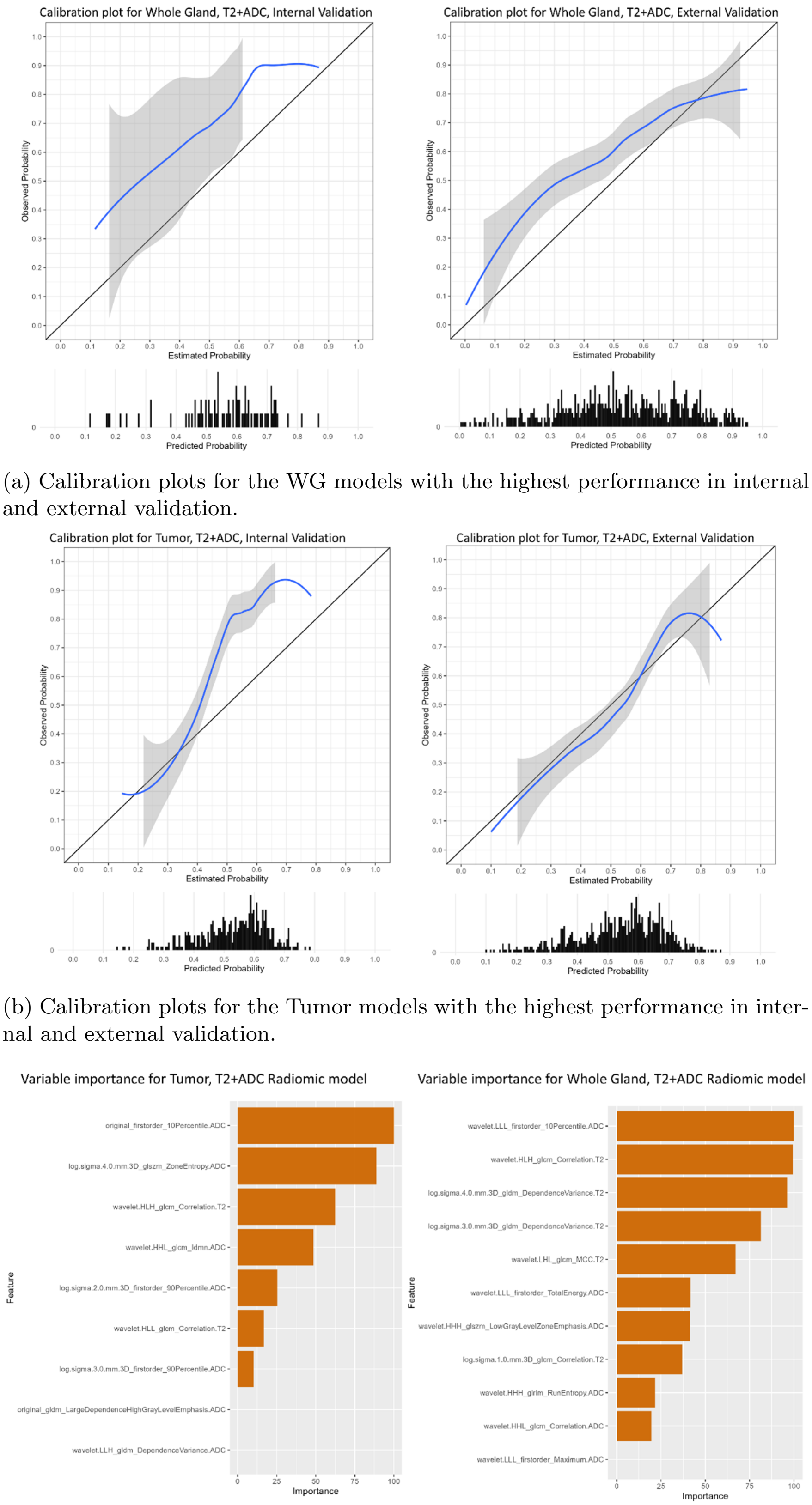
Table 4.89: raddeep_uc8_ADC_noERC_SGD model cross-validation performance on sub cohorts of the train set, divided by presence/absence of lesion in the transitional zone.

index_lesion _location_AS	accuracy	fbeta_2	precision	recall	Test counts	Test counts target_0	Test counts target_1	Train counts
0	0.7059	0.4125	0.1667	0.6667	82	74	8	82
1	0.8333	0	0	0	10	10	0	10

Table 4.90: raddeep_uc8_ADC_noERC_SGD model cross-validation performance on sub cohorts of the train 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	0.7037	0.4273	0.1769	0.6667	75	67	8	75
Portugal	0.7222	0	0	0	14	14	0	14
Turkey	1	0	0	0	3	3	0	3

Table 4.91: raddeep_uc8_ADC_noERC_SGD model cross-validation performance on sub cohorts of the train set, divided by country of origin of the data.



(a) Calibration plots for the WG models with the highest performance in internal and external validation.

(b) Calibration plots for the Tumor models with the highest performance in internal and external validation.

(c) Variable importance of the best radiomic models for the tumor and the WG.

Figure 4.17: Calibration plots for the (a) WG's best performing model in internal (left) and external (right) sets, (b) tumor's best performing model in internal (left) and external (right) sets and (c) The variable importance of tumor (left) and WG (right).

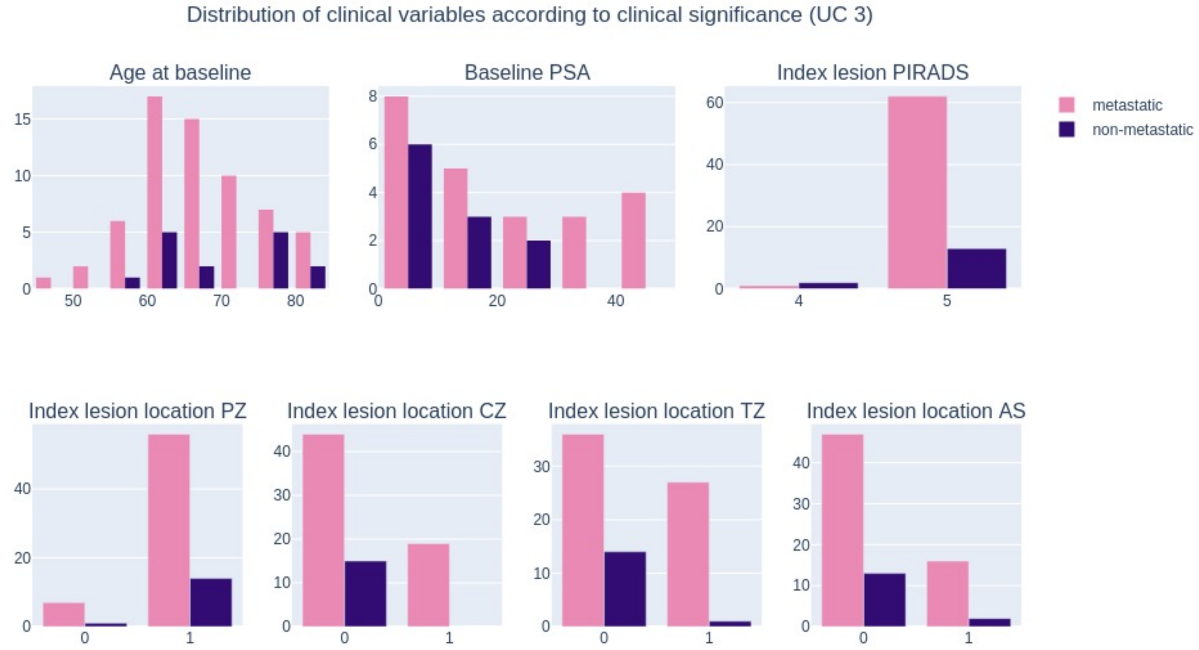


Figure 4.18: Distribution of clinical variables according to development of metastatic disease whithin 6 months.

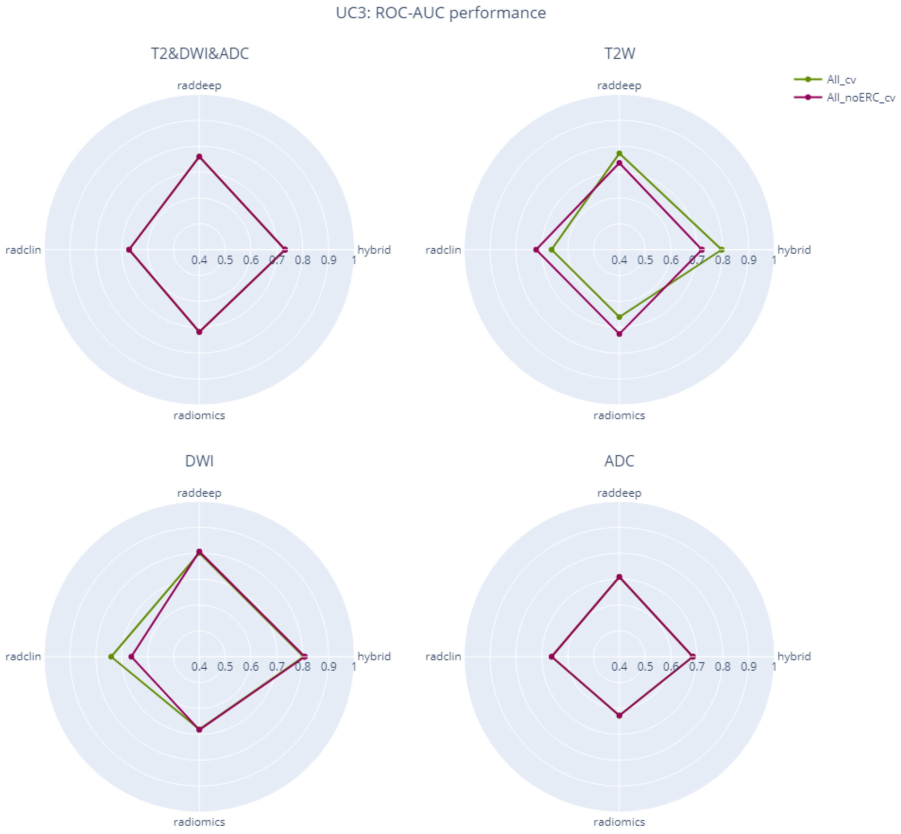


Figure 4.19: Cross-validation ROC-AUC model performance of 32 models trained to predict metastatic development in UC3. The observations are color coded according to the inclusion or exclusion of endorectal-coil exams: all patients in green; excluding ERC cases in pink.

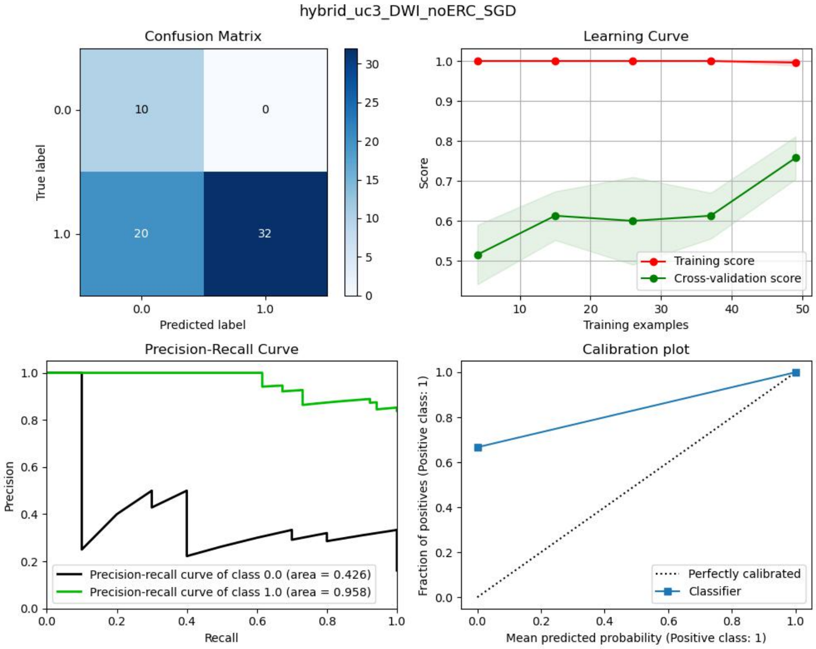


Figure 4.20: Analysis of the hybrid_uc3_DWI_noERC_SGD model in terms of cross-validation confusion matrix, learning curve, precision-recall curve and calibration plot.

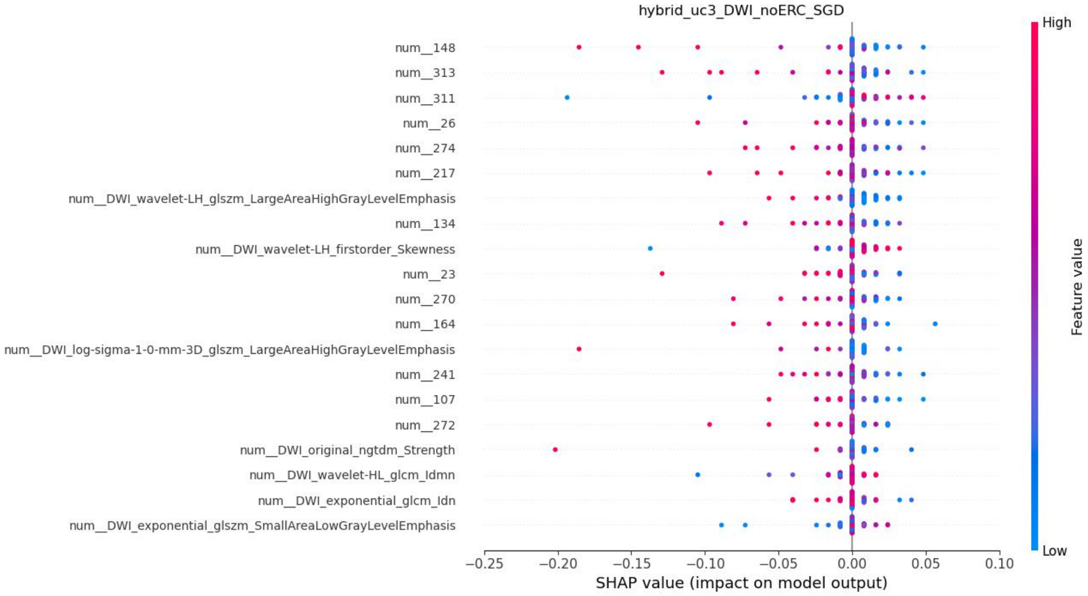


Figure 4.21: SHAP values contribution of features for hybrid_uc3_DWI_noERC_SGD model.

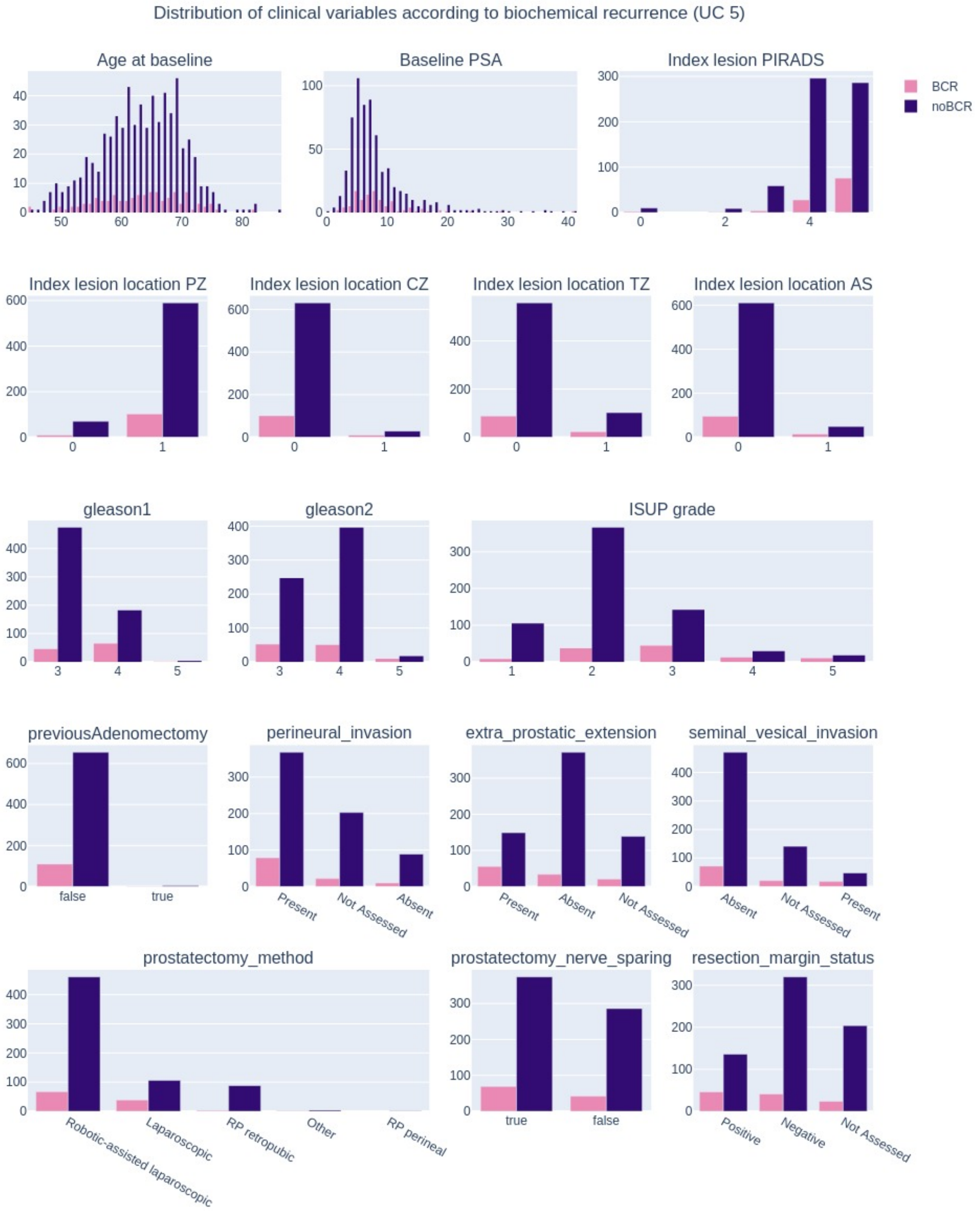


Figure 4.22: Distribution of clinical variables according to biochemical recurrence.

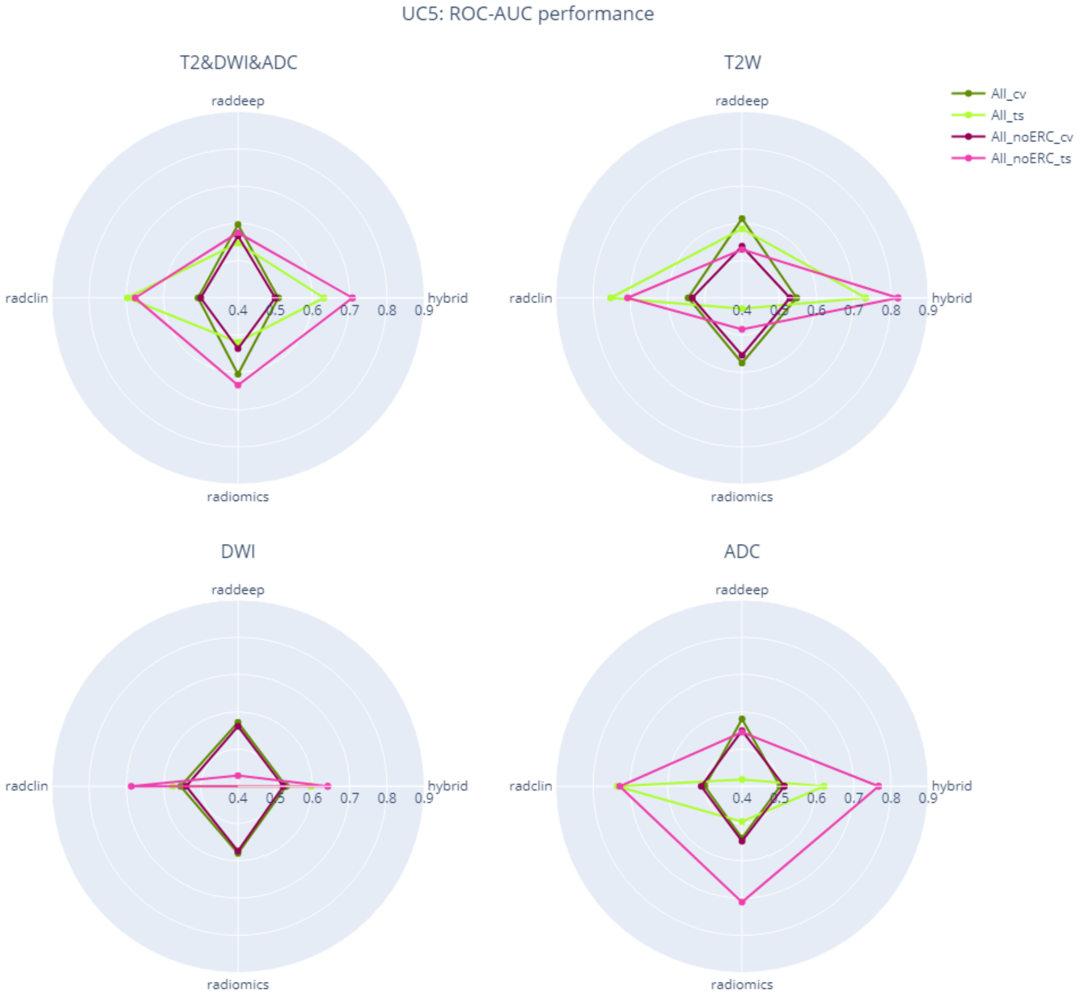


Figure 4.23: Cross-validation and hold-out test set ROC-AUC model performance of 32 models trained to predict biochemical recurrence in UC5. The observations are color coded according to the inclusion or exclusion of endorectal-coil exams: all patients in green; excluding ERC cases in pink.

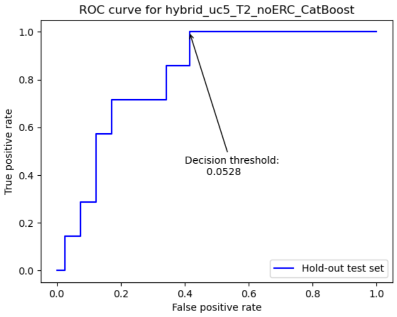


Figure 4.24: Hold-out test set receiver operator characteristics curve of the hybrid_uc5_T2_noERC_CatBoost model.

hybrid_uc5_T2_noERC_CatBoost	
AUC	0.8188
Sensitivity/Recall/TPR	0.8571
Specificity/TNR	0.5854
Precision/PPV	0.2609
F1	0.4000
F2	0.5882
Cohen's Kappa	0.2272

Table 4.55: Multi-metric performance of the hybrid_uc5_T2_noERC_CatBoost model at 0.0528 probability threshold.

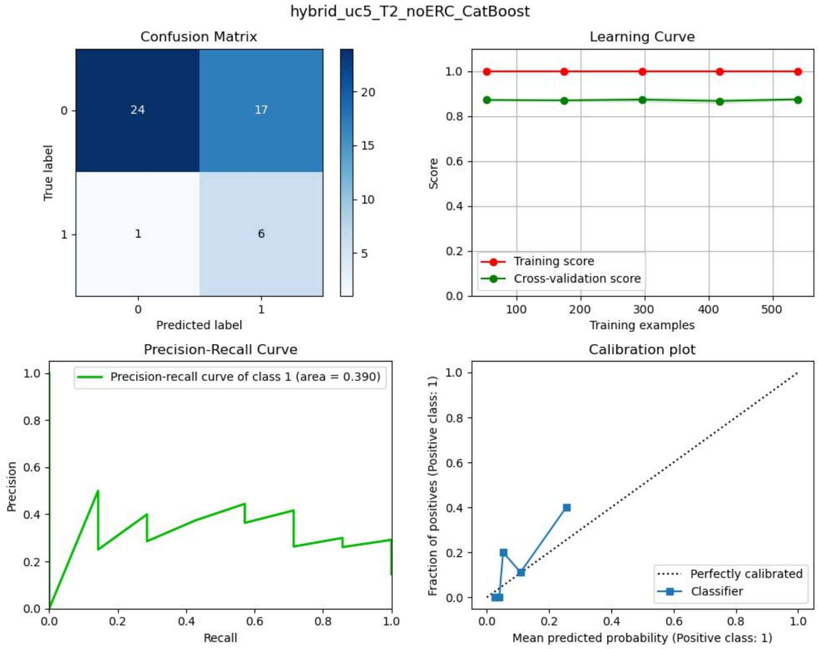


Figure 4.25: Analysis of the hybrid_uc5_T2_noERC_CatBoost model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

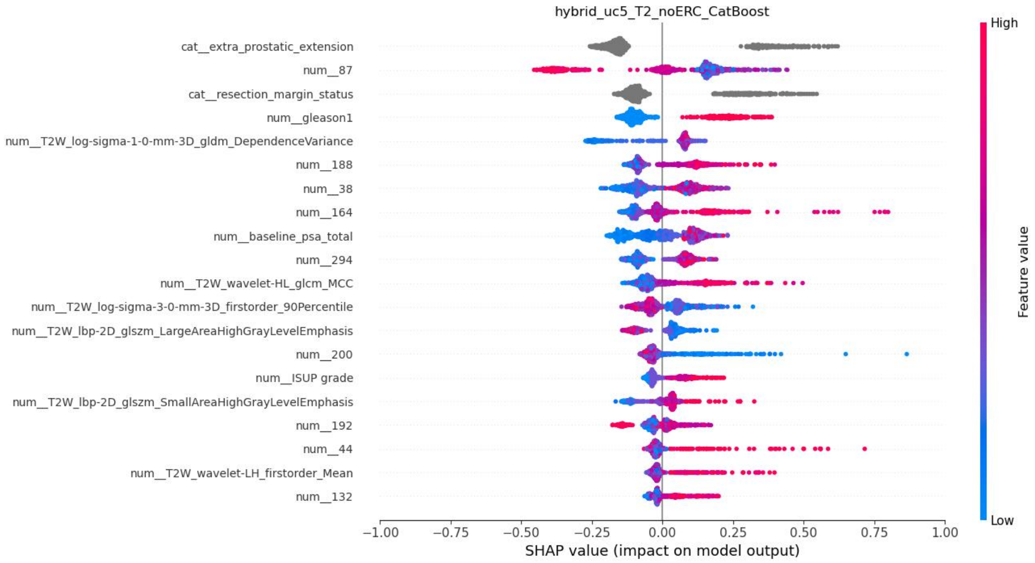


Figure 4.26: SHAP values contribution of features for hybrid_uc5_T2_noERC_CatBoost model.

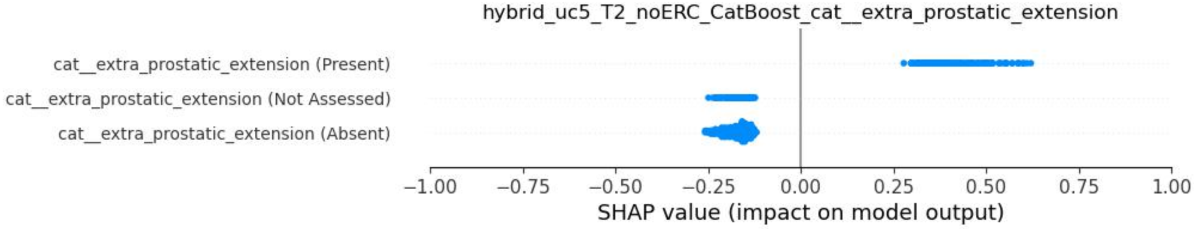


Figure 4.27: SHAP values contribution of the categorical feature extraprostatic extension for the hybrid_uc5_T2_noERC_CatBoost model output.

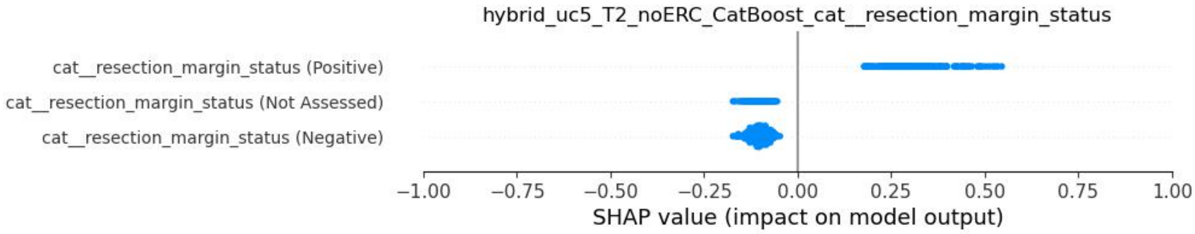


Figure 4.28: SHAP values contribution of the categorical feature resection margin status for the hybrid_uc5_T2_noERC_CatBoost model output.

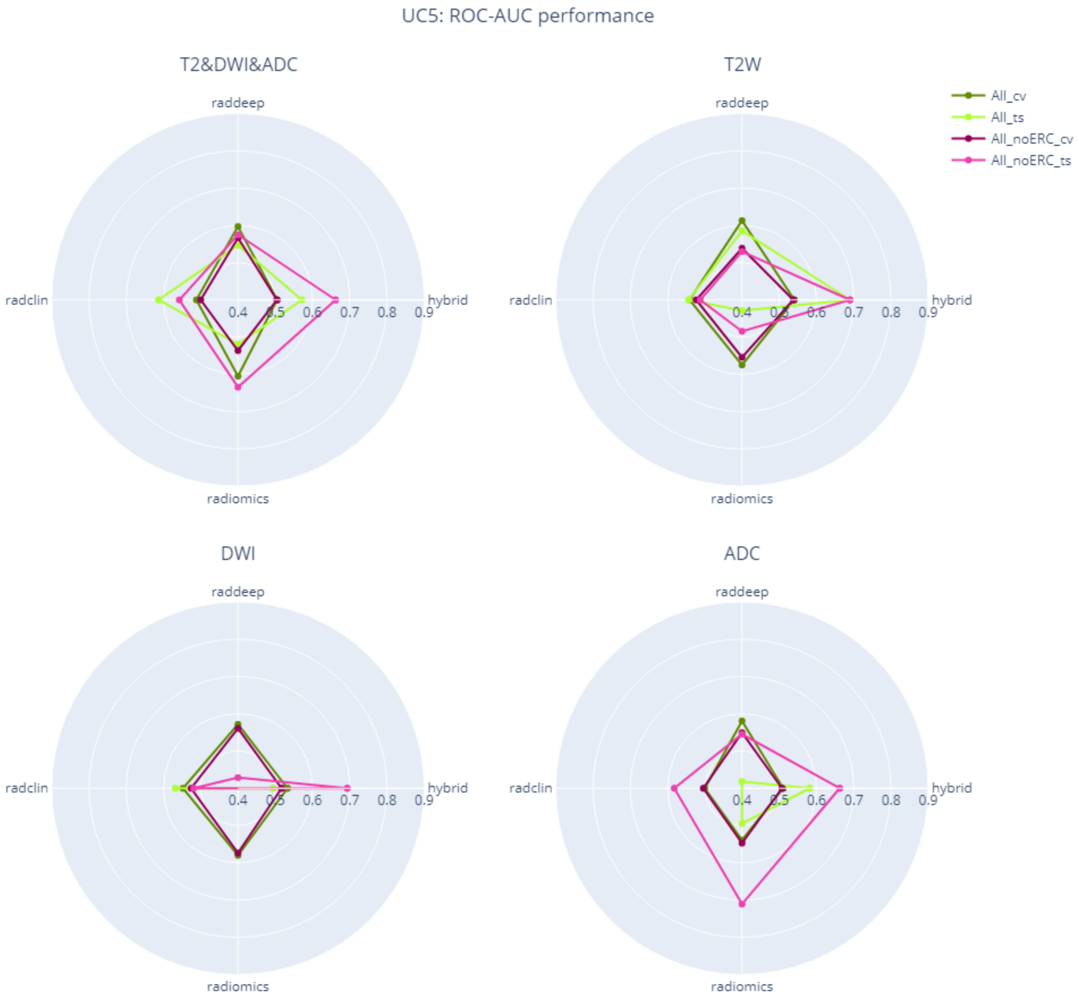


Figure 4.29: Cross-validation and hold-out test set ROC-AUC model performance of 32 models trained to predict biochemical recurrence in UC5 (pre-surgery context). The observations are color coded according to the inclusion or exclusion of endorectal-coil exams: all patients in green; excluding ERC cases in pink.

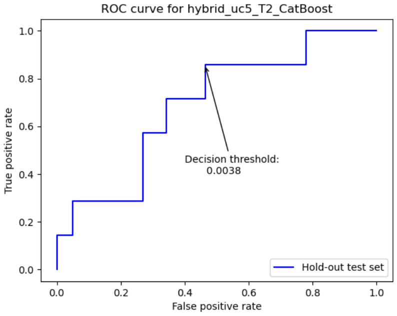


Figure 4.30: Hold-out test set receiver operator characteristics curve of the hybrid_uc5_T2_CatBoost model.

hybrid_uc5_T2_CatBoost	
AUC	0.6899
Sensitivity/Recall/TPR	0.7143
Specificity/TNR	0.5366
Precision/PPV	0.2083
F1	0.3226
F2	0.4808
Cohen's Kappa	0.1250

Table 4.61: Multi-metric performance of the hybrid_uc5_T2_CatBoost model at 0.0038 probability threshold.

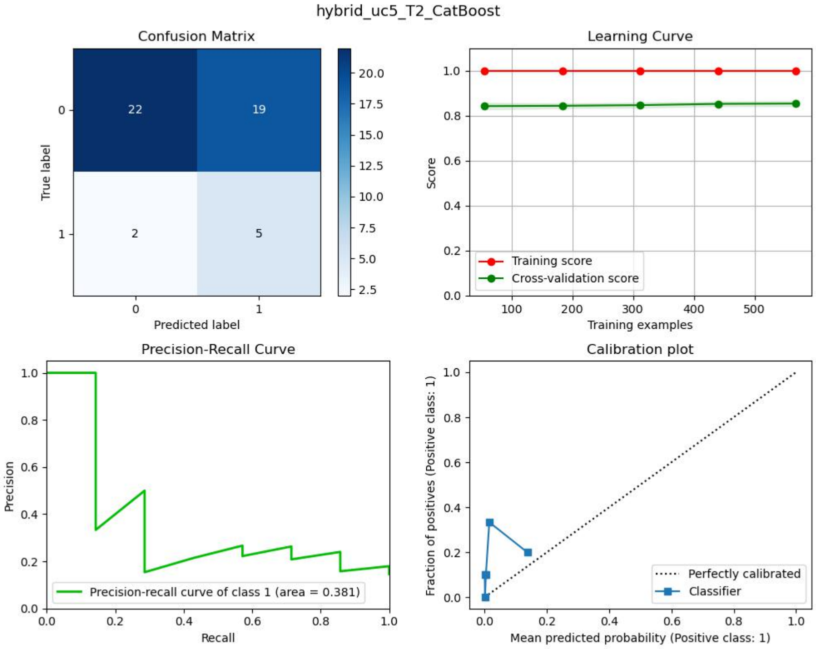


Figure 4.31: Analysis of the hybrid_uc5_T2_CatBoost model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

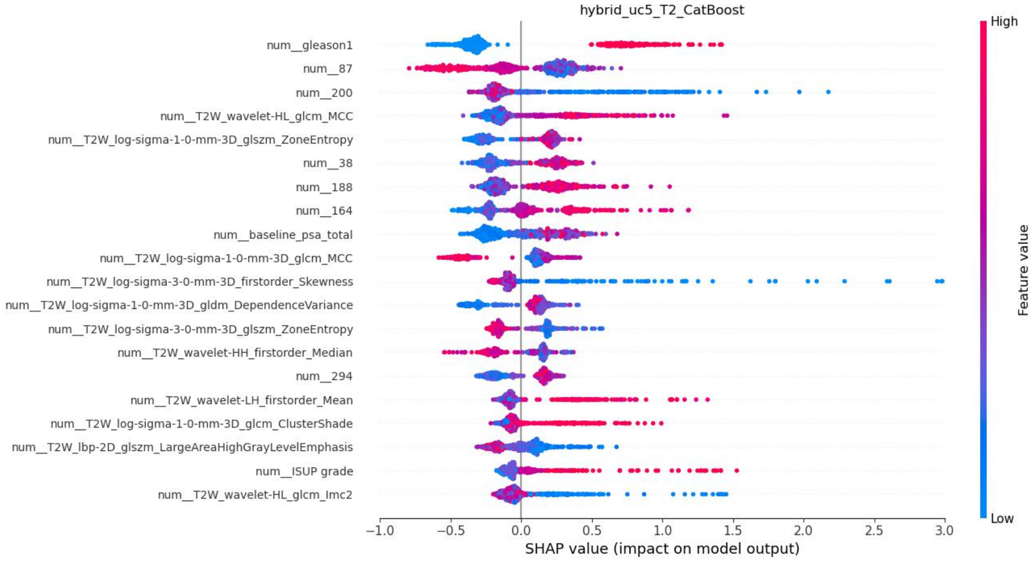


Figure 4.32: SHAP values contribution of features for hybrid_uc5_T2_CatBoost model.

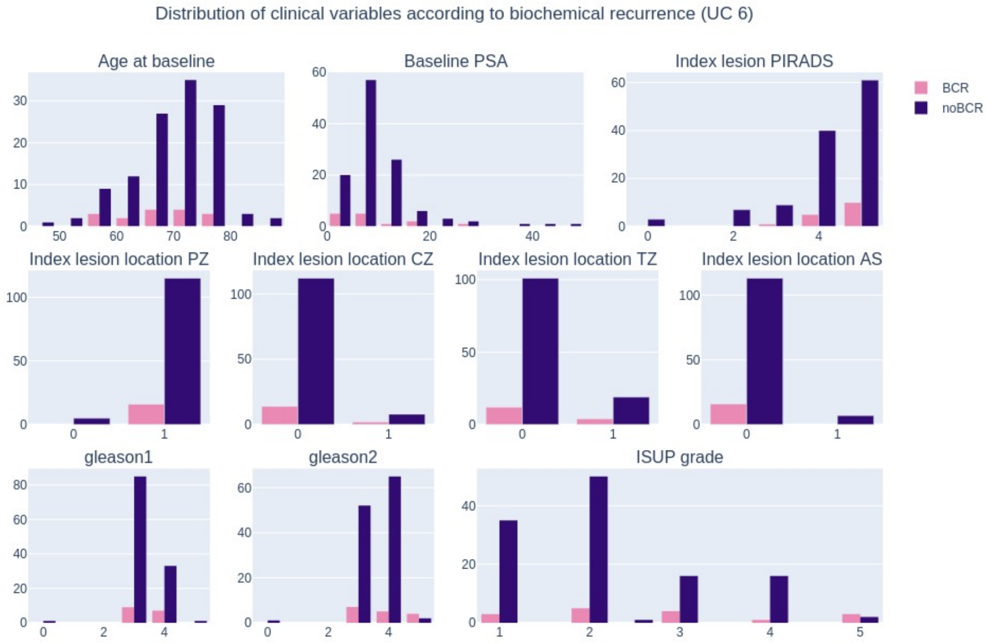


Figure 4.33: Distribution of clinical variables according to biochemical recurrence.

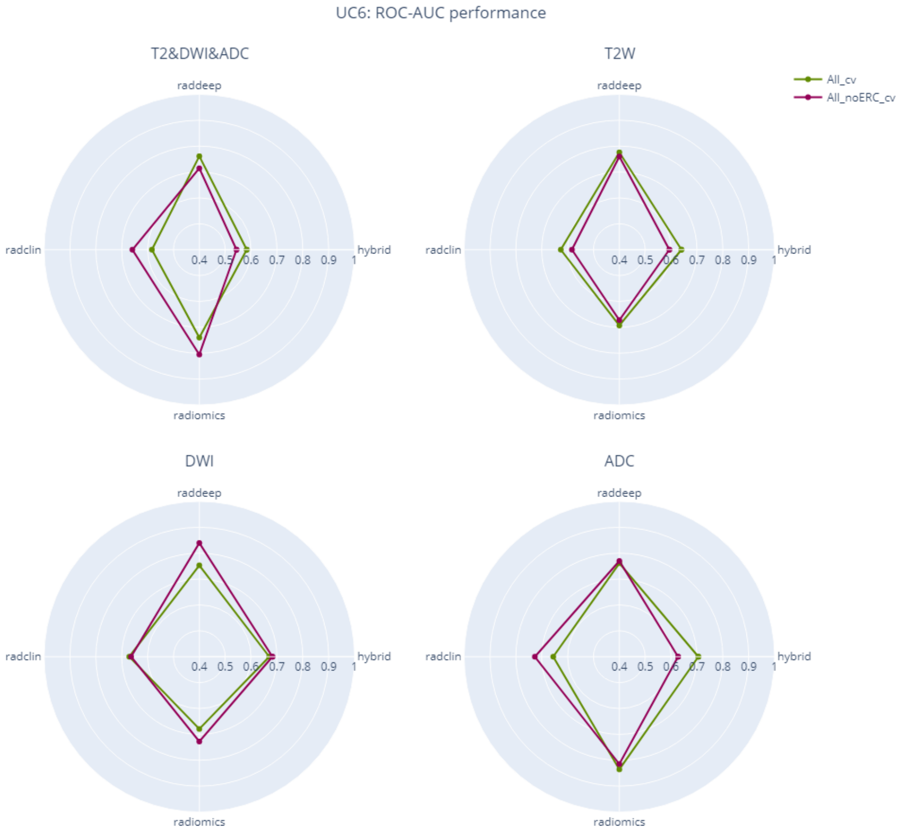


Figure 4.34: Cross-validation ROC-AUC model performance of 32 models trained to predict biochemical recurrence in UC6. The observations are color coded according to the inclusion or exclusion of endorectal-coil exams: all patients in green; excluding ERC cases in pink.

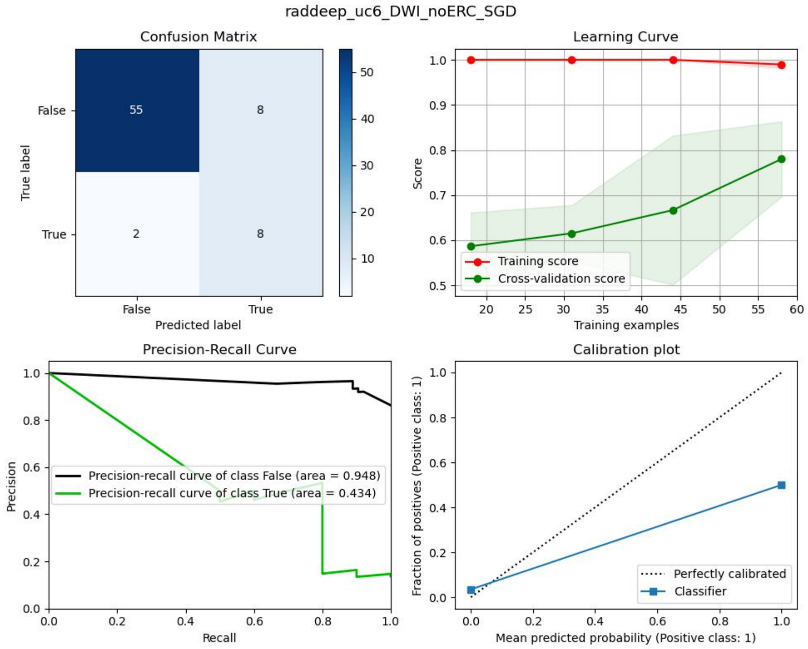


Figure 4.35: Analysis of the raddeep_uc6_DWI_noERC_SGD model in terms of cross-validation confusion matrix, learning curve, precision-recall curve and calibration plot.

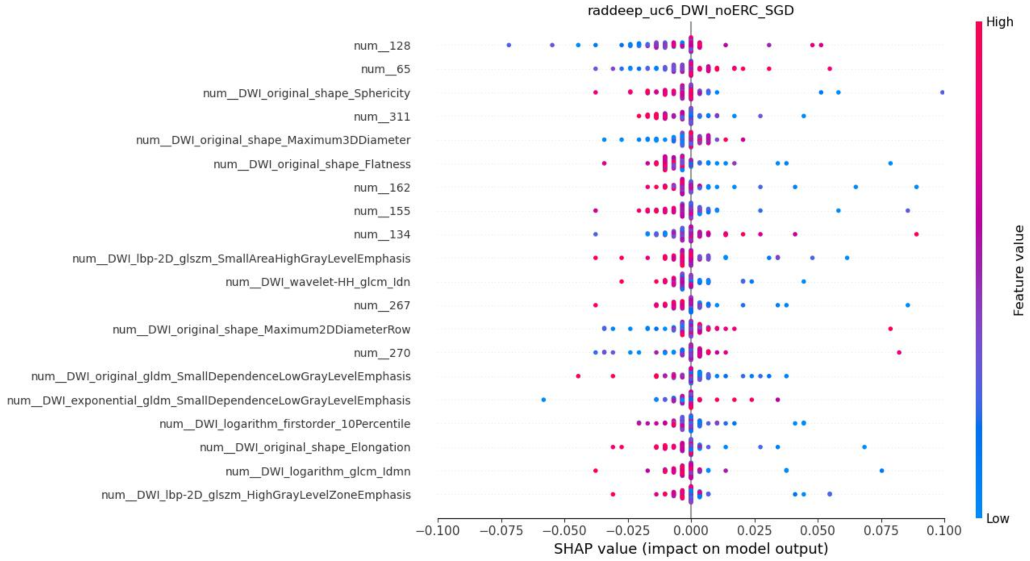
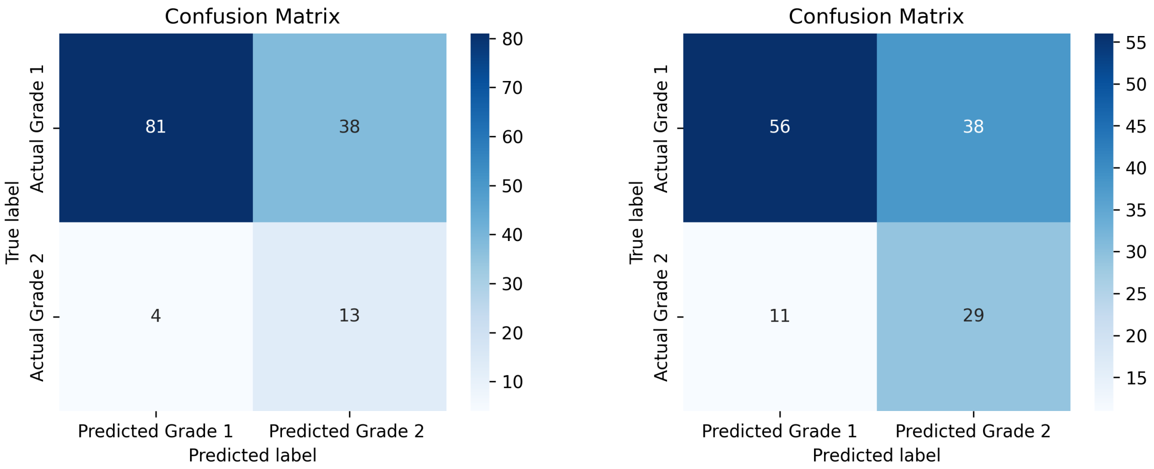


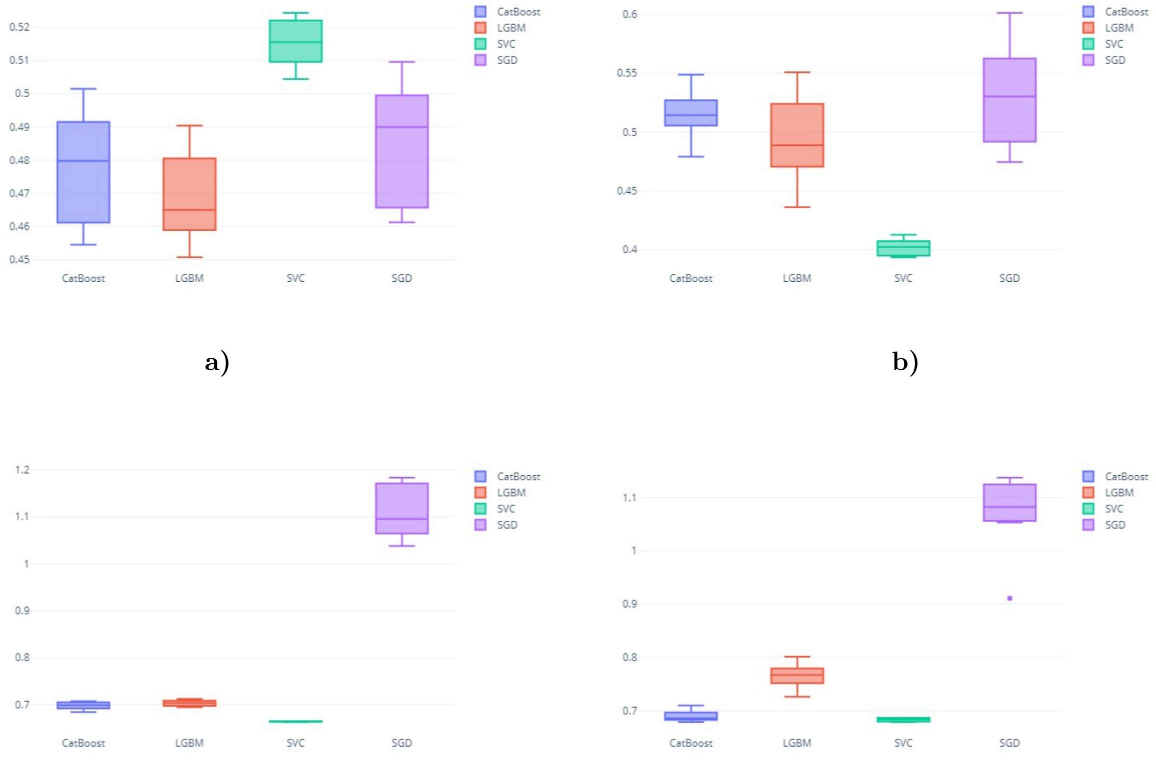
Figure 4.36: SHAP values contribution of features for raddeep_uc6_DWI_noERC_SGD model.



(a) Rectal Toxicity

(b) Urinary Toxicity

Figure 4.37: Confusion Matrices from the optimal models for Rectal and Unrinary Toxicity.



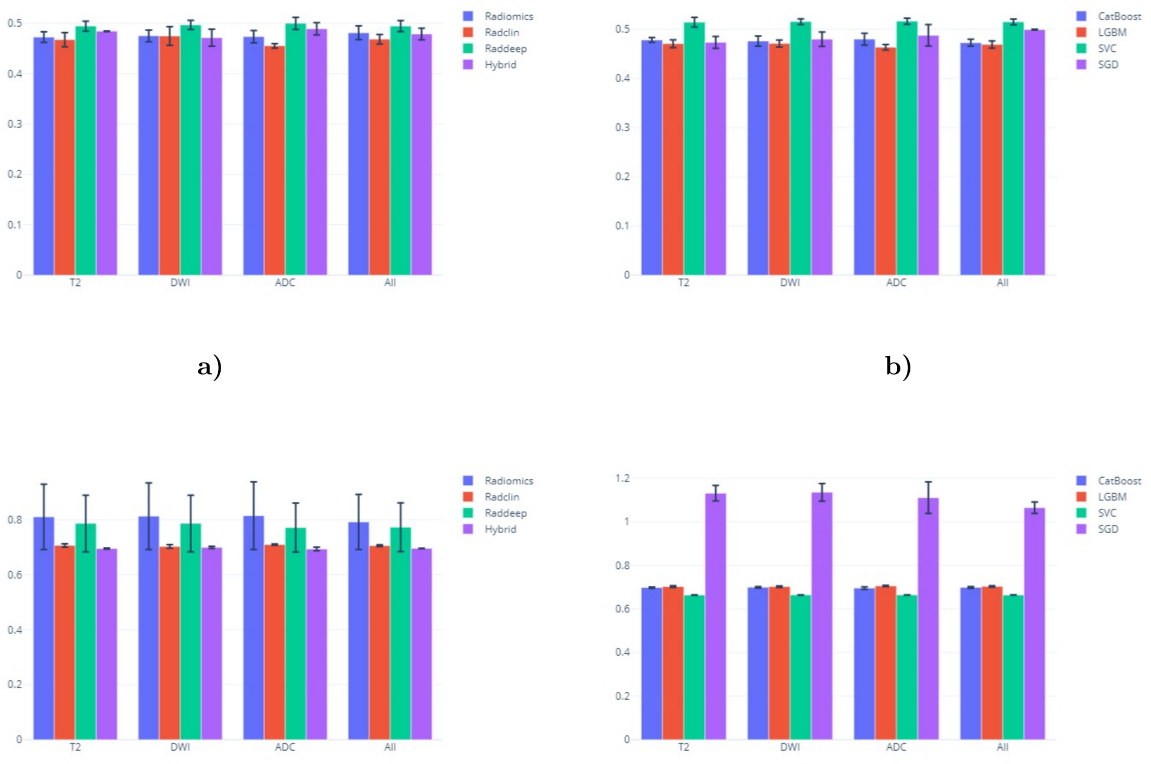
a)

b)

c)

d)

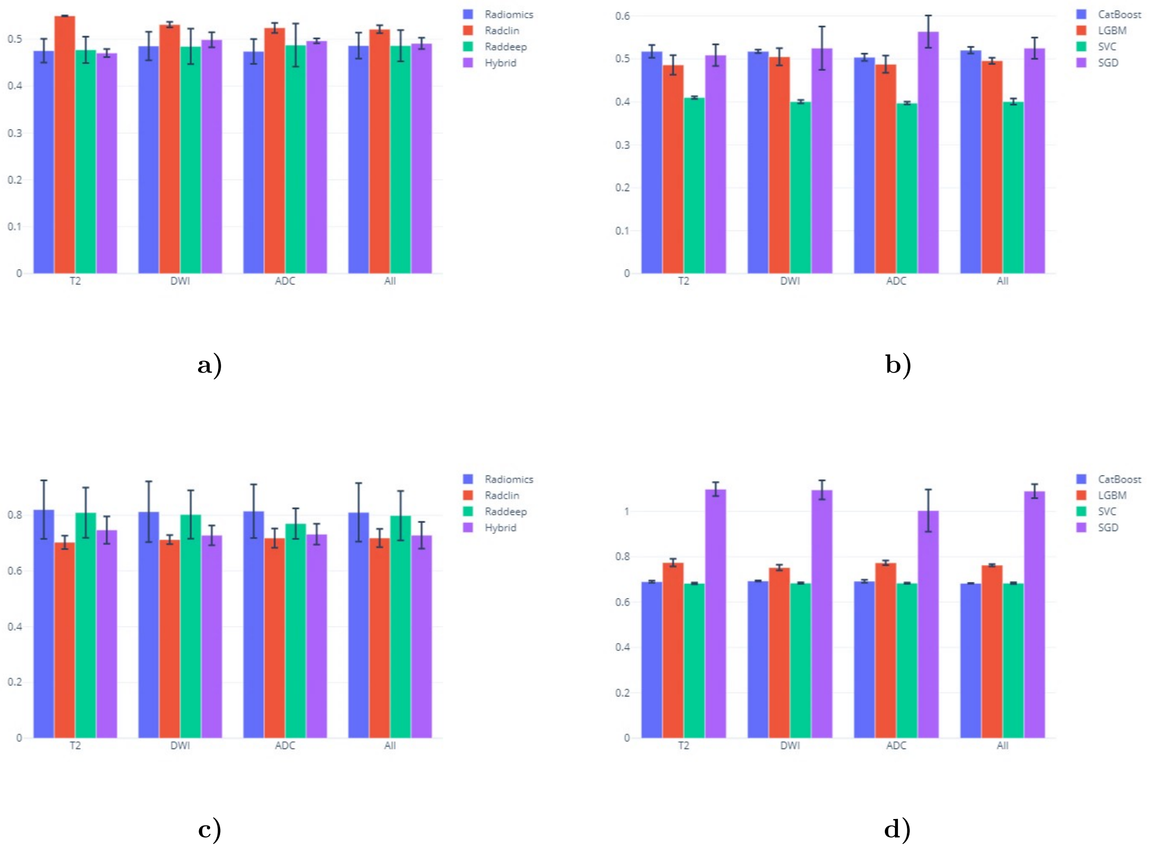
Figure 4.44: General overview of ROC-AUC (**a** and **b**) and Brier scorer (**c** and **d**) performance for the different predictive models. **a** and **c** refer to the cross-validation results, while **b** and **d** refer to the holdout results.



a) **b)**

c) **d)**

Figure 4.45: Mean cross-validation ROC-AUC (**a** and **b**) and brier scorer (**c** and **d**) performance for 48 models trained to predict quality of life in UC7b. Results are stratified by dataset and predictive model



a)

b)

c)

d)

Figure 4.46: Mean holdout ROC-AUC (**a** and **b**) and brier scorer (**c** and **d**) performance for 48 models trained to predict quality of life in UC7b. Results are stratified by dataset and predictive model

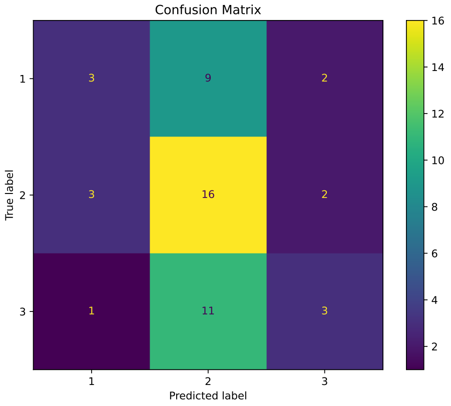


Figure 4.47: Multiclass confusion matrix.

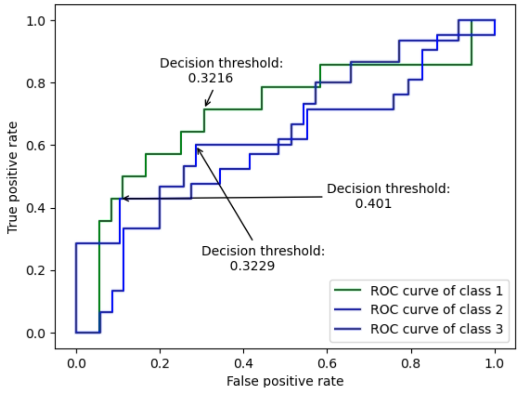


Figure 4.48: Multiclass ROC-AUC curves threshold.

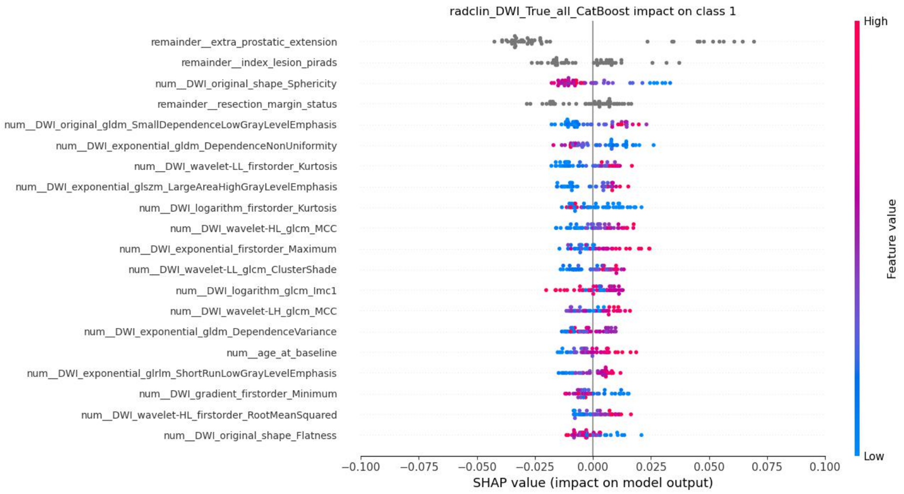


Figure 4.49: SHAP values contribution of features for CatBoost model using Radclind data and DWI features class1 output.

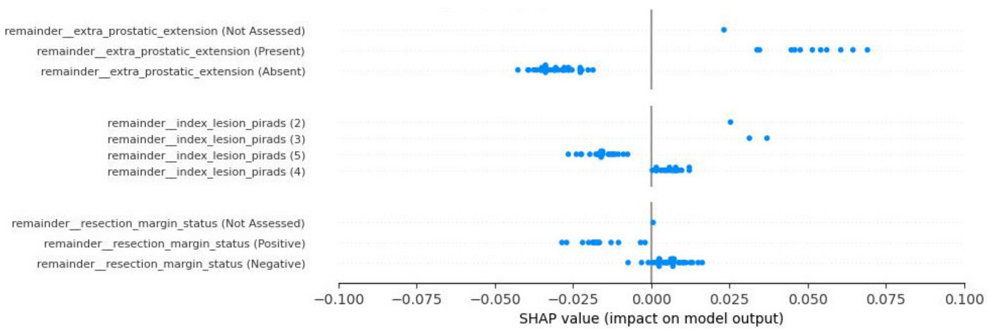


Figure 4.50: SHAP summary plot of the impact of categorical variables on the prediction of class 1.

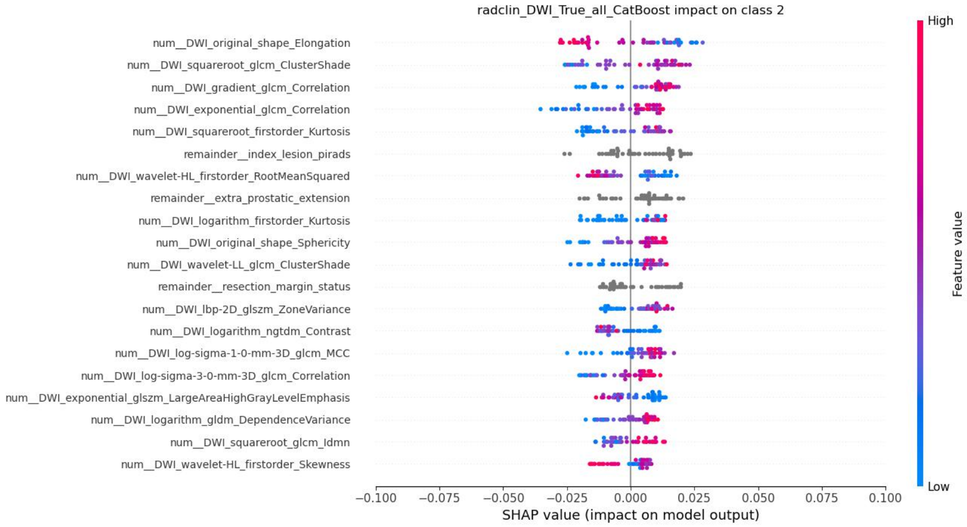


Figure 4.51: SHAP values contribution of features for the CatBoost model using Radclind data and DWI features class2 output.

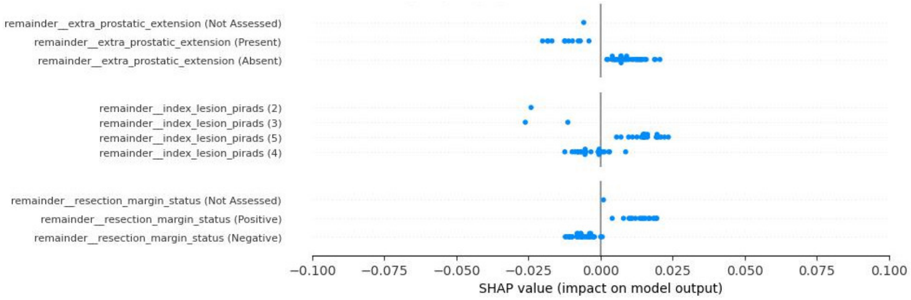


Figure 4.52: SHAP summary plot of the impact of categorical variables on the prediction of class 2.

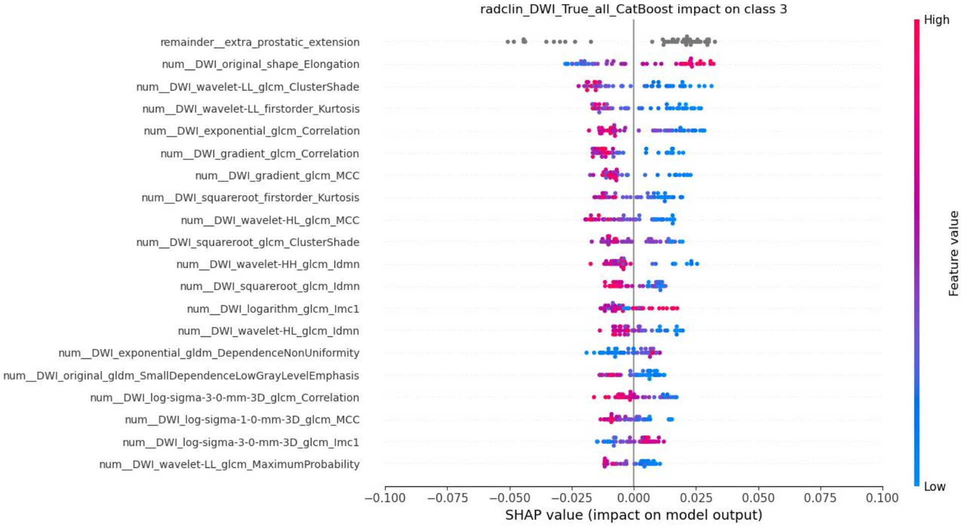


Figure 4.53: SHAP values contribution of features for the CatBoost model using Radclind data and DWI features class3 output.

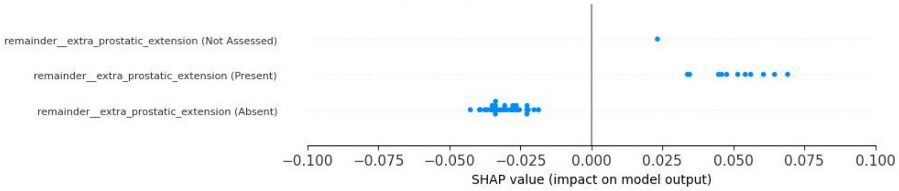


Figure 4.54: SHAP summary plot of the impact of categorical variables on the prediction of class 3.

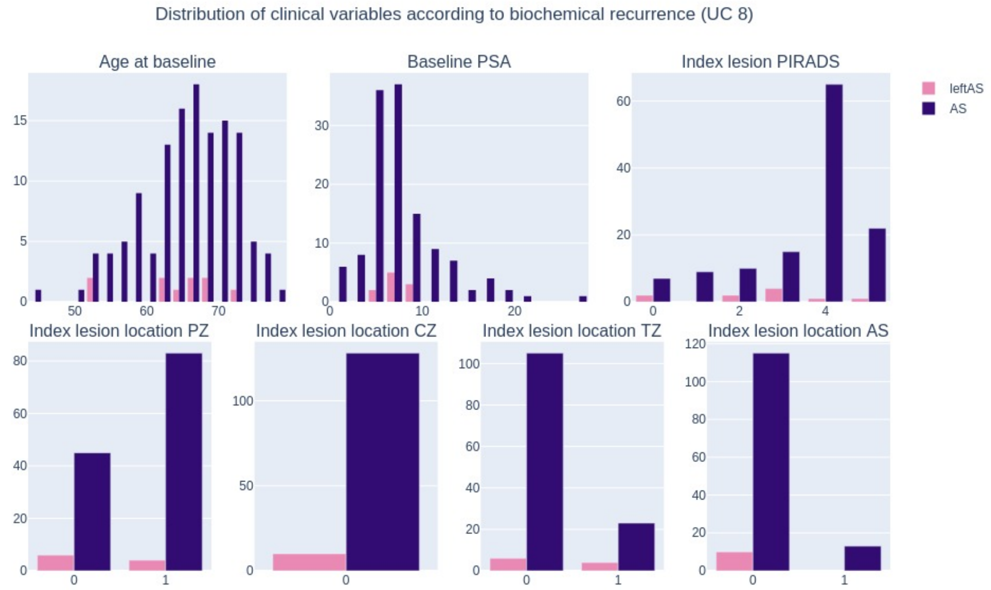


Figure 4.55: Distribution of clinical variables according to the stay in the active surveillance program

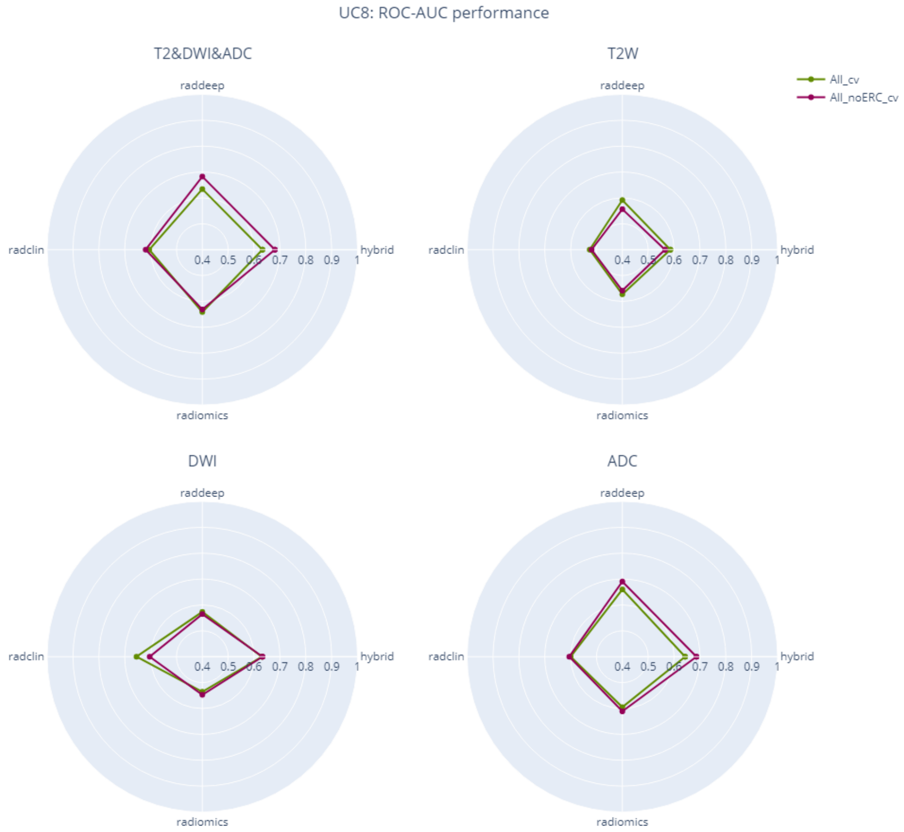


Figure 4.56: Cross-validation ROC-AUC model performance of 32 models trained to predict early withdrawal from the active surveillance program in UC8. The observations are color coded according to the inclusion or exclusion of endorectal-coil exams: all patients in green; excluding ERC cases in pink.

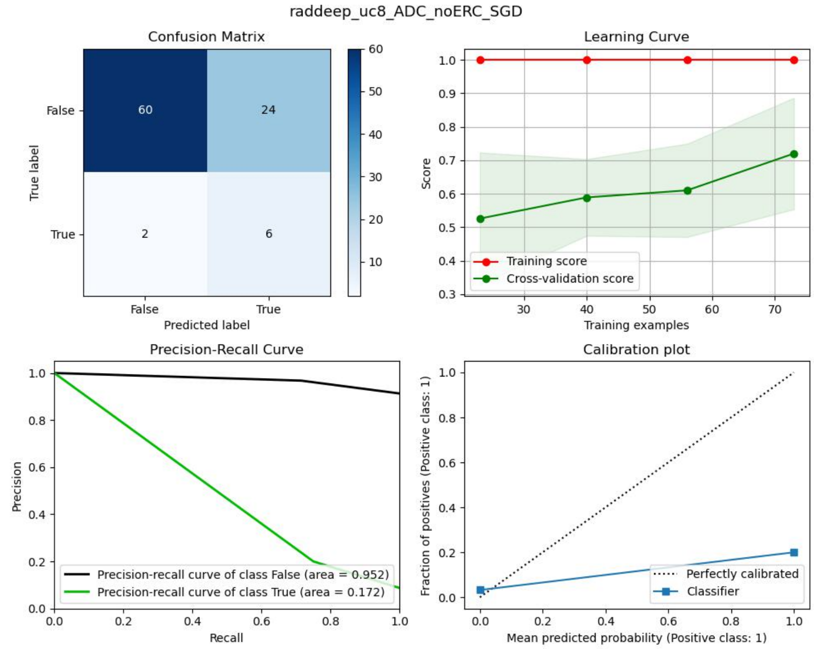


Figure 4.57: Analysis of the raddeep_uc8_ADC_noERC_SGD model in terms of cross-validation confusion matrix, learning curve, precision-recall curve and calibration plot.

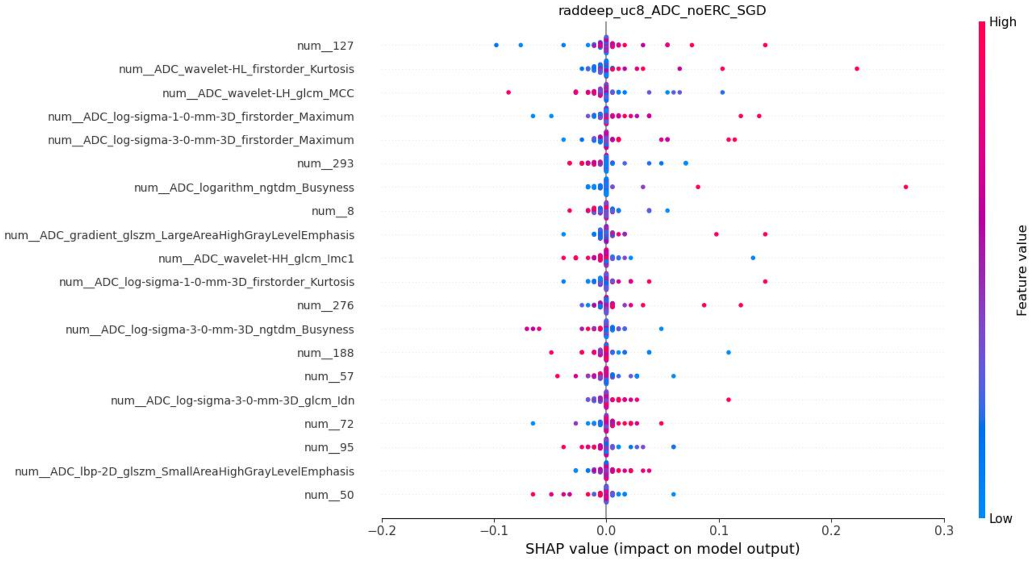


Figure 4.58: SHAP values contribution of features for raddeep_uc8_ADC_noERC_SGD model.

4.3 Discussion

In this work, we report the use of bi-parametric MRI whole-prostate-gland radiomics analysis for seven separate use cases: the prediction of disease aggressiveness (UC2), development of metastasis within 6 months (UC3), biochemical recurrence risk after prostatectomy or radiation therapy (UC5 and UC6, respectively), toxicity after radiation therapy (UC7a), quality of life after prostatectomy (UC7b) and early withdrawal from the active surveillance program (UC8).

As is well known, radiomics analysis require segmentations of the target volumes of interest. These can be performed manually, by an expert radiologist, in an automated fashion, by deep learning models, or semi-automatically, where the masks are AI-generated and corrected where needed by an expert radiologist. Due to the significant size of our dataset, the whole gland segmentation masks were generated automatically for T2W sequences by an inhouse trained software. For DWI, however, this was not possible so an alternative approach was selected: to coregister the T2W and DWI volumes and apply the transformation matrix to the previously automatically generated mask. The segmentation quality assessment revealed the T2W segmentations to be of extremely high quality, however, this was not the case for DWI. Further analysis revealed that the lower quality was mostly due to “mismatches”, defined by the radiologist as an “apparent translation of the mask from the gland location”, which revealed some imperfections in the coregistration stage of the pipeline. Despite this apparent lower quality, the radiomic features extracted proved useful, especially for use cases 3, 6, 7b and 8, where DWI or ADC models were ultimately preferred.

Regarding UC2, the known lack of concordance in the literature for the definition of clinically significant cancer (further discussed in Chapter 5 section 5.1.3), lead us to train three binary classification models (ISUP=1 vs ISUP=2,3,4,5; ISUP=1,2 vs ISUP=3,4,5; and ISUP=1,2,3 vs ISUP=4,5) and one multiclass classification model (ISUP=1 vs ISUP=2,3 vs ISUP=4,5. Interestingly, for all ground truth definitions, the highest performing models combined information from the three MRI volumes (T2W, DWI and ADC) to arrive at a decision. The preferred data type was pure radiomics features, with the exception of ISUP=1,2 vs ISUP=3,4,5, where the combination of radiomics and clinical data outperformed the pure radiomics model. This ground truth definition also proved to be the hardest to model (lowest performance out of the three classifiers), which was expected given the known mixture of clinically significant and insignificant cases in the ISUP classifications 2 and 3, corresponding to GS = 3 + 4 and GS = 4 + 3, respectively. The explainability analysis revealed that the three models relied heavily on radiomic features extracted from wavelet filters, which are known for their relative invariability across volumes with big differences in intensity values, such as those found in different scanner manufacturers. This was further supported by the fairness analysis, where it was found that the models performed the highest with PHILIPS cases, despite not being the most represented cohort seen during training. This was followed by SIEMENS (the largest fraction of the training examples) and GE (the smallest fraction of the training examples). Despite this, for the first ground truth definition (ISUP=1 vs ISUP=2,3,4,5), the model generalized very well to GE cases and had the lowest performance with exams from SIEMENS scanners. Still regarding the second ground truth definition (ISUP=1,2 vs ISUP=3,4,5), it is of note that the model relied heavily on clinical information, namely baseline PSA and age, both positively associated with the aggressive output, and imaging information, namely PIRADS and presence of index lesion in the peripheral zone, again both showing a positive association to the aggressive output.

Use case 3, prediction of metastatic development within 6 months, was the smallest use case, with only 78 complete cases. While for most use cases, the most significant reduction in the number of patients happened due to unavailability of DWI or ADC volumes, for UC3 the reduced dataset size is mostly attributable to ground truth unavailability. It was found that for around 30% of the original number of patients, the reported metastatic status was ”MX”, corresponding to an inability to evaluate the metastatic status by the radiologist. These cases were treated as having a missing ground truth and were removed from the analysis. Despite the small dataset, the selected model achieved a 3-fold cross-validation AUC of 0.8077. Surprisingly, the model was very good at identifying the minority class (specificity of 1.0), corresponding to no metastasis after 6 months, but not so great with the majority class (sensitivity 0.6155), corresponding to metastatic development within 6 months. Even though the model was trained with a hybrid dataset (radiomics, clinical and deep features), no clinical variables were found among the 20 most significant to the model’s output, relying heavily on deep features. Regarding the subcohort analysis, the model generalized well to GE exams, but the performance dropped around 15% for SIEMENS.

Use case 5 related to the prediction of biochemical recurrence after radical prostatectomy surgery. For this use case, two approaches were taken: pre-surgery prediction and post-surgery prediction. The latter including clinical variables reported during or after the radical prostatectomy surgery itself. Interestingly, for both scenarios, a hybrid T2 model was selected, emphasizing the importance of this sequence in biochemical recurrence prediction. Also in agreement were the explainability analyses, with both models relying heavily on clinical information such as gleason score of the most prominent histological pattern, baseline PSA level and ISUP grade, all positively associated with the biochemical recurrence output. The post-surgery classifier additionally utilized the variables presence of extraprostatic extension and resection margin status, also with a positive influence on the biochemical recurrence output, which can explain the higher performance of the post-surgery model (0.82 AUC), when compared with the pre-surgery classifier (0.69 AUC).

Use case 6 aimed to predict biochemical recurrence after radiation therapy. The selected model achieved a 3-fold cross-validation AUC of 0.8393, and sensitivity and specificity of 0.8056 and 0.8730, respectively. The explainability analysis, showed the model relied heavily on shape features, namely: Sphericity, Maximum3DDiameter, Flatness, Maximum2DDiameterRow and Elongation. This suggests that the shape of the prostate gland is highly associated with the success of radiation therapy. Namely, high values of gland sphericity, elongation and flatness, and low values of Maximum3DDiameter and Maximum2DDiameterRow are associated with recurrence.

Regarding UC7b, the epic26 continuous values were divided into 3 bins, representing good, average and high quality of life. With these new labels, a multiclass classification approach was followed. The preferred model, which tried to balance both AUC classification performance and the Brier score calibration error, was the CatBoost model which used Radclin data and DWI radiomic features. However, it was shown that due to the class imbalance, this model tended to classify the majority of samples as the average class. Taking this into account, possible paths forward include calibrating the models through conformal prediction, as well as turning this multiclass problem into an ordinal classification problem.

Use case 8 related to the prediction of an early withdrawal from the active surveillance program. This was the only use case where an ADC model was ultimately selected. Here, we achieved an AUC of 0.6905, and 0.6667 and 0.7143 sensitivity and specificity, respectively. In concordance with the other small use cases, the fairness analysis was mostly inconclusive due to the low representation of minority cohorts.

Overall, despite the significant data size of some use cases, the learning curves demonstrated there still was learning capacity. The discrepancy between training and cross-validation performance was as high as 28% for UC2 (ISUP=1,2 vs ISUP=3,4,5) and UC8, while the lowest discrepancy was found in UC5, where the cross-validation performance was only 12% below training performance. Consistently across the different use cases, the hold-out test set performance was higher than the corresponding cross-validation performance. This can be explained by the relatively low number of folds used in the cross-validation stage, which led to models trained with fewer data. In contrast, the hold-out test set performance reflects models trained on the full train set, which was beneficial for model performance. Another possible explanation would be hold-out test set selection bias. Even though the held-out patients were randomly selected from the entire pool of patients, it is possible that, by chance, patients that were ”easier” for the model to classify were selected. However, given that this phenomenon was observed in all use cases it is much more likely that the root cause was a methodological aspect, that all use cases have in common, such as the cross-validation strategy.

Several limitations can be identified in the body of work. The first and largest limitation was the very reduced dataset size of some use cases, the most extreme case being UC3. Secondly, the same methodology (or with only slight deviations) was applied to the different use cases. This is not ideal, as different classification problems might thrive with different learning strategies. With more time other methodologies could be explored. Thirdly, the lack of lesion segmentation masks made it impossible to explore lesion radiomics in addition to the whole-gland radiomics results presented.

.

Chapter 5

Deep Learning Master Models

Chapter summary

ProstateNet constitutes a sizeable dataset, providing ample opportunities in terms of developing modelling strategies. This section is thus split into several distinct sections, each of which is relatively self-contained and focuses on a specific modelling aspect. We note that each one of these sections was carried out by different partners (Table 5.1) and focuses on either classification ($\mathbb{C}$) or segmentation/detection ($\mathbb{S}$):

- **On the impact of target definition and additional variables on UC2 (FCHAMPALIMAUD; section 5.1)** — in this section, we investigated how different models and features contribute towards the performance of aggressiveness classification models. Additionally, we consider how detailing different targets — as specified in chapter 4 — could have an impact on this. Particularly, the main question was one concerning the definition of aggressive prostate cancer. We consider three distinct alternatives: ISUP=1 vs. ISUP=2,3,4,5, ISUP=1,2 vs. ISUP=3,4,5 and ISUP=2,3 vs. ISUP=4,5. We train different DL models (VGG, ConvNeXt, ViT and factorized ViT) and test whether the inclusion of clinical and demographic variables (PSA, age, PI-RADS). Finally, a fairness analysis was performed to identify possible biases in our modelling approach. Following this study, we finally created a model ensemble to predict PCa risk between low (ISUP=1), intermediate (ISUP=2,3) and high (ISUP=4,5) and tested whether conformal prediction could be used to improve prediction ($\mathbb{C}$);
- **On the impact of cropping strategies in UC2 and UC5 (CNR; section 5.2)** — here, we investigated how different cropping strategies (central crop and adaptive whole prostate gland cropping) affect the performance of deep learning models in both UC2 and UC5. Additionally, we also investigated how the amount of training data and different DL models impact performance ($\mathbb{C}$);
- **On the differences between supervised and unsupervised learning strategies for UC1 (FORTH; section 5.3)** — here, the performance of different learning paradigms (supervised and unsupervised) was inspected as potential solutions for UC1 (determining whether an individual has a lesion), while also investigating different architectures. Additionally, lesion segmentation models with different architectures were also tested in order to determine how these specifications have an impact on segmentation and detection performance ($\mathbb{C} + \mathbb{S}$);
- **On the performance of 2D & 3D models on index lesion segmentation with a curated dataset (ADVANTIS; section 5.4)** — here we tested how the use of two- or three-dimensional inputs to neural network models impacts the performance of segmentation models; this helped us better understand more precisely whether there are important relationships between different slices in the same series that can help DL models better predict lesion segmentation maps ($\mathbb{S}$);
- **On the impact of mpMRI sequence combination to automatically detect prostate cancer (FPO; section 5.5**) — here, we investigated how different ways of combining mpMRI data can lead to improved performance in lesion segmentation and detection models by considering different feature fusion paradigms — using the mpMRI sequences as the input or processing each sequence separately

and fusing their features at later learning stages. This helped us understand how fusing features at different levels can impact performance (𝕊);

- **Validation of ProCAncer-I prostate segmentation tool (QUIBIM; section 5.6** — here, the performance of the QUIBIM prostate zone segmentation tool was assessed on novel ProstateNet data. Additionally, new 2D and 3D models were trained and tested to further improve on the performance of prostate zone segmentation, considering the presence of different types of data (i.e. with and without endorectal coil) (𝕊);

- **Effect of dataset characteristics on segmentation performance (FCHAMPALIMAUD; section 5.7** — here, we used different datasets — ProstateX [5], Prostate158 [3] and ProstateNet — to determine how using ProstateNet could impact the performance of whole prostate gland, prostate zone and lesion segmentation models. In particular, we were interested in showing how annotations which are both more diverse and more numerous — as is the case of ProstateNet — can be determining in improving the performance of segmentation and lesion detection models (𝕊);

- **ProstateNet Lesion Segmentation with Deep Learning (HULAFE; section 5.8** — here, the focus is on performing lesion segmentation on T2-weighted axial images sourced from the ProstateNet dataset using deep learning techniques with Tensorflow. The primary objective involves developing and refining models to accurately identify and delineate lesions within the prostate gland. The data extraction process involves selecting 419 T2-Ax segmented monochromatic series from the ProstateNet dataset, which underwent further processing and refinement. Preprocessing steps included analyzing spatial resolutions, standardizing dataset spacing, and implementing image cropping, denoising, and intensity normalization. For the 2D model, a sequence of three frames was used to predict a final 2D segmentation mask, while for the 3D model, a frame depth of 16 was chosen to ensure consistent input size. The architecture and specifics of the 2D and 3D prediction models are detailed as well as the issues regarding prediction in the 2D models and over-fitting, and the strategies employed to mitigate it, such as regularization techniques and adjustments in layer complexity (𝕊).

In more concrete terms, we present here work comprising seven distinct and self-contained analyses, each of which presents a series of models. In total, six partners were responsible for developing the work presented here.

Section	Partner
On the impact of target definition and additional variables on UC2	FCHAMPALIMAUD
On the impact of cropping strategies in UC2 and UC5	CNR
On the differences between supervised and unsupervised learning strategies for UC1	FORTH
On the performance of 2D & 3D models on index lesion segmentation with a curated dataset	ADVANTIS
On the impact of mpMRI sequence combination to automatically detect prostate cancer	FPO
Validation of ProCAncer-I prostate segmentation tool	QUIBIM
Effect of dataset characteristics on segmentation performance	FCHAMPALIMAUD
ProstateNet Lesion Segmentation with Deep Learning	HULAFE

Table 5.1: List of sections in this chapter and the responsible partners.

5.1 On the impact of target definition and additional variables on UC2

5.1.1 Methods

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/demographic data – prostate specific antigen (PSA) and age at baseline – we further calculate the amount of sequences with clinical/demographic 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/demographic 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 and endorectal coil usage as stratifying variables.

Three different ISUP-based target variables were considered:

- **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 5.2. 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 5.2: 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 [66]. 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). Additional models were trained using a 192x192x24 voxel-size crop to confirm that a smaller size crop contains the relevant predictive signal (this assessment consists of comparing the performance of the 128x128x24 crop with that of the 192x192x24 crop).

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 5.3: 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) [80], a ConvNeXt model [51], a 3D vision transformer (ViT) model [21] and a variation of the 3D ViT that separates within and between slice processing (factorized ViT). General training details are provided in Table 5.3. 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 [51] 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 [21] 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 [61], 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 (this is a protocol similar to that proposed as TrivialAugment [63]).

Optimization. Models were trained using the AdamW optimizer [52], a modification to the Adam optimizer that corrects the application of weight decay with a standard cross-entropy loss. To account for the class imbalance, we calculate a weight for the loss such that each positive instance is multiplied by pos/neg, where pos and neg are the number of positive and negative cases, respectively. This is performed at the beginning of each fold.

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

- Retraining the models and concatenating the standardized age and PSA – 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/radiological features. Each of these 16 combinations is trained using 5-fold cross validation for a total of 80 training runs.

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

To understand how crop size impacts the performance of each model, we train the best performing model using a larger crop size (192 × 192 × 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; this allows us to build learning curves, which describe how the amount of data has an impact on performance.

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 [88] 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.

Model ensembling

To study how binary models can be combined into an ensemble to produce a multiclass model, we ensemble the low vs. possibly high and the intermediate vs. high models to produce a three class classifier — predicting ISUP=1 vs. ISUP=2,3 vs. ISUP=4,5. To do so, we use the same folds and the mpMRI VGG model, freeze the encoders from these tasks and concatenate their outputs ($512 + 512 = 1024$ total features). Then, a GeLU activation with linear normalization and a 25% dropout is followed by a linear layer to reduce these features to 512. This final set of features is used to classify instances between the three aforementioned classes.

As extensions to this, two minor adaptation modules are tested:

- Low rank adaptation module (LoRA) — a module which linearly converts features before and after the application of the last VGG block (loosely inspired on the work on LoRA in large language models [32] and on the batch ensemble operator [93])

- Squeeze and excite (SAE) — a CNN-specific attention mechanism which reduces features both in terms of space and feature space in order to learn which parts of the volume/feature space are the most important. The SAE module is applied to the output of the frozen binary class VGG encoders [33].

These models are compared against a more naive approach — a simple multiclass VGG — using one-versus-one multiclass AUC. Additionally, the inclusion of sensitive variables (lesion location, age, PSA, PI-RADS) is also tested in a simple post-hoc linear model. Finally, a conformal prediction method — adaptive prediction sets (APS) [74] — is tested to demonstrate how these methods, while reducing the prediction coverage, lead to an improvement in performance. Conformal prediction methods introduce a notion of uncertainty into ML models, making it possible to reject predictions if high-uncertainty is detected. APS, in particular, defines, for each sample, a set of predictions (in this case one of ISUP=1, ISUP=2,3 or ISUP=4,5) by considering the cumulative (and descending order-sorted) output probabilities (COP): the prediction set is defined with the COP index at which it crosses an inferred threshold.

A nominal example for APS may be helpful, considering probabilities $[0.1, 0.5, 0.4]$ for classes $[0, 1, 2]$ and a threshold $t = 0.8$. First, the COP is calculated as COP $= [0.5, 0.9, 1.0]$. It is easily verifiable that 1 (considering a 0-index system) is the minimum index at which the COP crosses the threshold, corresponding to a prediction set $= [1, 2]$.

5.1.2 Results

Performance of sequence-only models

In general, CV performance highlights two specific trends — T2W+DWI+ADC (mpMRI) models outperform models using only T2W sequences, and VGG-based models generally perform better than other architectures (Figure 5.1A), trends which remain consistent upon evaluation on a hold-out test set for the low vs. possibly high ({1} vs. {2,3,4,5}) and possibly low vs. high ({1,2} vs. {3,4,5}) targets (Figure 5.1B; Table 5.4). Interestingly, the intermediate vs. high target ({2,3} vs. {4,5}) shows relatively good generalization except for the VGG mpMRI models, which end up performing comparably to the VGG T2W model or to the ConvNeXt mpMRI model (Figure 5.1C). Indeed, while the better performance of mpMRI models is to be expected, it was surprising to observe that a simple model architecture (VGG) outperformed more modern and complex architectures (ConvNeXt, ViT-based models). Additionally, the superiority of CNN-based models was routinely observed when compared with ViT-based models (Table 5.5), possibly indicative of the latter requiring more data to achieve comparable performance.

Target	Model (other)	Mean VGG AUC	Mean 2nd best AUC	Sequences	p-value
1 vs. 2,3,4,5	ConvNeXt	0.6152	0.5858	T2W	0.0426
	ConvNeXt	0.7286	0.6774	T2W+DWI+ADC	0.0354
1,2 vs. 3,4,5	ConvNeXt	0.6167	0.5954	T2W	0.0566
	ConvNeXt	0.6827	0.6651	T2W+DWI+ADC	0.1606
2,3 vs. 4,5	ConvNeXt	0.6547	0.5969	T2W	0.0159
	ConvNeXt	0.6476	0.6437	T2W+DWI+ADC	0.7956

Table 5.4: t-test results comparing sequence-only VGG models with the second best sequence-only model for each target and sequence input. p-values are highlighted in black if differences in mean are statistically significant and in grey otherwise.

Performance of hybrid models

Hybrid deep models. Hybrid models (deep-learning models trained using both volumes and age and PSA as inputs) show comparable trends to sequence-only models, including the lack of generalization for VGG mpMRI models (Figure 5.2). However, specific trends, particularly those comparing architectures, are not as clear (Table 5.6; Table 5.7). When directly comparing sequence-only and hybrid models, there is little evidence that the inclusion of age or PSA contributes positively towards prediction (Figure 5.3; Table 5.8), suggesting that either these models are not capable of learning how to combine information from volumes and from age and PSA, or that there is little else to be learned from these images.

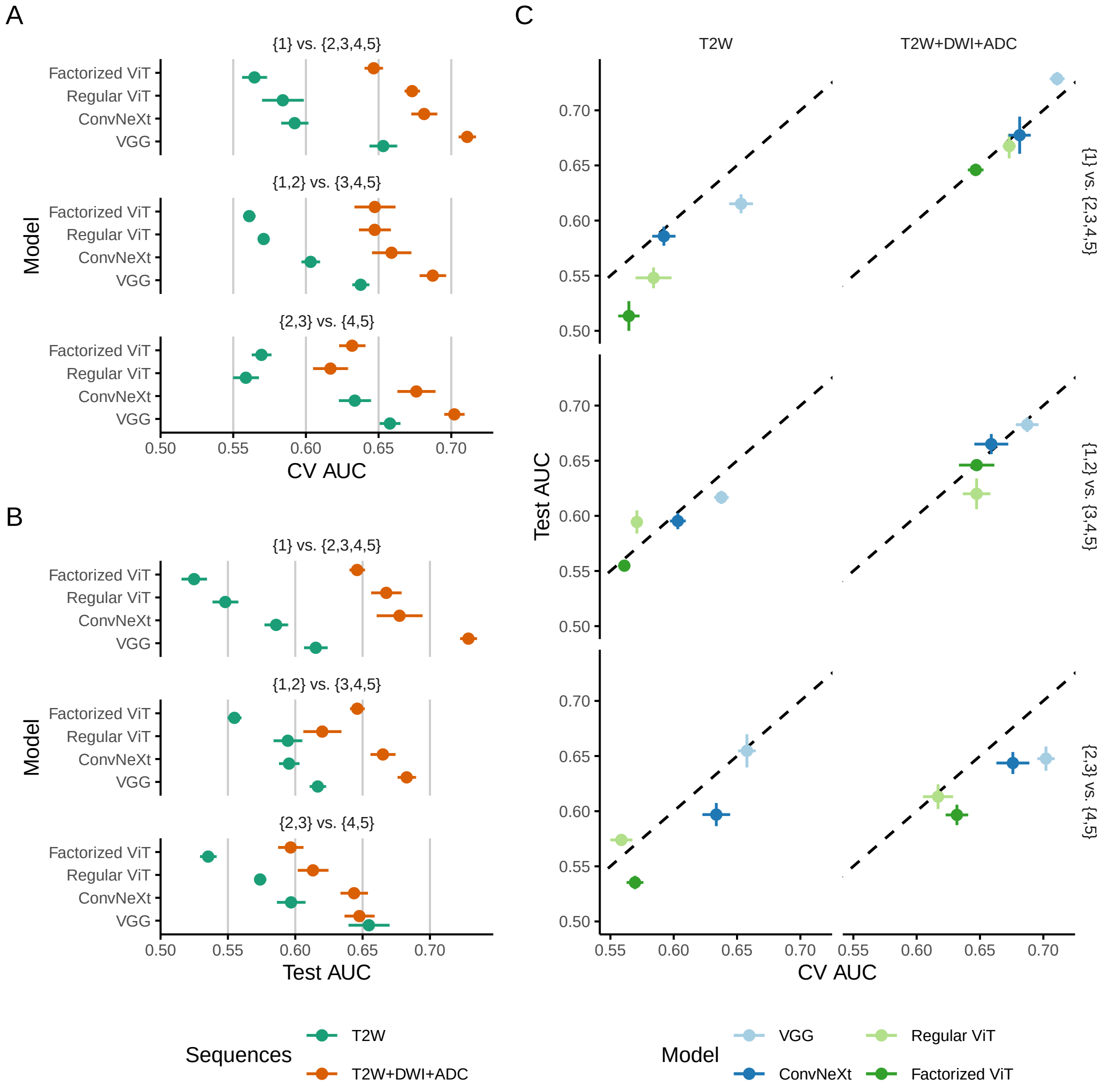


Figure 5.1: Performance (AUC) for models trained only on sequence data. A: Cross-validated (CV) performance. B: Hold-out test set performance. C: Comparison between CV and hold-out test set performance. In all panels, points represent the average, while lines (horizontal or vertical) represent the standard error around the mean.

Target	Mean CNN AUC	Mean ViT AUC	Sequences	p-value
1 vs. 2,3,4,5	0.6227	0.5743	T2W	0.0039
	0.6961	0.6598	T2W+DWI+ADC	0.002
1,2 vs. 3,4,5	0.6204	0.5660	T2W	0.002
	0.6731	0.6473	T2W+DWI+ADC	0.0645
2,3 vs. 4,5	0.6457	0.5640	T2W	0.002
	0.6890	0.6243	T2W+DWI+ADC	0.0039

Table 5.5: t-test results comparing sequence-only CNN-based models with sequence-only ViT-based models each target and sequence input. p-values are highlighted in black if differences in mean are statistically significant and in grey otherwise.

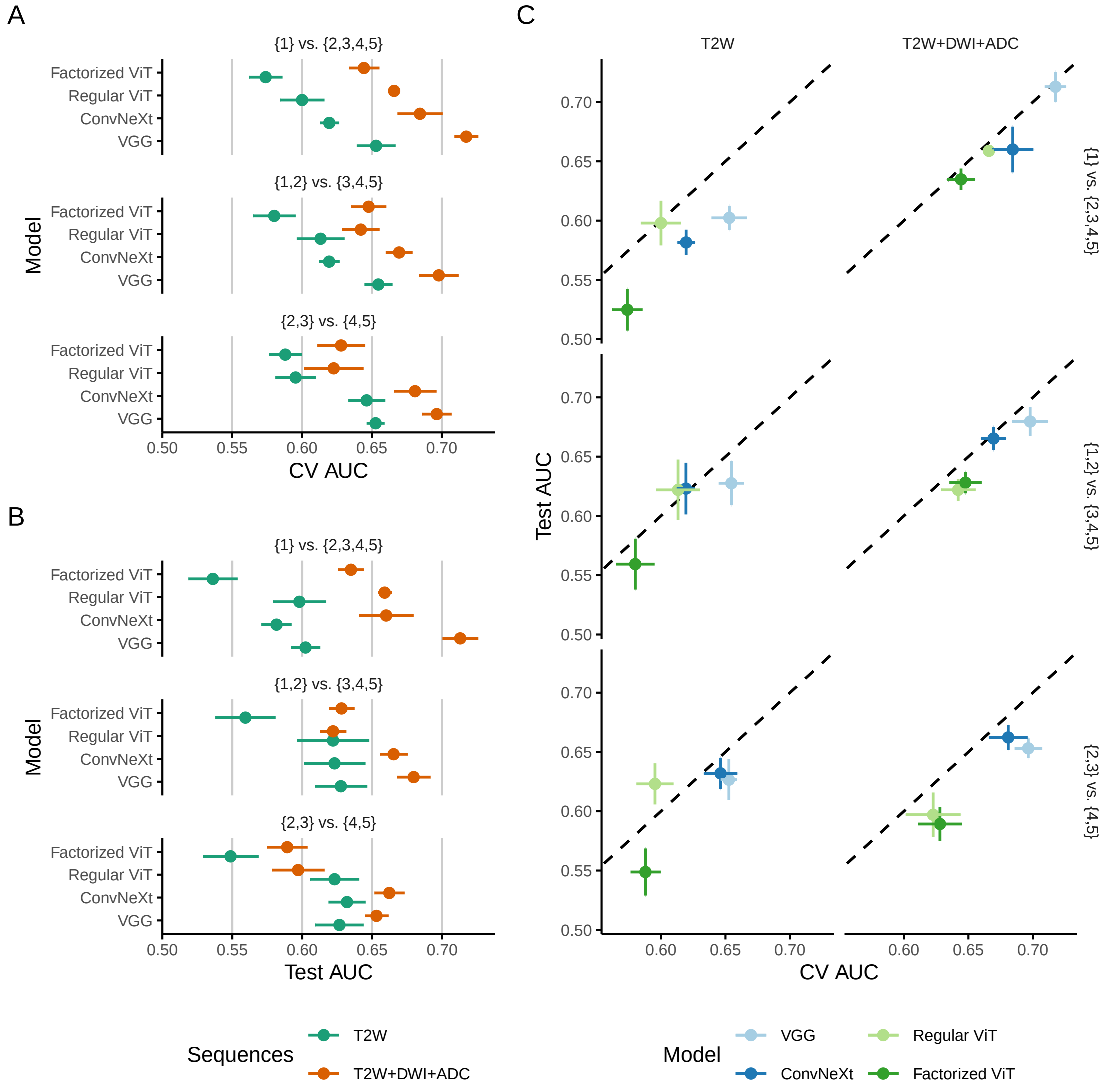


Figure 5.2: Performance (AUC) for models trained on sequence and MRI and age and PSA. A: Cross-validated (CV) performance. B: Hold-out test set performance. C: Comparison between CV and hold-out test set performance. In all panels, points represent the average, while lines (horizontal or vertical) represent the standard error around the mean.

Target	**Model (other)**	**Mean VGG AUC**	**Mean 2nd AUC**	**Sequences**	**p-value**
1 vs. 2,3,4,5	Regular ViT	0.6023	0.5979	T2W	0.8444
	ConvNeXt	0.7129	0.6600	T2W+DWI+ADC	0.056
1,2 vs. 3,4,5	ConvNeXt	0.6276	0.6231	T2W	0.8787
	ConvNeXt	0.6796	0.6653	T2W+DWI+ADC	0.3834
2,3 vs. 4,5	ConvNeXt	0.6266	0.6320	T2W	0.8115
	ConvNeXt	0.6530	0.6622	T2W+DWI+ADC	0.5168

Table 5.6: t-test results comparing sequence-only VGG models with the second best sequence-only model for each target and sequence input. p-values are highlighted in black if differences in mean are statistically significant and in grey otherwise.

Target	Mean CNN AUC	Mean ViT AUC	Sequences	p-value
1 vs. 2,3,4,5	0.5920	0.5614	T2W	0.1055
	0.6864	0.6468	T2W+DWI+ADC	0.0059
1,2 vs. 3,4,5	0.6253	0.5907	T2W	0.1602
	0.6725	0.6250	T2W+DWI+ADC	0.0039
2,3 vs. 4,5	0.6293	0.5859	T2W	0.1055
	0.6576	0.5932	T2W+DWI+ADC	0.002

Table 5.7: t-test results comparing sequence-only CNN-based models with sequence-only ViT-based models each target and sequence input. p-values are highlighted in black if differences in mean are statistically significant and in grey otherwise.

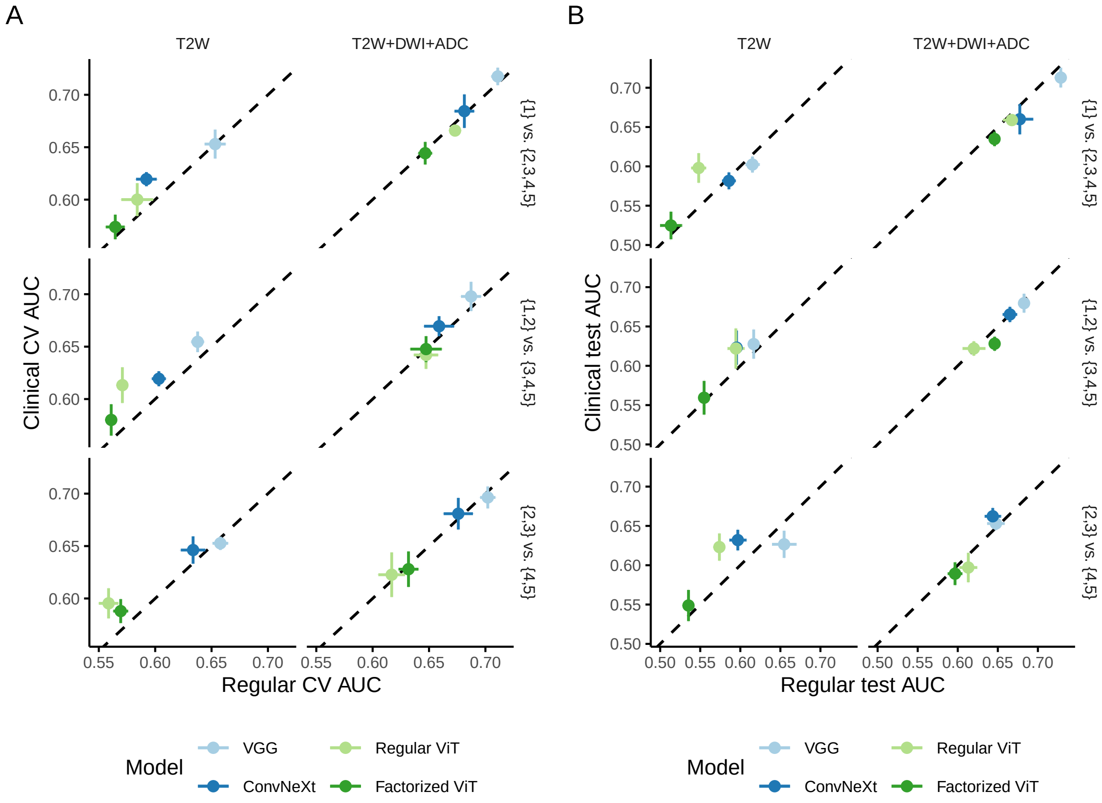


Figure 5.3: Comparison of sequence-only and hybrid models. A: Comparison of cross-validated performance. B: Comparison of hold-out test set performance. In all panels, points represent the average, while lines (horizontal or vertical) represent the standard error around the mean.

Hybrid linear models. To better understand the impact of age and PSA, we calculated the non-normalized probability output (logit) from the deep-learning models and used it as a feature in a simple linear model with age and PSA. Given the relative flexibility of this modelling strategy, we also tested how the inclusion of PI-RADS could improve performance. We focused specifically on mpMRI VGG models and show that these simpler linear models may lead to slightly improved performance when compared with sequence-only models; however, these differences are all statistically non-significant with relatively wide performance values, possibly indicating that, while clinical/demographic/radiological features may lead to improved performance, this should be assessed and determined for each use case.

Target	Sequences	Model	SO avg. AUC	Hybrid avg. AUC	p-value
1 vs. 2,3,4,5	T2W	VGG	0.615	0.602	0.365
		ConvNeXt	0.586	0.582	0.7696
		Regular ViT	0.548	0.598	0.0567
		Factorized ViT	0.513	0.525	0.6224
	T2W+DWI+ADC	VGG	0.729	0.713	0.3078
		ConvNeXt	0.677	0.660	0.5157
		Regular ViT	0.668	0.659	0.5034
		Factorized ViT	0.646	0.635	0.3359
1,2 vs. 3,4,5	T2W	VGG	0.617	0.628	0.6023
		ConvNeXt	0.595	0.623	0.2841
		Regular ViT	0.594	0.622	0.3629
		Factorized ViT	0.555	0.559	0.8454
	T2W+DWI+ADC	VGG	0.683	0.680	0.8288
		ConvNeXt	0.665	0.665	0.9877
		Regular ViT	0.620	0.622	0.9092
		Factorized ViT	0.646	0.628	0.1328
2,3 vs. 4,5	T2W	VGG	0.655	0.627	0.2564
		ConvNeXt	0.597	0.632	0.0725
		Regular ViT	0.574	0.623	0.0453
		Factorized ViT	0.535	0.549	0.5474
	T2W+DWI+ADC	VGG	0.648	0.653	0.7072
		ConvNeXt	0.644	0.662	0.2379
		Regular ViT	0.613	0.597	0.4874
		Factorized ViT	0.597	0.589	0.6829

Table 5.8: Comparison of hybrid and sequence-only (SO) model hold-out test set performance.

Target	R	p-value
{1} vs. {2,3,4,5}	0.774	5.8e-06
{1,2} vs. {3,4,5}	0.520	0.008
{2,3} vs. {4,5}	0.070	0.740

Table 5.9: Association between training data fraction and hold-out test set performance. The p-values correspond to the Pearson correlation coefficient (R).

Learning curve analysis

Deep-learning models are highly data-dependent. To understand this dependency on our data, we trained mpMRI VGG models on fractions of the total trainign data (0.1, 0.25, 0.5, 0.7) and evaluated them afterwards on the same validation set. In general, we observe a distinct positive trend when analysing how the amount of data impacts performance (Figure 5.5). However, this trend is not as evident when considering the test performance for the intermediate vs. high target (Figure 5.5B) — indeed, while CV performance improves, test-set performance remains constant for this target, suggesting that more data may not lead to improved generalisation. For other targets, this trend is clearly positive and there should be a clear performance benefit in collecting more data (Table 5.9).

Sensitivity analysis to crop size

To estimate if the crop size could have a negative impact on performance by accidentally excluding the lesion, mpMRI VGG models were trained with a larger crop size (192x192 rather than 128x128). This analysis shows that performance is nearly identical (Figure 5.6; tbl:deep-crop) — the central crop used for all models has no significant impact on either target, suggesting that a 128x128 central crop is sufficient to contain the signal relevant for classification.

Feature representation

To better understand feature distributions, a t-distributed stochastic neighbour embedding (t-SNE) projection was calculated for the bottleneck features (Figure 5.7A). Upon stratification by risk, it is possible to see some minimal clustering (Figure 5.7B); however, stratifying by dataset reveals that the high-dimensional

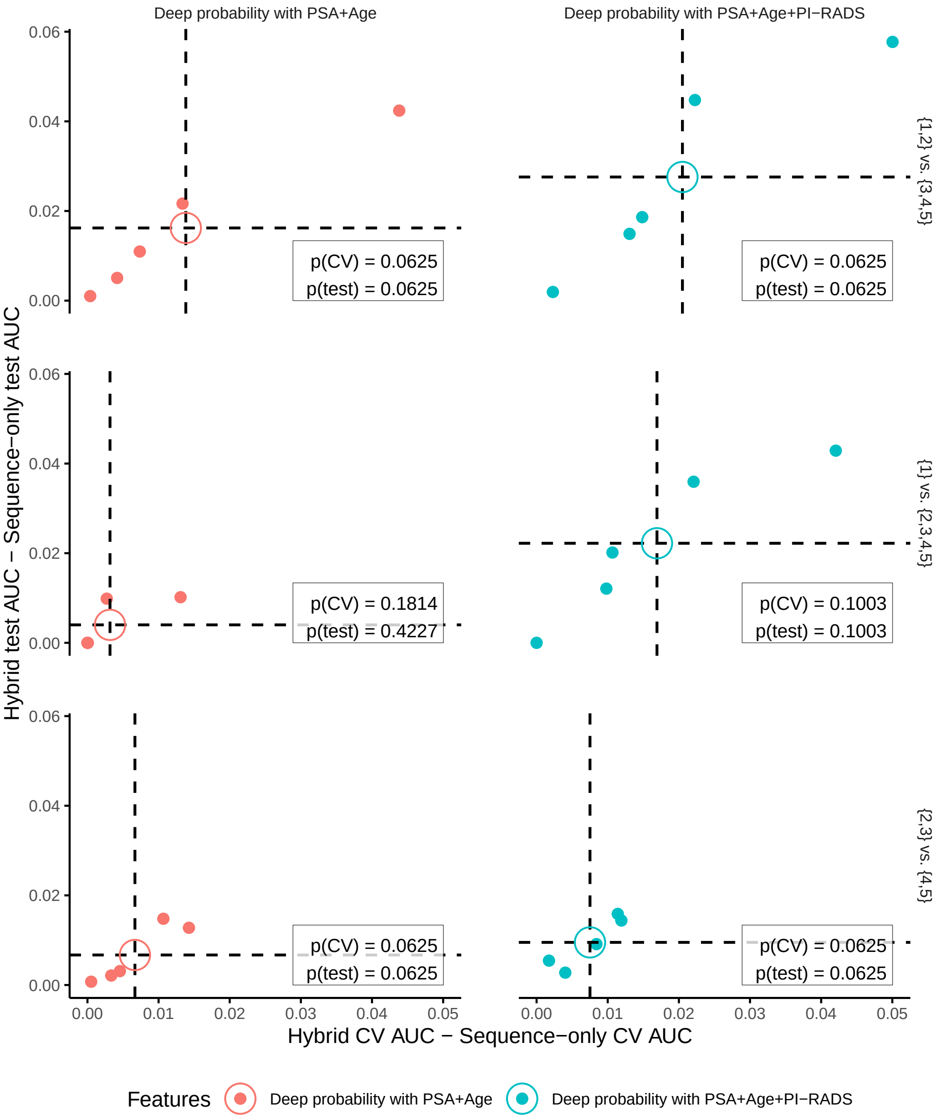


Figure 5.4: Difference between CV (x-axis) and test (y-axis) performance for hybrid linear models and sequence-only models for all targets. The first column refers to models incorporating logits, PSA and age, whereas the second columns refers to models incorporating logits, PSA, age and PI-RADS. Individual points represent the metric value for each fold, while the large hollow circles represent the average performance.

Target	128x128 avg. AUC	192x192 avg. AUC	p-value
{1} vs. {2,3,4,5}	0.729	0.713	0.256
{1,2} vs. {3,4,5}	0.683	0.675	0.480
{2,3} vs. {4,5}	0.648	0.642	0.755

Table 5.10: Comparison of model hold-out test set performance when trained with different crop sizes (128x128 and 192x192).

structure tends to be more strongly influenced by dataset than by classification ((Figure 5.7B vs. Figure 5.7C). An additional factor contributing towards the structure of the feature space is the use of an endorectal coil — in particular, studies performed with endorectal coils tend to have very few neighbouring studies using no endorectal coil. This highlights an important aspect of these models — not only is there a

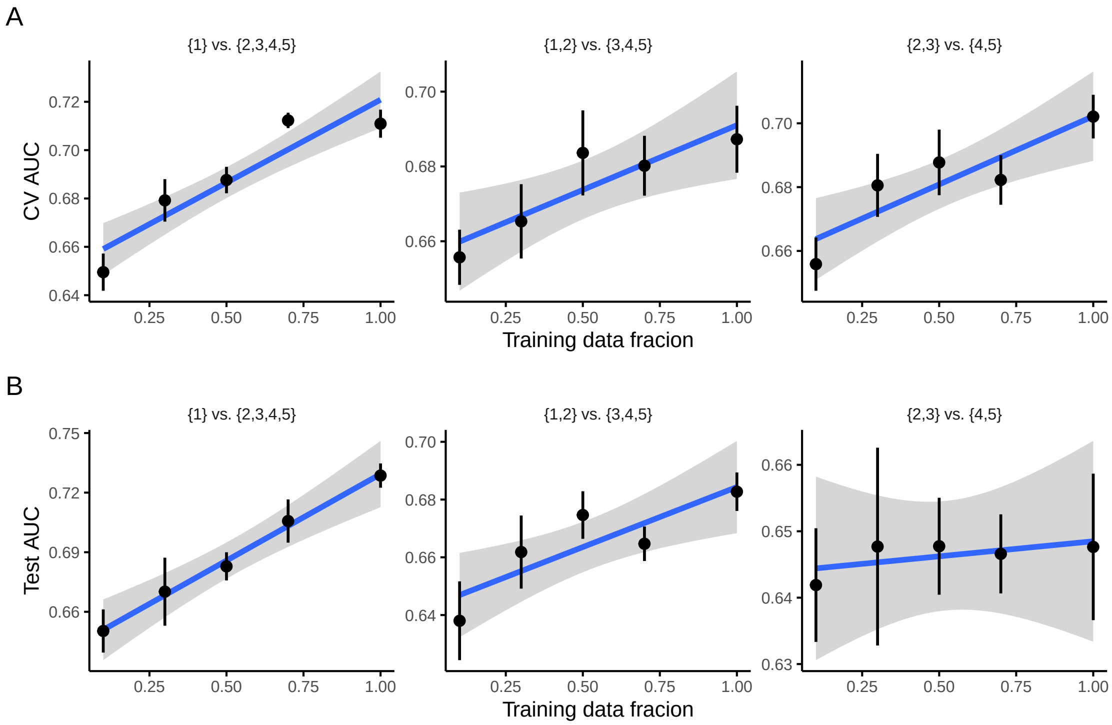


Figure 5.5: Learning curve analysis. A: Learning curve for cross-validated performance. B: Learning curve for hold-out test set performance. The black points/lines represent the mean/standard error around the mean, the blue line/grey area represent a linear fit to the plotted data/95% confidence interval for the fit, respectively.

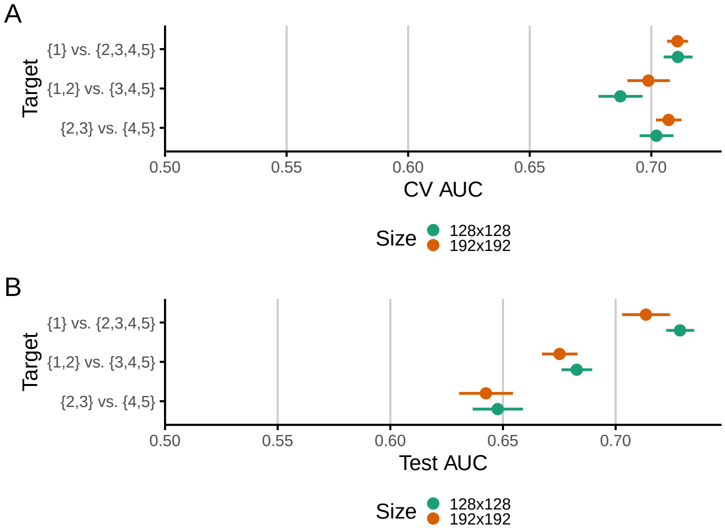


Figure 5.6: Sensitivity analysis to size. A: Cross-validation performance. B: Hold-out test set performance. Points/lines represent the mean/standard error around the mean.

significant association between features and data providers, the use of endorectal coil also leads to significant changes in the feature representation of these studies.

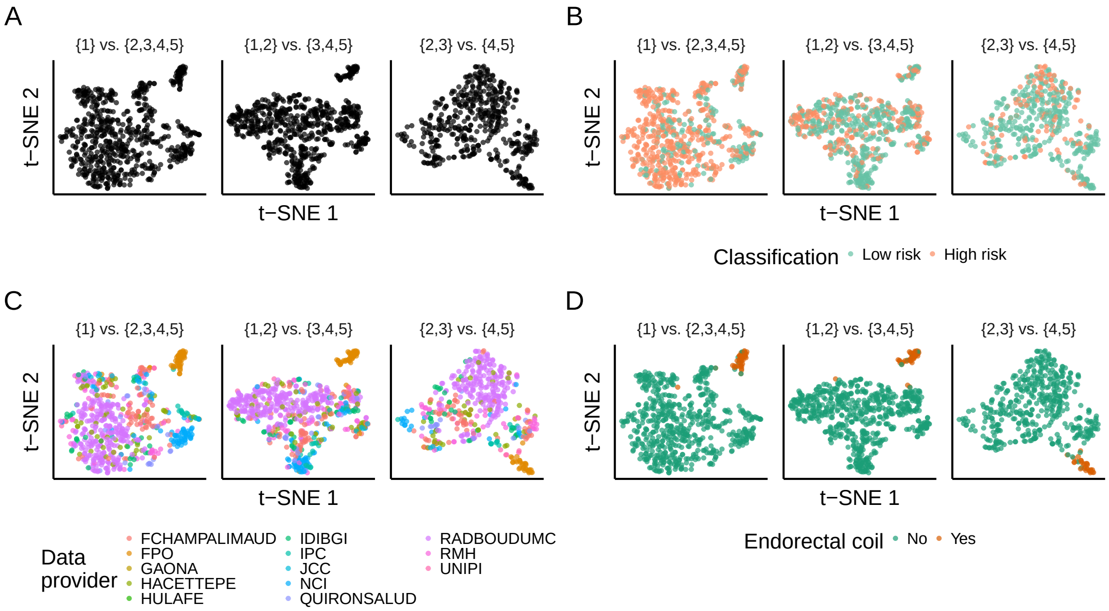


Figure 5.7: t-distributed stochastic neighbour embeddings for all targets. A: Original embeddings. B: Embeddings stratified by classification (low/high risk is ISUP={1}/{2,3,4,5}, ISUP={1,2}/{3,4,5} and ISUP={2,3}/{4,5} for each column). C: Embeddings stratified by data provider. D: Embeddings stratified by use of endorectal coil.

Fairness analysis

Models may perform differently depending on different forms and levels of stratification and analysing this is of great importance — by understanding where different models perform better, it is possible to better define the cases where they can be applied and generate value. In this section, different stratifying variables — dataset provider, PI-RADS, PSA quartiles, age quartiles and lesion location (in terms of transversal zone, peripheral zone, central zone and anterior fibromuscular stroma (TZ, PZ, CZ and AS, respectively) and apex, middle and base (A, M and B, respectively)) — are analysed for all three target definitions. These results are displayed in Figure 5.8. To calculate each AUC, the average predictions of the 5 CV folds were calculated for the hold-out test set for all three targets. 95% confidence intervals for the AUC were calculated using 2,000 bootstrap samples.

Dataset provider. In general, a wide range of performances is observed across datasets, most likely in association with the relative abundance of cases — as is the case for RADBOUDUMC. While some cases with a relatively small number of cases show considerably good performance (IPC, IDIBGI for the low vs. possibly high target), we note that their small numbers (no more than 30 cases) should prevent us from making overarching claims about performance in these centers. While other instances are harder to explain, the relatively poor performance for FPO is easily explained — FPO is a centre using almost exclusively endorectal coils in their examinations.

PI-RADS. The best performance is consistently observed in PI-RADS=5. This is not unreasonable — the expectation is that malignancy is more evident in these cases, which would make the definition of a decision threshold easier for an automated approach. While unexpected, the low performance of the intermediate vs.

high risk model in PI-RADS=4 is likely associated with the relatively small number of positive cases (23 out of 188).

Endorectal coil. Performance is significantly worse in cases where endorectal coil is used. While it may be due to changes in contrast (endorectal coil leads to changes in contrast, particularly in T2W images), there is also the possibility that the relatively small amount of studies with endorectal coil skews the performance negatively; however, it should be noted that as evidenced in Figure 5.7C, the feature representations are relatively different between studies using and not using endorectal coil. For this reason, we posit that a mixture of both — few instances and changes in volume contrast — lead to the very poor performance in studies using endorectal coil.

Age (quartiles). Performance remains relatively stable across different age groups for the low vs. possibly high and possibly low vs. high targets (with the exception of the thid quartile for the possibly low vs. high risk target). For the lowest and the highest age quartile there is a stark drop in performance for the intermediate vs. high risk models.

PSA concentration (quartiles). As with age, prediction is worse for half of the quartiles when considering the intermediate vs. high risk target models. There is also a clear drop in performance for the low vs. possibly high risk for the highest quartile — this would be indicative that these models are more likely to fail at higher PSA concentrations, when the prevalence of possibly high risk cases is considerably higher than for the rest of the PSA quartiles.

Location. Performance for different locations appears to be largely influenced by the amount of data when considering TZ, PZ and CZ. However, for AS, performance appears to be consistently good across targets (it should be noted nonetheless that there are only 32 test cases for AS).

Location (apex, middle, base). Performance for lesions in the middle of the prostate appears to be more variable than for lesions in the apex despite existing in similar numbers. A marked increase in the variability of the performance is observed for lesions in the base of the prostate.

Multiclass ensembling

Ensemble performance To test how these models can be further repurposed in a multiclass setting, additional experiments are performed using the low vs. possibly high and the intermediate vs. high models to construct an ensemble model. As demonstrated in Figure 5.9 (top), CV performance is best for ensemble models when compared with the baseline. While some alterations to the ensemble were tested, given that improvement was minimal, the more basic ensemble model (with no LoRA or SAE) was used and considered for further experimentation, particularly through the addition of clinical features and conformal prediction. Further testing these models shows that their performance generalises well as was the case with earlier target definitions.

Inclusion of clinical variables To test how the inclusion of clinical variables to prediction, three different elastic net-regularized linear classification models are trained — one with all clinical variables (incl. PI-RADS), one with multiclass probabilities and PI-RADS, and one with multiclass probabilities and all clinical variables. As visible in Figure 5.10, models incorporating DL are marginally better than those using clinical information derived through careful inspection by a set of medical experts. This has been, in fact, a trend throughout these results — PI-RADS has, in general, no additional predictive power once features extracted through DL methods are considered, implying that, for the tasks considered, DL methods are extracting the information that radiologists would generally consider without being explicitly guided to do so (i.e. predict PI-RADS).

Conformal prediction Conformal prediction methods have the potential to improve performance, but it is hard to consider them to be particularly useful in this scenario — as illustrated in Figure 5.11, the main improvement lies in the detection of High risk cases for models using DL, with a considerable drop in coverage (from 100% to 77.36%), a drop which becomes considerably more significant once High risk cases are considered (100% to 46.8%). So, while using conformal prediction can have a positive impact by removing cases where the model is uncertain (and patients should simply follow a standard and ”traditional” care routine), it does not provide a predictive performance adequate for a clinical care model. Nonetheless, it should be noted that as master models — models which can be used as foundational building blocks for other approaches — these models should promising discriminatory performance.

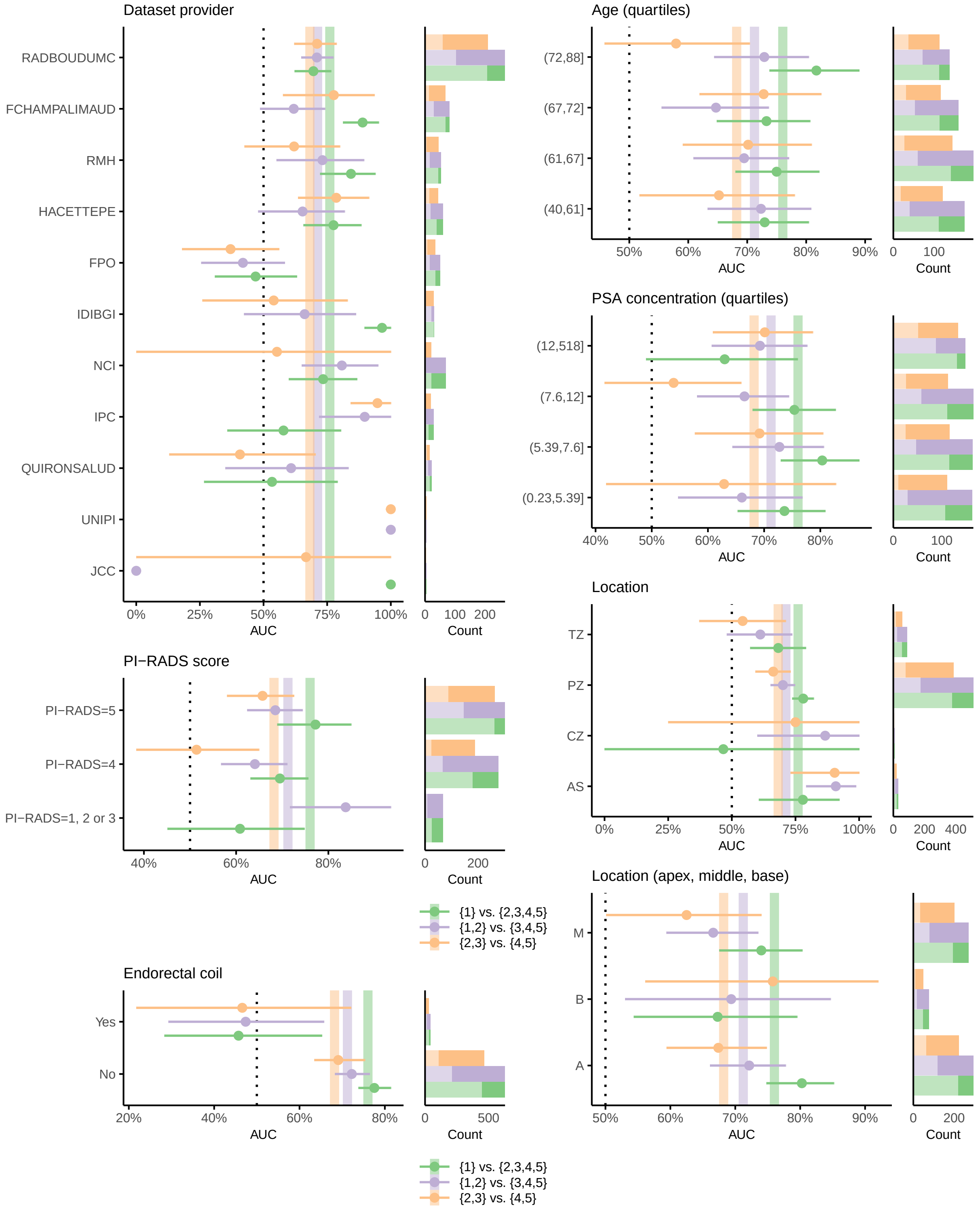


Figure 5.8: Fairness analysis for all binary target definitions for the deep-learning models. Point-and-range plots represent the performance (mean and bootstrap 95% confidence interval) and the coloured vertical lines represent the expected performance on the whole dataset. Horizontal bar plots represent the counts in each target and stratum. The lighter fraction of the bars represents the positive cases.

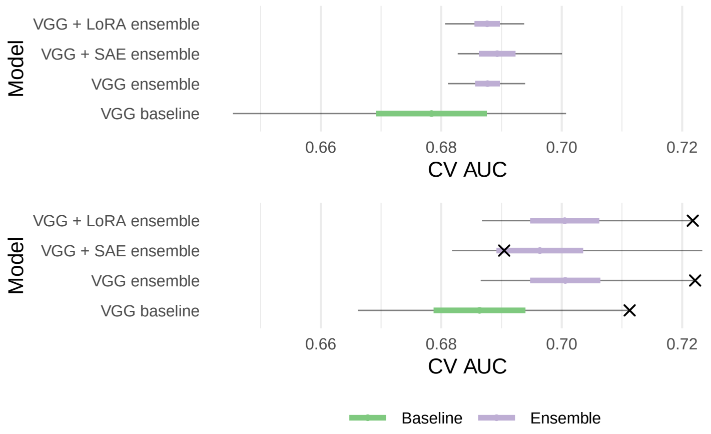


Figure 5.9: Performance of different ensemble models and comparison with baseline VGG model. Top panel refers to CV performance and bottom panel refers to testing performance. Crosses in the bottom panel refer to the best-performing model (highest CV AUC). For both panels, the points refer to the mean, the thick lines represent the standard error range centred around the mean and the black lines represent the minimum and maximum performance.

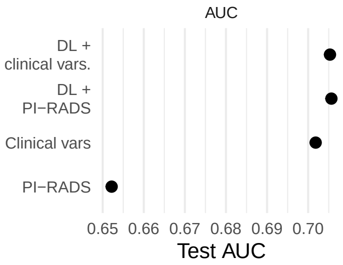


Figure 5.10: Hold-out test set AUC for linear models using PI-RADS, clinical variables, ensemble probabilities (DL) and PI-RADS and DL and clinical variables (including PI-RADS).

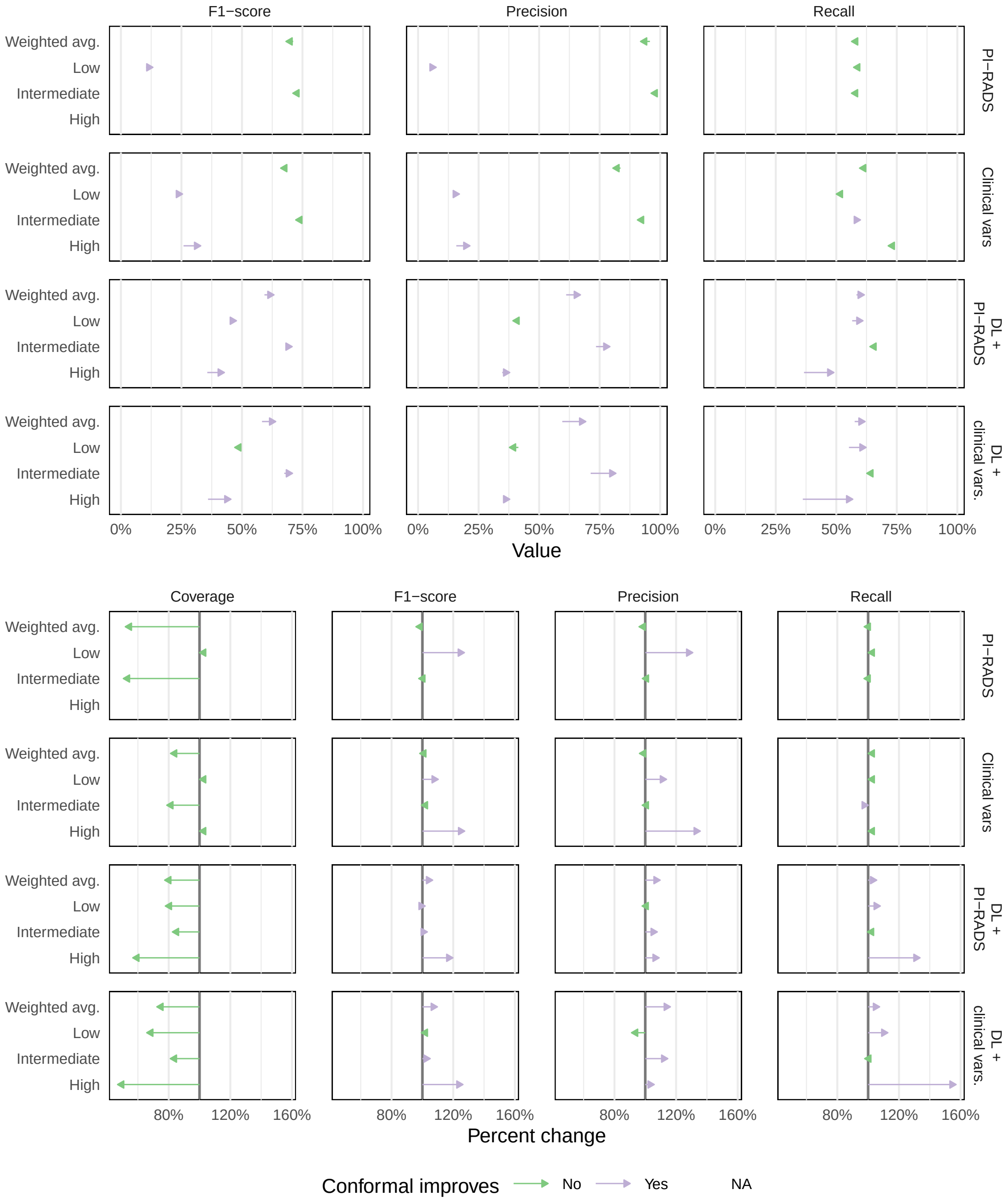


Figure 5.11: Comparison of absolute (top) and relative (bottom) performance with and without adaptive prediction sets (APS). Arrows point in the direction of the APS results, while the base of the arrow represents the performance without APS. It should be noted that the PI-RADS-only model does not detect any High risk cases and, as such, no performance is displayed

5.1.3 Discussion

Direct comparisons of performance between targets may not be particularly informative. Nonetheless, we attempt here to summarize the more relevant performance differences when considering different targets.

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 [7]. 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 is relatively stable between CV and hold-out test set.** In general, we observe that generalizability — the ability of models to perform as well on a hold-out test set — is good for the low vs. possibly high and possibly low vs. high target definitions, while being relatively poor for mpMRI intermediate vs. high models. We posit that this may be due to the relatively smaller amounts of data which can lead to more dramatic cases of overfitting [62]

4. **Clinical data (age, PSA, PI-RADS) does not appear to improve the performance of DL models.** While multiple different models were assessed, we failed to see consistent gains by using PSA, age or PI-RADS as additional predictors in a model. While this complicates future additions to these models as other types of data are 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 — the performance of intermediate vs. high models appear to show no association with the amount of data, suggesting that other studying alternative approaches may be more useful

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 [66]. 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. **Data provider and endorectal coil use overpowers the feature landscape.** Finally, we note an important aspect of this analysis — the effect that data provider and endorectal coil use 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 either of the first two targets — **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

For the classification tasks, we focus primarily on aggressiveness classification (use case 2). To define aggressiveness, conflicting evidence arises — to the best of our knowledge, the most recent definitions rely either on the clinical categorisation of prostate cancer (PCa) into T1-4 (depending mostly on the size of the tumour), N1-2 (depending on the infiltration of lesions in nearby lymph nodes) and M1 (depending on the detection of metastases) or on the International Society of Urological Pathology (ISUP) grading, which is itself derived from Gleason scores [37]. While both are valid, we have, primarily, access to ISUP grading derived from Gleason scores, causing us to choose this as our preferred target. However, there is still an issue as ISUP grading is an ordinal value, with 5 possible values — the integers between 1 (least aggressive) and 5 (most aggressive).

There are clear differences in overall survival between each individual ISUP score with the exception of the comparison between 1 and 2, where differences in survival are not statistically significant [84], and there is evidence that ISUP=1,2,3 vs. ISUP=4,5 is associated with the most discernible distinction in terms of clinical progression [37]. However, discerning between overall survival categories is not the only aspect which can be of interest — indeed, one can be interested in discerning between clearly low risk lesions (ISUP=1) from other lesions (ISUP=2,3,4,5) as this would prevent unnecessary biopsies as suggested by Schoots and others [79], or follow the ISUP guidelines [94], determining that high risk PCa corresponds to ISUP=3,4,5. In either case, this requires distinct models — while a categorical or ordinal multiclass classification is always possible, we note that this is a highly unbalanced problem with little precedent and to the best of our knowledge there is always a step of discretization involved [85] (we note here that the absence of a precedent for such a multiclass approach is likely explained by reporting bias which, tendentially, leads researchers to avoid publishing results were performance is not satisfactory [58]).

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 [84], the evidence for stratification in recurrence-free survival is mixed [84, 82, 60]. 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 [47], whereas another showed that reevaluation of Gleason scores leads to a different grading in approximately 20% of instances [91]. Indeed, ISUP is a useful, albeit noisy, grading and we believe this is consequential in terms of defining a target variable for prediction.

5.2 On the impact of cropping strategies

5.2.1 Methods

Data curation and preparation

We used the retrospective cases available through ProstateNet in May 2023, which consisted of 16921 T2W series belonging to 9582 patients. The following filtering steps were applied to select suitable data.

- We exclusively chose axial acquisitions by assessing the presence of the terms *ax* or *tra* in the **series_description** field. To further validate the axial nature of the acquisition, we calculated the cross product of the x, y, and z coordinates of the upper left hand corner of the image provided in the **image_orientation_patient** field. If the cross product equaled 2, it confirmed the acquisition as being axial.

- We excluded series that did not include *FS* in the **scan_options** field.
- We filtered out series containing *whole pelvis*, *bh*, *star* or *kidney* in the **series_description** field.
- We kept only the series where the slice thickness, as indicated in the **slice_thickness** field, was less than or equal to 4 mm.

These filters resulted in a filtered dataset consisting of 4903 UC2 series, corresponding to 4686 unique patients.

To further homogenise the dataset, additional filtering steps were applied, based on the information derived from DICOM tags. More precisely, we:

- required a minimum of 16 slices in the acquisition.
- set a gap of less than or equal to 1 mm.
- ensured that the slice distance, defined as the sum of slice thickness and gap, was less than or equal to 5 mm.
- ensured a vertical FOV of less than or equal to 150 mm.

By applying these additional criteria, we identified 4796 UC2 series, corresponding to 4613 unique patients, which met the specified conditions.

By further analysing the dataset, we identified cases with anomalous ground truth values, namely Gleason Scores ≤ 0. This value was assigned to patients that had no biopsy or with an erroneous selection of the index lesion during data upload. Also there are cases with PI-RADS = 0. We preferred not to consider all these cases, thus deleting 301 series (belonging to 297 patients), in order to be sure the cases we use have visible lesions of prostate tumor, along with a bioptic ground truth.

We conducted a final manual inspection to identify any remaining artifacts and endorectal coils in the dataset. As a result of this last filter, we obtained a cleaned T2W dataset referred to as **C_T2W_UC2**, which consists of 4145 UC2 series of 3984 patients.

Since UC5 is a subset of UC2, we derived the ultimate UC5 dataset, designated as **C_T2W_UC5**, by extracting all UC5 series from the final refined UC2 dataset. **C_T2W_UC5** comprises 694 series from 659 patients.

To train the models based on both T2W and ADC data, we paired each T2W acquisition from **C_T2W_UC2** with the corresponding ADC acquisition, by matching the same patient ID wtih the study. This pairing resulted in the creation of the cleaned T2W+ADC dataset, referred to as **C_T2W+ADC_UC2**, which comprises 3298 UC2 series involving 3294 patients. Notably, this indicates that 690 patients exclusively possess T2W series without any corresponding ADC data. Furthermore, the corresponding **C_T2W+ADC_UC5** dataset consists of 555 series from 554 patients, within the UC5 subset.

Finally, to train models that utilize T2W, ADC, and DWI data, we associated each T2W+ADC acquisition from **C_T2W+ADC_UC2** with the corresponding DWI acquisition. It is worth noting that multiple DWI acquisitions were collected for each patient, each corresponding to a different b value. However, for our dataset, only the DWI series corresponding to the maximum b value was included. We acknowledge that this approach may introduce biases into the model, as discussed in subsection 5.2.3.

The resulting cleaned T2W+ADC+DWI dataset, referred to as **C_T2W +ADC+DWI_UC2**, comprises 2979 UC2 series involving 2977 patients. Notably, this indicates that 317 patients exclusively possess T2W+ADC series without corresponding DWI data. Additionally, the corresponding **C_T2W +ADC + DWI_UC5** dataset includes 508 series from 507 patients within the UC5 subset.

We provide an overview of our filtering pipeline in Figure 5.12.

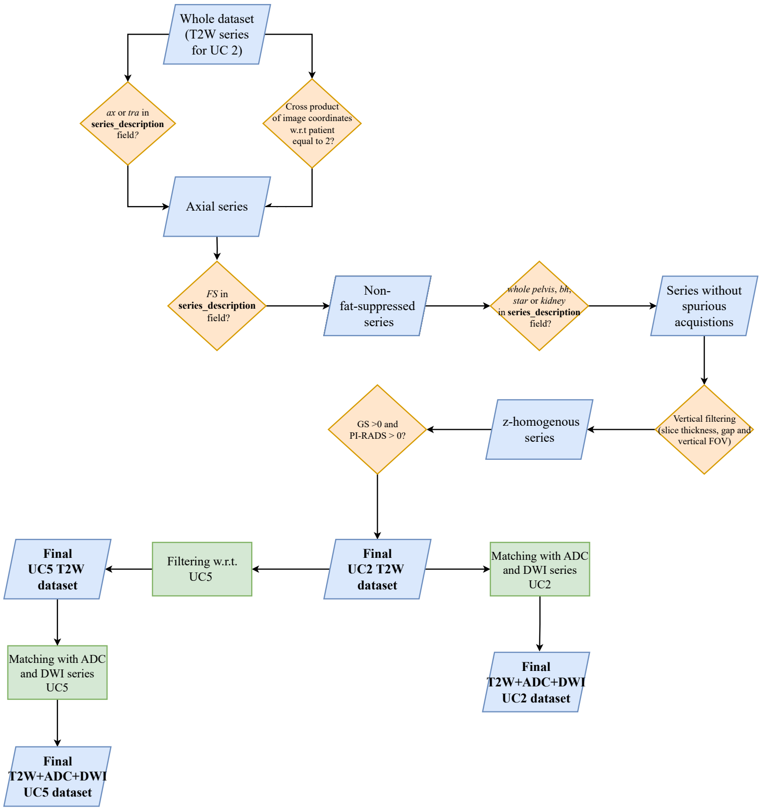

Whole dataset (T2W series for UC 2)
ax or tra in series_description field?
Cross product of image coordinates w.r.t patient equal to 2?
Axial series
FS in series_description field?
Non-fat-suppressed series
whole pelvis, bh, star or kidney in series_description field?
Series without spurious acquistions
Vertical filtering (slice thickness, gap and vertical FOV)
z-homogenous series
GS >0 and PI-RADS > 0?
Final UC2 T2W dataset
Filtering w.r.t. UC5
Final UC5 T2W dataset
Matching with ADC and DWI series UC5
Final T2W+ADC+DWI UC5 dataset
Matching with ADC and DWI series UC2
Final T2W+ADC+DWI UC2 dataset

Figure 5.12: Filtering pipeline flowchart.

Difference between central and adaptive crop

Generally, the prostate gland is located around the middle zone of the acquisitions. To avoid focusing on regions that do not contribute to the prediction, we decided to crop the images (using 64 × 64mm ×16 slices square FOV, followed by a resampling using linear interpolation to 256 × 256 pixels ×16 slices).

However, this FOV may cause some of the information from the prostate to be lost. In addition, the mismatch between T2 and ADC acquisitions is amplified by the use of a reduced FOV, which may decrease the predictive ability of the models.

Consequently we defined two cropping strategies. The first one is a Central Crop Strategy: a central crop of 64 × 64mm ×16 slices FOV) is extrapolated from T2W, ADC, and DWI. Without co-registration, these 3 crops are used to train the models.

The second strategy is an Adaptive Crop Strategy: an in-house segmentation model (i.e., based on an Attention-UNet [12, 11]) is trained on the T2W/ADC PICAI challenge [78] acquisitions.Consequently, the Adapative Crop Strategy is applied only to C_T2W and C_T2W_ADC datasets.

This model identifies the 3D prostate independently on both T2 and ADC.

For each slice, only the greater connected component is kept (if greater than $5mm^2$), whereas the other segmentations are removed. A 3D prostate mask is identified. The center (x,y,z) of the gland is defined as the baricenter of the mask. Consequently, both T2W and ADC acquisitions are cropped around their center (using a 64 × 64mm ×16 slices FOV). This allows the images to be co-registered and reduces the risk of losing information about the prostate.

Since the segmentation model was training on T2W and ADC data, the Adapative Crop Strategy is applied only to C_T2W and C_T2W_ADC datasets.

All the volumes are normalized by applying a Z-score normalization. The mean μ of the volume is moved to 128, whereas $[\mu - 2\sigma, \mu + 2\sigma]$ gray-scale values are mapped on $[0, 255]$, with σ the standard deviation of values of the whole volume.

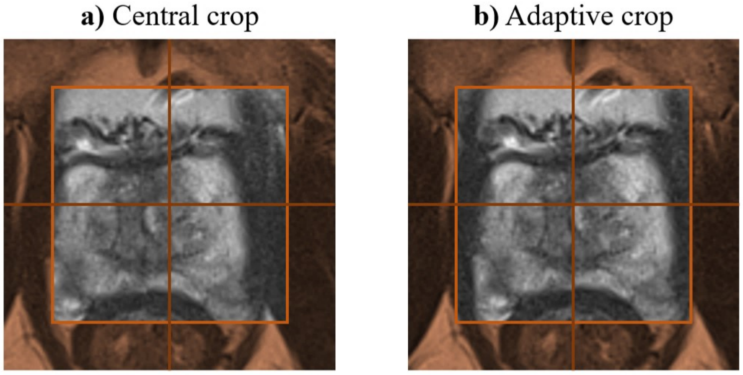


Figure 5.13: Example of improved crop using Attention-UNet segmentation on T2W acquisitions.

Deep learning model specification

Models description. We used two models for our experiments: 3D Vision Transformer (ViT) and a 3D modified VGG, introduced in the previous section, as this gave the best results. The ViT model was originally designed for handling two-dimensional data. We modified the original model [21] to handle three-dimensional input, i.e., each embedding is obtained by flattening a 3D patch rather than a 2D one. We define our input as $x \in \mathbb{R}^{H \times W \times Z \times C}$, where (H, W, Z) represents the resolution of the volumetric input, and C denotes the number of channels. The ViT divides the input volume into (P, P, Z) patches and flattens them into a one-dimensional vector. As a result, the encoder receives a sequence of flattened patches $x_p \in \mathbb{R}^{Nx(P^2 \cdot Z \cdot C)}$ as input for each input volume, where $N = HW/P^2$ represents the number of patches. In this case we considered P=16 and Z=4, i.e., each volumetric input is splitted into $16 \times 16 \times 4$ volumes from which computing the embeddings. For our 3D ViT model architecture we considered the following hyperparameters: MLP dimension of 2048; number of encoders equal to 32; number of attention heads of 8 and embedding size of 32.

For the description of the VGG model, please refer to the Section above.

Initially, both ViT and VGG models were trained exclusively using T2w images. To incorporate the ADC and DWI modalities into both the 3D ViT and 3D VGG models, we developed a multi-branch architecture. We began by training a two-branch model that handled the T2 and ADC modalities and subsequently expanded it into a three-branch model to accommodate the DWI modality as well. In both scenarios, each branch received one modality as input and extracted its relevant features. Finally, the features extracted from each branch were concatenated and forwarded to a common linear layer, which generated the ultimate prediction. We present our multi-branch 3D ViT in Figure 5.14.

Experiments. We utilized the Binary Cross Entropy loss function and the Adam optimizer during the training of our models. To mitigate the effects of class imbalance, we implemented batch-weighting, a technique that dynamically adjusts the weights assigned to each sample within a batch based on the class distribution observed in that particular batch. We did not employ any data augmentation technique.

Throughout our experiments, we maintained consistent hyperparameters for both the ViT and VGG models. Specifically, all models were trained for 100 epochs, with an initial 10 epochs designated for warmup. The ViT model started with an initial learning rate of 5e-5, while the VGG model began with a learning rate of 5e-4. Additionally, we applied a weight decay value of 0.1 to the ViT and 0.005 to the VGG model. Finally, we used a batch size of 30 for the ViT and of 16 for the VGG.

In our study, we trained a total of six distinct models, encompassing T2W, T2W+ADC, and T2W+ADC+DWI combinations, for both the ViT and VGG architectures on the center cropped dataset. Additionally, for the

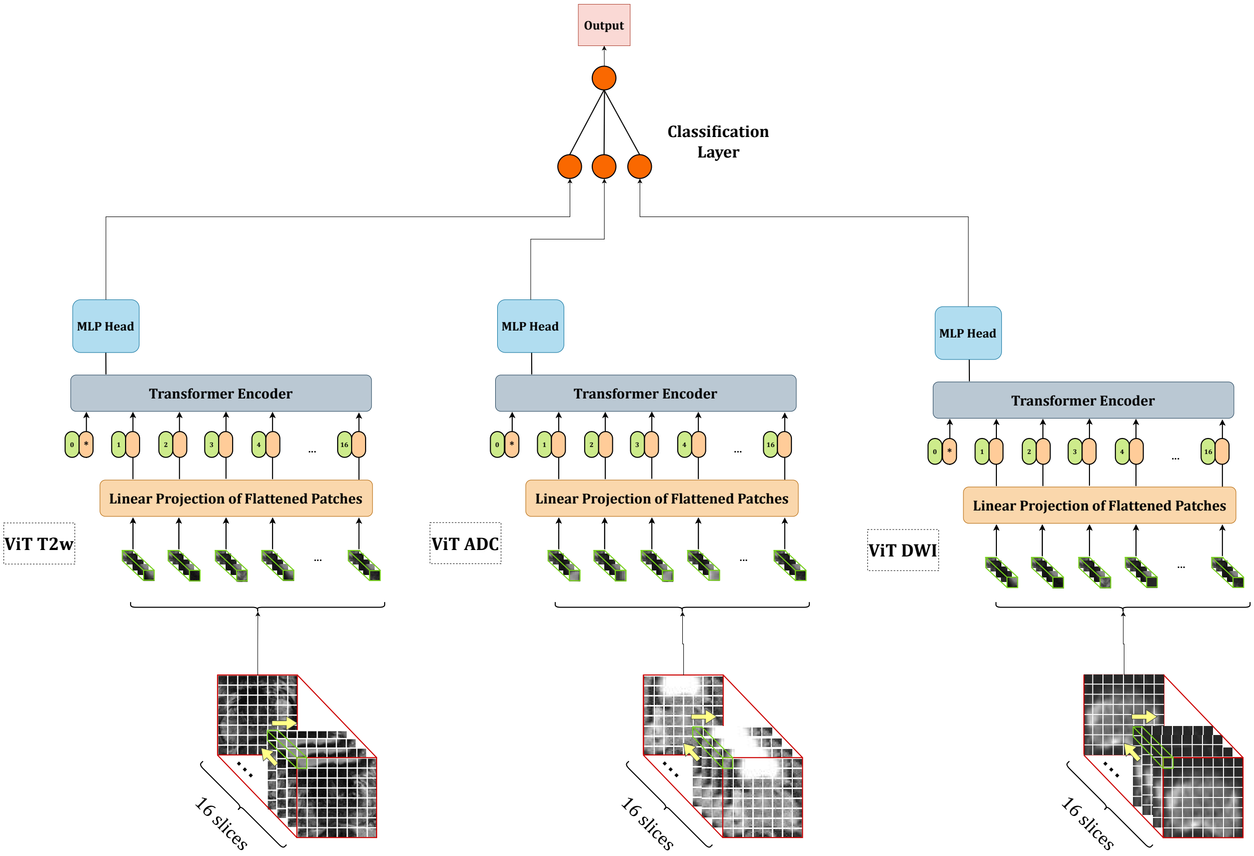


Figure 5.14: Multi-branch 3D ViT

adaptive cropped dataset, we trained four models, covering T2W and T2W+ADC combinations for both ViT and VGG models. This sums up to a total of 10 models for each specific use case.

Models evaluation. For each combination of use case (UC2/UC5) and model (ViT/VGG) we applied a 5-fold cross validation procedure, tested on an external independent test set. The dataset is stratified for both vendor ('Philips', 'Siemens', 'GE', 'Other') and ground truth ('HG', 'LG') and randomized to each other patient characteristic to reduce bias. This stratification ensures consistency across the 5 folds.

The dataset is subdivided in 6 folds (16.6%). One fold is kept as an external test set, while the other 5 are used to train 5 models in a 5-fold cross-validation approach. Each time, 4 folds are used as training set and one as validation set (to define a stopping strategy to prevent overfitting). The selected model is the one minimizing the validation loss for each of the folds. The 5 models trained in this manner are applied on the external test set and the AUROC scores are reported as median and [0%,100%] percentiles.

Learning curve analysis

After training the models, we evaluate the learning capacity of the VGG and ViT models for UC2 and UC5 with respect to the dataset size.

The data set were resampled using the same stratification (vendor+pathology) to obtain 3 subsets of cardinality 25%, 50%, and 75% of the full set. For each subset, the same training/validation procedure is applied (6 folds: 1 test, 5 used as training+validation) and the scores are reported as median performance of the models on the test set and percentiles on the 5 folds.

5.2.2 Results

UC2

In this section, we present the results achieved for UC2 for both 3D ViT and 3D VGG. These models were trained using two datasets, namely Central crop and Adaptive crop datasets. Using the Central crop dataset, the models were trained using T2w modality alone, a combination of T2w and ADC, and a combination of T2W, ADC, and DWI. For the Adaptive dataset, instead, only T2W alone and the combination of T2W and ADC modalities were used for training. The results we provide are computed as averages across the 5-fold for both the validation and test datasets. Specifically, we provide statistics including the mean, median, 0th percentile, and 100th percentile.

We provide results for Adaptive and Central in Table 5.11 and Table 5.12, respectively.

Regarding the Adaptive Crop dataset, both the ViT and VGG models exhibit comparable performance levels. This holds true for both the T2W-only modality and the combination of T2W and ADC. Notably, the inclusion of the ADC modality appears to yield only marginal improvements in terms of AUROC performance for both models. The VGG model, in particular, achieves the highest performance, with a mean AUROC of 55.6% when trained on the combined T2W and ADC modalities. Across all experiments, we consistently observe minimal variation among the 5-folds, as evidenced by the closely aligned values for the 0th and 100th percentiles. This demonstrates a high level of model stability across different dataset splits.

Adaptive crop	**Dataset**	**ViT**	**VGG**
T2W	Validation	53.28%/52.4% [51.4%-56.1%]	59.1%/59.3% [57.1%-61%]
	Test	53.6%/53.7% [52.7%-54.4%]	54.3%/56.1% [50.6%-57.1%]
T2W+ADC	Validation	53.02%/52.1% [51.4%-56.5%]	57.6%/57.5% [55.4%-59.8%]
	Test	54.5%/54.3% [54.2%-55.2%]	55.6%/55.4% [53.7%-57.6%]

Table 5.11: For each experiment, we present the validation and test performances in terms of mean, median, 0th percentile, and 100th percentile.

In the case of the Central crop dataset, it exhibits slightly better but comparable performance compared to the Adaptive Crop dataset. Once again, all experiments yield fairly consistent results, with the inclusion of the ADC modality resulting in only a modest improvement in performance. Specifically, there is an approximate 2% increase in AUROC for the ViT model and slightly over a 1% improvement for the VGG model on the test set when ADC is added. It's worth noting that, similar to the Adaptive Crop dataset, we still observe model stability across the 5 folds in most cases. However, there is an exception when considering the VGG model trained solely on T2W images, where we observe an almost 10% variation between the 0th and 100th percentiles. Once again, the VGG model trained on the combination of T2W and ADC achieves the best performance, with a mean AUROC of 56.2% on the test set.

Central crop	**Dataset**	**ViT**	**VGG**
T2W	Validation	51,7%/51,9% [49,2%-53,6%]	56.7%/56.6% [55.4%-58.7%]
	Test	53,8%/53,8% [53,5%-54%]	54.9%/52.9% [51.9%-61.1%]
T2W+ADC	Validation	59%/58.9% [57.5%-60%]	61.6%/61.5% [60.6%-63.3%]
	Test	55.8%-56.8% [52.4%-57.5%]	56.2%/56% [55%-57.7%]
T2W+ADC+DWI	Validation	59.4%/59.9% [55.4%-62.1%]	64.2%/64.4% [60.7%-67.6%]
	Test	59.3%/59.3% [58.8%-59.7%]	63.1%/63.2% [61.3%-65.2%]

Table 5.12: For each experiment, we present the validation and test performances in terms of mean, median, 0th percentile, and 100th percentile.

UC5

In contrast to UC2, in the case of the Adaptive Crop dataset, we observe an improvement in performance on the test set w.r.t the Central crop dataset. This improvement ranges from 2% to almost 9% for the ViT model when trained on both modalities. However, there is an exception for the VGG model trained on T2W-only images, where the Adaptive Crop dataset leads to worse performance. Additionally, it's noteworthy that for the Adaptive Crop dataset, the inclusion of the ADC modality results in a significant 5% improvement when using the VGG model. Specifically, the VGG model trained on both T2W and ADC modalities achieves the highest performance, boasting a mean AUROC of 58.7%.

Adaptive crop	**Dataset**	**ViT**	**VGG**
T2W	Validation	51%-54.8% [38.4%-56.4%]	64.9%64.5% [62.6%-67.8%]
	Test	57.48%-59.1% [51.4%-60.1%]	53.7%-52.3% [46.8%-61.7%]
T2W+ADC	Validation	53.8%-53.1% [35.2%-72.5%]	66.7%/67.4% [63.4%-68%]
	Test	55.4%-55.8% [37.6%-73%]	58.7%-60.8% [53.4%-60.8%]

Table 5.13: For each experiment, we present the validation and test performances in terms of mean, median, 0th percentile, and 100th percentile.

Central crop	**Dataset**	**ViT**	**VGG**
T2W	Validation	48.1%/48.2% [42%-54.3%]	64.3%/64.8% [61.1%-66.8%]
	Test	52.8%/53% [51.5%-53.6%]	56.9%/57.4% [49.3%-63.4%]
T2W+ADC	Validation	42.9%/44.5% [34.3%-47.6%]	69.1%/68.3% [65.5%-72.4%]
	Test	46.7%/47.3% [37.2%-58.2%]	56.2%-56.9% [51.6%-60.7%]
T2W+ADC+DWI	Validation	66.3%-62.7% [53.1%-69.8%]	65.7%-68% [65.3%-73.4%]
	Test	62.4%-63.2% [58.1%-64.3%]	56.02%-57% [48%-60.3%]

Table 5.14: For each experiment, we present the validation and test performances in terms of mean, median, 0th percentile, and 100th percentile.

Learning curve analysis

The results of the curve analysis are reported in Figure 5.15.

The learning curves indicate that, concerning UC2, the performance trends for both VGG and ViT do not follow the expected pattern of improvement as the dataset size increases. Instead, there is a slight decline in performance (even if not statistically significant). This phenomenon could be attributed to the fact that the expansion of the training dataset introduces more complex and challenging cases (e.g., increasing number of vendors) for the neural network to classify, consequently affecting its overall performance negatively. Alternatively, one could consider the possibility of the networks being undersized. This notion gains support from the UC5 scenario, where we observe a contrasting performance trend. In the case of UC5, there is a notable increase in performance as the dataset size grows, particularly evident in the case of VGG, where the improvement is consistent with dataset expansion. Indeed, the UC5 task stands out due to its considerably smaller training dataset, approximately one-sixth the size of the UC2 dataset. Additionally, it consists of a more challenging task, also because of a higher degree of data imbalance. Consequently, any further reduction in the training data has a more pronounced impact on the model's performance when compared to the UC2 scenario. This distinctive nature of the UC5 task results in both the ViT and VGG models delivering comparatively better performance, on average, for UC5 in comparison to UC2. However, it's important to note that these differences do not attain statistical significance, as the confidence intervals exhibit substantial overlap.

5.2.3 Discussion

We report here some considerations that might be useful to analysis and doublecheck results of models using DWI data.

BY analysis ProstateNet dataset and the DWI data, we think it is worth noting that heterogeneous b values may have a short-cut learning effect on the the trained deep learning models. More precisely, the models may learn a short-cut rule based on the b value "meta-information": the maximum b-values of the DWI series determines the prediction of tumor aggressiveness. This, of course, could potentially redirect the network's attention away from the features to be extracted and toward an attempt to reconstruct a simpler relationship.

To provide a numerical example, for UC2, 1044 patients with ADC+T2W+DWI meet the inclusion criteria. Out of these: 810 are LGs and 234 are HGs. The distributions of maximum b values are close (mean LG 1468, mean HG 1527, median 1400).

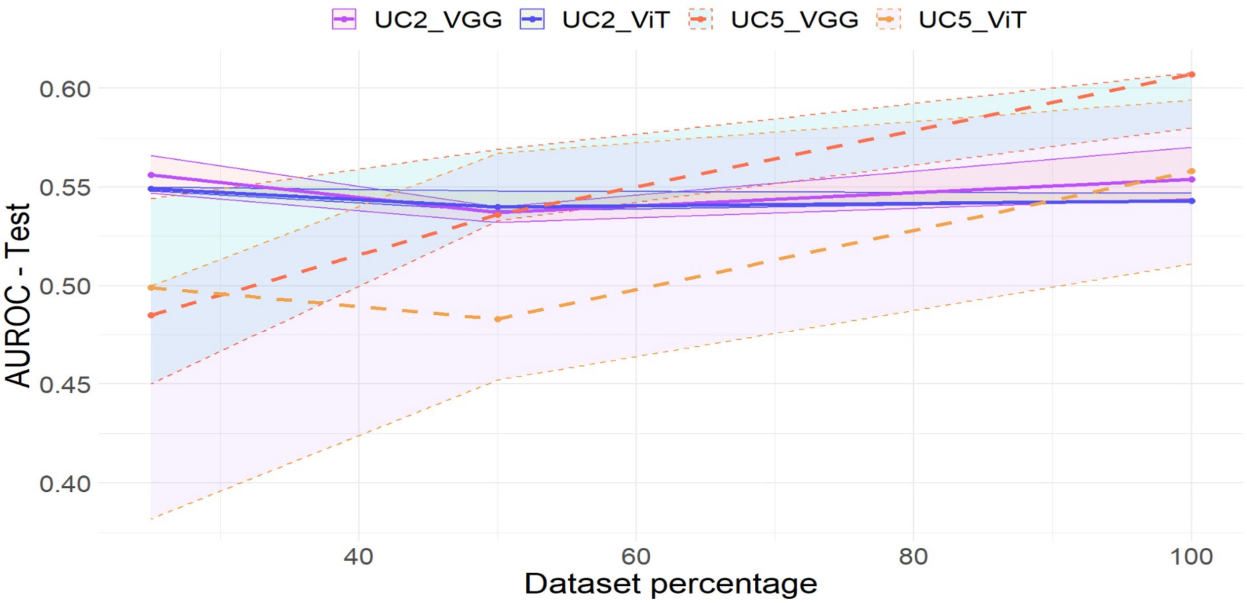


Figure 5.15: Learning curves for UC5 and UC2

However, 92.8% of LG values (in contrast to 79.4% of HG values) are below a b value of 1500. Similarly, 51.1% of LGs are below a threshold of 1400 while 60.3% of HGs are below the same threshold. Wanting to define a trivial classification: (HG if ¡1400 or ¿1500) results in the following confusion matrix, corresponding to an F1 score of 84% and a sensitivity of 87%.

	Actual LG	**Actual HG**
Predicted LG	710	170
Predicted HG	100	64

Table 5.15: Classification model based on an elementary b value-Ground Truth relationship.

Therefore, it is necessary to analyze the dependence of the model on the provided b value by applying an appropriate correction strategy. Examples include:

- Homogenize data to provide the same b value to each patient.
- Evaluate the network's ability to learn the b value (and not the requested ground truth).
- Determine whether, given the same ADC and T2W inputs, changing the DWI input (in terms of the b value) changes the prediction of the model.

5.3 On the differences between supervised and unsupervised learning strategies for use case 1

Use Case 1 (UC1) of the ProCAncer-I Project, *detection of prostate cancer*, can be thought of as comprehending two distinct tasks: a) Binary Classification of Prostate Cancer Presence denoted as *UC1—tA*; b) Prostate Index Lesion Segmentation denoted as *UC1—tB* as seen in Figure 5.16. Specifically for *UC1—tA* two different sub-tasks were designed on the basis of two well known machine learning techniques, *unsupervised* and *supervised* learning denoted as *UC1—tA1* and *UC1—tA2* respectively. The deep learning models developed (FORTH contributor) for those tasks can be used either as standalone models or dependently operating.

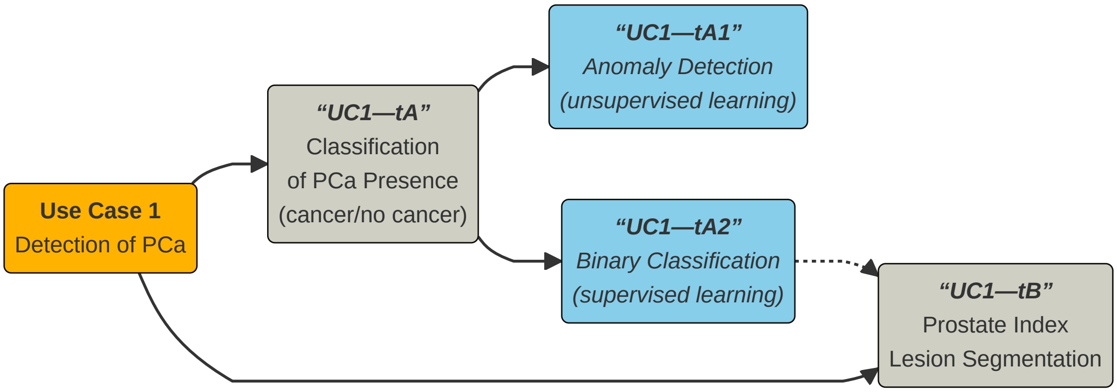


Figure 5.16: Tasks for addressing ProCAncer-I's Use Case 1, detection of prostate cancer. Different colours correspond to different level of tasks.

5.3.1 Binary Classification of PCa Presence (UC1—tA)

Data Curation

We make use of the retrospective cases from the ProstateNet imaging archive. We start by obtaining a total of 9,095 T2-weighted (T2w) sequences consisting of negative PCa (Use Form 1) and positive PCa (Use Form 1+2) cases, as described in [8]. Then, patients where an endorectal coil had been used were excluded alongside those cases that were obtained with fat-suppression technique. We end up with 6,843 cases of T2w sequences. 43% (2,946) are patients without confirmed PCa by pathology (e.g. positive MRI but negative biopsy) or men with no PCa findings on MRI and confirmed negative at follow-up (at least 1 year). The words *normal* and *negative PCa* are used interchangeably for these cases. Oppositely, 57% (3,897) are patients with confirmed prostate cancer at biopsy. For these cases the words *abnormal* and *positive PCa* are used interchangeably. The dataset that contains all Use Case 1 cases is denoted as *UC1-T2w*. Additionally, for our experiments we consider two different routes for collecting negative PCa cases as seen in Figure 5.17. By following the left branch of the flowchart, we are able to assemble a subgroup of 1,864 negative PCa cases solely with men that had no PCa findings on MRI and confirmed negative at follow-up (at least 1 year). This subgroup alongside the positive PCa cases from the original *UC1-T2w* dataset is denoted as *UC1-T2w-LeftBranch*. The complete composition of all different datasets used in our expirements is provided in Table 5.16. For all of them, 15% of the cases are used for validation, 15% are used as holdout-test, and the remaining 70% for training. Additionally, and in order to understand how the amount of data impacts model performance we split the training data into different fractions of the total amount of data – 0.1, 0.5 and 0.75; this allows us to build learning curves, which describe how the amount of data has an impact on model's performance.

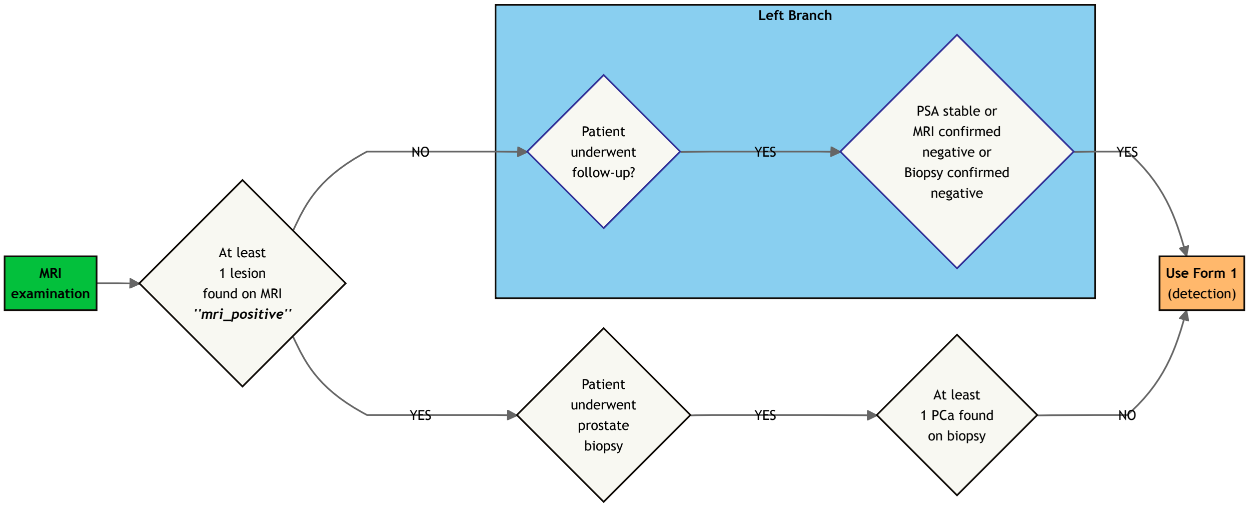

Left Branch
MRI examination
At least 1 lesion found on MRI "mri_positive"
NO
Patient underwent follow-up?
YES
PSA stable or MRI confirmed negative or Biopsy confirmed negative
YES
Use Form 1 (detection)
YES
Patient underwent prostate biopsy
YES
At least 1 PCa found on biopsy
NO

Figure 5.17: Stratification of negative PCa cases for Use Case 1 by PCa findings.

UC1 subset	train		val		test		total
	neg.	pos.	neg.	pos.	neg.	pos.	
UC1-T2w	2062	2728	442	584	442	585	6843
UC1-T2w-LeftBranchRaw	1305	2357	279	505	280	505	5231
UC1-T2w-LeftBranch	1305	1491	279	319	280	320	3994

Table 5.16: T2w data curation for Use Case 1. Note that for *UC1-T2w-LeftBranch*, the number of positive cases has been purposely reduced in order to maintain the true ratio of positive vs negative PCa cases (57% vs 437%) found in the full-size ProstateNET retrospective data. On the other hand, *UC1-T2w-LeftBranchRaw* includes the cases maintaining their original distribution, without any reduction to alter the true ratio of positive to negative cases.

Data Preprocessing

We utilized a pre-trained nnUNet [40] model to automatically segment the prostate gland for all ProstateNet cases. This model was validated on all 600 available ProstateNet ground truth prostate gland masks, and reached a Mean DSC of 0.8751 with a standard deviation of ±0.094. Additionally, we trained from scratch the v2 of the nnUNet model from MONAI on those 600 ground truth masks, and it produced a 5-fold Mean DSC of 0.9252, with a standard deviation of ±0.0012. Training for Use Cases 1 and 7a was conducted on the whole gland predicted from the pre-trained nnUNet model. For the subsequent phase, the volumes were resampled to a pixel spacing of 0.5mm x 0.5mm and a slice thickness of 3.0mm. This decision was based on the average spatial spacings observed in the cohort. To align with the model's requirements, a cropping and padding strategy was adopted, resulting in each volume being adjusted to dimensions of 32 slices x 224 pixels x 224 pixels. Additionally, for the cases where the field of view was wide an interpolation strategy was used.

Anomaly Detection (UC1—tA1)

Anomaly detection is the problem of recognising abnormal inputs based on observed examples of normal data and has been well-studied within diverse research areas and application domains. It arises from a common need when analysing real-world datasets to identify which instances stand out as being dissimilar to all others. Such instances are known as anomalies, and the goal of anomaly detection (also known as outlier detection) is to spot the 'abnormal' data samples knowing the 'normal' ones in a data-driven fashion [13]. This trait is especially important in high consequence applications, such as medical decision support

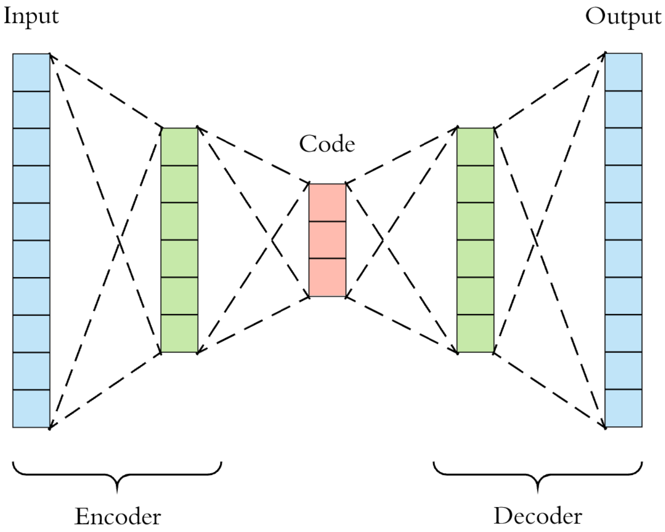


Figure 5.18: A generic architecture of an autoencoder. An autoencoder employs an encoder-decoder structure, where the encoder maps the input data to a low-dimensional latent representation and the decoder interprets the code and reconstructs the input.

systems, where it is vital to know how to recognise the anomalous data.

Auto-encoders (AEs). Auto-encoders are a type of neural network where the output layer has the same dimensionality as the input layer as seen in Figure 5.18. An auto-encoder looks to replicate the data from the input to the output in an unsupervised manner. In particular, AEs look to project the original input $x_i \in \mathbb{R}^d$, where $d << D$, to a *lower-dimension space* $\mathbb{R}^d$ and obtain a reconstruction of the original input x_i from the compressed version of the input. The two other main components of AEs are the *encoder* and the *decoder*. The encoder usually takes the form of a deep convnet (*i.e.*, FFNN, CNN or Transformer) and aims to compress the input into a latent space representation, obtaining a lower dimension representation. The decoder, typically "mirrors" the structure of the encoder and is responsible for reconstructing the input back to the original dimensions from the reduced representation obtained by the encoder. Two commonly known AEs are the convolutional auto-encoders (cAE) [56] and the variational auto-encoders (VAE) [69]. cAE are based on encoder-decoder structures that exploit convolutional layers, which allows an optimal encoding to be learnt for imaging-related tasks.

Representation Learning. Typically, in annotated medical imaging datasets, cases that demonstrate abnormalities (otherwise called positive) are scarce. By exploiting the ability of AEs to extract useful representations from the input data by reconstructing it from a compressed representation of it, and abundant negative (controls, healthy or normal cases) data, we hypothesise that AEs will be able to learn the 'concept of normality' and discriminate in an unsupervised way those cases that significantly differ from the "learnt normality". Specifically, AEs are trained solely with normal cases to learn their representations (features). Then, during testing, both normal and abnormal cases will be discriminated based on the divergence of the testing representations from the ones learnt during training (normal population) [18]. This process can also be defined as *outlier detection*, which has been extensively studied in other areas [15].

Task Definition. We assume an input $X_h = (x_{h1}, x_{h2}..., x_{hN})$, where $x_{h1}, i = 1...N$ are either normal (healthy, negative PCa) T2w volumes. The objective here is to learn the distribution $p(X_h)$ of normal cases through an auto-encoder architecture as seen in Figure 5.19. Our hypothesis is that a trained auto-encoder model will not be able to reconstruct abnormal images accurately due to the fact that it has only been trained with control ones, hereby learning the non-anomalous data distribution as a prior $p(X_h)$. To quantify the quality of the reconstruction the following two metrics are used:

$$MSE = \sum_{k=1}^{N} (x_{h_k} - x_{\hat{h}_k})^2 \tag{5.1}$$

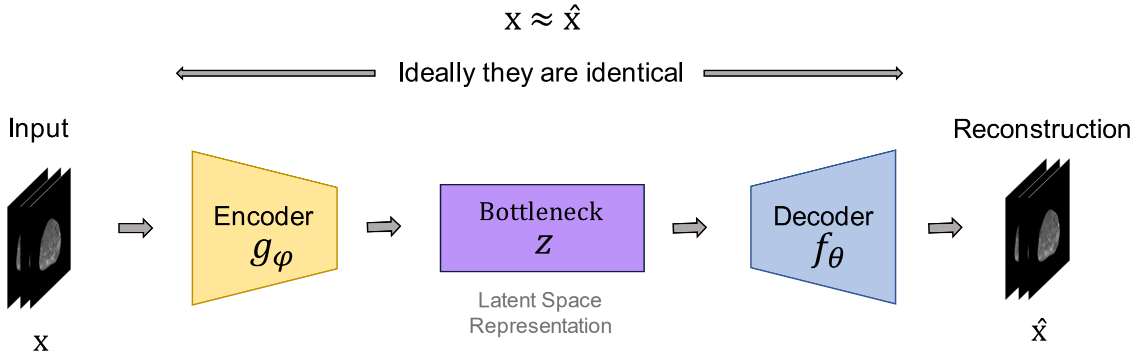


Figure 5.19: Pipeline of the proposed framework to learn the distribution of normal prostate MRI volumes.

$$SSIM = \sum_{k=1}^{N} l(x_{h_k}, x_{\hat{h}_k})c(x_{h_k}, x_{\hat{h}_k})s(x_{h_k}, x_{\hat{h}_k}) \tag{5.2}$$

Where c represents the contrast, l the luminance, s the structure of the images, $x_{\hat{h}_k}$ represents the kth reconstructed image and N is the number of images in the batch under consideration.

Proposed Architecture. One of the most prominent cAE architectures follows a U-Net like structure [75], including 3D convolution filters followed by a rectified linear unit (ReLU), pooling operations and dense layers, as seen in Figure 5.20. All architectures are trained using Mean Squared Error (MSE) loss function for all image modalities, with a learning rate of $1e^{-5}$, 200 epochs and batch size of 8. Solely negative PCa volumes are used during the training phase whereas both neagative and positive PCa volumes are used for testing purposes. The "base channel size" is another network parameter of our development and refers to the number of channels used in the first convolutional layer. The hyperparameters of the developed models are shown in Table 5.17.

Hyperparameter	Value
Activation Function	ReLU
Output Function	tanh or sigmoid
Loss Function	MSE
Optimizer	Adam
Batch Size	8 or 4
Epochs	200
Learning Rate	0.00001 (1e-5)

Table 5.17: Hyperparameter settings of the cAE models.

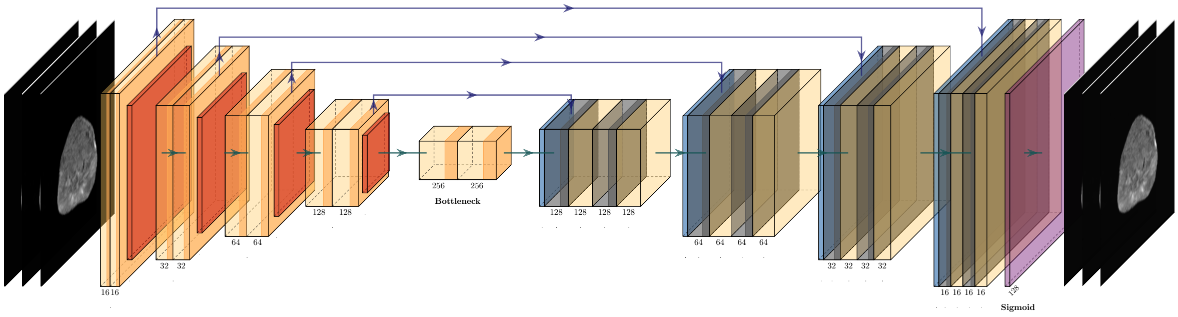


Figure 5.20: U-Net-based convolutional auto-encoder architecture for reconstructing prostate MRI volumes.

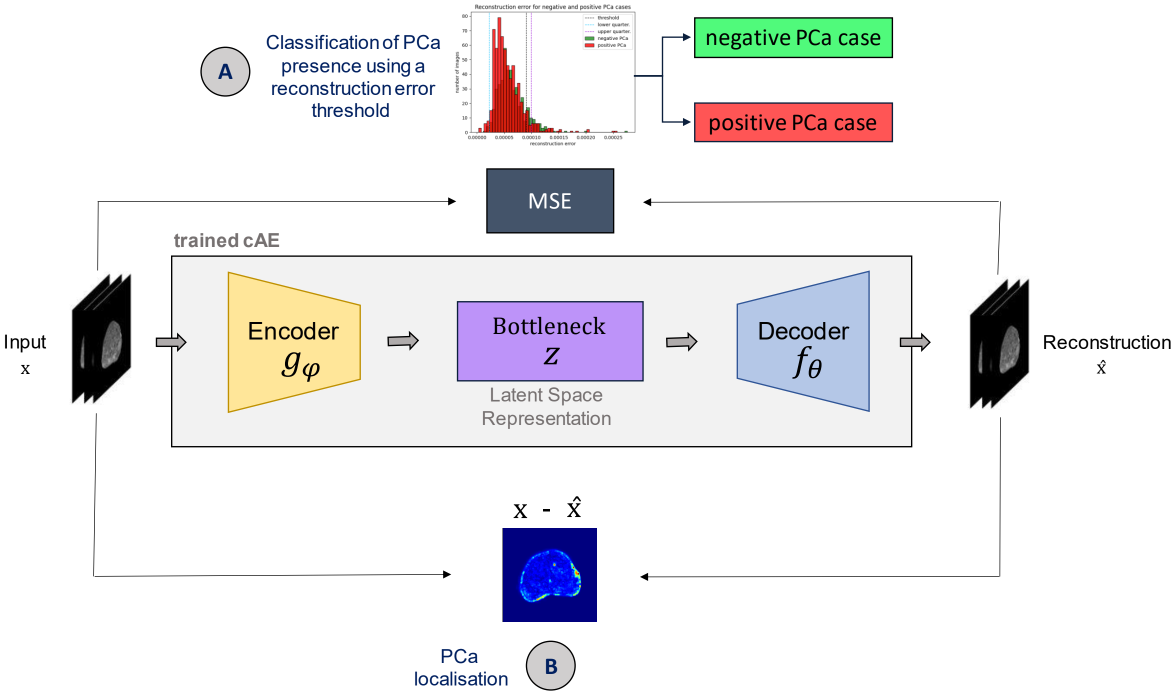


Figure 5.21: Inference phase for (A) classification and (B) anomaly detection (lesion localisation) tasks using an unknown MRI slice as input and the prior distribution of normal prostate MRI slices learned by the autoencoder model.

Once the prior distribution $p(X_h)$ has been learnt, a mix of 3D prostate MRI volumes are used as an input to the trained models.

Binary Classification of Volumes. The validation set is composed by a mix of normal and abnormal cases (denoted as $X_a = (x_{a1}, x_{a2}..., x_{an})$). This process is required in order to obtain an estimate of the mean squared error (MSE) or structural similarity index (SSIM) distribution of X_a and X_h, which will be later utilised to determine a classification threshold t_{MSE} or t_{SSIM}. Using this classification threshold, we are able to turn the problem of detecting tumour presence into a binary classification task (distinguish between negative PCa and positive PCa MRI volumes) as seen in section A of Figure 5.21. All images with associated $MSE > t_{MSE}$ or $SSIM > t_{SSIM}$ are deemed as abnormal.

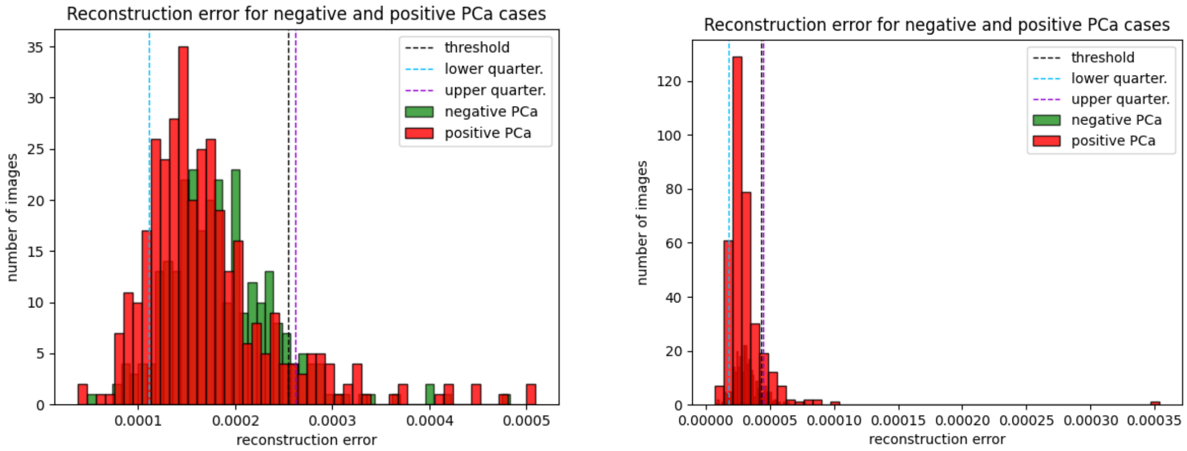
Residual Maps: PCa Localisation without Supervision. The trained cAE is capable of detecting potential anomalous regions on slice-level without being trained explicitly on this task. This is accomplished by highlighting the poorly reconstructed regions of the image. These regions are identified by the reconstruction error between the input image and it's reconstructed counterpart as seen in section B of Figure 5.21. Specifically, during inference the difference between the input image and the reconstructed one is calculated to derive an anomaly (residual) map, which is then binary divided based on a specific threshold to detect the anomalous region.

Results and Discussion. We report results on the discrimination ability of four variants of the U-Net-based cAE model in terms of AUC. The results presented in Table 5.18 refer to models trained on the *UC1-T2w-LeftBranch* dataset. We have also trained the same cAE models on the *UC1-T2w-LeftBranchRaw* where we observed similar performance. As it can be observed from Table 5.18, models trained with either output function reach similar performance. It is evident from Figure 5.22 that for the hold out test set, the reconstruction error corresponding to the two mutually exclusive classes, negative PCa and positive PCa cases, is largely similar. Ideally for our setting, reconstruction errors for those two classes should be linearly separable. Almost identical reconstruction errors do not allow our models to discriminate between negative PCa cases and positive PCa cases as seen in Table 5.18. This occurs because the models failed to extract useful representations from the input data and/or because the positive PCa cases do not significantly

output function	base channel	0.1	0.3	0.5	0.7	1	Trainable Params.
sigmoid	16	0,4591	0,4473	0,4634	0,4733	**0,4861**	6,561,793
	8	0,4975	0,4459	0,4651	0,4651	0,4758	1,641,729
Tanh	16	0,4607	0,4895	0,4648	0,4744	**0,489**	6,561,793
	8	0,3999	0,4436	0,4696	0,4722	0,4761	1,641,729

Table 5.18: Comparison between different U-Net-based models trained on prostate MRI volumes with sigmoid and Tanh output functions for various *UC1-T2w-LeftBranch* training subsets. Volumes with an associated $MSE > threshold$ are deemed as positive PCa (abnormal).

differ from their negative counterparts for this cohort. The shortcomings of the models are evident from the confusion matrices in Figure 5.21. Both models regularly fail to classify positive PCa cases (labelled as 1) correctly. Instead, the majority of those cases are classified as negative PCa (false negative). Finally, It is evident from Figure 5.24 that there is correlation between the amount of training data (the fraction of available training data) and performance for the cAE models.



(a) base channel = 8

(b) base channel = 16

Figure 5.22: Threshold calculation using MSE error distribution for U-Net-based cAE trained with the sigmoid output function.

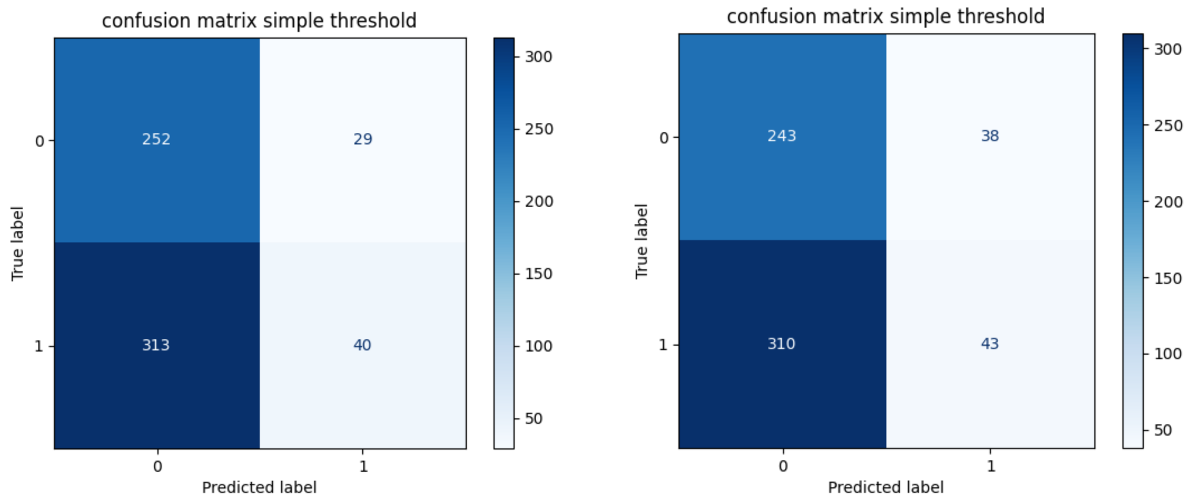
Supervised Binary Classification (UC1—tA2)

Supervised learning involves training models to classify data into distinct classes, which can extend beyond just two classes. In our context, the task at hand is the differentiation between 'normal' and 'abnormal' cases.

Common Image Classification Architectures.

Residual Networks: ResNet [31] is characterized by its residual or skip connections, which bypass one or more layers. The core idea behind these connections is to solve the vanishing gradient problem on deep learning architectures. By incorporating these connections, the network can learn identity functions that ensure the higher layers perform at least as well as the initial.

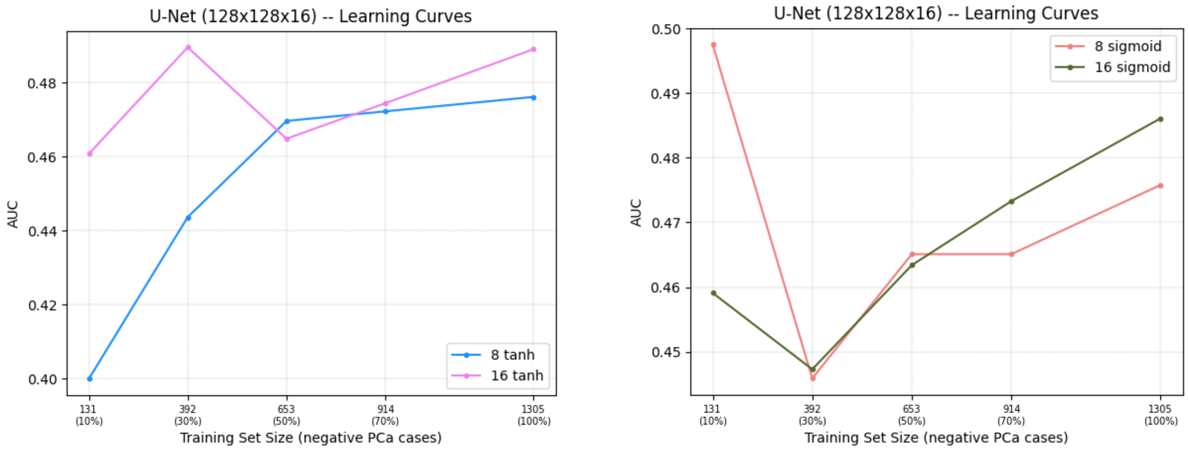
Densely Connected Convolutional Networks: DenseNet [35] stands out due to its dense connectivity pattern. In this architecture, each layer receives feature maps from all preceding layers, ensuring a compounded feature integration process. This dense feature fusion leads to improved gradient flow, more diversified features, and requires fewer parameters than a traditional CNN, enhancing the model's efficiency.



(a) base channel = 8

(b) base channel = 16

Figure 5.23: Confusion matrix for the optimal classification threshold for U-Net-based cAE trained with the sigmoid output function.



(a) tanh output function

(b) sigmoid output function

Figure 5.24: Learning curves for the anomaly detection task.

VGG (Visual Geometry Group): VGG [81] employs 3x3 convolutional layers stacked upon one another, followed by max-pooling. The model's depth is significant, with versions containing up to 19 layers (VGG19). This depth and the small filter size allow the network to learn hierarchical features from the data, capturing intricate patterns effectively.

ViT (Vision Transformer): The Vision Transformer [21] diverges from conventional convolutional approaches by leveraging the self-attention mechanism from the Transformer architecture. An image is divided into fixed-sized patches, and these patches are then linearly embedded into a sequence of vectors. These vectors are then processed through Transformer blocks, allowing the model to capture both global and contextual information. The self-attention mechanism enables the model to weigh the importance of different patches in relation to each other, granting the ability to understand long-range dependencies and complex relations across the image.

Hyperparameter Selection. Numerous experiments were conducted to find the optimal hyperparameters, shown in Table 5.19. Initially, the models quickly overfitted, leading to significant differences between training and validation results. However, the introduction of pooling, dropout, L1 & L2 regularization, and

Hyperparameter	Value
Input size	224x224x32
Activation Function	ReLU
Normalization	Batch Norm
Augmentation	Random Flip p=0.4
Loss Function	BCEwithLogits
Optimizer	Adam
Regularization	L1 & L2
Dropout	p=0.5
Pooling	AverageMax()
Batch Size	2
Epochs	200
Learning Rate	0.00001 (1e-5)

Table 5.19: Hyperparameter settings of the Classification models.

augmentation using random flip greatly improved training consistency. The Adam optimizer outperformed SGD, which did not surpass 60% accuracy. Increasing the depth of the model did not aid the classification performance. In fact, it caused more challenges. Only when the batch size was adjusted to 16-18, the models seem to perform better with added layers. The overarching observation was that simpler models yielded better outcomes. The models also displayed a keen sensitivity to normalization and learning rates, making batch normalization and a learning rate of 1e-5 ideal. Additionally, various tests were carried out concerning the volumetric dimensions of the input data. Combinations involving 128x128, 224x224 (height, width), and depths of 32, 30, 28, and 26, focusing on 16 middle slices, were examined. The specific depth values were chosen because the prostate might be absent or minimally present in the final slices. However, better results were obtained using the entire volume.

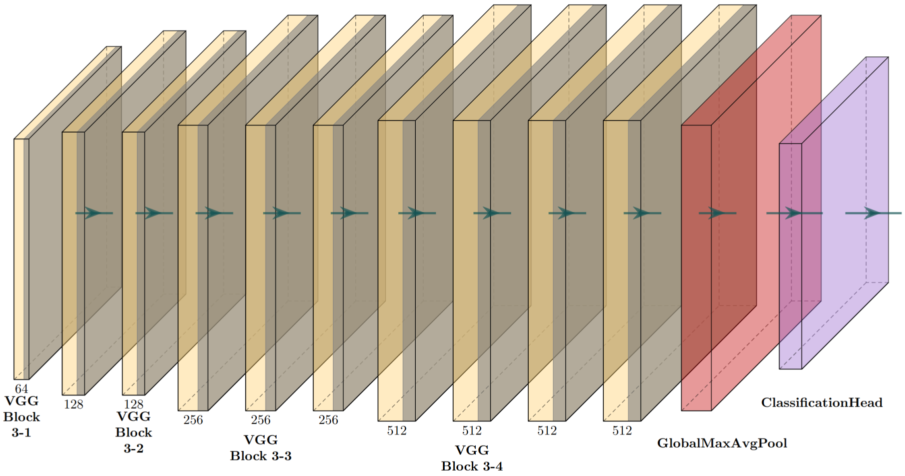


Figure 5.25: The VGG architecture comprises four stages with increasing filters: 64, 128, 256 and 512. Each stage contains multiple VGG blocks, with each block having a convolution, batch normalization, ReLu, and dropout. At the last layers, a combined Global Max and Average pooling is applied, followed by a binary classification head.

Results Among the aforementioned models, VGG (Figure 5.25) was superior in performance but notably in the stability of the training process.

UC1-T2w. As previously stated, the *UC1-T2w* dataset is comprised of both negative PCa cases (as per Form 1) and positive PCa cases (incorporating Forms 1+2). Consequently, all the related experiments in this section have been performed utilizing this dataset. In Table 5.20, we present a comparative evaluation of different classification models, specifically DenseNet, ResNet50, ViT, and VGG, trained on prostate MRI volumes. These models were analyzed with fine-tuned hyperparameters to ascertain their performance.

	DenseNet	**ResNet50**	**ViT**	**VGG**
ACC	0.7047	0.7274	0.6654	**0.7479**
AUC	0.7758	0.7920	0.7021	**0.8121**
F1 score	0.7137	0.7682	0.7089	**0.7860**
Trainable Params.	11,243,649	46,157,121	82,250,754	29,868,353

Table 5.20: Comparison between different classification models trained on prostate MRI volumes with the fine-tuned hyperparameters. The presented results are based on a hold-out test set.

	UC1-T2w	***UC1-T2w-LeftBranchRaw***
ACC	0.7479	**0.7956**
AUC	0.8121	**0.8242**
F1 score	0.7860	**0.8455**

Table 5.21: Comparison across the initial dataset and derived cases from the left branch.

Each model exhibited varying degrees of accuracy (ACC), Area Under the Curve (AUC), and F1 score, reflecting their unique capabilities in handling the dataset. Among the models analyzed, VGG demonstrated the highest accuracy of 0.7421 and a AUC of 0.8036, indicating its superior performance in distinguishing between the classes effectively. It also secured an F1 score of 0.7787, reflecting a balanced precision and recall.

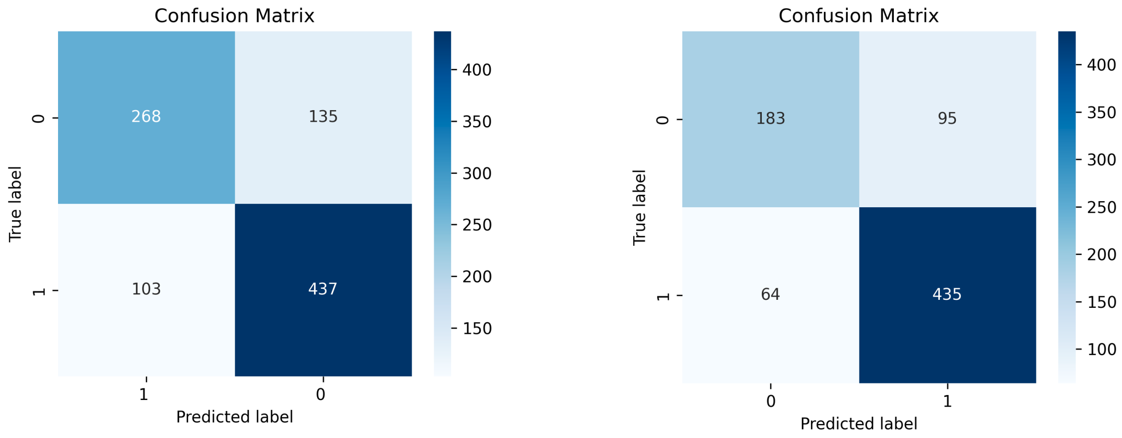
ViT, having the highest number of trainable parameters, showed relative underperformance, suggesting possible overfitting for this dataset. Conversely, ResNet50, with significantly fewer parameters, demonstrated a balanced performance. DenseNet, with the least number of parameters, also exhibited comparable effectiveness.

UC1-T2w-LeftBranchRaw. In this section, we conduct experiments with the VGG architecture using a subset of the initial dataset, specifically excluding the cases from the right branch, those instances where a positive MRI was followed by a negative biopsy. Table 5.21 illustrates a comparative analysis between experiments on *UC1-T2w* and *UC1-T2w-LeftBranchRaw* datasets. A marked enhancement is observed in the *UC1-T2w-LeftBranchRaw* dataset, evidenced by its higher ACC, AUC, and F1 score. This indicates a superior performance when excluding MRI scans initially annotated as cancerous but later determined as normal through biopsy.

Discussion A key insight derived from our unsupervised methodology is the substantial similarity between negative PCa and positive PCa cases. Additionally, Figure 5.26, depicting the confusion matrices, reveals a tendency of the model to misclassify normal 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 labelled 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. Therefore, an augmentation strategy with synthetic normal or abnormal MRI cases with generative networks [27] could potentially increase the discrimination ability of our models. Finally, the learning curves depicted in Figure 5.27 for both datasets clearly illustrate that reducing the volume of training data correlates with a decline in performance on both the training and testing sets.

5.3.2 Prostate Index Lesion Segmentation (UC1—tB)

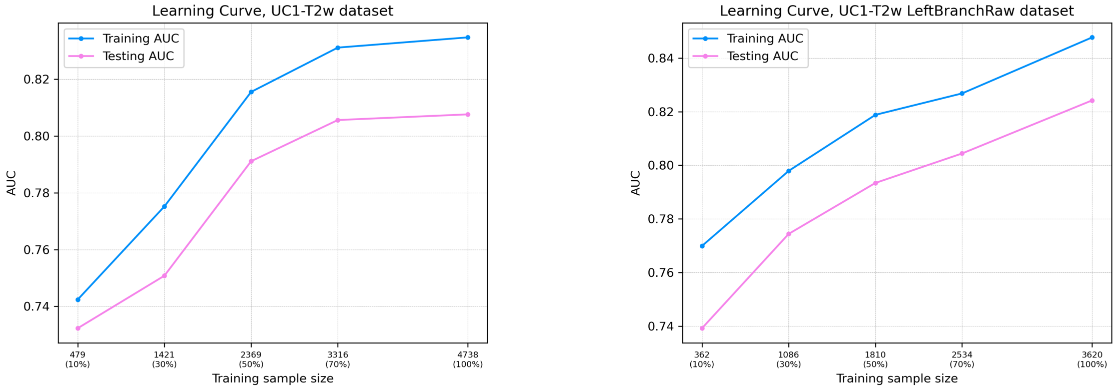
In this section, a variety of DL models have been implemented to address the task of lesion segmentation in a fully-supervised manner (FORTH contributor). More specifically, the models have been trained to to accurately identify, at a pixel level, if a particular region is likely to be associated with a prostate lesion. In addition, the binary ground truth masks (GT) exclusively included the index lesions while any other lesions were excluded.



(a) *UC1-T2w dataset* dataset

(b) *UC1-T2w-LeftBranchRaw* dataset

Figure 5.26: Confusion matrices from the optimal model trained on initial dataset and cases from left branch, respectively.



(a) *UC1-T2w* dataset

(b) *UC1-T2w-LeftBranchRaw* dataset

Figure 5.27: Learning curves from the optimal model trained on initial dataset and cases from left branch, respectively.

Data Curation

For index lesion segmentation, data originated from 12 clinical centers and 4 manufacturers were used. The initial number of cases were 440 and after the preprocessing stages a cohort of 419 cases were used for patient train and validation. More specifically in Table 5.22 the total number of cases for each clinical center and MR vendor is provided extensively. The analysis made on 3 available sequences, namely T2-Weighted (T2W), Apparent Diffusion Coefficient maps (ADC) and Diffusion-weighted imaging (DWI) which represented the input of the Deep learning models for each case.

Data Preprocessing

Initially, the cohort contained 440 patient cases. For each case, an affine registration between pairs of T2W, ADC and DWI in respect to T2W has performed. Index Lesion ground truth masks were outlined on T2W and that was the underlying reason of selecting T2W as the baseline sequence to register ADC

	Philips	Siemens	GE Medical System	Toshiba
FCHAMPALIMAUD	44	6	-	-
NCI	50	-	7	-
HACETEPPE	34	55	22	-
QUIRONSALUD	27	-	-	-
RMH	7	37	13	1
RADBOUDUMC	-	21	-	-
IDIGBI	33	-	10	-
HULAFE	1	1	28	-
IPC	-	-	24	-
JCC	12	-	-	1
GAONA	3	1	1	-
FPO	-	-	1	-
Total	211	121	106	2

Table 5.22: Number of Cases per clinical center & MR vendor.

& DWI. However in a couple of cases the registration failed and they were kept apart. Sequentially, the analysis performed on the Prostate's Whole Gland (WG). These masks obtained from a pre-trained nnU-Net [40] and dilated accordingly to extend the boundaries of the WG. The underlying cause of the dilation operation is that it has been observed that prostate lesions tend to invade in other structures outside the WG. The performance of the nnU-Net for WG segmentation was between of $88 - 94\%$ Dice Score. As a next step, a resample of $0.5mmX0.5mmX3.0mm$ for pixel spacing and slice thickness for the volumes performed while the selection criteria for that included the mean spatial spacings from the cohort. Consequently, cropping/padding strategy was chosen to meet models requirements and transform each volume to $24\ slices\ X\ 192\ pixels\ X\ 192\ pixels$. The percentage of cropping, if needed, for a specific volume can be set accordingly to the original work of Aldoj et al., [4], where they cropped the top half 25% and bottom half 25%. For our cases the margins were lower and hence no WG region was taken apart. Ultimately, the normalization strategy utilized was the minimun-maximum strategy, for model's ease of convergence, and therefore the voxels for each sequence normalized to have values at the interval of $voxel\ value\ \in\ [0, 1]$. The selection of number of floating points for the voxels included 2 configurations, 2^{16} floating points and 2^{8} integer points. The latter seemed to produce worse results than the former.

Deep Learning Segmentation Models

The following section provides a concise overview over the Deep Learning models that were employed for the Index Lesion Segmentation - Lesion Detection Use Case. In our analyses, we transformed several well-known deep learning networks to their 3D Variants to analyse cases in a voxel-based manner. **U-Net 3D**. U-Net 3D [19] is a CNN architecture designed for semantic segmentation of 3D volumetric images. Modifications were made to the original U-Net architecture to incorporate 3D data. U-Net 3D consists of two primary components (a) a contracting path and (b) an expanding path. The contracting path downsamples the input image to extract features, whereas the expansive path upsamples the extracted features and combines them with the extracted features from the contracting path to generate a mask containing segmented pixels.

Attention U-Net 3D. Attention U-Net 3D model is a modified version of the U-Net 3D architecture that incorporates attention mechanisms. Attention mechanisms assist the model to focus on specific regions of the image that are more relevant especially for lesion segmentation tasks [1]. The attention mechanism is placed at the contracting path of the U-Net model and its mathematical expression is given by Equation 5.3

$$WeightedMap = \sigma(x_{contracting} + x_{expanding}) * W_{expanding} * x_{expanding} \tag{5.3}$$

Where $x_{contracting}$ and $x_{expanding}$ represent the features from the contracting and expanding path respectively, $W_{expanding}$ represents the weight matrix, and σ is the sigmoid function.

VNet. VNet model [59] consists of an encoding and decoding pathway, which is identical to U-Net. However it incorporates residual connections as an additional component. The V-Net model utilizes a sequence of 3D convolutions, batch normalization, and non-linear activation functions, such as ReLU, GeLU etc., in both the encoder and decoder components. The usage of residual connections in V-Net is seen as a significant advancement, since it effectively addresses the issue of vanished gradients and facilitates the training of more complex networks. The residual connections are given by the expression Equation 5.4

$$X_{residual} = X_{input} + X_{input} * Tr_{Conv->BN->ReLU} \tag{5.4}$$

where $X_{residual}$ is the outcome of the layer, X_{input} are the input features and $Tr_{Conv->BN->ReLU}$ are the sequence of transformations, in the case of VNet are convolutional operations, Batch Normalization and ReLU non linear activation units.

USE-Net 3D. USE-Net 3D [77] is also and encoder-decoder architecture with squeeze & Excitation (SE) attention mechanisms [34] placed at the end of each encoder layers. SE improves the representational power of the initial U-Net 3D model by enabling it to perform dynamic channel-wise feature recalibration. The SE block aims to adaptively adjust the importance of each channel in the feature maps, thereby allowing the network to focus more on the most informative features. The squeeze operation is given by Equation 5.5

$$X_{squeeze} = \frac{1}{H \times W \times D} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{d=1}^{D} x_{hwd}^{channel} \tag{5.5}$$

Where H, W, D are the spatial dimensions of the volume representing the depth, weight and height while *channel* are the number of channels. The excitation operation is given by Equation 5.6

$$X_{exc} = \sigma(W_2 ReLU(W_1 X_{squeeze})) \tag{5.6}$$

where $X_{squeeze}$ is the outcome of Equation 5.5, W_2, W_1 are the weights of fully connected neural networks, σ is the sigmoid function and $ReLUistheRectifiedLinearUnit$. Finally the outcome of the SE layer is given by

$$SE = X_{exc} * x_{hwd}^{channel} \tag{5.7}$$

where X_{exc} is the outcome of Equation 5.6 and $x_{hwd}^{channel}$ is the input of the SE layer.

Dense2U-Net 3D. Dense2U-Net 3D [4] is a variant of 3D U-Net architecture. However instead of using typical CNN layers, it utilizes Densely Connected CNNs. More specifically, in typical CNNs each layer receives input only from its immediate predecessor while in Densely connected CNNs, each layer receives input from all preceding layers. This layer architecture facilitates more efficient gradient flow during backpropagation, mitigates the vanishing gradient problem, and encourages feature reuse, thereby making it possible to train deeper networks with fewer parameters. The expression that define Densely connected CNNs is given by Equation 5.8

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{5.8}$$

where $[x_0, x_1, \ldots, x_{l-1}]$ are the concatenations of each layer $l[0, length(l)]$ sequentially.

TransU-Net 3D. TransU-Net 3D [14] incorporates the Vision Transformer (ViTs) [21] architecture as a bottleneck for intensive feature extraction at the very last step of the contracting path. The idea is to analyse feature maps as sequences of patches and apply self-attention mechanisms to capture both local and global dependencies. More importantly, Unlike CNNs, which capture local features through small receptive fields, the self-attention mechanism in ViTs allows for capturing global context from the entire image. The equation that describes self-attention mechanism which is the main component of ViTs are given by Equation 5.9

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5.9}$$

Initially, the feature maps are divided into N overlapping patches where N is a hyperparameter. Eq.5.10 gives the expression for the patchification of the volumetric feature maps.

$$\begin{aligned} patch_{WHD} &= Convolution(X_{input}) \\ features = embedding\ dimension,&\ kernel\ size = strides \end{aligned} \tag{5.10}$$

Where *embedding dimension* are the number features in the embedding space, *kernel size* and *strides* are set accordingly based on the number of patches for each axis for each volumetric patch $patch_{WHD}$ with dimensions W, H, D for weight height and depth, respectively. After that operation each volumetric patch is flattened in a 1-dimensional sequence and positional encodings assigned to each sequence to provide the model with information about the position of each element in the sequence. these are given by Equation 5.11

$$embs = \sum_{posenc=1}^{length(patches)} \text{Flatten}(patch_{WHD}^{posenc}) + posenc \tag{5.11}$$

where *patches* are the total number of patches and *posenc* is a value related to the position of each patch.

$$\begin{aligned} trans_l &= \text{LayerNorm}(x + \text{MultiHeadAttention}(embs, embs, embs)) \\ trans &= \text{LayerNorm}(trans_l + \text{FeedForward}(trans_l)) \end{aligned} \tag{5.12}$$

where MultiHeadAttention($embs, embs, embs$) is the multiple self-attention obtained from Equation 5.3, $trans_l$ is the outcome of each transformer layers, FeedForward is the fully connected neural network and LayerNorm is the layer normalization component.

nnU-Net. The nnU-Net [40] is considered a state of the art in medical image segmentation tasks while its novelty comes to the automatic hyperparameter configuration to provide strong performance across a wide range of 2D and 3D segmentation. nnU-Net contains 3 architectures namely, U-Net 3D, Cascaded U-net and U-Net 2D while for the experiment the 3D full resolution U-net was chosen.

Hyperparameters Selection

The training hyperparameters utilized for each individual model training are presented in Table 5.23. Initially, the models were trained using the Sigmoid Focal Crossentropy loss function [49]. However, the obtained results for the same architectures were significantly lower. Consequently, a combination of Dice and Binary Crossentropy loss functions employed. The same approach is implemented within the plans configured by nnUnet. Moreover, the utilization of a cyclical learning rate, characterized by a comparatively low minimum learning rate, is favored to facilitate the optimization of models towards attaining optimal local minima. Simultaneously, a high maximum learning rate is employed to aid models in escaping local minima and exploring the global ones. The cyclical strategy assists the models to diverge the learning rates values from maximun to minimum learning rate periodically.

Hyperparameters / Models	Loss Function	Activation Function	Batch Size	Epochs	Training Strategy	Optimizer	Callbacks
U-Net 3D	Weighted Combination of DC & BCE (a:0.7, b:0.3)	Sigmoid	4	250	Cyclical Learning Rate Max Learning Rate: 0.01, Min Learning Rate: 0.0001	Adam	Model Checkpoint
Attention U-Net 3D	Weighted Combination of DC & BCE (a:0.7, b:0.3)						
VNet	Weighted Combination of DC & BCE (a:0.7, b:0.3)						
USE-Net 3D	Weighted Combination of DC & BCE (a:0.7, b:0.3)						
Dense2U-Net 3D	Weighted Combination of DC & BCE (a:0.7, b:0.3)						
TransU-Net 3D	Weighted Combination of DC & BCE (a:0.7, b:0.3)						
nnU-Net	Combination of Dice Loss & Binary Crossentropy		18	1000	Polynomial LR Init LR: 0.01, Momentum: 0.99, Nesterov: True, Decay: 0.00003	SGD	

Table 5.23: Model Training Hyperparameters.

DL Models Configuration

In this analysis various model configurations were used to find the optimal ones. For instance, different kernel sizes for each convolutional operation and pool sizes for each layer tested with the best proven ones to be those of Table 5.24. Throughout the experiments, the worst configurations proved to be those with isotropic kernel size of 3 across each encoding and decoding layer and isotropic pool sizes of 2, respectively. This may attributed to low spatial resolution of the axial plane. On the contrary, identity pool sizes of 1 for the axial plane for the 2 first layers of the encoding path proved to be the most effective ones. The same applies for the kernel sizes where the a kernel size of 1 for the axial plane at the 2 first layers was the best, as indicated by the automatic configuration of nnU-Net.

	U-Net 3D	Attention U-Net 3D	VNet	USE-Net 3D	Dense2U-Net 3D	TransU-Net 3D	nnU-Net
Kernel Sizes	(1,3,3), (1,3,3), (3,3,3), (3,3,3), (3,3,3)						
Pool Sizes	(1,2,2), (1,2,2), (2,2,2), (2,2,2)						(1,1,1), (1,2,2), (2,2,2), (2,2,2)
Filters	(32,64,128,256,512)						(32,64,128,256,320)
Patch Size	-	-	-	-	-	(1,3,3)	-
Transformer Layers - Heads	-	-	-	-	-	8 -8	-
Dense Size	-	-	-	-	8	-	-
Growth Rate	-	-	-	-	4	-	-
Dropout	0.3						-

Table 5.24: Model Architecture Hyperparameters.

Results

The following section showcases the segmentation performance of DL models to identify primary lesions located within the WG. The experiments performed included 350 cases for train-validation & 69 cases as a hold-out set. The models' and training hyperparameters were tuned in the train-validation experiment while the hold-out set kept apart for evaluation purposes. Table 5.25 highlights the performance of each model in the hold-out set. In terms of purely segmentation metrics like volumetric dice score (VDS) and Hausdorff distance (HD) nnU-Net seems to outperform all the compared models with a large margin. For instance, nnU-Net achieves mean VDS 36.1 % while the second best performing (TransU-Net 3D) reaches, at most, 31.1 %. Same applies for the HD where nnU-net is significantly better than the second one, which in this case is Attention U-net 3D. However, regarding Recall, TransU-net 3D is the most dominant one, by a significant difference of 14% from nnU-Net while it retains a relatively comparable precision with the other models. On the other hand, nnU-net reaches a precision of almost 50% outperforming VNet, which is the second best by 5.5%. Sensitivity analysis also included as a part of this experiment to investigate the optimal thresholds to flatten the sigmoid outcomes from each model and produce the final binary masks. The most notable part is that for all the models, the optimal threshold resulted to be that of 0.05 where this threshold may take continuous values from $SigmoidOutcome \in [0,\ 1\ ]$.

	Volumetric Dice Score (%)	HD (mm)	Precision (%)	Recall (%)	Optimal Threshold (Outcomes Flatten)
U-Net 3D	27.6	22.8	31	36	0.05
Attention U-Net 3D	27.1	20.1	35.3	27.9	
VNet	29.1	21.5	43.8	30.1	
USE-Net 3D	28	23.2	41.5	28.2	
Dense2U-Net 3D	19.5	33.2	20.3	39.5	
TransU-Net 3D	31.1	21	36.1	46.7	
nnU-Net	36.1	14.1	49.3	32.1	

Table 5.25: Mean Results for each model and metric.

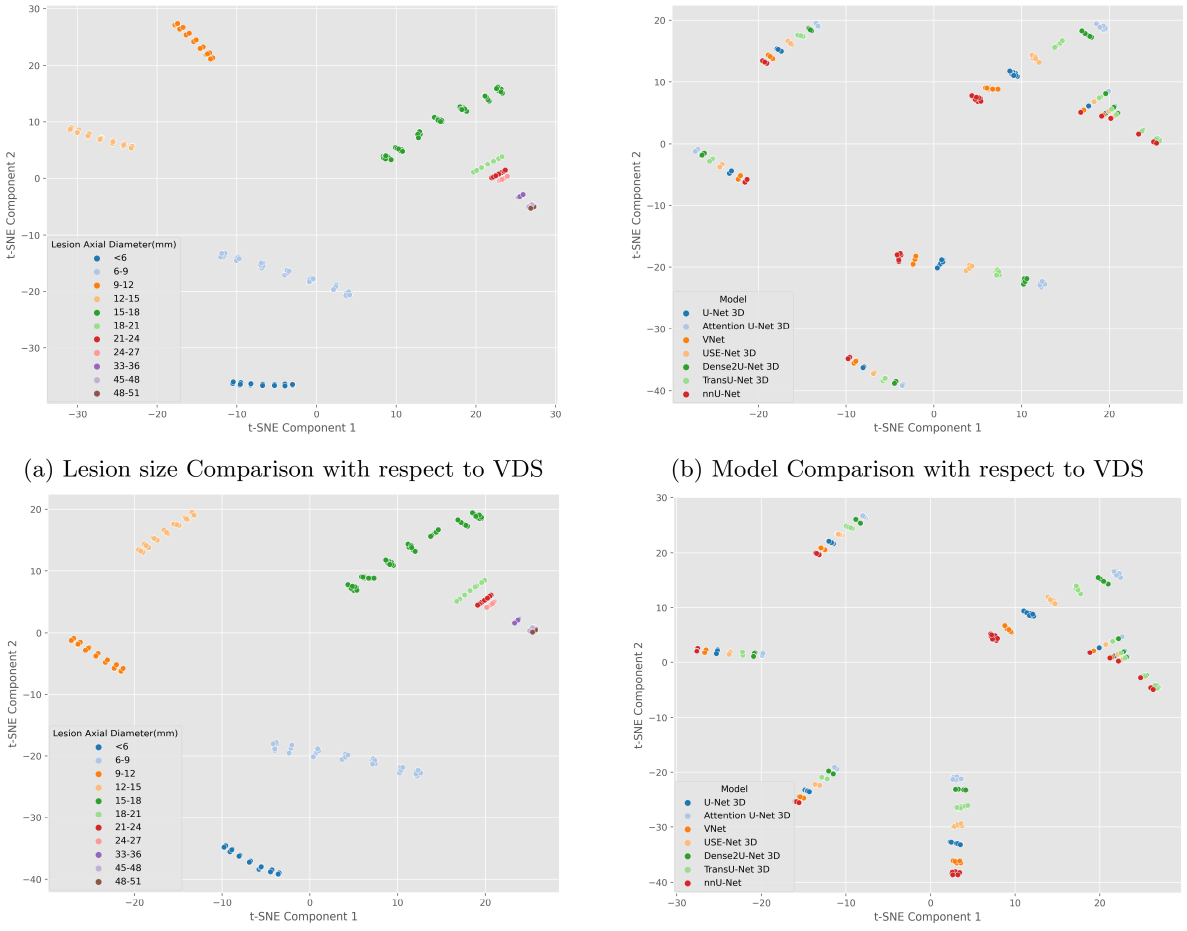
Vendor Specific Results In Table 5.26, the mean VDS is presented with respect to each MR Vendor for the hold-out set. It is clearly evident that the majority of the models indicate a significant VDS variability across vendors. On the contrary, TransU-Net 3D proved to be less prone to variations across vendors with a standard deviation(std) of 3.91% while the second less prone to vendor model, namely Atention U-Net 3D, achieves an std of 6.19%. The best performing model, nnU-Net, has an std of 8.55% . Noteworthy, the compared models have their performance significantly degraded on Siemens Vendor.

Correlation with Lesion Axial Diameter In this analysis t-SNE utilized for reducing the dimensionality of a subset of features for projection in a lower dimensional space. It allows us to visualize these features in a scatter plot by mapping it into two dimensions as components 1 and 2. The algorithm behind t-SNE calculates the probability of relationships between pairs of data points, in the vector space. Essentially t-SNE aims to maintain the relationships of the features when projecting them onto a vector space. More specifically, VDS & Recall correlated with DL models and Lesion Axial Diameter in mm to identify possible clusters. Figure 5.28 present the t-SNE components originated from VDS distribution with respect to (a) lesion sizes and (b) DL models and those originated from Recall with respect to (c) lesion sizes and (d) DL models. Regarding VDS, Figure 5.28a and Figure 5.28b highlight that Lesion axial Diameters > 15 mm are

	Philips	Siemens	GE Medical System
U-Net 3D	28.5%	21.2%	39.5%
Attention U-Net 3D	31%	20.7%	31.8%
VNet	31.2%	18.3%	32.2%
USE-Net 3D	29.3%	19.9%	41.6%
Dense2U-Net 3D	26%	8.8%	23.8%
TransU-Net 3D	34.7%	26.9%	31.4%
nnU-Net	36%	30.9%	47.6%

Table 5.26: Mean VDS results for each MR Vendor.

close to each other for each model's performance while lesions with axial diameter < 15 mm are much more distant between each other & between lesion with axial diameter > 15 mm. Especially for lesions with a diameter > 18 mm models' outcomes seem to be less variant with each other and with different diameter categories indicating similar performance.



(a) Lesion size Comparison with respect to VDS

(b) Model Comparison with respect to VDS

(c) Lesion size Comparison with respect to Recall

(d) Model Comparison with respect to Recall

Figure 5.28: t-SNE Analysis with respect to VDS & Recall, Axial Lesion Size in mm and DL model.

Discussion

Figure 5.29 shows the overlap of GT binary masks and predicted binary masks for 4 cases and the compared models. The most notable component is that of case 3, wherein the dimensions of the lesions exhibit a significant degree of overlap with the WG. However, it should be noted that out of all the models evaluated, only two models, namely TransU-Net 3D and nnU-net, were able to provide predictions. Notably, the TransU-Net 3D model demonstrated a high level of efficiency by accurately predicting the entire lesion area. Prostate lesion segmentation pose a particular difficulty due to variations in sizes and textures. Although Dense2U-Net 3D demonstrated strong prediction capabilities for Case 1 and 4, it exhibited a complete inability to generate a binary mask for Case 2 and 3. These findings further indicate the presence of diversity in the results obtained by the VDS across different networks.

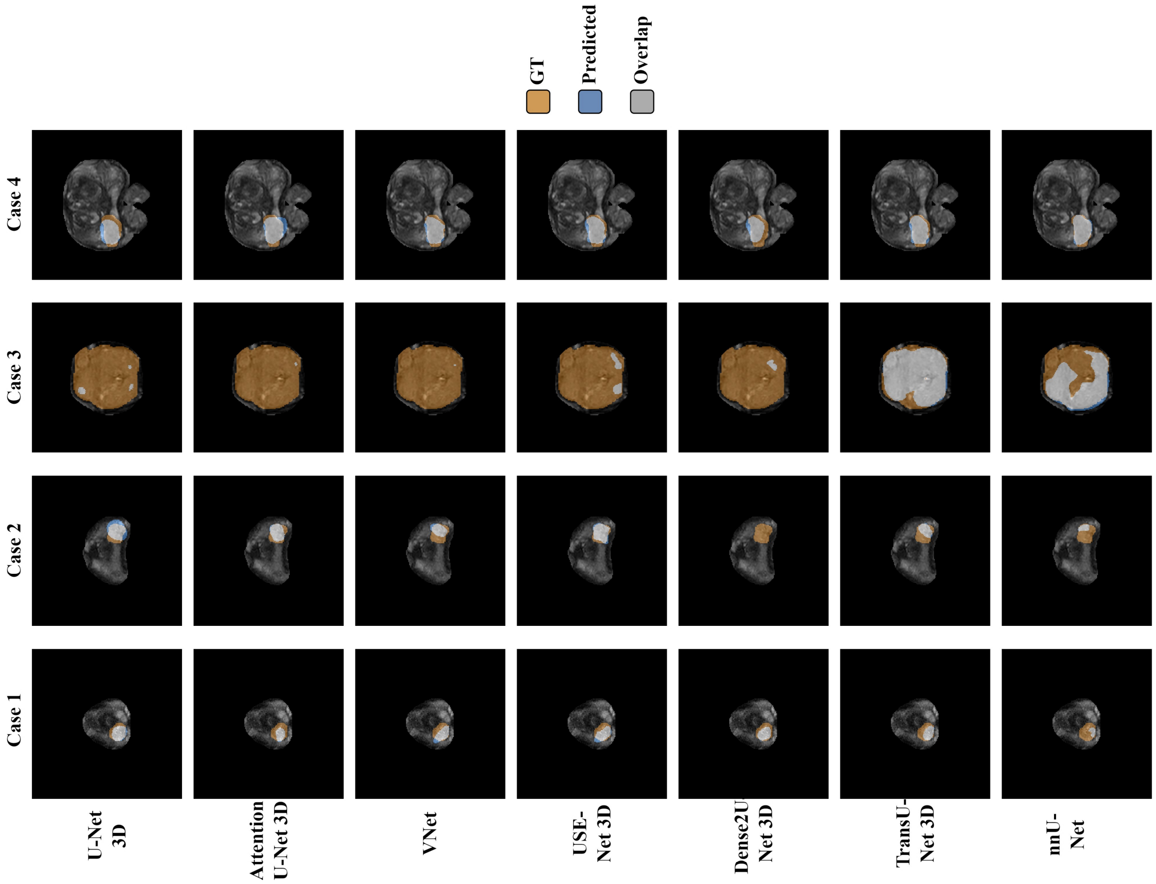


Figure 5.29: Overlap between GT and predicted binary masks for 4 cases.

5.4 On the performance of 2D & 3D models on index lesion segmentation with a curated dataset

5.4.1 Introduction

In this section, we detail the evaluation of the nnUNet framework for segmenting index lesions using a specific subset of the ProstateNet data. Only the T2-weighted axial images along with the lesion segmentations were used for this study.

5.4.2 Methods

Data Curation. The primary dataset consisted of 440 cases from 12 clinical centers and 4 manufacturers (as of 01/08/2023). This number was narrowed down to 419 after excluding cases where the T2-weighted sequences were fat-suppressed or when an endorectal coil was employed. A thorough visual review of the segmentations was undertaken to determine their appropriateness for deep learning modeling. During this phase, we identified several categories of concern. While some were addressed and included in the dataset, cases related to unresolved issues were omitted (refer to Table 5.27 for details on these issues). The final dataset comprised 301 cases. Of these, 241 were designated for training and 60 for testing, following an 80/20 training/testing split. Table 5.28 displays the final distribution based on manufacturer and provider.

Issue	Count
1 or more slices missing from the annotation	9
Challenging annotation	59
2 lesions with the same annotation label	10
Mixup of Whole Gland and Lesion annotations	21
Conversion to nifti issues	21

Table 5.27: List of unresolved issues and their respective counts. Note that a case may have more than 1 issue.

Manufacturer Provided by	GE Medical System	Philips	Siemens	Toshiba
FCHAMPALIMAUD	0	26	1	0
FPO	3	0	0	0
HACETTEPE	14	24	42	0
HULAFE	18	0	2	0
IDIBGI	9	25	0	0
IPC	17	0	0	0
JCC	0	0	10	1
NCI	1	26	0	0
QUIRONSALUD	0	11	0	0
RADBOUDUMC	0	0	18	0
RMH	12	7	33	1
Total	74	119	106	2

Table 5.28: Distribution of cases based on the manufacturer and provider.

Data Preprocessing. The images were center-cropped, as needed, using a fixed-size box with dimensions (115.0, 110.0, 130.0) mm for the x (L-R), y (P-A), and z (I-S) axes, respectively. The dimensions were selected to strike a balance between including all potential lesions and minimizing the input space's dimensionality. Segmentations were enhanced by retaining the largest connected component and filling any gaps. This step helped eliminate unrelated single pixels distant from the lesion and any lone-pixel gaps within segmentations observed during the review. Subsequently, the data were processed with nnUNet to define the normalization function, target spacing, and training/network parameters. The chosen normalization was z-score, and the target spacings were (0.462,0.462) for in-plane dimensions and 2.999 for slice-thickness (used only in 3D configuration).

Network Configurations. Both the 2D and 3D configurations of UNet supported by nnUNet were examined. For the 2D setup, the patch size was set to (256, 256), while for the 3D setup, it was set to (28, 256, 256). The design of the 3D network adheres to the classic encoder-decoder paradigm seen in U-Net architectures. The encoder is composed of seven stages, with channels increasing from 32 to 320. Each stage is made up of stacked blocks containing two 3D convolutional layers. The convolution kernel size starts at (1,3,3) and shifts to (3,3,3) as the channel count grows. The decoder reflects the encoder's structure but also incorporates skip connections. It begins by expanding the channel count and then gradually decreasing it in the later stages. Throughout the architecture, 3D data undergoes processing via 3D convolutions. Both instance normalization and LeakyReLU activations are uniformly used across the network. The 2D variant of the network is quite similar in design, but it can extend to as many as 512 channels or features.

Training. We followed the default nnUNet training procedure, which involves 1000 epochs, with each epoch defined as 250 iterations using a specified batch size of 49 for 2D and 2 for 3D configurations. The optimizer chosen was the stochastic gradient descent with an initial learning rate of 0.01, a nesterov momentum of 0.99, and a linear polynomial LR schedule. The loss function was a balanced combination of the Dice coefficient and cross-entropy with deep supervision. nnUNet's default data augmentation, which includes rotations and noise, was applied. For both 2D and 3D configurations, an ensemble of five models was trained in a 5-fold cross-validation manner.

Inference. During the evaluation phase, the T2-weighted sequence of each test case was fed into each model in the ensembles, encompassing both 2D and 3D configurations, via the nnUNet API. nnUNet derives the final prediction by averaging a collection of predictions for each model. This collection comprises predictions from various patches (if needed) and reflections across every dimension. Consequently, for each ensemble, there were five probability maps, each representing predictions from models trained on distinct folds. These maps were then averaged over the five folds. Segmentations were subsequently derived by applying two separate thresholds, 0.05 and 0.5, leading to a unique segmentation for each ensemble at both thresholds. No further post-processing steps were undertaken.

Analysis We used counting metrics like Dice, Precision, and Recall to evaluate the models' semantic segmentation and object detection capabilities. Additionally, the 95th and 100th percentiles of the Hausdorff Distance (HD95 and HD, respectively) were computed for a more detailed performance analysis.

5.4.3 Results

The 5-fold 2D and 3D UNet ensembles were trained using 241 cases and tested on the remaining 60. nnUNet generates both a best and a final checkpoint for each model. The results presented here pertain to the best checkpoint, as it showed marginally superior performance. Table 5.30 presents the average metric values across the two ensemble configurations, evaluated at two distinct thresholds. The inclusion of these thresholds in our analysis stems from the observation that model-generated probabilities often gravitate towards extreme values. As a result, a lower threshold effectively shifts the ensemble's interpretation strategy: rather than averaging the predictions of its constituent models, the ensemble determines if any single model positively predicts a voxel.

Model	Threshold	Dice (%)	Recall (%)	Precision (%)	HD (mm)	HD95 (mm)
2d	0.05	83.0	86.1	-	-	-
	0.5	80.4	75.1	-	-	-
3d	0.05	91.9	92.9	91.6	2.180	0.775
	0.5	89.3	85.0	-	-	-

Table 5.29: Macro-average metrics of the ensemble of models on training dataset. Empty metrics are due to blank predictions

Model	Threshold	Dice (%)	Recall (%)
2d	0.05	35.9	33.0
	0.5	27.8	22.3
3d	0.05	44.2	42.3
	0.5	40.1	35.6

Table 5.30: Macro-average metrics of the ensemble of models on the full testing dataset (N=60).

Upon examining the Dice coefficients from the ensemble results presented in Table 5.30, it becomes evident that the 3D ensemble consistently surpasses its 2D counterpart. More specifically, for the 0.05 threshold, the 3D ensemble achieves a Dice score of 44.2% as opposed to the 2D ensemble's 35.9%. Similarly, at the 0.5 threshold, the 3D ensemble attains a Dice score of 40.1% compared to the 2D ensemble's 27.8%. Evidently, the performance is amplified when adopting the lower 0.05 threshold in both ensemble configurations.

Regarding Recall, the pattern is analogous to the Dice coefficient pattern. However, it's notable that Recall scores are consistently lower than their corresponding Dice scores. The disparity is more accentuated at the 0.5 threshold, where the differences amount to 5.5% for the 2D ensemble (27.8% Dice vs. 22.3% Recall) and 4.5% for the 3D ensemble (40.1% Dice vs. 35.6% Recall). When the threshold is set at 0.05, these discrepancies shrink to 2.9% for the 2D ensemble (35.9% Dice vs. 33.0% Recall) and 1.9% for the 3D ensemble (44.2% Dice vs. 42.3% Recall). Such observations underscore that the enhancements in Dice scores predominantly stem from the models' heightened sensitivity.

Certain metrics, including Precision, HD, and HD95, were not computable across the entire testing dataset due to instances in every configuration where predictions were void or blank. Specifically, as detailed in Table 5.31, when considering the 0.05 threshold, 10 and 10 subjects from the 2D and 3D ensembles respectively had such blank predictions. For the 0.5 threshold, these counts increased to 17 and 13 for the 2D and 3D ensembles, respectively. In these scenarios, Precision becomes undefined, and the Hausdorff Distance (HD) would be infinitely large, skewing any averages.

To provide a clearer picture of model performance, Table 5.31 presents metrics computed only on cases with non-blank predictions. As anticipated, Dice and Recall scores in this table surpass those from the full dataset, given the exclusion of blank predictions. Notably, the 3D ensemble with a threshold of 0.05 emerges as the top performer, recording a Dice score of 53.0% and a Recall of 50.8%.

In all configurations, Precision consistently exceeds Recall. The most significant gap appears in the 2D ensemble with a threshold of 0.5, where Precision is 73.9% against a Recall of 31.1%, leading to a difference of 42.8%.

Assessing the Hausdorff Distances, the 3D ensemble consistently reports lower HD and HD95 values than the 2D ensemble across both thresholds. Specifically, for the 3D ensemble, the HD reduces from 15.158 mm at a 0.5 threshold to 13.375 mm at a 0.05 threshold, while the HD95 diminishes from 12.020 mm to 10.395 mm. Conversely, in the 2D ensemble, both HD and HD95 increase when the threshold is lowered: HD values change from 16.562 mm (0.5 threshold) to 18.022 mm (0.05 threshold), and HD95 values shift from 13.356 mm to 14.241 mm.

Model	Threshold	N Cases	Dice (%)	Recall (%)	Precision (%)	HD (mm)	HD95 (mm)
2d	0.05	50	43.1	39.6	65.1	18.022	14.241
	0.5	43	38.7	31.1	73.9	16.562	13.356
3d	0.05	50	53.0	50.8	70.1	13.375	10.395
	0.5	47	51.1	45.4	74.9	15.158	12.020

Table 5.31: Macro-average metrics of the ensemble of models on testing dataset, excluding blank predictions.

5.4.4 Discussion

In our investigation, we trained and tested 2D and 3D ensembles on a meticulously selected subset of the ProstateNet dataset's index lesion annotations. The data revealed a consistent trend: the 3D ensemble surpassed the 2D ensemble across all evaluated metrics, underlining the potential advantage of 3D modeling in this context.

A notable observation was the disparity between Recall and Precision scores. The lower Recall, juxtaposed with notably higher Precision, suggests that the primary challenge for the models lies in accurately detecting the presence of lesions. The significant count of blank predictions further bolsters this assertion.

Considering these findings, there may be merit in exploring ensembles composed of diverse models. These models could leverage varied strategies, such as incorporating different input modalities, and prioritize precision. The saturated probabilities produced by the models suggest two potential ensemble strategies. When adopting a threshold of 0.5, the ensemble tends towards a "majority with equal vote" strategy, seeking a collective consensus among the models. On the other hand, a threshold of 0.05 leans towards a strategy where even a few models' positive predictions can influence the overall ensemble decision. Given the observed improvement in model performance with reduced thresholds, further exploration of these combination strategies might prove beneficial.

5.5 On the impact of mpMRI sequence combination to automatically detect prostate cancer

5.5.1 Introduction

AI-based systems have been proposed to support radiologists to overcome important issues of prostate cancer (PCa) diagnosis pathway, by automatically detecting and characterizing PCa on magnetic resonance imaging

(MRI). These systems have contributed to the improvement of the detection rate while reducing reading time and inter-reader variability. Thanks to the increased availability of medical image datasets, deep learning (DL) algorithms have been preferred over traditional machine learning (ML) techniques for their ability to learn directly from data/images, without the need for extracting predefined parameters from suspected areas. However, one advantage of ML techniques is that they can easily combine different MRI sequences into a single vector that can be fed to the classifiers, mimicking the radiologist's behavior when reporting prostate multiparametric MRI (mpMRI). Indeed, among mpMRI sequences, T2-weighted (T2w) images, diffusion-weighted (DWI) sequence and the associated apparent diffusion coefficient (ADC) map are equally important to detect PCa in both peripheral (PZ) and transition (TZ) zones. Therefore, assessing the impact of different ways to combine different MRI sequences is very important to improve performances in detection and segmentation of PCa. We evaluated three different ways of combining mpMRI using two different DL architectures.

5.5.2 Methods

Networks architectures

We developed two different structures based on the U-Net architecture:

1. Unet-Resnet50 encoder (UR_50): in which the encoder of the U-Net has been replaced with a a ResNet-50, characterized by 50 layers (48 Convolution layers along with 1 MaxPool and 1 Average Pool layer). ResNet differs from the other CNNs for the presence of the skip connections that link the original input to the output of each convolutional block.

2. Unet-standard encoder with Attention Gate (UR_ATL): attention gates are incorporated into the standard U-Net architecture to highlight salient features that are passed through the skip connections, in order to highlight only the relevant activations during training. This reduces the computational resources wasted on irrelevant activations, providing the network with better generalization power.

Development and validation of the segmentation models

We compared three different strategies to combine complementary information coming from the different mpMRI sequences, i.e., T2w, ADC, and high b-value DWI (hbDWI). The following configurations were trained and validated:

- 3-channels input (model 1): T2w, ADC map, and hbDWI were concatenated and given as a single input of the UR_50 net.

- Multi-output (model 2): each sequence (T2w, ADC, and hbDWI) was used as input of a single channel UR_50 and the output image was created by averaging the results of the single networks.

- Multi encoders-single decoder with Attention Gate (model 3): we used the UR_ATL with three branches (encoders), each having a different combination of the three MR sequences (i.e., T2w-ADC-T2W/T2W-ADC-hbDWI/T2-hbDWI-ADC etc).

All models have been trained with the Adam optimizer and a learning rate of 0.001, $\beta1$ of 0.9, and $\beta2$ of 0.999. Data Augmentation was used during the training, to reduce the overfitting of the model. Among the different data editing techniques, we decided to use *Flipping* in which vertical or horizontal rotations are applied, and *Random Rotation*, in which rotations are applied randomly by 30°. The loss function was a combination of two metrics: Dice similarity coefficient and the Binary Focal loss, as follow:
$LossFunction = DiceLoss + BinaryFocalLoss$

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, each sequence was cropped around the automatically segmented prostate area using a bounding box of 224x224 pixels to ease the network training

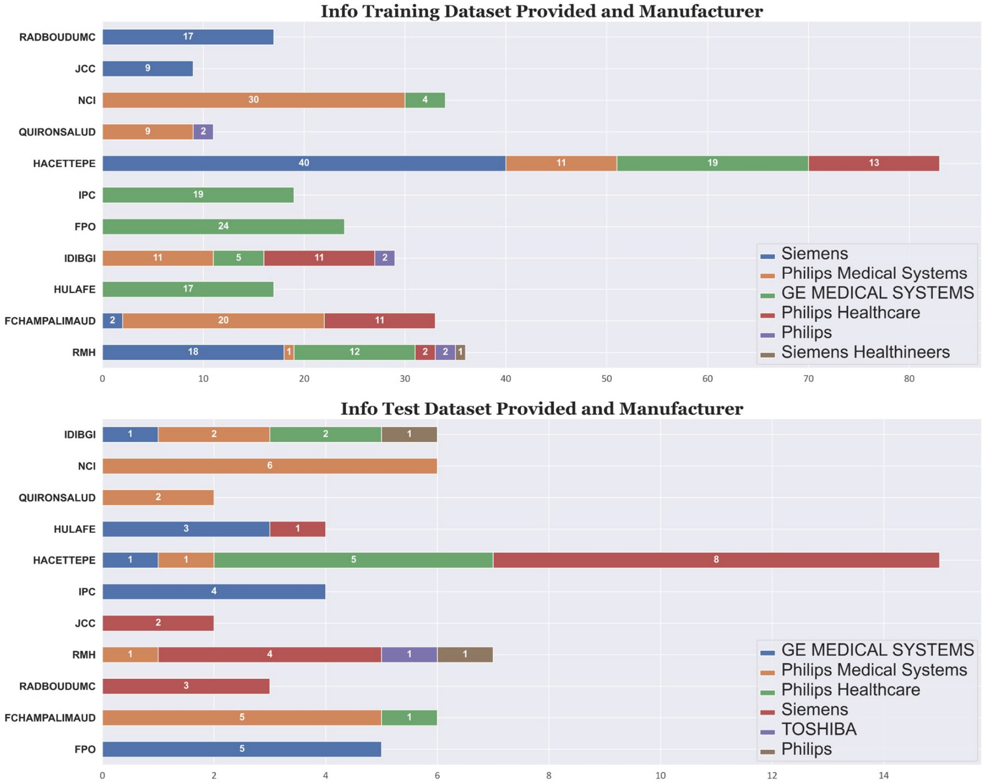


Figure 5.30: Distribution of centers and scanners among training and test set

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 since CNNs are mostly designed for natural images which usually have three channels. Since each model requires different input configurations, RGB images were obtained differently according to the model used: for models 1 and 4, each RGB channel is represented by a different sequence (T2w, ADC, and hbDWI), while for models 2 and 3 each channel was fed with a single sequence converted to RGB image, i.e., T2W-Grayscale to RGB, ADC-Grayscale to RGB, and hbDWI-Grayscale to RGB.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.
Each model was trained by transfer learning using weights obtained during a pre-training performed on other cases from one partner (FPO) with endorectal coil.

5.5.3 Results

Patients

For this task, we used all patients having a biopsy confirmed PCa and a manual segmentation of at least one tumor. A total of 371 tumors were collected, 312 used to train and tune the models, and 59 to test the networks on a subset of unseen tumors. To avoid bias, the random selection of tumors was made based on patients. Fig. 5.30 shows the distribution of centers and scanners among training and test set.

Results on training and validation set

Table 5.32 and table 5.33 show results obtained by the 6 networks (3 models with two different initialization methods) on both the training and the validation set. Detection rate (DR)-patient refers to the number of patient in which a lesion was detected (regardless of whether it matches exactly the position of the manually segmented tumor) over the total number of positive patients. DR-lesion refers to the total number of detected lesions over the total number of lesions. In this case, a PCa was considered detected if the Dice Similarity Coefficient (DSC) between manual and automatic masks was >0.10. Model 2 reached the highest DR per patient, however DR per lesion was extremely low. Moreover, these models produced a higher number of false positive (FP) either in terms of volumes and as number of voxels. Model 1 and model 3 obtained good performances in terms of detection rate and showed a good compromise between detection rate and number of FPs, in particular considering volumes of FPs.

Figure 5.31 shows the violin plots of Dice Similarity Score (DSC) considering the whole automatic mask (ALL) and only detected lesions, i.e., DSC >0.10 (TP). From the graphics, we can infer that model 1 and 3 are those reaching the highest accuracy in segmenting PCas, reaching a median DSC of 0.53 and 0.57, respectively, when considering TP lesions. From the distributions of the violin plot, we can also observe that when the tumor is detected, model 3 has a higher number of lesions better segmented (i.e., the curve is more concentrated around the mean value).

Model	DR-patient [rate] (%)	DR-lesion [rate] (%)	DSC-Patient	False negatives/N voxels
M1	195/250 (78%)	184/250 (74%)	0.7	5/465
M2	246/250 (98%)	213/250 (85%)	0.5	7/2937
M3	211/250 (84%)	180/250 (72%)	0.62	6/656

Table 5.32: Results of the three models on the training set. DR=Detection Rate, DSC=Dice Similarity Score, False negatives and N_voxels are reported as the median number for patient.

Model	DR-patient [rate] (%)	DR-lesion [rate] (%)	DSC-Patient	DSC-Lesion	False negatives/N voxels
M1	52/62 (84%)	41/62 (66%)	0.47	0.53	1/575
M2	62/62 (100%)	50/62 (81%)	0.27	0.44	3/3747
M3	50/62 (81%)	37/62 (60%)	0.50	0.57	0/455

Table 5.33: Results of the three models on the validation set. DR = Detection Rate, DSC = Dice Similarity Score, False negatives and N_voxels are reported as the median number for patient.

Figure 5.33 show the distribution of FNs of the training set, in terms of percentage,among centers and scanners.

Figure 5.32 shows the median number of false positives per patient stratified for vendors (Fig. 5.32a) and for centers (Fig. 5.32b). Considering FPs per patient, model 2 obtained the worst results, since for some centers there were, in median, more than 4 FPs findings per patient. Conversely, model 1 and 3 obtained comparable results in terms of FPs, with no clear differences among centers and scanners.

Figure 5.33 shows the percentage of patients in which the index lesion was not detected separately for the cases in which the slice thickness between the sequences was equal (orange) and different (blue). Except for model 2, most FNs were in patients in which the slice thickness was different, i.e., 39% and 45%, respectively in model 1 and 3, of index lesion in patients with different slice thickness were FNs vs 28% and 34% of index lesions in patients with the same slice thickness. This might be due to the fact that the coregistration between different sequences introduces reconstructed voxels that might no correctly reflect the correct signal intensity.

Results on test set

Table 5.34 reports the results of the three networks on the test set in terms of detection rate and number of FPs. As anticipated during the training, model 2 can't be used for segmentation purposes, due to the low DCS and number and volumes of FPs, while model model 1 and model 3, obtained comparable results.

Figure 5.34 shows the violin plots of Dice Similarity Score (DSC) on the test set considering the whole automatic mask (ALL) and only detected lesions, i.e., DSC >0.10 (TP).

Figure 5.35 confirms what we observed in terms of FPs of the three models among centers and vendors. In this case, model 3 seems to be that obtaining the highest precision in segmenting the tumor, at the expense of a slight decrease of detection rate.

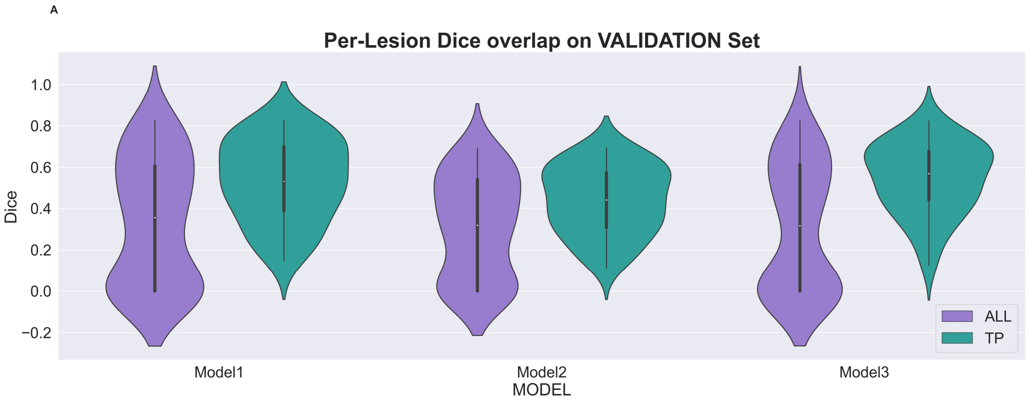
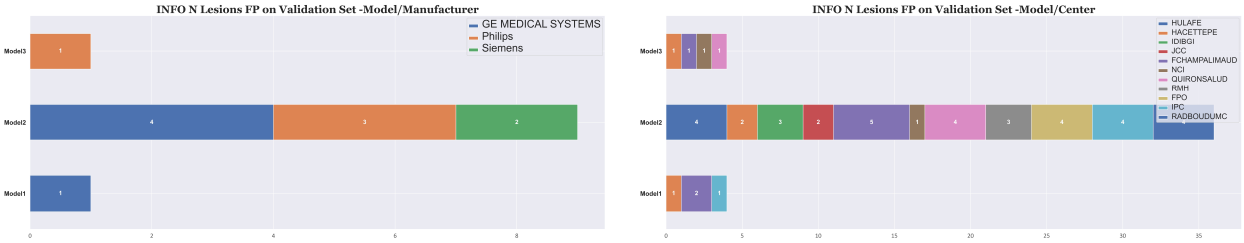


Figure 5.31: Violin Plot of Dice Similarity Score (DSC) obtained on the training set considering the whole automatic mask (ALL) and only detected lesions, i.e., DSC>0.10 (TP)



(a) Median number of false positive lesions on the validation set, grouped by vendors

(b) Median number of false positive lesions on the validation set, grouped by vendors

Figure 5.32: Median number of false positives

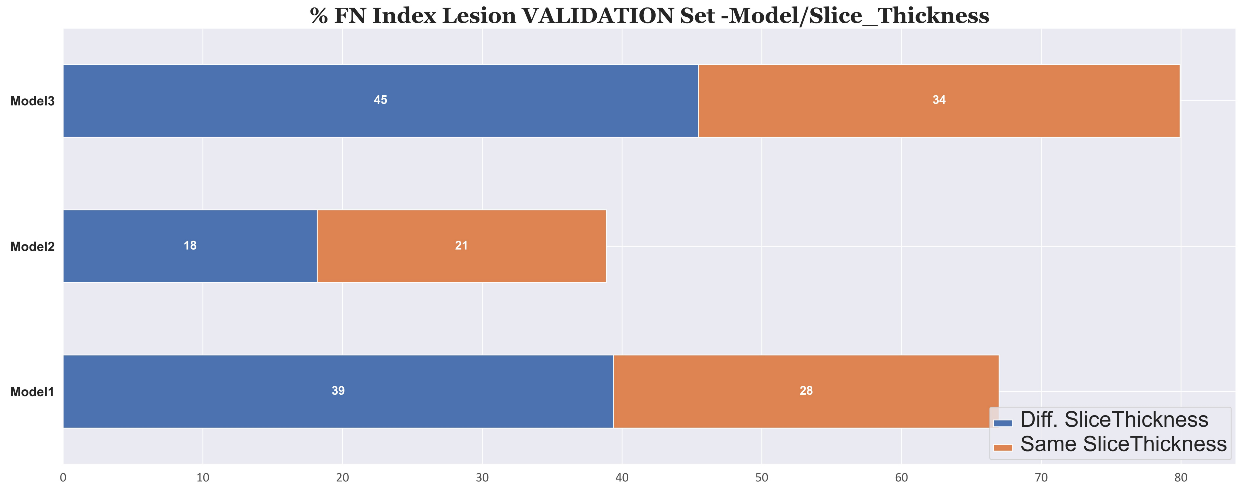


Figure 5.33: Percentage of false negatives index lesions on the validation set among patients having same slice thickness between sequences (orange) and a different slice thickness (blue).

Discussion

In our investigation, we trained and tested 3 different architectures for PCa detection and segmentation using a multi-parametric input. We observed that using networks that combine images at input-level seems to perform better in terms of segmentation's precision. However, the model that combined three different networks each fed with a different sequence was the one obtaining the highest detection rate, either at patient-level and at lesion-level. Considering these results, there may be worthy to combine different networks to increase either detection rate and number of false positives. Results obtained on the master model were promising, however it must be noticed that a number of issues should be faced. First, the dataset comprised images from 4 different vendors, therefore different normalization techniques should be exploited. Moreover, some images were acquired with endorectal coil (ERC) and others without. The ERC introduces strong artifacts (either in signal intensity and geometrical), therefore the pre-processing should be differentiated according to the presence of the ERC. Also, a larger number of cases with endorectal coil might be useful to increase performances. From our results, we also pointed out the need of perform further analysis while co-registering different sequences (when the slice thickness is different between sequences) to reduce artifacts that might be introduced by the registration itself.

Model	DR-patient [rate] (%)	DR-lesion [rate] (%)	DSC-Patient	DSC-Lesion	False negatives/N voxels
M1	52/59 (75%)	41/59 (64%)	0.50	0.58	0/537
M2	62/59 (88%)	50/59 (83%)	0.28	0.40	2/4167
M3	50/59 (75%)	37/59 (58%)	0.44	0.53	0/476

Table 5.34: Results of the three models on the test set. DR=Detection Rate, DSC=Dice Similarity Score, False negatives and N_voxels are reported as the median number for patient.

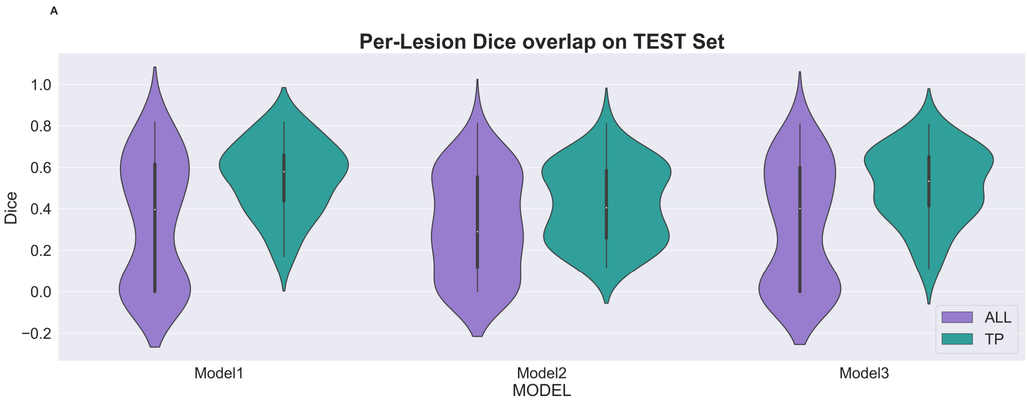
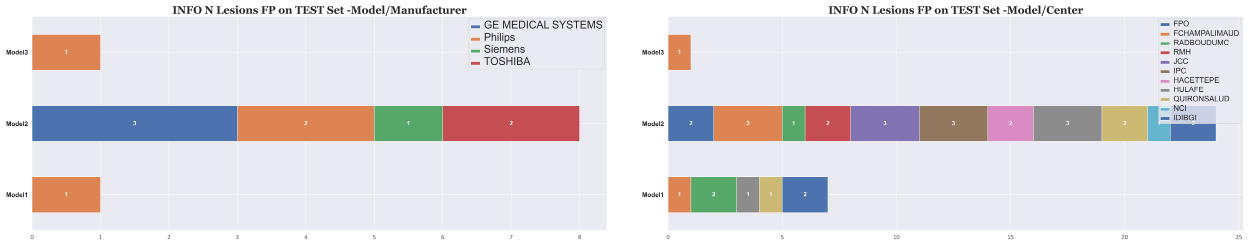


Figure 5.34: Violin Plot of Dice Similarity Score (DSC) obtained on the test set considering the whole automatic mask (ALL) and only detected lesions, i.e., DSC >0.10 (TP)



(a) Median number of false positive lesions on the validation set, grouped by vendors

(b) Median number of false positive lesions on the validation set, grouped by vendors

Figure 5.35: Median number of false positives on test set.

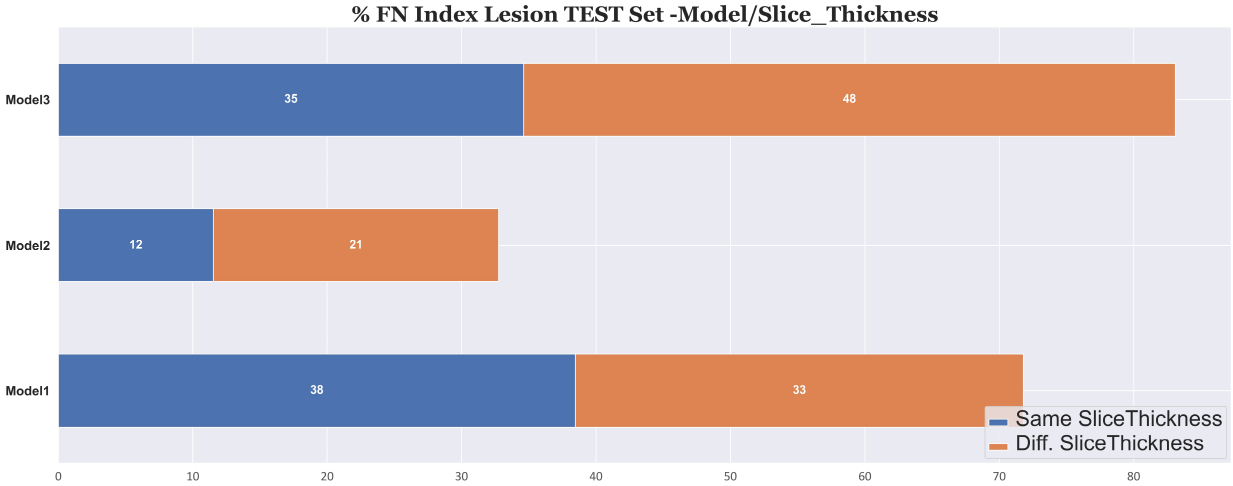


Figure 5.36: Percentage of false negatives index lesions on the validation set among patients having same slice thickness between sequences (orange) and a different slice thickness (blue)

5.6 Validation of ProCAncer-I prostate segmentation tool

In the scope of ProCAncer-I, there is the goal of developing algorithms for automatic segmentation of both prostate regions and lesions. For prostate regions Quibim provided an already integrated first version of an automatic segmentation tool in the ProstateNet platform. As mentioned earlier in the deliverable, the tool was used by the radiologists during the process of segmentation, helping them by non starting from scratch as it is known that manual segmentation is a time consuming task. The resulting final segmentations were used as ground truth for developing a new version of the automatic segmentation tool that will be trained and tested using the ProCAncer-I data.

In this section we present the results of the validation performed on the first version of the tool with the segmentations performed by the clinical partners.

5.6.1 Data

To validate the version 1 of the tool (V1) for automatic prostate region segmentation, we utilized a dataset comprising 552 cases, each with radiologist-validated segmentations. These cases were sourced from 12 different providers and acquired using equipment from three distinct manufacturers. The segmentation process exclusively employed MRI T2-weighted (T2w) images.

The dataset encompassed four image categories based on visual characteristics Figure 5.37:

1. **Normal T2w Images**: These images featured an enlarged prostate region and constituted the majority of the dataset.

2. **Endorectal Coil Images**: In some cases, patients underwent imaging with an endorectal coil, resulting in darker images.

3. **Whole Pelvis Field of View**: Images in this category provided a view of the patient's entire pelvic region.

4. **Fat Suppression Sequence**: A subset of images included a fat suppression sequence.

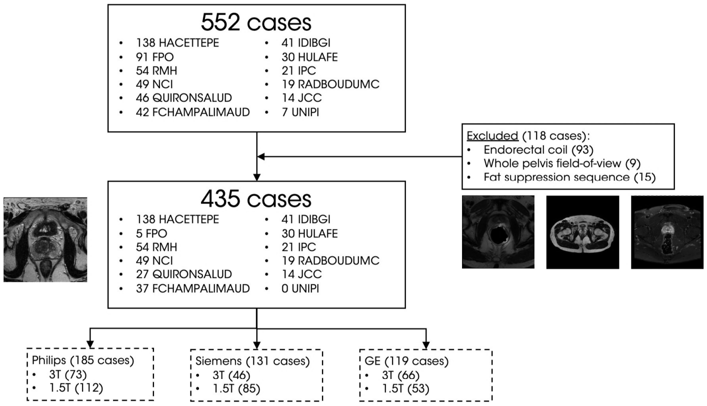


Figure 5.37: Description of the dataset used for V1 validation. 118 cases where excluded due to the image category and the 435 rest were used.

5.6.2 Validation of the tool Version 1

In the earlier stages of the project, Quibim integrated an automatic segmentation tool for prostate regions. This tool was developed using deep learning techniques, specifically employing a UNET architecture. It segments three distinct regions within the prostate: the Transitional + Central zone, the Peripheral Zone, and the Seminal Vesicles [42].

However, it's important to note that the training dataset used for this tool did not include images with Endorectal Coil, Whole Pelvis Field of View, or fat suppression sequences. Consequently, these specific image types were not considered in the subsequent analysis of the results, where the final number of cases was 435 Figure 5.37.

The ground truth segmentations used for evaluation were primarily obtained by refining the masks generated by the automatic segmentation tool. It is worth mentioning that this approach may introduce variations in the results compared to a scenario where radiologists manually performed the segmentation from scratch. These differences in segmentation origin could potentially impact the outcome of the validation analysis. The results show that the algorithm performed very well Figure 5.38. The results were grouped by manufacturer for each of the prostate regions to explore potential variability in the outcomes. For the majority of the groups the mean dice score consistently exceeded 0.9, even surpassing 0.95 in some of the groups, specially in Phillips cases Table 5.35. However, it's worth noting that some of the ground truth segmentations exhibited issues during the analysis, resulting in exceedingly low scores. These concerns have been duly communicated to the data providers, and corrective measures have been performed.

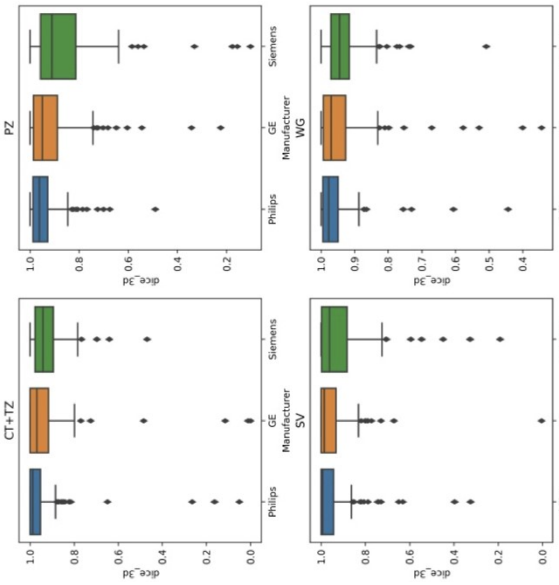


Figure 5.38: Dice Score results for the segmentations of the V1 tool. CZ+TZ: Central Zone and Tranzitional Zone; PZ: Periferal Zone; SV: Seminal Vesicles; WG: Whole Gland (CZ+TZ and PZ

DC	Phillips			GE			Siemens		
CZ+TZ	0.95 ± 0.12	1.5T 3T	0.96 ± 0.12 0.94 ± 0.12	0.92 ± 0.16	1.5T 3T	0.93 ± 0.09 0.92 ± 0.20	0.92 ± 0.08	1.5T 3T	0.92 ± 0.08 0.91 ± 0.07
PZ	0.94 ± 0.07	1.5T 3T	0.94 ± 0.08 0.94 ± 0.05	0.91 ± 0.12	1.5T 3T	0.89 ± 0.13 0.92 ± 0.12	0.86 ± 0.16	1.5T 3T	0.87 ± 0.16 0.85 ± 0.15
SV	0.95 ± 0.09	1.5T 3T	0.96 ± 0.10 0.94 ± 0.07	0.95 ± 0.11	1.5T 3T	0.93 ± 0.14 0.96 ± 0.07	0.92 ± 0.13	1.5T 3T	0.92 ± 0.14 0.92 ± 0.11
WG	0.96 ± 0.06	1.5T 3T	0.96 ± 0.07 0.96 ± 0.04	0.94 ± 0.11	1.5T 3T	0.93 ± 0.10 0.94 ± 0.11	0.93 ± 0.06	1.5T 3T	0.93 ± 0.07 0.93 ± 0.04

Table 5.35: Dice Scores for the V1 tool grouped by manufacturer and magnetic field strength

5.6.3 Development of the tool Version 2

Within the framework of Task T6.2 in the ProCancer-I project, new automatic segmentation tool is aimed to be developed using the project's available data. As previously highlighted, the considerable variability in the data arising from different manufacturers, magnetic field strengths, sites, and image acquisition protocols presents a valuable opportunity to create a robust tool. As mentioned before, the initial version (V1) of the algorithm was not trained with Endorectal Coil images, which consequently led to suboptimal segmentations in cases involving these cases Figure 5.39. Given the substantial presence of Endorectal Coil cases in the dataset (93 in total), one of the primary objectives for the forthcoming V2 of the automatic segmentation tool is to demonstrate improved performance on these specific cases.

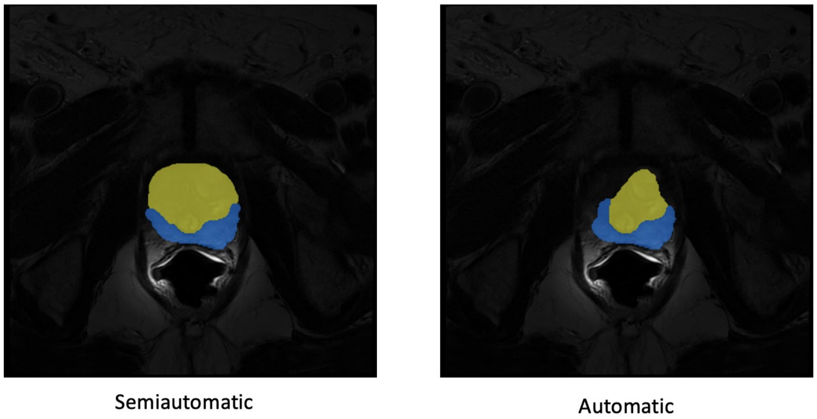


Figure 5.39: Example of bad segmentation performed by the V1 tool (Automatic) compared to the corrected segmentation performed by radiologists (Semiautomatic) when Endorectal Coil is present in the image

Methodology

Similar to the previously described approach, the experiments to development of a new version of the tool (V2) have been done using the nnUNet framework. nnUNet short for Neural Networks UNet, stands as a cutting-edge framework at the forefront of the medical image segmentation. Built upon the foundation of the well-known UNet architecture, nnU-Net represents an innovative deep learning segmentation approach that autonomously tailors itself to new tasks within the biomedical domain. This self-configuration encompasses preprocessing, network architecture, training, and post-processing, effectively streamlining the entire process [41].

Different approaches have been tested for finding the best results in both types of images: without and with Endorectal Coil.

1. Baseline model based on the 2D configuration trained with both types of images.
2. Model based on the 2D configuration trained only with "normal" images (without Endorectal Coil).
3. Model based on the 2D configuration trained only with Endorectal Coil images.
4. Baseline model based on 3D configuration trained with both types of images.

Results

2D vs 3D baselines

When examining the outcomes from the two baseline models configured in both 2D and 3D, with both types of images, no definitive superiority of one over the other emerged, as evident in Figure 5.40. In broad terms, the Dice score exhibited fluctuations within the range of 0.85 to 0.91. Considering that the 3D approach demands more time and resources, a strategic decision was made to proceed with the 2D configuration for each image type.

Moreover, when narrowing down the dataset to either of the two image types, the overall number of available images for training diminishes. This reduction could pose limitations when opting for the 3D configuration, as it would necessitate a larger dataset to maintain robust performance.

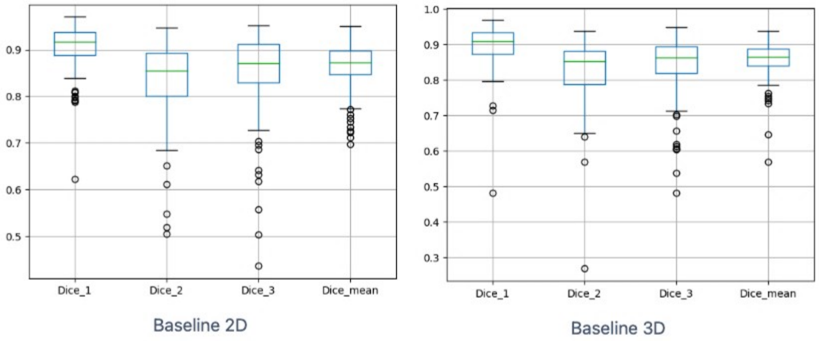


Figure 5.40: Dice scores for TZ+CZ (Dice1), PZ (Dice2), SV (Dice3) and mean for both 2D and 3D approaches.

Baseline vs non-EC model

Following the creation of the 2D baseline model, a dedicated model was developed specifically for images without the Endorectal Coil (EC). This model was trained using a dataset consisting of 304 cases for training and 66 cases for testing. Importantly, the data partitioning aligns with the same data split utilized in the baseline approach. In other words, the 304 cases featuring non-EC images in the non-EC model approach were the same used in the baseline, and this consistency extended to the test set as well to make the results comparable.

The outcomes of the analysis shows good results over all, being the mean Dice scores for each of the regions higher than 0.85 and even reaching 0.9 in the TZ+CZ Figure 5.41. Again not major differences are found between approaches. The results show that having also the EC cases in the training dataset didn't change the overall results.

When analyzing individual results it can be seen that both models tend to be consistent in the predictions. Only for some specific cases, there are significative differences such as in the case that can be seen in Figure 5.42, where the non-EC model was able to segment better an abnormal PZ region.

Baseline vs EC model

Same approach than before was carried out for the 2D configuration for the EC model where the training set contained 66 cases and the test set 28. As expected, the overall results were lower compared to the non-EC

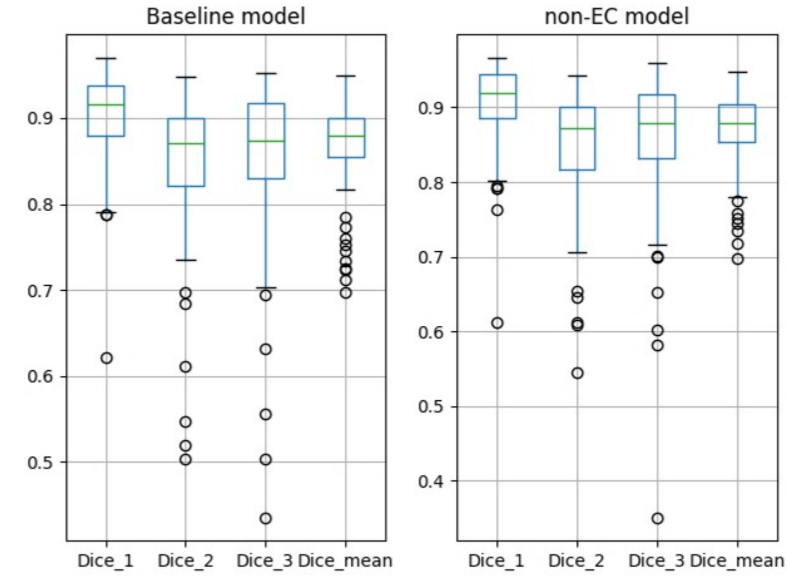


Figure 5.41: Dice scores for TZ+CZ (Dice1), PZ (Dice2), SV (Dice3) and mean for both 2D baseline and 2D non-EC model.

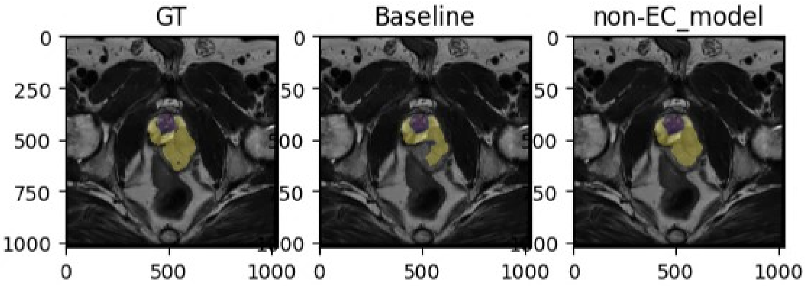


Figure 5.42: Example of non-EC model performing better than the 2D baseline for an abnormal case.

cases, but still good results Figure 5.43. The EC model performed slightly worse than the baseline for the Endorectal Coil images, in our opinion, due to the fact that the final model is trained with much less data than the baseline which could take advantage of using also the non-EC data to identify the morphology of the prostate and its regions.

In the Figure 5.44 we can se an example of how the baseline model was able to segment better the PZ (yellow) compared to the EC model.

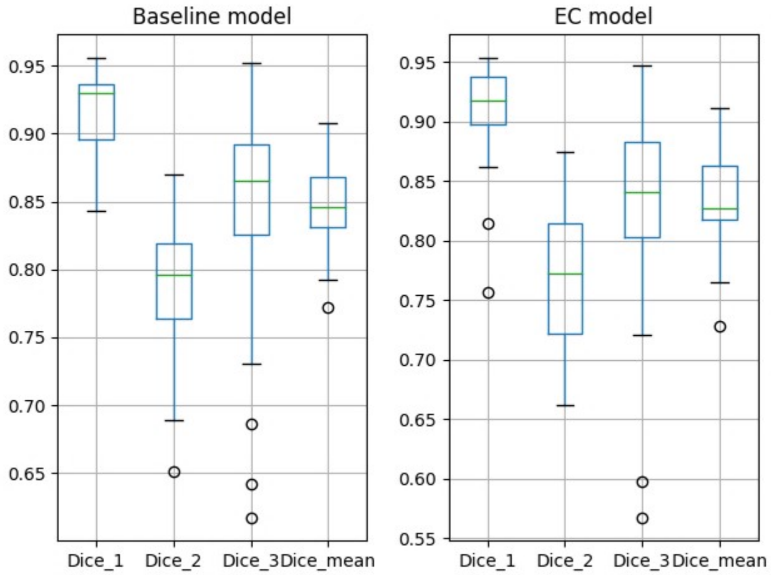


Figure 5.43: Dice scores for TZ+CZ (Dice1), PZ (Dice2), SV (Dice3) and mean for both 2D baseline and 2D EC model.

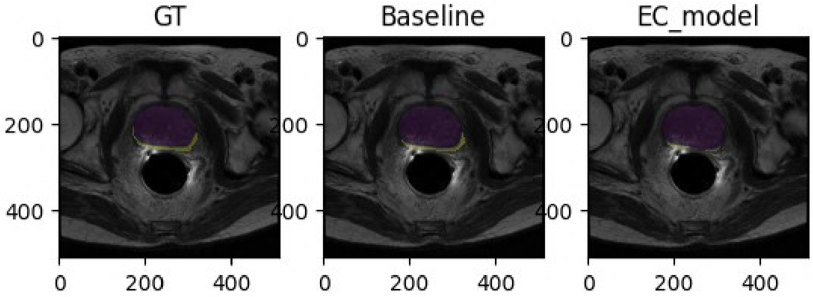


Figure 5.44: Example of baseline model performing better than the EC-model in the PZ.

Discussion

We have presented different approaches for the whole prostate gland segmentation master model where the main objective was to improve the overall results and, specially, in the cases where Endorectal Coil was present in the image. The 2D baseline model trained with both non-EC images and EC images shows good results for both image categories. Furthermore, as demonstrated, specific models for each of them were not able to improve significantly its performance. In summary, the 2D baseline shows good results in both non-EC and EC images, showing that the model new version of the model could target EC images which was an exlusion criteria in the V1 of the tool.

5.7 Effect of dataset characteristics on segmentation performance

In the scope of the ProCAncer-I project, FCHAMPALIMAUD developed several segmentation models for the tasks of whole gland, prostate zones (Peripheral and Transitional+Central) and lesion segmentation. The models were evaluated not only on the ProCAncer-I dataset, but also in additional ones and combinations between those and ProCAncer-I. The best models were then used to provide ground truths for the Radiomics models, in cases where no previous ground truths were available. In this section, we present the methodology behind the development of the models, as well as the extensive evaluation performed.

5.7.1 Methodology

Data description

Four different datasets were used in this study:

- **Prostate158** is a collection of biparametric MRI volumes that include T2W, DWI and ADC modalities. These volumes were obtained by the German university hospital - Charité University Hospital Berlin, using Siemens 3T MR scanners (VIDA and Skyra). Regarding the acquisition of the images, the following description was provided by the authors of the paper:"T2w sequences were acquired with the following parameters: slice thickness 3 mm, no interslice gap, in-plane resolution 0.47 × 0.47 mm, field of view (FOV) size 180*180mm, time to echo (TE)/repetition time (TR) 116 ms/4040 ms, turbo factor 25, flip angle 160º, acquisition time 3 min and 56s" [2]
- **ProstateX** is a collection of prostate MRI volumes that include T2W, DWI and ADC modalities. These volumes were obtained by the Prostate MR Reference Center — Radboud University Medical Centre (Radboudumc) in the Netherlands, using two Siemens 3T MR scanners (MAGNETOM Trio and Skyra). Regarding the acquisition of the images, the following description was provided by the challenge's organizers: "T2-weighted images were acquired using a turbo spin echo sequence and had a resolution of around 0.5 mm in plane and a slice thickness of 3.6 mm" [5]
- **ProstateNet** is a collection of multiparametric MRI volumes that include T2W, DWI and ADC modalities. These volumes were obtained by the 12 clinical partners of the ProCAncer-I project. These partners used Siemens (Aera, Skyra, Sola, Avanto, VIDA, Tim, Prisma, Veri, Symphony, Osirix), Philips (Ingenia, Achieva, Multiva) and GE scanners (Optima, Signa, DISCOVERY). Given that each center has specific acquisition protocols, no single one was used across all mpMRI studies used. Given that labels could be defined automatically (using the QUIBIM-developed ProCAncer-I prostate region segmentation tool), manually corrected but not validated and manually corrected and validated, we define a hierarchy of annotations, selecting the whichever one is available first: 1) manually corrected and validated (n=610), 2) manually corrected but not validated (n=30), 3) automatically generated (n=65).
- **ProstateAll** is a combination of all previous datasets. It was created with the purpose of increasing available data, as well as add even more heterogeneity, so that we could train more robust master models which would not suffer from the caveats of institution, protocol and scanner manufacturer variability.

Table 5.36 shows the entire composition of our entire dataset. From these values, 15% were used as holdout-validation, and the remaining was used for training, following a 5-fold cross-validation strategy.

Deep learning model specification

Three distinct 3D deep-learning (DL) segmentation models were trained - a simple 3D U-Net model [76] (U-Net), a 3D U-Net model with deep supervision [99] (U-Net + D.S.), and a 3D high-resolution nnUNet model [41] (nnUNet).

Deep supervision is a technique that utilizes intermediate predictions, generated at each step of the decoder to produce semantically meaningful features that will enhance the model's performance, instead of relying solely on the final output of the model.

	Gland			
	Total	Siemens	Philips	GE
Prostate158	139	139	-	-
ProstateX	182	182	-	-
ProstateNet	638	152	245	239
ProstateAll	959	473	245	239
	Zones			
Prostate158	139	139	-	-
ProstateX	181	181	-	-
ProstateNet	638	152	245	239
ProstateAll	958	472	245	239
	Lesions			
Prostate158	82	82	-	-
ProstateX	190	190	-	-
ProstateNet	461	136	184	136
ProstateAll	733	408	184	136
ProstateNet mpMRI	417	131	178	107

Table 5.36: Stratification of samples by manufacturer for all four segmentation datasets per task. Both the ProstateNet and ProstateAll datasets also include a very residual amount (≤ 5) of Toshiba scanners, which were accounted for in the Total values.

All networks are implemented in Pytorch [67], and trained for 1000 (250 mini-batches per epoch) or 200 epochs, for the nnUnet and U-Net/U-Net + D.S., respectively. Training for U-Net and U-Net + D.S. was performed using Lightning [24], a low-code and heavily customisable framework for neural network training and testing in PyTorch.

To train the nnUNet models, we used the provided 3D Full Resolution architecture. This framework uses stochastic gradient descent with Nesterov momentum ($\mu = 0.99$), a maximum initial learning rate of 0.01, and polynomial [16] learning rate policy which reduced the learning rate by a factor of $(1 - \frac{epoch}{epoch_{max}})^{0.9}$. The loss function is a combination of Dice and cross-entropy losses.

nnUNet applies automatic preprocessing based on the dataset fingerprint, thus the models for each dataset worked on data with slightly different spatial structures:

- ProstateX - spacing = 0.5x0.5x3.0mm and crop size = 320x320x16 voxels
- Prostate158 - spacing = 0.4x0.4x3.0mm and crop size = 256x256x28 voxels
- ProstateNet - spacing = 0.5x0.5x3.0mm and crop size = 256x256x28 voxels
- ProstateAll - spacing = 0.5x0.5x3.0mm and crop size = 256x256x26 voxels

Additionally, similar to the U-Net + D.S., nnUNet also employs deep supervision. All models were trained with a batch size of 2.

To train the U-Net and U-Net + D.S. models, we used a stochastic gradient descent optimizer with 0.99 momentum with a maximum learning rate of 0.01, weight decay of 0.005 [52] and cosine decay with a minimum learning rate of 0. The encoder is composed of five regular convolutional layers with kernel size [3,3,3] and increasing depths (32, 64, 128, 256, 320) intercalated with 2x2 max-pooling operations with strides [2,2,1], [2,2,1], [2,2,1], [2,2,2], [2,2,2], similar to what is used in nnUNet [41], while the decoder replicates the encoder but replaces the max-pooling operations with transpose convolutions. In any case, Swish activation functions [71] and instance normalization operations are used after each convolution, with a dropout [83] probability of 0.1. The anysotropic max-pooling allows for the preservation of a minimal resolution of 4 in the slice dimension. Using a batch-size of 2, we sampled 256x256x16 patches from the image such that patches with and without positive samples (i.e. voxels belonging to the prostate gland) are sampled equally. A combo loss [86] — the addition of the generalized Dice [98] and weighted focal losses [49] with alpha=0.5 — was used to train both U-Net and U-Net + D.S. models.

Deep supervision [99] for U-Net + D.S. was implemented using an additional classifier after each layer decoder layer that classifies voxels at a decreased resolution. To calculate the loss for each deep supervision output, the ground truth was downsampled to match the resolution at each decoder layer and the loss was calculated. All losses (for the original resolution and for the deep supervision) were combined using a weighted average where the weight is parameterized as $(\frac{1}{2})^{ds-1}$, where ds is the downsampling level. For instance, for the full resolution, this weight evaluates to 1, whereas for the lowest resolution (ds=4) this evaluates to 1/8.

During U-Net and U-Net + D.S. training, we augment data in real time to increase the variability of observed data by our model using MONAI [61], with each transform having a 0.1 probability of being applied. Particularly we use:

- Random contrast adjustments (gamma multiplier between 0.5 and 1.5)
- Random intensity standard deviation shift (multiplier between 0.9 and 1.1)
- Random intensity shift (multiplier between 0.9 and 1.1)
- Random addition of Rician noise (with a standard deviation of 0.02)
- Random addition of Gibbs noise (alpha between 0.0 and 0.6 and standard deviation of 0.25)
- Random affine transform (translation range in voxels [(0,4),(0,4),(0,1)], rotation range in radians [(0,π/16),(0,π/16),(0,π/16)])
- Random flipping along all axes
- Random shearing (shearing factor between 0.9 and 1.1 for all axes)
- Gaussian blurring (sigma between 0.25 and 1.5)

Model evaluation

Each model is evaluated by its Dice score (DS) using 5-fold cross-validation according to the best observed DS 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. Additionally, since DS only provides an overlap score, we also include the Hausdorff Distance (HD), Average Symmetric Surface distance (ASSD), and Relative Absolute Volume Difference (RAVD) during quality assessment of the model, as these metrics provide a quantitative measure of the spatial accuracy by considering the shape and volume of the segmented regions [97] (both distance metrics were calculated using MedPy [55]).

5.7.2 Results

Whole gland segmentation

For whole gland segmentation, we observe a relatively small range of performances in CV (Fig. 5.45 and Tab. 5.37). For the hold-out test set, we observe the same but only when models were trained and tested on data from the same distribution (i.e. same dataset; Fig. 5.46 and Tab. 5.38).

After training nnUNet models and simpler U-Net models (Figs. 5.45 and 5.46), we observed that performance for the former was almost always better than for the latter - both for CV and for holdout - which is on par with previous studies [73, 39]; this holds up for both in- and out-of-distribution data. Taking this into consideration, we did not train any other models apart from nnUNet models for the remaining tasks.

When looking at the performance of the nnUNet models trained on Prostate158 and then evaluated on ProstateX, it is possible to observe that it produces errors (HD) up to 100× higher than all other models (Fig. 5.49 and Tab. 5.38). These relatively large shifts are considerably smaller when models are trained on ProstateNet or ProstateAll, hinting that **model-specific effects on performance decrease as the amount and variety of data increases**. Indeed, the model-specific effects on performance become increasingly negligible as the amount and variability of data increases. Table 5.38 further shows this, as it can be observed that the nnUNet model trained on the ProstateAll data produces far lower HD, RAVD and ASSD scores over all datasets, which means overall lower errors.

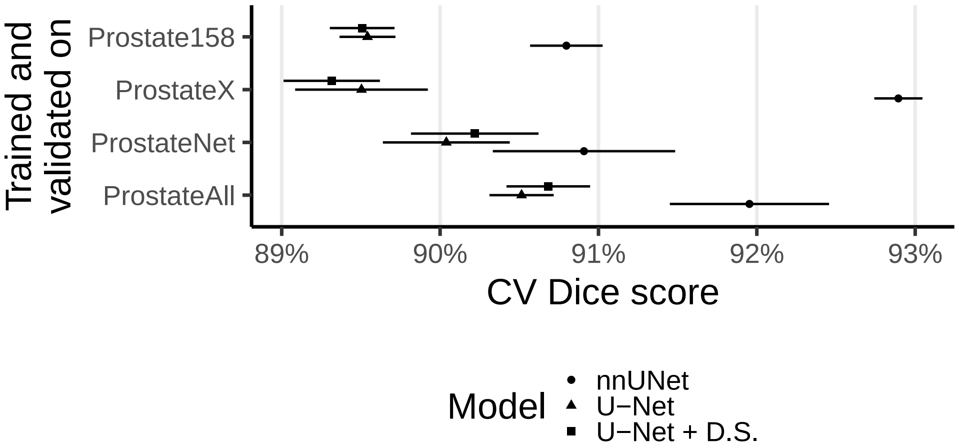


Figure 5.45: Whole gland segmentation CV results.

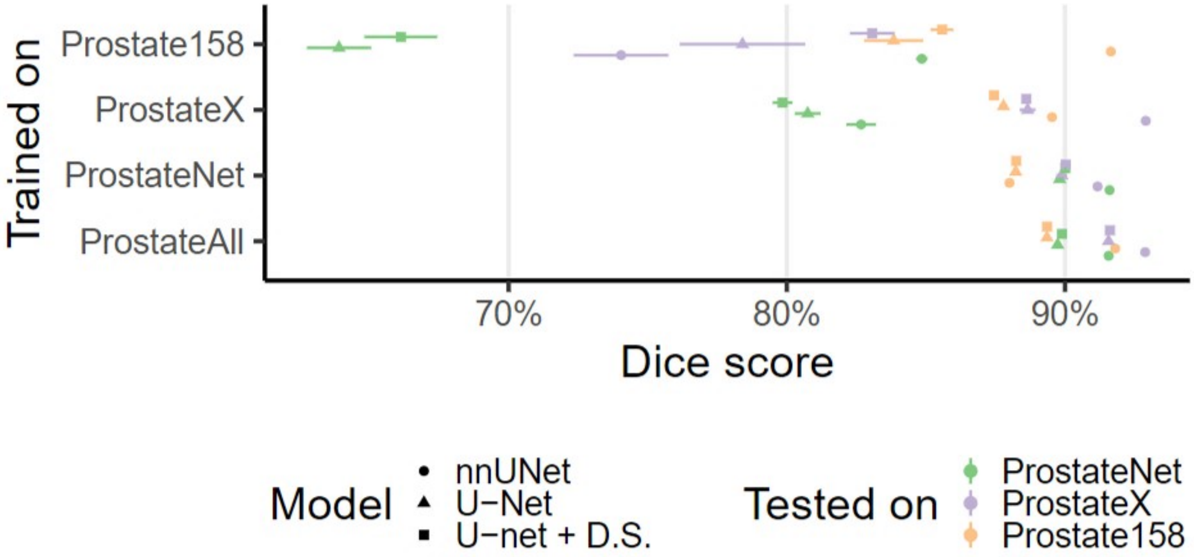


Figure 5.46: Whole gland segmentation hold-out test set Dice scores for the different models. Left: performance stratified by dataset. Right: performance stratified by manufacturer.

Prostate zone segmentation

During CV, zonal segmentation models present diverse levels of performance, both in terms of the segmented zone (peripheral (PZ) and transitional (TZ)), and in terms of the different sets of data used during training

	Dice	Hausdorf	RAVD	ASSD	Recall	p-value	
			Gland				
ProstateX	0.93 ± 0.02	8.01 ± 4.36	0.01 ± 0.07	0.32 ± 0.1	1.0 ± 0.0	0.1745	
Prostate158	0.91 ± 0.03	14.02 ± 11.81	0.0 ± 0.1	0.35 ± 0.19	1.0 ± 0.0	0.0758	
ProstateNet	0.91 ± 0.09	12.48 ± 23.27	0.09 ± 1.78	0.43 ± 0.72	1.0 ± 0.0	0.2506	
ProstateAll	0.92 ± 0.08	12.71 ± 27.82	0.06 ± 1.42	0.45 ± 1.79	1.0 ± 0.0	-	
			Zones PZ TZ				
ProstateX	0.8 ± 0.08	16.45 ± 9.75	−0.0 ± 0.24	0.64 ± 0.39	1.0 ± 0.0	0.009	
	0.88 ± 0.05	14.97 ± 9.31	0.02 ± 0.17	0.55 ± 0.22	1.0 ± 0.0	0.0758	
Prostate158	0.76 ± 0.09	20.95 ± 18.13	−0.01 ± 0.26	0.64 ± 1.16	1.0 ± 0.0	0.009	
	0.88 ± 0.06	22.09 ± 16.84	0.03 ± 0.14	0.44 ± 0.35	1.0 ± 0.0	0.0758	
ProstateNet	0.81 ± 0.11	15.7 ± 22.22	0.08 ± 1.45	0.56 ± 0.8	1.0 ± 0.0	0.4647	
	0.89 ± 0.08	13.13 ± 11.9	0.14 ± 2.2	0.5 ± 0.62	1.0 ± 0.0	0.0758	
ProstateAll	0.82 ± 0.1	15.43 ± 19.97	0.06 ± 1.24	0.5 ± 0.81	1.0 ± 0.0	-	PZ
	0.9 ± 0.08	14.61 ± 13.34	0.09 ± 1.78	0.44 ± 0.54	1.0 ± 0.0		TZ
			Lesions				
ProstateX	0.17 ± 0.24	100.45 ± 89.78	2.22 ± 13.35	32.9 ± 51.24	0.4 ± 0.04	0.009	
Prostate158	0.25 ± 0.27	95.44 ± 87.01	0.04 ± 1.17	24.5 ± 47.58	0.5 ± 0.06	0.1172	
ProstateNet	0.38 ± 0.3	66.45 ± 66.83	0.33 ± 2.79	18.43 ± 35.56	0.7 ± 0.02	0.4647	
ProstateAll	0.36 ± 0.3	77.3 ± 73.68	0.9 ± 9.75	22.74 ± 39.54	0.65 ± 0.02	-	
ProstateAll Cascade	0.36 ± 0.3	81.03 ± 73.59	0.68 ± 8.39	23.98 ± 40.4	0.65 ± 0.02	0.9168	
ProstateNet mpMRI	0.4 ± 0.28	71.3 ± 69.18	0.61 ± 3.16	15.37 ± 27.46	0.76 ± 0.02	0.1172	

Table 5.37: nnUNet CV results stratified by segmentation task. For each dataset, the average Dice, Hausdorf, RAVD and ASSD performance, along with their respective standard deviations, are presented. p-values for Kruskal-Wallis significance test comparing the Dice score between ProstateAll results and each other model are also shown, with significant results (p-value < 0.01) are marked as green.

		Tested on				
		ProstateX	Prostate158	ProstateNet	ProstateAll	
Trained on	ProstateX	0.93 ± 0.02	0.9 ± 0.04	0.84 ± 0.12	0.86 ± 0.11	Dice
		8.8 ± 4.61	19.12 ± 35.05	47.99 ± 76.07	36.56 ± 65.85	Hausdorf
		−0.03 ± 0.06	0.05 ± 0.11	0.24 ± 0.56	0.17 ± 0.48	RAVD
		0.32 ± 0.08	0.48 ± 0.45	2.91 ± 5.75	2.08 ± 4.86	ASSD
	Prostate158	0.66 ± 0.18	0.92 ± 0.04	0.86 ± 0.11	0.83 ± 0.15	
		364.2 ± 87.76	13.04 ± 12.98	22.37 ± 42.1	85.29 ± 143.78	
		0.95 ± 0.84	0.03 ± 0.08	0.19 ± 0.62	0.31 ± 0.7	
		49.91 ± 24.9	0.31 ± 0.18	1.74 ± 5.76	10.59 ± 22.29	
	ProstateNet	0.91 ± 0.02	0.88 ± 0.04	0.92 ± 0.04	0.91 ± 0.04	
		8.64 ± 3.75	13.23 ± 6.88	9.88 ± 9.94	10.12 ± 8.8	
		−0.07 ± 0.07	−0.03 ± 0.08	−0.0 ± 0.08	−0.02 ± 0.08	
		0.39 ± 0.09	0.42 ± 0.14	0.34 ± 0.15	0.36 ± 0.14	
	ProstateAll	0.93 ± 0.02	0.92 ± 0.03	0.92 ± 0.04	0.92 ± 0.04	
		9.01 ± 4.52	10.71 ± 8.28	9.62 ± 9.76	9.66 ± 8.81	
		−0.03 ± 0.06	0.02 ± 0.07	0.01 ± 0.08	0.0 ± 0.08	
		0.31 ± 0.07	0.27 ± 0.1	0.34 ± 0.14	0.33 ± 0.13	

Table 5.38: nnUNet whole gland segmentation hold-out test set results. For each pairwise evaluation, the average Dice, Hausdorf, RAVD and ASSD performance, along with their respective standard deviations, are presented.

(Fig. 5.47 and Tab. 5.37) — both ProstateX and Prostate158 models yield Dice scores significantly worse than those of ProstateAll in the context of PZ segmentation. Apart from the lower Dice, it can be noted that Prostate158 produces maximum errors (Hausdorf) $5 - 8mm$ bigger than the remaining models.

Considering the hold-out test performance, and similarly to what was shown for the whole gland seg-

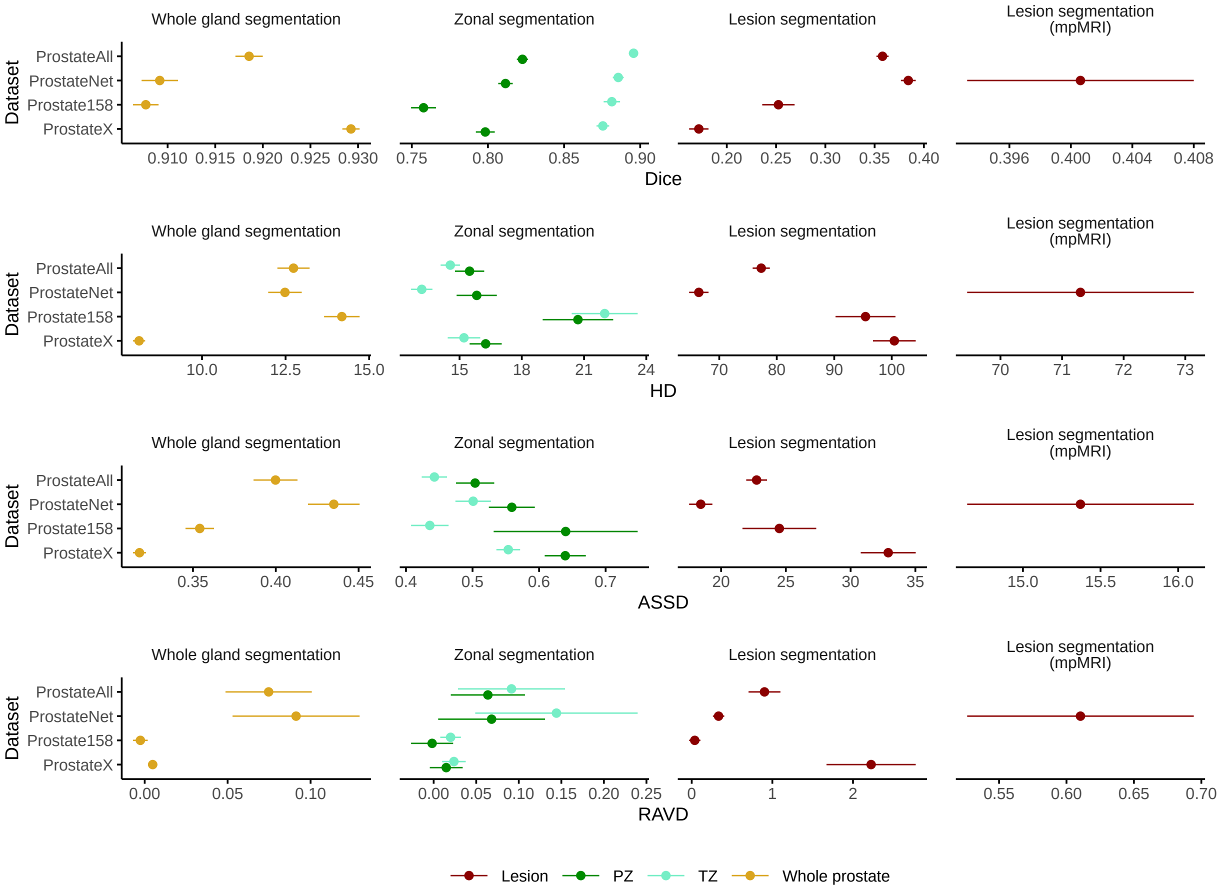


Figure 5.47: Multi-metric analysis of the CV performance of the nnUNet models, stratified by segmentation task. Each circle represents the mean and the horizontal lines represent the standard error around the mean.

mentation, it can be seen that the performance is similar between CV and hold-out test set (Tab. 5.37 and 5.39). On par with what was observed during whole gland segmentation, it can be seen that the trend of producing very large maximum errors (HD) - on average 100× (PZ) and 200× (TZ) larger than all other models (Fig. 5.49) - is kept, when evaluating the Prostate158 models on ProstateX data.

Learning curve analysis

To better understand how data variety and size impact performance, we conducted a simple learning curve analysis by training the nnUNet model on different proportions of the ProstateAll training dataset ($[0.1, 0.5, 0.7, 1.0]$). Initially, our conceptual understanding focused on data variability, rather than on the amount of data — i.e. having data from more diverse sources would lead to improved performance. However, it was also possible that simply increasing the amount of data could lead to better performance. The CV results allow us to validate the expected outcome of this analysis (Fig. 5.48 **B**) — performance increases as the amount of data increases. Extending this analysis to the hold-out test set and stratifying by testing dataset (Fig. 5.48 **C**) while comparing the learning curve performance with that of models trained and tested on the same dataset reveals something striking — for starters, even at relatively low amounts of ProstateAll data (0.1 — corresponding to 65 cases) the performance is at least comparable to that of models trained and tested on the same dataset, suggesting that data variability does indeed play a role in increasing performance. The additional insight is that increasing the size of the training dataset — which inevitably increases the variability of the data as well — also leads to improved performance which is oftentimes better than that

		Tested on								
		ProstateX		Prostate158		ProstateNet		ProstateAll		
Trained on	ProstateX	0.82 ± 0.06	0.87 ± 0.07	0.72 ± 0.1	0.84 ± 0.05	0.72 ± 0.17	0.77 ± 0.19	0.74 ± 0.15	0.8 ± 0.17	Dice
		15.66 ± 8.74	15.73 ± 11.06	17.73 ± 8.4	33.21 ± 20.62	20.46 ± 18.72	56.17 ± 85.67	19.2 ± 16.3	45.56 ± 73.0	Hausdorf
		−0.04 ± 0.11	0.01 ± 0.31	0.32 ± 0.24	−0.1 ± 0.17	0.12 ± 0.54	0.54 ± 1.23	0.12 ± 0.47	0.35 ± 1.06	RAVD
		0.61 ± 0.48	0.57 ± 0.35	0.59 ± 0.3	0.71 ± 0.22	1.14 ± 1.65	4.28 ± 9.5	0.97 ± 1.4	3.11 ± 8.01	ASSD
	Prostate158	0.7 ± 0.1	0.68 ± 0.2	0.79 ± 0.11	0.89 ± 0.06	0.7 ± 0.15	0.79 ± 0.18	0.71 ± 0.14	0.78 ± 0.18	
		296.17 ± 101.81	353.73 ± 115.13	19.48 ± 13.78	20.96 ± 18.05	23.92 ± 35.72	27.65 ± 51.67	73.94 ± 118.74	87.37 ± 143.28	
		−0.08 ± 0.28	0.96 ± 1.46	0.07 ± 0.16	0.06 ± 0.16	−0.12 ± 0.41	0.59 ± 1.6	−0.09 ± 0.37	0.59 ± 1.49	
		19.69 ± 15.82	51.68 ± 34.68	0.51 ± 0.48	0.36 ± 0.15	1.16 ± 2.13	2.78 ± 9.16	4.51 ± 10.11	11.55 ± 25.48	
	ProstateNet	0.81 ± 0.05	0.85 ± 0.08	0.74 ± 0.09	0.84 ± 0.05	0.8 ± 0.15	0.86 ± 0.16	0.79 ± 0.13	0.86 ± 0.13	
		16.73 ± 8.43	16.52 ± 8.18	17.49 ± 6.84	31.48 ± 22.19	15.02 ± 11.23	14.77 ± 19.17	15.67 ± 10.3	17.34 ± 18.97	
		−0.09 ± 0.1	−0.09 ± 0.21	0.27 ± 0.23	−0.17 ± 0.1	0.03 ± 0.37	0.1 ± 0.73	0.04 ± 0.34	0.03 ± 0.62	
		0.59 ± 0.3	0.66 ± 0.41	0.58 ± 0.26	0.66 ± 0.29	0.63 ± 1.15	1.24 ± 7.34	0.62 ± 0.96	1.06 ± 6.06	
	ProstateAll	0.83 ± 0.05	0.88 ± 0.07	0.81 ± 0.08	0.9 ± 0.04	0.8 ± 0.15	0.86 ± 0.16	0.81 ± 0.13	0.87 ± 0.14	
		15.19 ± 7.79	15.17 ± 8.3	17.05 ± 6.94	23.98 ± 19.29	14.69 ± 11.55	14.64 ± 20.05	15.1 ± 10.44	15.99 ± 18.6	
		−0.04 ± 0.12	−0.01 ± 0.25	0.08 ± 0.13	−0.01 ± 0.1	0.03 ± 0.37	0.11 ± 0.7	0.03 ± 0.31	0.07 ± 0.59	
		0.56 ± 0.45	0.55 ± 0.35	0.43 ± 0.25	0.36 ± 0.13	0.62 ± 1.13	1.31 ± 7.51	0.58 ± 0.96	1.04 ± 6.21	
		PZ	TZ							

Table 5.39: nnUNet PZ and TZ segmentation hold-out test set results. For each pairwise evaluation, the average Dice, Hausdorf, RAVD and ASSD performance, along with their respective standard deviations, are presented.

observed in models trained and tested on the same data. For this reason, we suggest here that this increase in performance is driven by a combination of increased data size and increased variability, highlighting the importance of initiatives such as ProstateNet in training algorithms that can be clinically deployable and robust to large shifts in performance at test and inference time.

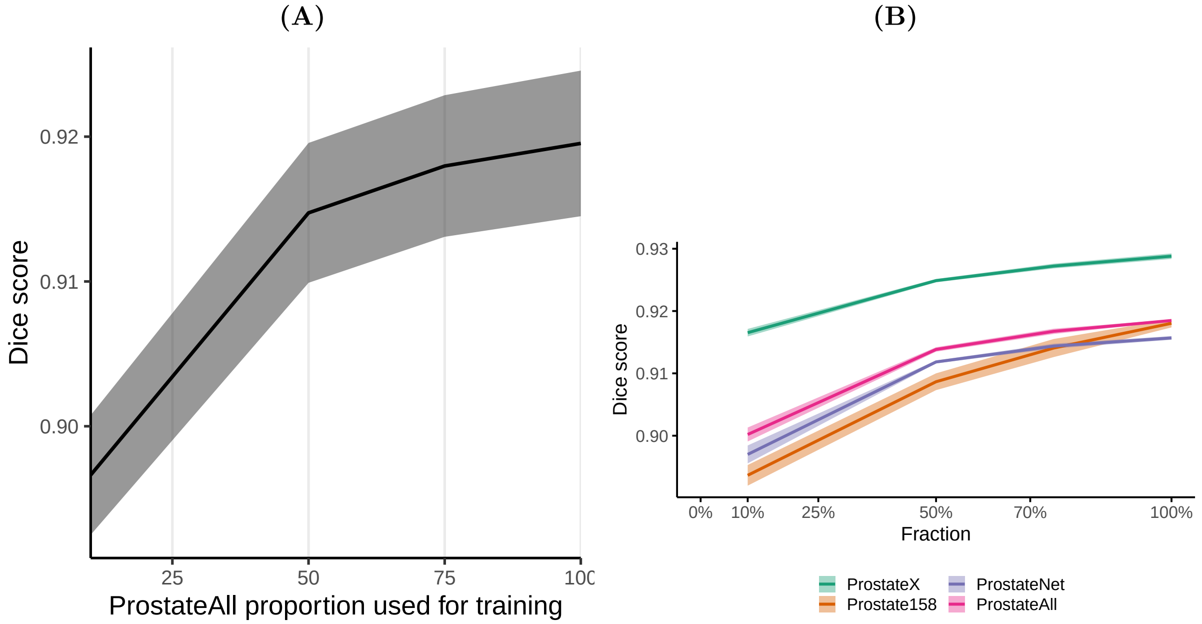


Figure 5.48: Learning curves performance scores for the data partitions of ProstateAll. **A**: Cross-validation performance. **B**: Test set performance stratified by dataset.

Lesion detection improves with larger/more diverse datasets

The performance of lesion segmentation models is generally worse when compared with that of other prostate segmentation tasks, mostly due to the small nature and highly irregular shape of the ROIs (Fig. 5.47 and

Tab. 5.37). Using data which is both more abundant and more diverse — ProstateNet/ProstateAll — leads to improved results when compared with ProstateX and Prostate158, two small, single-institution datasets. This is true not only of Dice scores, but also of HD and ASSD (both generally lower for both ProstateNet and ProstateAll). Additionally, as demonstrated by the precision-recall curves, ProstateNet and ProstateAll models are better at detecting lesions.

Apart from the standard single modality full-resolution models, we also trained a cascade model for the ProstateAll T2W data (in theory more capable of capturing small objects) and a full-resolution model using ProstateNet T2W, DWI and ADC (mpMRI) images (DWI and ADC are more appropriate than T2W, as they hyper- and hypo- saturate in the areas where lesions exist). While the cascade model provided minimal differences - with a p-value of almost 1 - apart from slightly smaller errors, the ProstateNet model using the three modalities showed some improvements, most notably in terms of dice and recall, albeit these improvements were not statisticaly significant.

Similarly to what was shown for whole gland and zone segmentation models, the majority of in-distribution hold-out results are similar to those obtained during CV. Regarding out-of-distribution performance, both ProstateX and Prostate158 show a steep performance drop, while ProstateAll provides consistent results all throughout. Additionally, the performance drops significantly when testing the Prostate158 model on the ProstateX dataset. Contrary to what was hypothesised, the mpMRI ProstateNet and the Cascade ProstateAll models showed a poor generalization performance, showing a greater decrease when compared to its T2W-only counterpart, in both Dice and recall. Overall, the best model turned out to be the ProstateAll T2W-only model, particularly when considering the average precision for each model (Fig. 5.50). To calculate this, we first identified the lesion candidates as described in [9] at different IoU thresholds ($[0.1, 0.25, 0.50, 0.75]$) and calculated the precision-recall curve for each of these thresholds (Figs. 5.51, 5.52, 5.53).

Here, we can again observe that the ProstateAll model is the one that better generalizes, achieving relatively high AP scores up until IoUs of 0.50. For higher values (@75) however, the AP becomes lower or even 0 simply due to the fact that very few, or even none, of the samples achieve such high IoU values. Fig. X shows some of the heatmaps for lesion detection, showcasing that despite not being accurate in regards to segmentation, they are capable of gathering information from the area where the lesions are located.

These results further strengthen our findings that increasing variability within leads to models that are far more robust, not only for segmentation but also for detection tasks.

		Tested on				
		ProstateX	Prostate158	ProstateNet	ProstateAll	
Trained on	ProstateX	0.17 ± 0.25	0.28 ± 0.28	0.12 ± 0.21	0.15 ± 0.23	Dice
		90.33 ± 70.94	69.41 ± 41.86	89.95 ± 59.44	88.11 ± 62.67	Hausdorf
		1.49 ± 5.37	-0.19 ± 0.6	13.98 ± 100.1	8.22 ± 74.71	RAVD
		35.34 ± 43.29	21.98 ± 26.6	33.03 ± 35.23	32.78 ± 37.76	ASSD
		0.39 ± 0.07	0.58 ± 0.14	0.33 ± 0.06	0.38 ± 0.04	Recall
	Prostate158	0.04 ± 0.13	0.21 ± 0.34	0.14 ± 0.23	0.11 ± 0.22	
		207.98 ± 171.3	30.94 ± 53.34	69.03 ± 96.71	114.29 ± 143.71	
		12.97 ± 78.73	0.04 ± 0.65	0.06 ± 2.14	4.6 ± 47.14	
		107.81 ± 123.17	2.97 ± 6.7	24.09 ± 48.69	51.53 ± 91.71	
		0.11 ± 0.05	0.33 ± 0.14	0.3 ± 0.06	0.24 ± 0.04	
	ProstateNet	0.1 ± 0.2	0.25 ± 0.28	0.33 ± 0.3	0.24 ± 0.29	
		98.97 ± 122.66	69.69 ± 57.29	54.52 ± 62.22	71.62 ± 90.4	
		0.31 ± 2.21	-0.25 ± 0.77	0.26 ± 1.59	0.23 ± 1.79	
		50.86 ± 88.74	21.07 ± 36.2	14.32 ± 35.78	27.83 ± 62.42	
		0.23 ± 0.06	0.58 ± 0.14	0.62 ± 0.06	0.48 ± 0.04	
	ProstateAll	0.21 ± 0.26	0.45 ± 0.3	0.39 ± 0.3	0.33 ± 0.3	
		76.23 ± 54.67	65.81 ± 43.74	67.3 ± 65.2	67.55 ± 59.25	
		1.24 ± 4.54	-0.07 ± 0.56	1.39 ± 5.96	1.18 ± 5.22	
		22.91 ± 28.61	11.69 ± 15.09	18.73 ± 34.38	19.15 ± 30.3	
		0.45 ± 0.08	0.83 ± 0.11	0.7 ± 0.06	0.63 ± 0.04	
	ProstateAll Cascade	0.21 ± 0.26	0.45 ± 0.32	0.4 ± 0.3	0.33 ± 0.3	
		68.42 ± 52.75	52.56 ± 53.6	78.81 ± 70.46	75.38 ± 64.22	
		1.21 ± 4.61	0.06 ± 0.71	2.92 ± 21.38	2.06 ± 16.14	
		21.85 ± 25.89	6.97 ± 10.84	22.25 ± 39.89	21.02 ± 34.62	
		0.48 ± 0.08	0.75 ± 0.12	0.7 ± 0.06	0.62 ± 0.04	
	ProstateNet mpMRI	–	–	0.34 ± 0.28	–	
		–	–	80.16 ± 80.62	–	
		–	–	0.16 ± 1.32	–	
		–	–	28.52 ± 47.19	–	
		–	–	0.71 ± 0.03	–	

Table 5.40: nnUNet lesion 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.

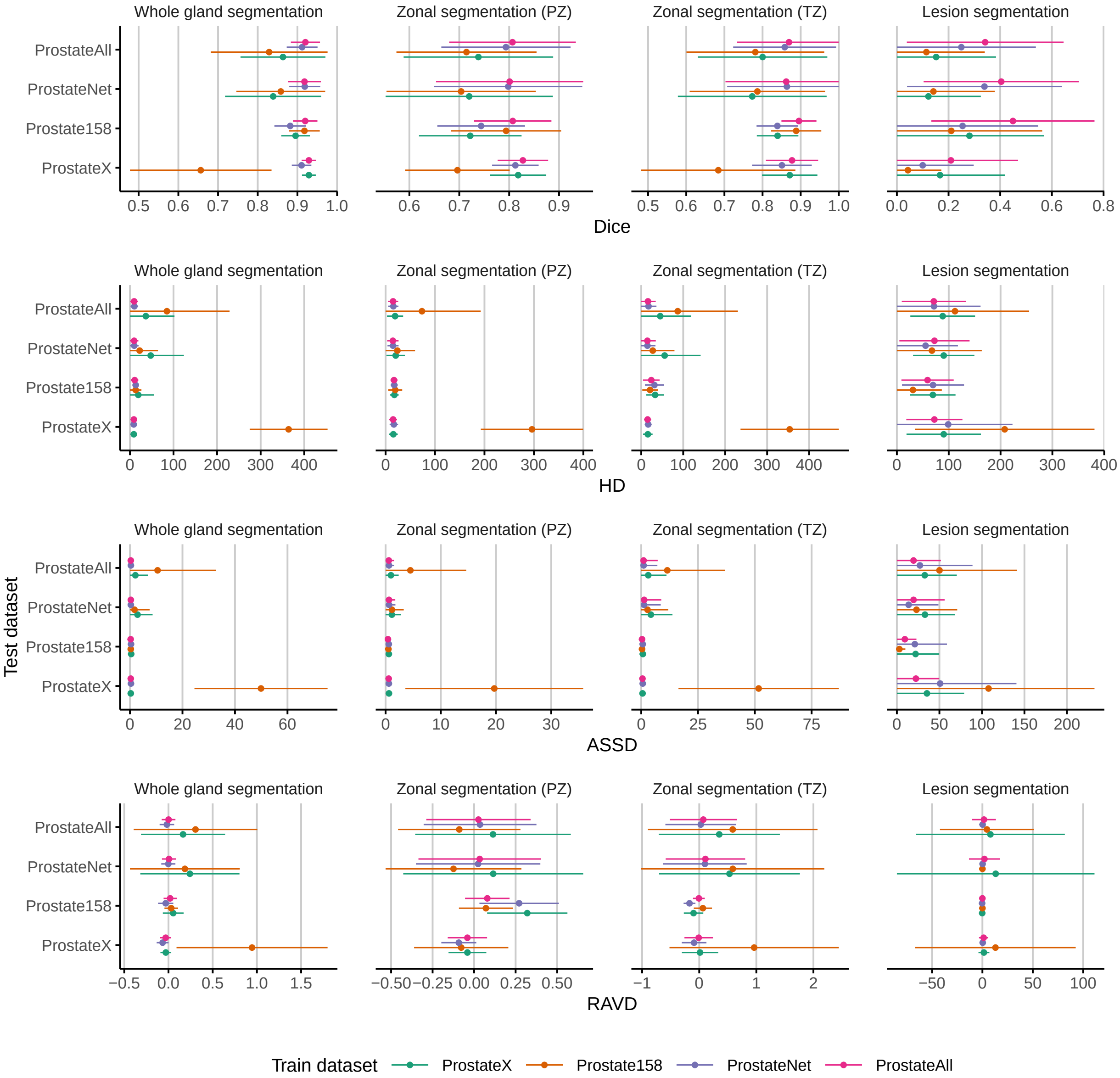


Figure 5.49: Multi-metric analysis of the hold-test performance of the nnUNet models, in- and out- of distribution, stratified by segmentation task.

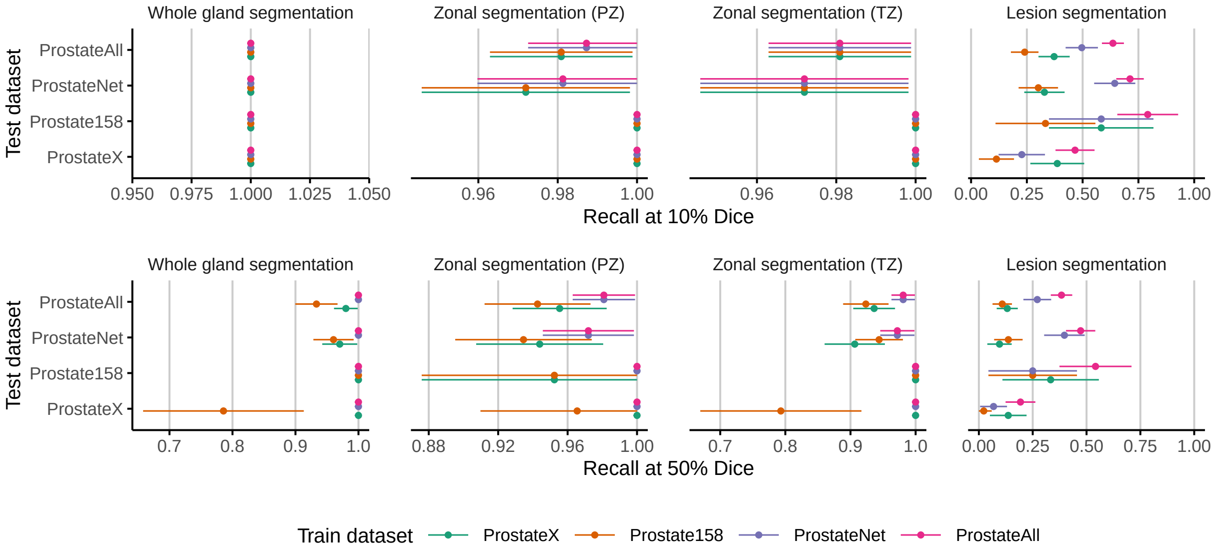


Figure 5.50: Test set ROI detection analysis based on Recall at different Dice levels (10% and 50%)

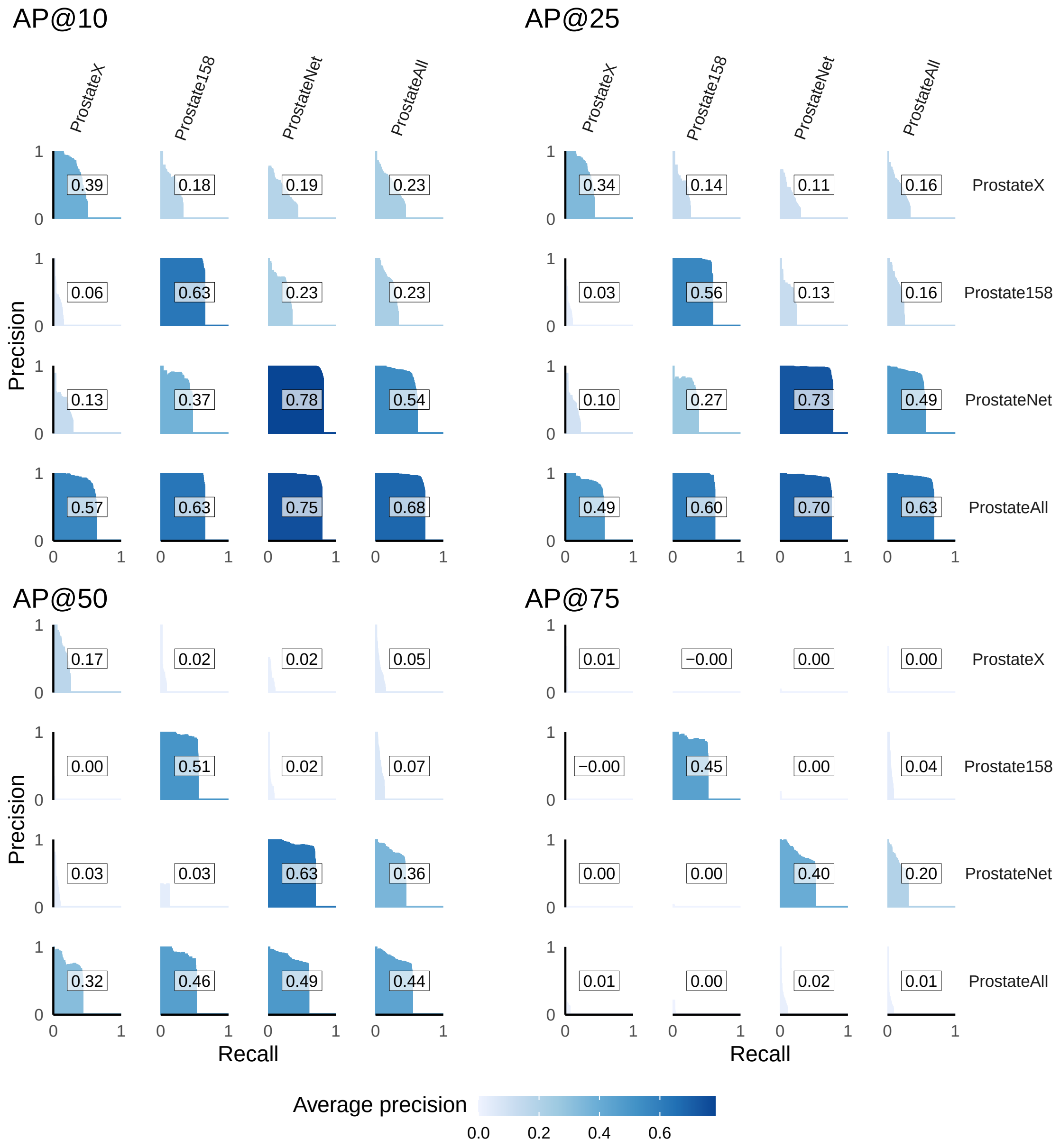


Figure 5.51: Test set precision-recall scores @ different IoU thresholds percentages ($[0.1, 0.25, 0.50, 0.75]$) for the full-resolution nnUNet models. Right: Models trained on those datasets; Top: Models tested on those datasets.

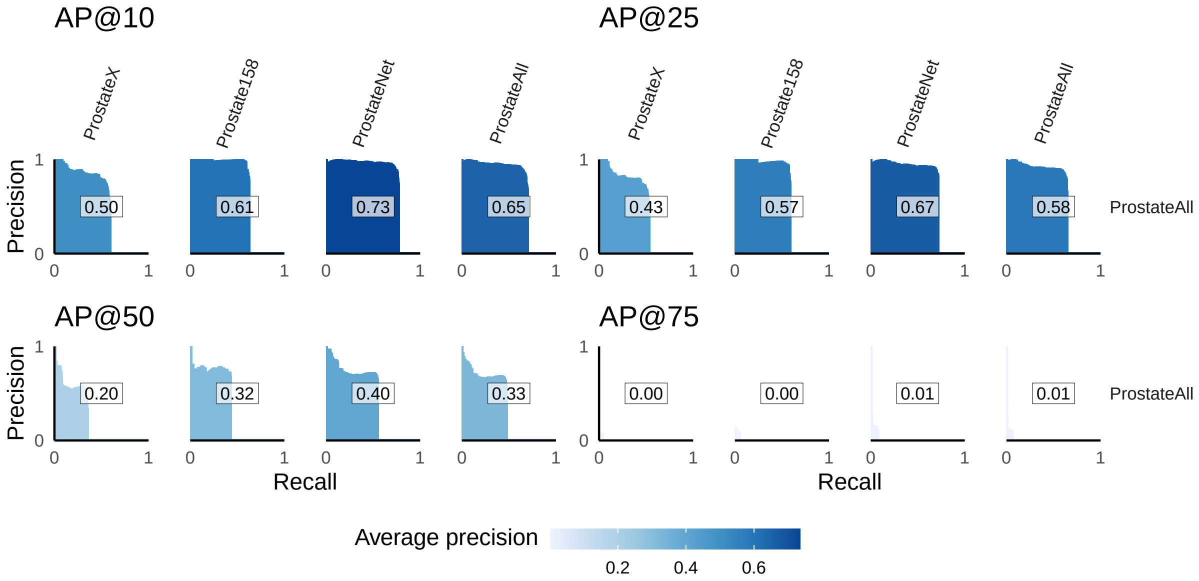


Figure 5.52: Test set precision-recall scores @ different IoU thresholds percentages ($[0.1, 0.25, 0.50, 0.75]$) for the Cascade ProstateAll nnUNet model.

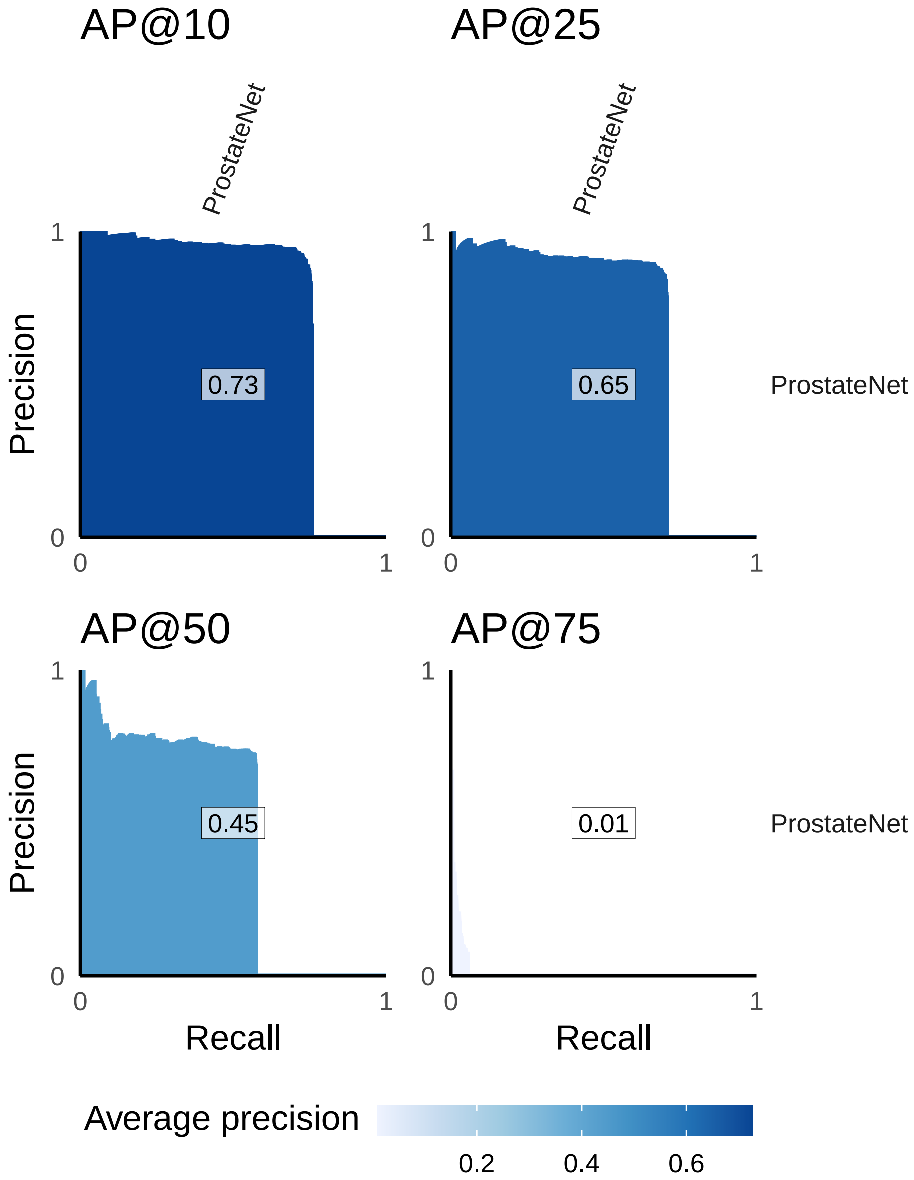


Figure 5.53: Test set precision-recall scores @ different IoU thresholds percentages ($[0.1, 0.25, 0.50, 0.75]$) for the mpMRI full-resolution ProstateNet nnUNet model.

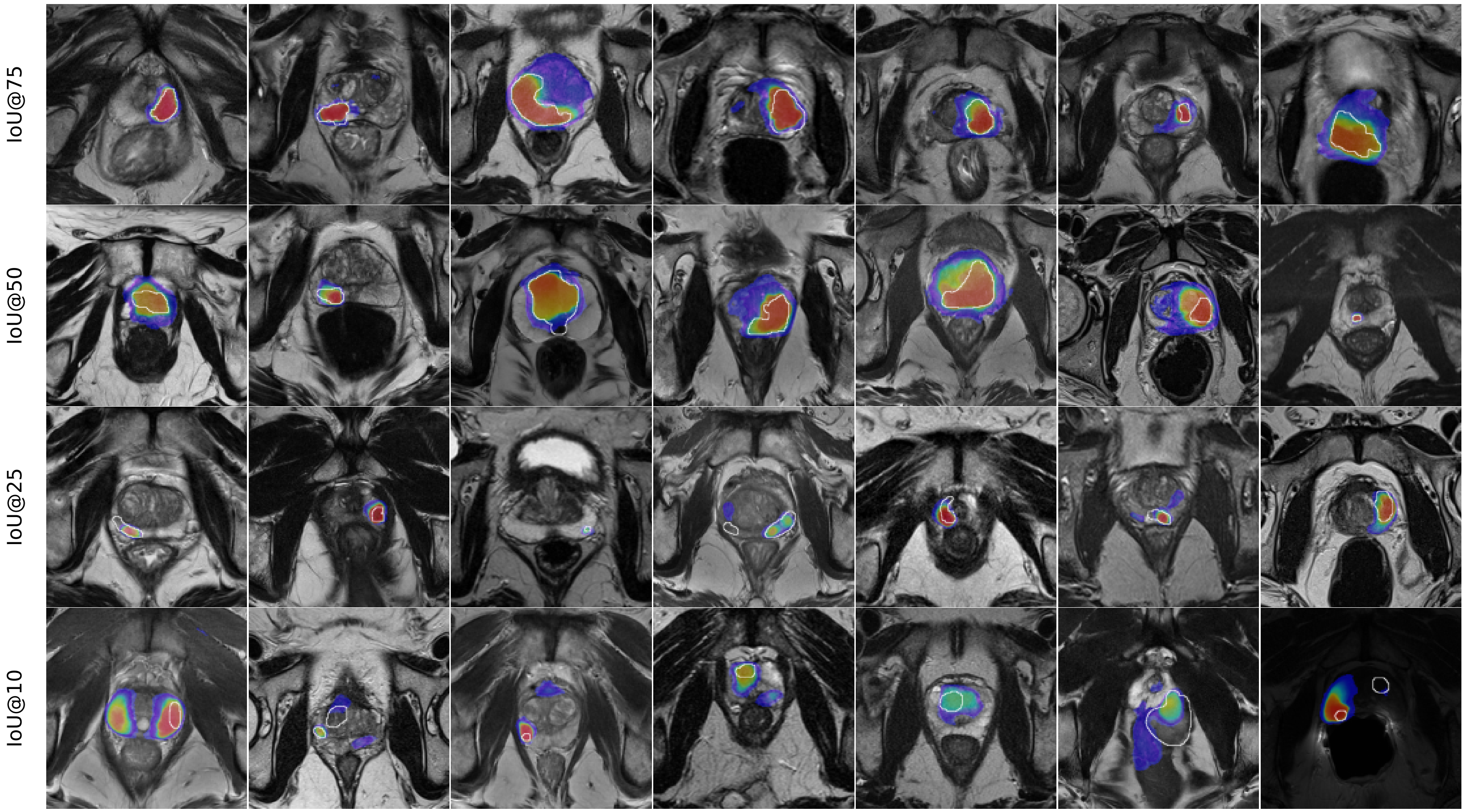


Figure 5.54: Lesion detection probability maps at different IoU thresholds. Green hues denote zones of lower probability, while red hues denote areas of higher probability.

Qualitative analysis

Finally, to better understand failure cases — here defined as cases where the Dice score was inferior to 90% — we individually inspected each case that fit this criterion in ProstateNet with the assistance of a radiologist with 6.5 years of experience (RM). Interestingly, the outcome of this analysis is not associated with the failure of the model — rather, it is associated with low quality labels. Particularly, this is associated with cases where labels were automatically generated using the ProCAncer-I tool or when the corrections provided by clinicians contained mistakes. Additionally, through the analysis or large ($\approx > 20$) Hausdorf errors, an issue became apparent — some of the errors stemmed from the existence of more than one connected component (given that the prostate gland is a single continuous object in 3 dimensions, there can be no more than one component corresponding to the prostate gland). To understand this quantitatively, we isolated the largest connected component for all masks and calculated the IoU score between the largest connected component and the totality of the ground truth (if there is no more than one connected component, the IoU score should be 100%). As shown in Table 5.41, approximately 1% of ground truths have a large spurious object not belonging to the prostate gland, while most detected abnormalities (74%) are relatively small. In other words, there are cases where the calculated IoU score will be relatively worse than expected due to the quality of the labels as shown by our visual inspection and annotation. Taking the aforementioned aspects into account, it becomes evident that this approach for prostate gland segmentation — training a nnUNet on the ProstateAll dataset — is of high value and can be safely deployed across several different centres.

Conclusively, the segmentations inferred by our model were of considerable quality and the failure cases were typically associated with poor annotation.

IoU interval	[0%,90%[	[90%,99%[	[99%,100%[	100%	Total
Number of cases	7	21	82	524	637

Table 5.41: Number of ground truths for different IoU scores between the largest connected component and the entire ground truth.

5.8 ProstateNet Lesion Segmentation with Deep Learning

The emphasis of this report revolves around the delineation of lesions within the prostate gland via the segmentation of T2-weighted axial images sourced from the ProstateNet dataset, employing advanced deep learning methods. The initial phase of this research involves a data curation process, selecting 419 T2-Ax segmented monochromatic series from the ProstateNet dataset. Subsequently, preprocessing was performed consisting in analyzing spatial resolutions, standardizing dataset spacing, and employing techniques like image cropping, denoising, and intensity normalization. The report offers insights in the distinct approaches adopted for the 2D and 3D models, outlining their architectures, and the solutions applied to counter issues of prediction discrepancies observed in the 2D model and overfitting, such as adjustments in layer complexity and the application of regularization techniques.

5.8.1 Methods

Data extraction

From all the ProstateNet series, an initial selection of compatible 420 T2-Ax segmented monochromatic series was used, in which 1 was discarded due to insufficient pixels in the segmentation. No further manual evaluation was done for the images. From those cases, 47 with a second segmented lesion were found and 10 with a third segmented lesion as well, in all those cases the segmentation masks were added together in a final binary mask.

Data Preprocessing

Initially, an analysis of the spatial resolution and spacing of the images was performed (see Fig. 5.55). Values close to the median were selected to achieve a uniform dataset spacing, using the values $[0.5, 0.5, 3.4]$.

Considering the size and position of the organ within the images, a strategy of central cropping was adopted to standardize and reduce the dataset's resolution to 320 × 320 pixels.

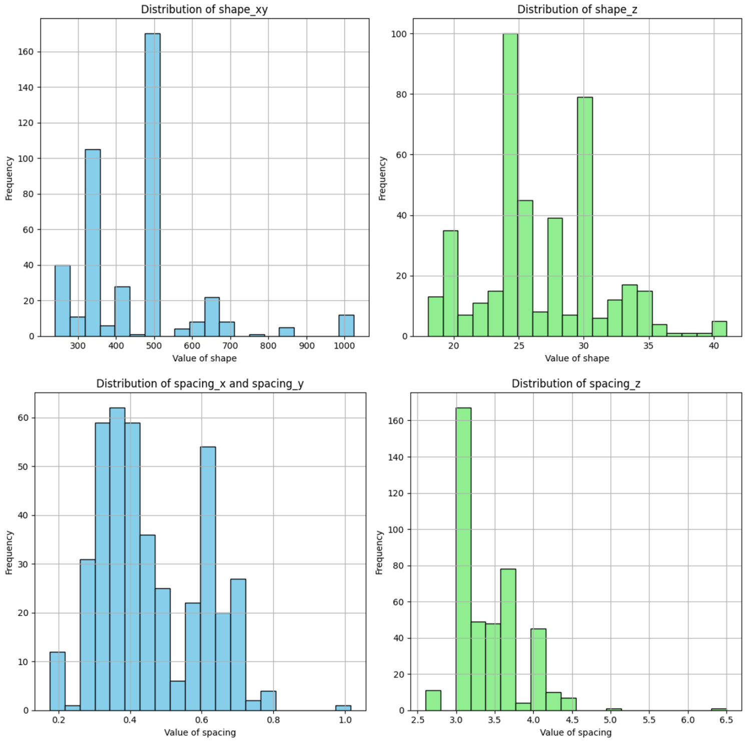


Figure 5.55: Dataset resolution (top) and spacing distribution (bottom) for x, y (left), and z coordinates (right).

Two additional preprocessing steps were applied to the images: a denoising process utilizing anisotropic diffusion with 2 iterations and a conductance value of 1.5, followed by final Z-score intensity normalization.

The images used for the 2D model were not subjected to a denoising process. The input configuration comprised a sequence of 3 frames corresponding to a resolution of 320 × 320 × 3: the preceding frame, the frame of interest, and the subsequent frame. These frames were overlapped, and this sequence was utilized to predict a final 2D segmentation mask corresponding to the frame of interest. Due to over-representation of the background labels a ratio of 0.6 of background outputs and their respective input were removed.

Regarding the input configuration for the 3D model, a frame depth of 16 was selected, aligning with the smallest number of frames among the 419 cases. Utilizing an overlapping approach, this frame depth produced an input size of 320 × 320 × 16 elements.

The division of cases within the dataset was structured as follows:

- 346 patients were allocated for training (constituting 81.3% of the dataset).
- 42 patients for validation (9.9% of the dataset).
- 31 patients for testing (8.8% of the dataset).

5.8.2 Models

2D prediction

In the 2D prediction model, two main parts were joined together: a Conv3D encoder and a U-Net model for predicting masks. The Conv3D encoder starts with smaller 3D data in a $32 \times 32 \times 3 \times 1$ shape and uses layers that focus on 3x3x3 filters to transform the data into a more detailed form. It builds up to a global max-pooling layer, which gathers and refines the key features before output.

On the other side, the U-Net model for mask prediction starts with larger $160 \times 160 \times 1$ input data. It works through a classical U-Net style with Conv2D layers using 3×3 filters and 64 filters in each layer. After moving through a bottleneck, it expands through the network with convolutional and upscaling layers.

The output from the Conv3D encoder gets reshaped and adjusted to fit the U-Net model's requirements, now at $160 \times 160 \times 1$. Then, the U-Net's upsampling layers finally produce the 320×320 prediction mask. This combined model brings together both the Conv3D and U-Net models to create predictions for binary masks.

3D prediction

Two 3D mask prediction models were employed:

- **Model 1**: A 3D Unet architecture model that employs a Leaky ReLU activation function with an alpha value of 0.1 to prevent gradient banishing. The U-Net structure consists of an encoder-decoder layout, employing convolutional blocks from 6 to 30 filters for feature extraction, batch normalization, and Leaky ReLU activation. The encoder segment downsamples the spatial dimensions via max-pooling, while the decoder section uses up-sampling to restore the spatial resolution and integrates high-resolution features from the encoder through skip connections. The output is generated through a final 3D convolutional layer with a sigmoid activation function, producing a segmentation mask.

  This U-Net architecture, configured to accept input shapes of (320, 320, 16, 1), is crafted to generate segmentation masks of the same dimensions as the input. The number of units and parameters was kept low due to a consistent observation of over-fitting across the models as it will be discussed later.

- **Model 2**: This model follows the same structure as Model 1 but adding complexity in the convolutional layers using 16 to 80 filters and adding three multi-head attention layers between the middle convolutional blocks of the model. L2 kernel regularization of 0.01 and drop out of 0.15 in the multi-head attention layers were added to reduce over-fitting.

5.8.3 Results

For the evaluation of the models the following metrics and loss was used:

- **2D and 3D Dice Coefficient**: Employed to measure overlap between predicted and ground truth masks in both 2D and 3D models.
- **Jaccard Distance**: Used to measure dissimilarity between predicted and ground truth masks, emphasizing its usefulness for unbalanced datasets.
- **Binary Cross Entropy**: Employed to measure the dissimilarity between the predicted and actual masks in the models.
- **Precision**: Used to evaluate the model's accuracy in predicting positive instances (correctly identified lesions) among all predicted positive instances using a 0.5 threshold.
- **Dice Jaccard Cross Loss**: The combined loss function integrating Dice coefficient, Binary Cross Entropy, and Jaccard distance to optimize the model predictions.

2D prediction results

The 2D model was trained for 100 epochs, and the loss and Dice metric data can be found in Figure 5.56.

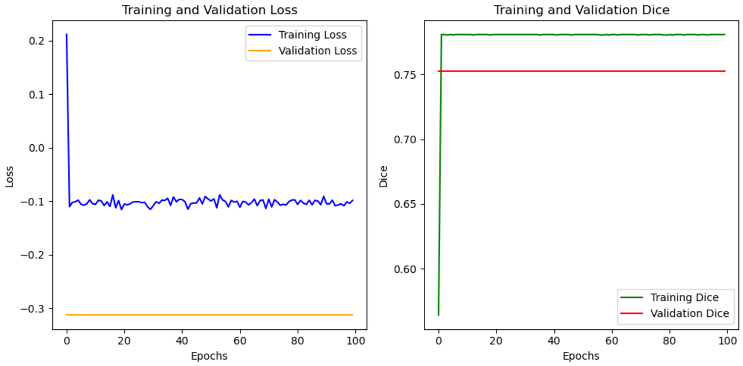


Figure 5.56: Training and validation loss and Dice coefficient metric for the 2D model.

An important observation is the consistent values for the validation loss and metrics, attributed to void predictions made by the model. The seemingly high Dice coefficient values can be traced back to the class imbalance within the dataset. Notably, precision across training, validation, and testing remained null. Addressing class imbalance through an expanded dataset and data augmentation would be needed for further testing of 2D models.

3D prediction results

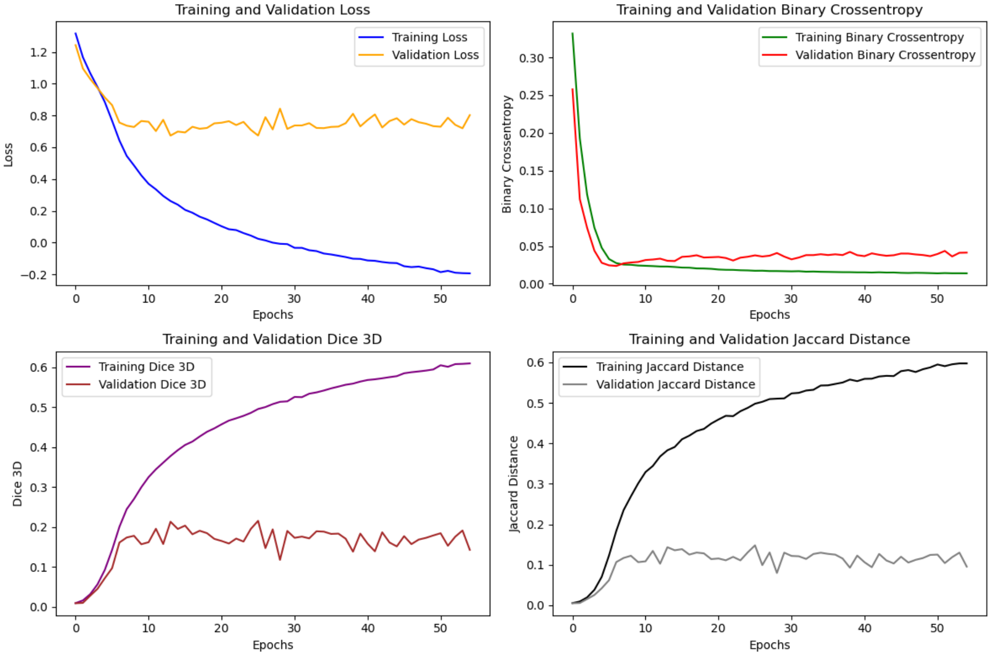


Figure 5.57: Training and validation loss and metrics for the 3D Model 1 for 55 epochs.

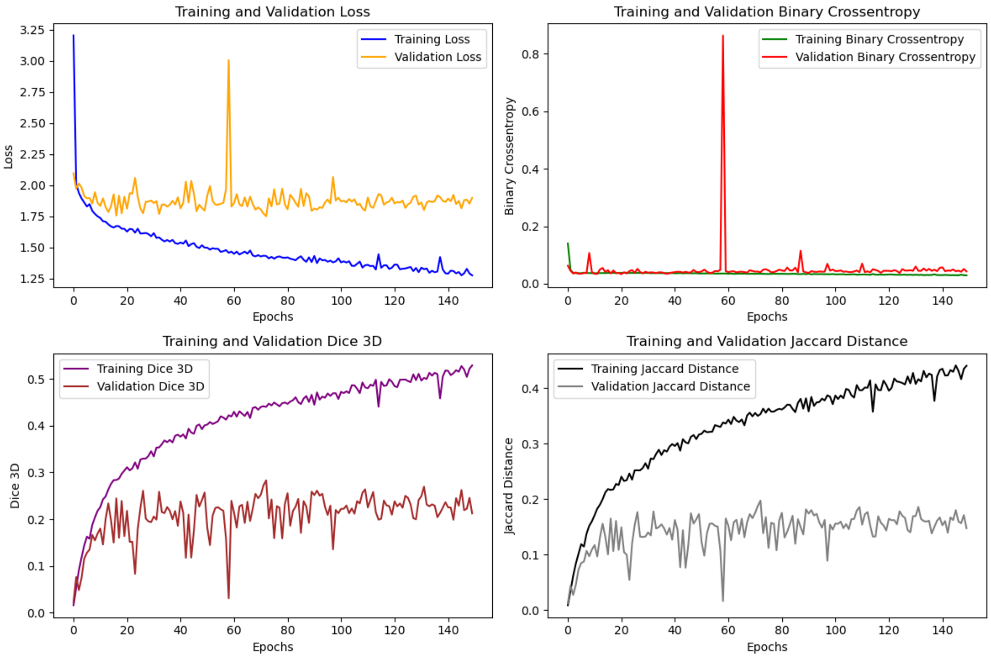


Figure 5.58: Training and validation loss and metrics for the 3D Model 2 for 150 epochs.

Figure 5.57 and 5.58 depict the learning curves for Model 1 and Model 2, respectively. Both models exhibit signs of overfitting. Despite rapid learning observed in background regions indicated by the steep decline in loss and BCE, this swift learning does not translate into favorable validation results.

Model 2, designed with additional regularizers and dropout to address the overfitting witnessed in Model 1, displays lower training metrics and improved validation outcomes. However, it's apparent that the model keeps struggling beyond the training dataset, especially from the initial epochs.

This behavior becomes more evident during the final evaluation of the training and test sets (5.42). While both models effectively learn the lesion features and segmentation within the training set, as evident from the Dice, Jaccard, and Precision metrics, this proficiency is not mirrored in the test set. Instead, the models seem to primarily learn the background.

		Loss	Dice	BCE	Jaccard	Precision
Train	**Model 1**	-0.1939	0.6101	0.0139	0.5976	0.8147
	Model 2	1.2767	0.5289	0.0291	0.4406	0.7116
Test	**Model 1**	1.0407	0.0920	0.0702	0.2007	0.6168
	Model 2	2.1692	0.1168	0.1512	0.0812	0.3706

Table 5.42: Loss and metrics of the training and testing set for Model 1 and 2.

The added complexity and regularizers added in Model 2 did not help the inference for the datasets outside training. The outcomes strongly suggest the necessity for more data and consideration of data augmentation techniques.

5.8.4 Discussion

In this investigation we focused on lesion segmentation within the ProstateNet's T2-weighted axial images using deep learning, and several key findings have emerged. Both 2D and 3D models were developed and tailored to accurately identify and delineate lesions within the prostate gland. However, despite initial promise, several challenges were encountered during the development and evaluation stages.

The preprocessing steps, including image denoising, resolution standardization, and cropping, laid the groundwork for the model development. Notably, the 2D model's configuration and training resulted in seemingly good losses and dice validation metrics, yet this masked a significant issue related to class imbalance. The over-representation of background labels skewed the model's predictions, highlighting the need for dataset expansion and careful consideration of class balance for improved testing accuracy.

The 3D models showed better results but still encountered challenges beyond the training dataset. Both 3D models demonstrated proficiency in learning lesion features during training but struggled to replicate this performance in the test set. Overfitting remained a persistent issue, even with the introduction of regularization techniques and dropout in Model 2.

For further development of the models with this dataset, there are several next steps and considerations to make:

- **Dataset expansion:** To explore data augmentation techniques and additional image sources to build a more robust model.
- **Addressing class imbalance:** Explore various techniques like oversampling, undersampling, or advanced methods such as SMOTE to rectify class imbalance issues.
- **Model improvements:** Investigate more complex and deeper 3D architectures, advanced regularization techniques, or hyperparameter tuning to improve model performance.
- **Ensemble learning and transfer learning:** Introducing ensemble learning methodologies and transfer learning from pre-trained models help inference in limited datasets such as this one.

Finally, it's also important to note that the current models were exclusively trained with T2 axial images. There exists potential in exploring additional modalities like Diffusion Weighted Imaging (DWI) or Apparent Diffusion Coefficient (ADC) channels found in a subset of the segementations. These additional channels might contribute to a more comprehensive understanding of prostate lesions and potentially mitigate overfitting observed in the current models. Thus, future work could explore and incorporate these modalities to enhance the segmentation performance and robustness of the models.

Chapter 6

Repurposing orphan data with self-supervision

Chapter summary

Conceptually, orphan data is data which does not necessarily have the necessary metadata for its correct cataloguing. For instance, a prostate mpMRI study which does not have an associated ISUP could be considered orphan for use case 2. In ethical and organizational terms, it may also be complicated to have access to clinical features which would allow computational models to be trained on these data, thus ”creating” large volumes of data which lack appropriate annotations and/or metadata for supervised learning. However, self-supervised learning (SSL), which does not require annotations or metadata, could be a helpful solution for this as it enables the training of rich feature extraction methods without the need for any form of image annotation. Here, we show how having relatively large amounts of ”orphan data” (here we simulate the orphan data by assuming that only part of the images in ProstateNet are annotated) can lead to models which outperform fully supervised classification learning models. We also show how SSL can be trained on 2D data and then applied to downstream supervised tasks using 3D volumes using multiple instance learning, making it ideal for radiology data.

6.1 Methods

6.1.1 Dataset

All retrospective ProstateNet cases avaliable on July 28th, 2023 were retrieved, constituting a total of 8,891 studies. 6,798 studies were used for SSL (1,722,978 DICOM files across 49,808 series). The remaining 2,093 studies were used for empirical validation of the SSL models with supervised learning and split into training (75%) and hold-out test sets (25%), and the training data was further divided into 5 folds for cross-validation. The empirical validation set contained all UC5 (biochemical recurrence) cases.

Supervised learning tasks for empirical validation

For the validation of self-supervised learning models, three binary tasks are considered (counts in Table 6.1):

- **PCa prediction** — confirmatory biopsy following mpMRI (a soft proxy for use case 1 — UC1)
- **PCa aggressiveness** — International Society of Urology Pathology (ISUP) grading [23] was split between ISUP=1 (clinically non-significant) and ISUP=2,3,4,5 (clinically significant; use case 2 — UC2)
- **Biochemical relapse** — prostate specific antigen concentration greater than 0.05ng/mL six months after radical prostatectomy (use case 5 — UC5)

Task	Sequences	Classification	Count
PCa prediction	T2	False	378
		True	1647
	mpMRI	False	343
		True	1467
csPCa prediction	T2	ISUP=1	397
		ISUP=2	667
		ISUP=3	314
		ISUP=4	126
		ISUP=5	126
	mpMRI	ISUP=1	324
		ISUP=2	614
		ISUP=3	286
		ISUP=4	113
		ISUP=5	114
Biochemical relapse prediction	T2	False	776
		True	127
	mpMRI	False	736
		True	96

Table 6.1: Case count for binary classification tasks.

DICOM processing

During SSL training (details below) DICOM image intensity values are normalised to be between 0 and 1. A random crop with at least 64 × 64 pixels is then extracted and resized to 128 × 128.

Volume conversion and processing for supervision tasks

DICOM files were converted to T2/DWI/ADC files as in chapter 5. A central crop with size $[128, 128, 20]$ is then used in training all models.

6.1.2 Self-supervised learning

Two different SSL methods are used — simple framework for contrastive learning of visual representations (SimCLR, which is contrastive) [17] and image-based joint-embedding predictive architecture (I-JEPA, which is generative) [6]. SimCLR is a contrastive learning approach trained by minimising the distance between the features of two views of the same image (the same image augmented in different ways). I-JEPA is a recent method based on the vision transformer (ViT) architecture [21] and the optimiziation is similar to that of the masked autoencoder (MAE) [29] — both I-JEPA and MAE train an encoder while trying to restore parts of the input which were masked; however, while the training protocol of MAE masks *parts of the image* and trains an auxiliary decoder model capable of predicting the masked parts, the I-JEPA training protocol masks *parts of the encoded feature space* and optimises an auxiliary model capable of predicting the masked parts in the feature space.

SimCLR

Encoder architecture. Two different architectures were tested with SimCLR — ResNet [30] and ConvNeXt [51].

The used ResNet architecture had 4 residual blocks, each consisting of increasing depth ($[64, 128, 256, 512]$) and each having a set number of residual layers ($[3, 3, 3, 2] + 1$ input layer, 12 layers in total). At each block, the input is processed by two sequential convolutional layers which first duplicate and the reduce the number of features to the original input size (similarly to ConvNeXt). After each residual block the resolution is halved using a 2 × 2 maximum pooling operation. Each convolutional operation is followed by a batch normalization operation [38] and a leaky ReLU activation [95]. The ConvNeXt architecture comprised of 4 ConvNeXt blocks with the same depths as the ResNet and each containing a set of ConvNeXt layers ($[3, 3, 9, 3] + 1$ input layer, 19 layers in total). Within each block, the number of features is quadruplicated as suggested by [51]. A 2 × 2 maximum pooling operation was performed after each ConvNeXt block.

Predictor architecture. Encoder (ResNet/ConvNeXt) outputs are linearly transformed using a linear predictor with structure $[512, 4096, 4096, 512]$, with batch normalization and leaky ReLU activations before each layer (the output is not transformed). This large predictor architecture is motivated in large part by the work of Garrido *et al.* stating that larger predictors are more beneficial for downstream performance [26].

Training and DICOM image sampling and augmentation. Each SimCLR model is trained with an AdamW optimizer [52] with weight decay of 0.00001 and a batch size of 256 (128 *per* GPU) for 100 epochs. The learning rate (0.0005 for ResNet and 0.0001 for ConvNeXt) was linearly increased over 5 epochs and was decreased using cosine decay during training. To generate batches, series were randomly sorted and a single image was sampled from each series and used to produce batches — this constitutes a ”series iteration”, and each epoch has 4 series iterations, i.e. in each epoch the model ”sees” 4 images from each series. The NT-Xent loss was used as suggested in [17] with a temperature of 0.1. To reduce the memory requirements of these models, FP16 mixed-precision was used during training.

View generation for SimCLR is achieved through the independent augmentation of the same image patch twice, each with two random augmentations (specified in Table 6.2). Originally, SimCLR extracts two non-overlapping patches from the input image [17]. However, we avoid doing this as it may lead to cases where one view has a lesion and the other does not, which may lead to mismatched information during training.

Augmentation	**Parameters**
Gaussian noise	$\sigma = 0.5$
Intensity shifting	offset $= [-0.25, 0.25]$
Intensity scaling	scale $= [-0.25, 0.25]$
Random bias field	coefficient $= [0, 0.15]$
Contrast	maximum $\gamma = 1.5$
Gaussian smoothing (indep. axes)	maximum $\sigma = 0.15$
Gibbs noise	$\alpha = [0, 0.5]$
Spike noise	intensity $= [-0.25, 0.25]$
Rician noise	$\sigma = 0.1$
Coarse dropout	Max. number of holes $= 16;$ Max. size of holes $= [12, 12]$
Translation (indep. axes)	$[-15, 15]$
Rotation (indep. axes)	$[-\frac{\pi}{12}, \frac{\pi}{12}]$
Shearing (indep. axes)	$[-0.25, 0.25]$
Scaling (indep. axes)	$[-0.15, 0, 15]$

Table 6.2: List of augmentations used during SimCLR training. ”indep. axes” means that this transform is applied independently for each axis.

I-JEPA

Encoder and decoder architecture. We follow the original paper for the encoder and decoder architectures [6], which are based on the ViT architecture [21] — the encoder and predictor are composed of 8 and 4 transformer blocks, respectively, and each transformer block has an embedding size of 512, 8 multi-attention heads and a multi-layer perceptron (MLP) with structure $[512, 1024, 512]$. A patch size of $[16, 16]$ was used.

Training and DICOM image sampling and augmentation. I-JEPA was trained with hyperparameters and image sampling routine identical to that of SimCLR, with a few exceptions — the learning rate was 0.001, and given that one the advantages of I-JEPA is that no augmentations are necessary, we have maintained this when training our models. For masking, we note that $[16, 16]$ patches for $[128, 128]$ images yield a total of 64 blocks. With this in mind, we mask 1-3 parts of the image, where each part has sides with 1-2 patches in length for reconstruction. As recommended, we mask an additional patch which is not reconstructed but is simply representative of missing information in the input image.

6.1.3 Supervised learning with multiple instance learning

Model specification

Each study $S \in \mathbb{R}^{n,h,w,c}$ is characterized as having n slices, height h, width w and c channels. To obtain a prediction p for S (Equation 6.1), it is necessary to apply a given feature extraction function F (Equation 6.2) producing a vector of size f to each slice and aggregate each score with function G to produce a vector (Equation 6.4). Prior to aggregation, a function A (Equation 6.3) can be used to produce an attention vector which is used to change the impact of different elements of $F(S)$ when aggregating (this vector sums to 1: in other words it is a simplex in $\mathbb{R}^{n,1}$). Finally, a binary prediction can be obtained using the prediction function P on the aggregation output (Equation 6.5; while other target specifications — namely multi-target, multiclass and regression targets — are possible, here only binary classification problems are considered).

$$p = P(A(F(S)) \odot G(F(S))) \tag{6.1}$$

$$F : \mathbb{R}^{n,h,w,c} \rightarrow \mathbb{R}^{n,f} \tag{6.2}$$

$$A : \mathbb{R}^{n,f} \rightarrow \mathbb{R}^{n,1} \tag{6.3}$$

$$G : \mathbb{R}^{n,f} \rightarrow \mathbb{R}^{1,f} \tag{6.4}$$

$$P : \mathbb{R}^{1,f} \rightarrow \mathbb{R} \tag{6.5}$$

In this deliverable, F and P are parameterized as the fitted SSL models and an MLP followed by a sigmoid activation function, respectively, while A is parameterized in 4 distinct manners (we define here f as the output of $F(S)$, and f_i and a_i for a given slice S_i as the the output of $F(S_i)$ and $A(S_i)$, respectively):

- **mean** — the simplest formulation of G is $G_{\text{mean}}(f) = \frac{1}{\sum_i^n a_i} \sum_i^n a_i \odot f_i$. In other words, G_{mean} is the weighted mean of $F(S)$, where each weight is produce by $A(f)$;
- **max** — G_{max} is similar to G_{mean}, but rather than summing over all $a_i f_i$, the maximum value of each feature along the slice dimension is extracted;
- **vocabulary** — here, G is split into a two-step process: G_{class}, parameterized as an MLP followed by a softmax layer M, first classifies f_i into a term v_i in a k-sized vocabulary. All term predictions v_i are summed after being multiplied by a_i and divided by $\sum_i^n a_i$. In other words, $G_{\text{vocabulary}}(f) = \frac{1}{\sum_i^n a_i} \sum_i^n a_i \odot G_{\text{class}}(f_i)$, where $G_{\text{class}}(f_i) = \text{softmax} M(f_i)$. Here, k is a hyperparameter and 3 different values are tested: 10, 25 and 50.
- **transformer** — $G_{transformer}$ is a 2 block transformer which uses the 512-sized embeddings produced by the SSL models for each slice as tokens. The multi-head attention consists of 4 heads (128 features *per* head), a MLP with structure [1024, 512] and using a classification token.

For the mean, max and vocabulary models, F was an MLP with structure $[1024, 512]$. All models used an 512 layer MLP as P, and all MLP layers excluding the last are followed by layer normalization, GELU activations and a dropout layer with 25% probability.

Model training

Models were trained using AdamW [52] with a weight decay of 0.05 (mean, max, vocabulary) or 0.3 (transformer) and a maximum learning rate of 0.00001 for 50 epochs with a batch-size of 8. Learning rate was increased linearly for 5 epochs and decreased with cosine decay during training. Images were randomly augmented during training with random flips, affine, intensity and Rician noise transforms (Table 6.3). The best performing model according to its validation loss was used for evaluation on the hold-out test set.

Augmentation	Parameters
Rician noise	$\sigma = 0.02$
Affine transform	Translation range = $[4, 4, 1]$ Rotation range = $[\frac{\pi}{16}, \frac{\pi}{16}, \frac{\pi}{16}]$
Flip	Along all axes (x, y, z)
Contrast	maximum $\gamma = 1.5$
Intensity shifting	offset = $[-0.1, 0.1]$
Intensity scaling	scale = $[-0.1, 0.1]$

Table 6.3: List of augmentations used during fully supervised and multiple instance learning.

6.1.4 Fully supervised learning

To better understand the benefits of MIL-SSL, we train a VGG-based model [80] from scratch using the same training and validation protocol as other models. The architecture and training parameters of this model are identical with those specified in Cchapter 5 and a batch size of 16.

6.1.5 Model validation

The AUC of all models is estimated using the area under the receiver-operating characteristic (AUROC/AUC) as a metric and 5-fold cross validation. A 25% hold-out test set is then used to evaluate the generalisability of all models.

6.1.6 Learning curve analysis

To better understand the data requirements of each model, we subsample the training data during cross-validation to yield datasets with 10%, 25%, 50% and 70% of all training data. Models are then trained with these data subsets to determine how performance evolves as the amount of training data is increased.

6.2 Results

Here, the focus of this analysis will be on the empirical assessment (the three enunciated classification tasks).

6.2.1 Cross-validation performance

Cross-validation performance (Figure 6.1) — the AUC for the lowest training loss observed during training — shows that UC5 is a particularly complicated problem, with average AUC values of 66% and a considerable amount of variability in the performance (VGG model).

Regarding SSL models, it is clear the I-JEPA is not capable of learning aspects of the image which are relevant for classification (we further elaborate on this in the Discussion of this chapter). Nonetheless, performance for both SimCLR-based encoders appears to be consistent for both UC1 and UC2, where little differences between both can be detected. In terms of the performance of MIL methods, we note that few fluctuations are observed when it comes to the choice of the method. Indeed, much like what has been exposed in Chapter 5, performance is mostly determined by the sequences used in classification.

6.2.2 Hold-out test set performance

Given that CV performance is a biased estimate (the reported CV performance is calculated for the model corresponding to the best observed validation loss), a hold-out test set was used to produce more realistic estimates of performance and generalisability. We note here that, unlike VGG models, SSL-MIL-based methods are capable of generalization when compared with FSL methods across all use cases — as visible in Figure 6.3a and Figure 6.2, VGG models consistently suffer drops in performance (particularly for UC5), whereas SSL-MIL models are capable of maintaining their performance. To further understand the unbiased model performance of SSL-MIL methods, the performing model/fold from the set of SSL-MIL models and the best FSL fold was picked for each use case. As demonstrated in Figure 6.3b, the best performing

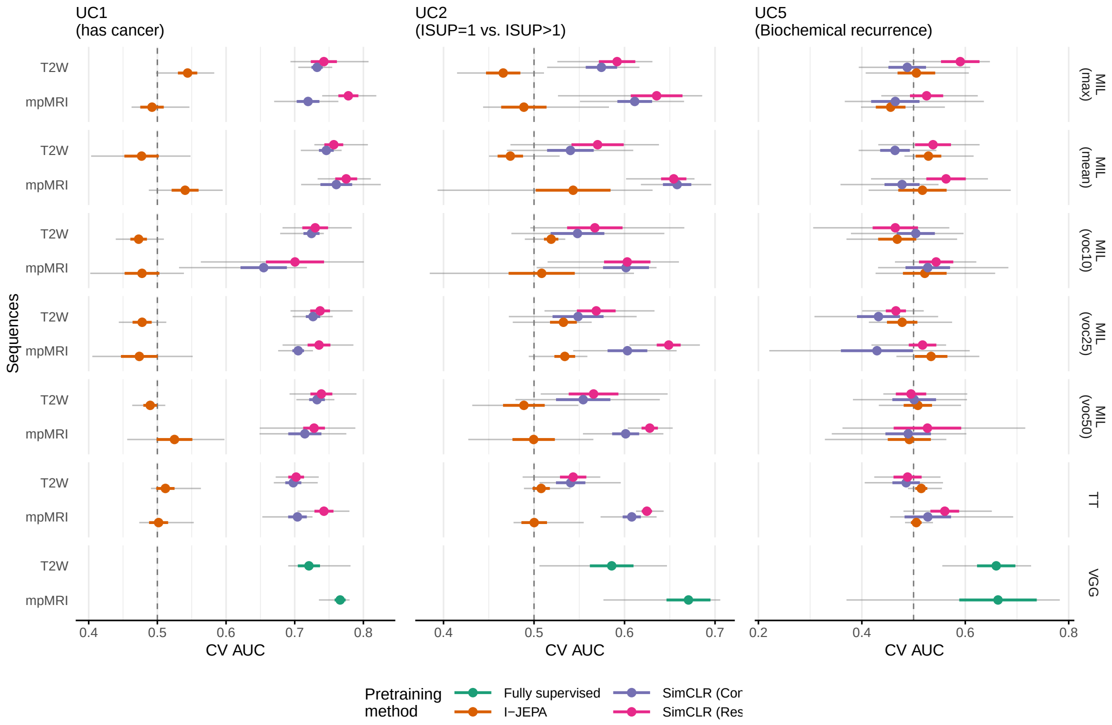


Figure 6.1: Cross-validation performance for different SSL-MIL models and comparison with a baseline VGG fully supervised model. Points represent the average AUC, colourful horizontal lines represent the standard error range, and the black horizontal lines represent the full performance range (minimum and maximum).

SSL-MIL models (TTable 6.4) are at least as good as the best FSL models, with superior AUC in both UC1-T2W models and UC2-mpMRI models (it should be noted that, despite being large, this difference is not statistically significant for the latter). This last case — selecting the best performing models — is a more realistic scenario and is closer to a real application scenario, where developers would select the best performing model and use it on their data.

Use case	Sequences	SSL method	MIL method	Test value	Fold
UC1	T2W	SimCLR (ResNet)	MIL (max)	0.7652	2
UC1	mpMRI	SimCLR (ConvNeXt)	MIL (mean)	0.7715	1
UC2	T2W	SimCLR (ResNet)	MIL (voc10)	0.5737	0
UC2	mpMRI	SimCLR (ConvNeXt)	MIL (mean)	0.6932	1

Table 6.4: Best performing SSL-MIL models.

6.2.3 Learning curve analysis

To better understand the data requirements of SSL-MIL and FSL, learning curves were calculated using different amounts of training data — 10%, 25%, 50% and 75%. For this we focused on two well performing models from previous tasks (SimCLR ResNet with MIL (mean) or MIL (max)). As visible in Figure 6.4a, the average performance of these models — particularly the MIL (mean) models — is generally better than the FSL VGG models. If the best folds are selected, as earlier, it is clear that MIL-SSL outperforms the FSL VGG models both in terms of data efficiency and performance and this is particularly true for T2W models (Figure 6.4b). Finally, it should be noted that in all cases the capacity of the models has not been achieve — in other words, more annotated data can be beneficial.

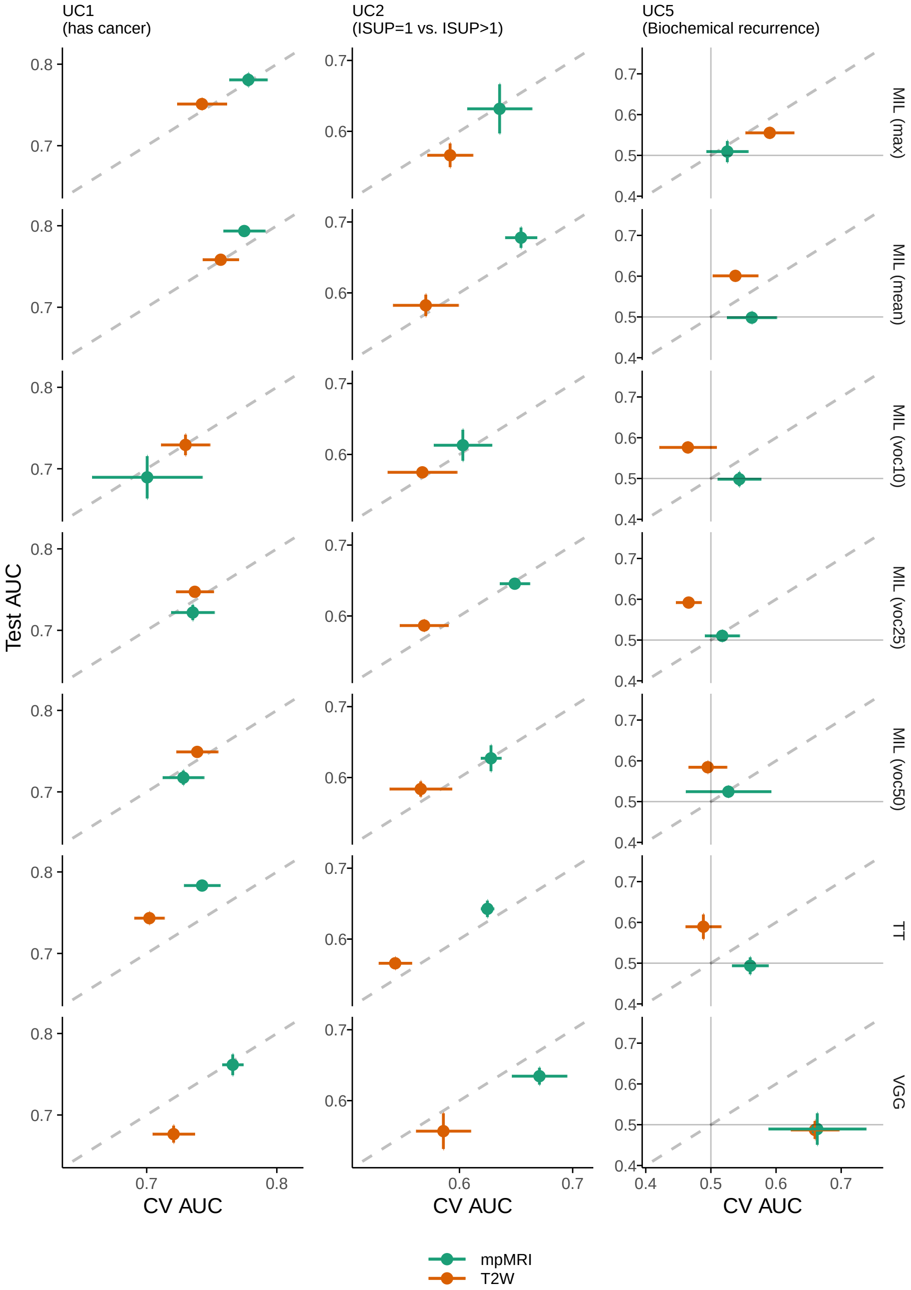


Figure 6.2: Comparison of CV and hold-out test set AUC for different folds. Both vertical and horizontal lines represent the standard error and points represent the average performance.

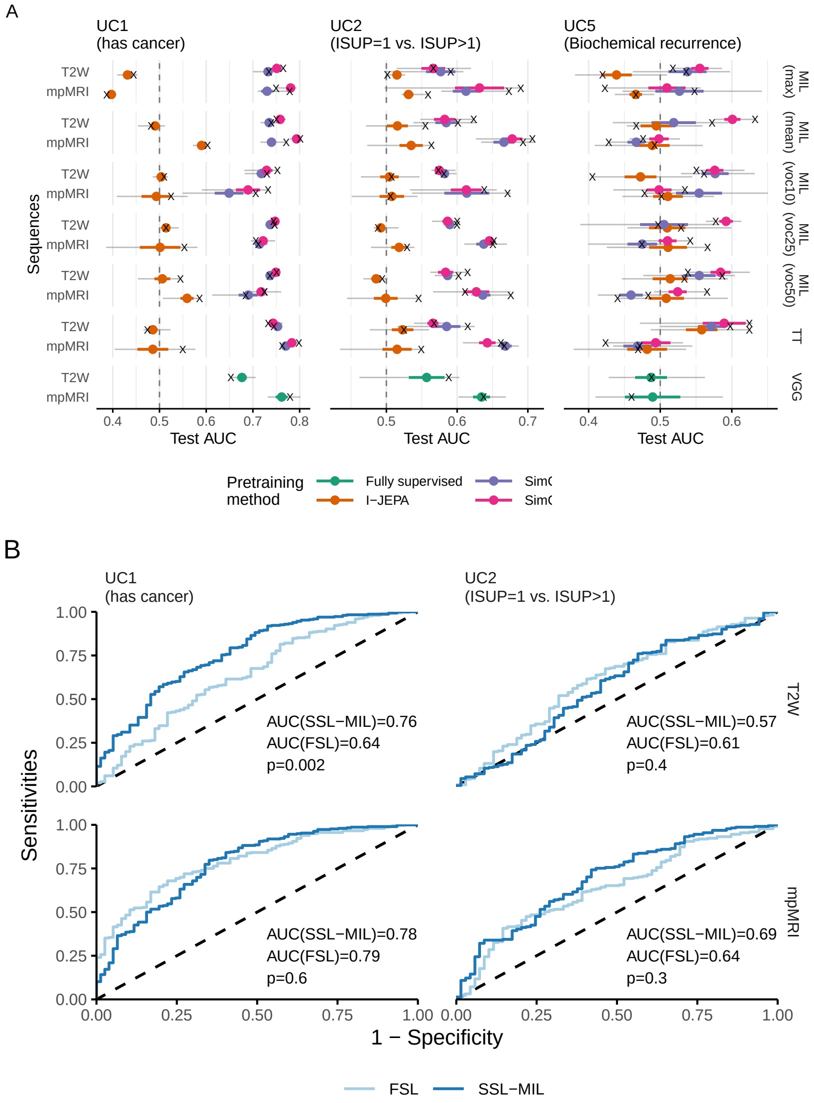


Figure 6.3: **A**. Hold-out test set performance for different SSL-MIL models and comparison with a baseline VGG fully supervised model. Points represent the average AUC, colourful horizontal lines represent the standard error range, and the black horizontal lines represent the full performance range (minimum and maximum). Crosses represent the test AUC for the best performing model during CV. **B** Receiver operating curves for the best performing SSL-MIL (dark blue) and FSL models (light blue). The p-values were calculated using a 2,000 sample bootstrap.

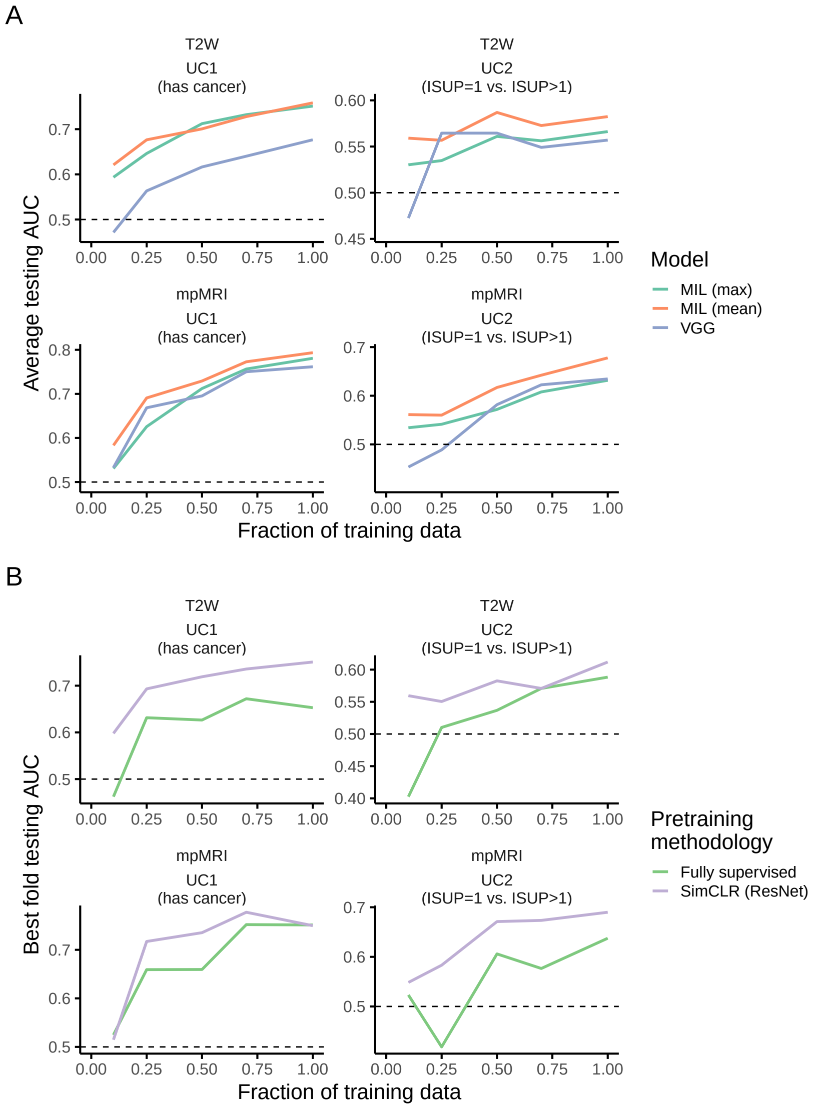


Figure 6.4: **A**. Learning curve for the average testing AUC for different use cases and sequence inputs.

6.3 Discussion

In this chapter, it is shown that self-supervised learning, as a paradigm, can lead to performance improvements even when annotated data is scarce.

In particular, we show how training models in 2D is sufficient to achieve good performance in 3D models. This creates new possibilities, as most of the available image data is two-dimensional. While an important assessment — showing that natural image models can be transferred in this setting — previous works, such as those developed around the Segment Anything Model [45] have shown that models trained on natural images can perform well on medical images but do not surpass models trained specifically for these tasks [54, 57]. This similar to what was shown previously for histopathology, where models pre-trained on histopathology outperformed those pre-trained on much larger natural image datasets [72]. In essence, while pre-training on large image datasets may be helpful, previous works do not point in this direction.

The case of UC5 — detection of biochemical recurrence from mpMRI data — remains a complicated challenge. While some demonstrations that there exists some discriminatory power are offered here (T2W + mean MIL models, for instance; Figure 6.2), the performance is remarkably low. From these data and the analysis here offered, it is unlikely that deep learning models will overcome this.

While both contrastive encoders worked relatively well, an important failure case is I-JEPA, a method based on masking. As suggested by Huang *et al.* and demonstrated here, it is possible that using models which assume redundancy in the feature space of different patches in the same image is not useful in medical imaging applications where lesions are relatively small [36] — if the lesion is masked, there is no good reason to assume that the rest of the prostate should have sufficient information to predict the presence of a lesion.

Chapter 7

Discussion

In this deliverable, we have outlined a number of predictive strategies which depend on different types and qualities of data and provided a comprehensive analysis of the failings of these models. We also detail how different unforeseen obstacles — particularly the lack of strict adherence to the data quality requirements initially defined, which resulted in several issues, as reported at the start of this deliverable — were handled and in which way this affected the expected course of this process.

Arriving at automatic data curation as an iterative and collaborative process. Starting from a disorganised dataset proved challenging for a number of reasons — firstly, the development of an automatic sequence classification algorithm was necessary, and the development of several heuristics was also a necessity to faithfully convert DICOM to volume files which are usable in downstream model development. This was possible mostly due to an iterative process (several different attempts had to be made to ensure that a robust protocol had been achieved) relying on inter-institutional collaboration (the identification of different issues with the images depended heavily on the pooled past experiences of different collaborators, allowing us to arrive to more concrete and usable solutions).

Both radiomics and deep-learning classification models generalise well but show relatively poor performance. We present here important trends and a reasonable starting point for more ambitious models, but it should be noted that, in their current state, there is very limited application to a clinical context due to the low exhibited performance. However, and as mentioned, we hope that this constitutes a good starting point for the development of better performing models by providing concrete research directions.

Self-supervsion can lead to more robust and data-efficient models. Self-supervision is a relatively recent paradigm and we show here that it can be used to create models which generalise better and are more data efficient. This should motivate the collection of orphan data, which could be used to create more robust models.

Prostate segmentation models are robust but lesion segmentation/detection is still problematic. While prostate segmentation (whole gland and by zone) is relatively easy, we show here that lesion segmentation/detection models underperform; this creates significant hurdles in terms of what can be achieved by other approaches which require the existence of relatively detailed segmentation models (i.e. radiomics using lesion segmentations; deep-learning models which use the lesion region-of-interest as input).

7.1 Limitations and setbacks

The main limitation we identified was the relatively low abundance of lesion segmentations, which prevented the development of good lesion segmentation models which could have greatly improved the performance of machine-learning models. Additionally, it should be noted that some of these annotations are of low quality and contain obvious mistakes, highlighting an unforeseen necessity — a laborious process of quality control which was not anticipated by our team. While we tried to overcome this through manual inspection of annotations, the time spent doing so was significant, similarly to the time spent creating an automatic curation pipeline.

7.2 Future steps

Moving forward, it should be noted that, while lesion segmentation models do not perform particularly well, lesion detection for some instances is possible — indeed, we are able to detect approximately 78% of all lesions in the testing data at an IoU of 10% as illustrated in section 5.7. Taking this into account, it should be possible to use a protocol similar to that described in Bosma *et al.* — in this work, the authors used a similar lesion detection model to propagate annotations in data which had no segmentation maps but had information on the number of lesions per case [10]. Given that we are particularly interested in the index lesion, it would be possible to keep the highest probability lesion on each weak prediction for all studies belonging to use case 2 and create a relatively large dataset of low quality segmentation annotations which could lead to the training of better radiomics and deep-learning models.

Concerning the self-supervised learning avenue, there should also be benefit to moving the pre-training from two-dimensions to three-dimensions as this would allow us to capture higher order relationships in these models. This would have the added benefit of also making the models more easy to transfer and would make more complicated multiple-instance learning-based approaches unnecessary.

Bibliography

[1] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 683–687. IEEE, 2019.

[2] Lisa C Adams, Marcus R Makowski, Günther Engel, Maximilian Rattunde, Felix Busch, Patrick Asbach, Stefan M Niehues, Shankeeth Vinayahalingam, Bram van Ginneken, Geert Litjens, and Keno K Bressem. Prostate158 - an expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. *Comput. Biol. Med.*, 148:105817, September 2022.

[3] Lisa C. Adams, Marcus R. Makowski, Günther Engel, Maximilian Rattunde, Felix Busch, Patrick Asbach, Stefan M. Niehues, Shankeeth Vinayahalingam, Bram van Ginneken, Geert Litjens, and Keno K. Bressem. Prostate158 - an expert-annotated 3t mri dataset and algorithm for prostate cancer detection. *Computers in Biology and Medicine*, 148:105817, 2022.

[4] Nader Aldoj, Federico Biavati, Florian Michallek, Sebastian Stober, and Marc Dewey. Automatic prostate and prostate zones segmentation of magnetic resonance images using densenet-like u-net. *Scientific Reports*, 10, 08 2020.

[5] Samuel G Armato, 3rd, Henkjan Huisman, Karen Drukker, Lubomir Hadjiiski, Justin S Kirby, Nicholas Petrick, George Redmond, Maryellen L Giger, Kenny Cha, Artem Mamonov, Jayashree Kalpathy-Cramer, and Keyvan Farahani. PROSTATEx challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *J Med Imaging (Bellingham)*, 5(4):044501, October 2018.

[6] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised learning from images with a Joint-Embedding predictive architecture. *arXiv*, January 2023.

[7] 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.

[8] Biotronics3D, FORTH, ADVANTIS, and QUIBIM. *ProCAncer-I Data Upload User Manual*. The ProCAncer-I Project, June 2022. Available at `https://prostatenet.eu/wp-content/uploads/2022/04/procancer-i-data-upload-user-manual_v1.2.1.pdf`.

[9] J. S. Bosma, A. Saha, M. Hosseinzadeh, I. Slootweg, M. de Rooij, and H. Huisman. Semisupervised Learning with Report-guided Pseudo Labels for Deep Learning-based Prostate Cancer Detection Using Biparametric MRI. *Radiol Artif Intell*, 5(5):e230031, Sep 2023.

[10] Joeran S Bosma, Anindo Saha, Matin Hosseinzadeh, Ivan Slootweg, Maarten de Rooij, and Henkjan Huisman. Semi-supervised learning with report-guided pseudo labels for deep learning-based prostate cancer detection using biparametric MRI. *Radiology: Artificial Intelligence*, page e230031, July 2023.

[11] Rossana Buongiorno, Danila Germanese, Leonardo Colligiani, Salvatore Claudio Fanni, Chiara Romei, and Sara Colantonio. Chapter 9 - artificial intelligence for chest imaging against covid-19: an insight into image segmentation methods. In Parag Chatterjee and Massimo Esposito, editors, *Artificial Intelligence in Healthcare and COVID-19*, Intelligent Data-Centric Systems, pages 167–200. Academic Press, 2023.

[12] Rossana Buongiorno, Danila Germanese, Chiara Romei, Laura Tavanti, Annalisa De Liperi, and Sara Colantonio. Uip-net: A decoder-encoder cnn for the detection and quantification of usual interstitial pneumoniae pattern in lung ct scan images. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *Pattern Recognition. ICPR International Workshops and Challenges*, pages 389–405, Cham, 2021. Springer International Publishing.

[13] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[14] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[15] Jinghui Chen, Saket Sathe, Charu Aggarwal, and Deepak Turaga. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM international conference on data mining*, pages 90–98. SIAM, 2017.

[16] 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.

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. Simclr: A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607, 2020.

[18] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

[19] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Cham, 2016. Springer International Publishing.

[20] Renato Cuocolo, Arnaldo Stanzione, Anna Castaldo, Davide Raffaele De Lucia, and Massimo Imbriaco. Quality control and whole-gland, zonal and lesion annotations for the prostatex challenge public dataset. *European Journal of Radiology*, 138:109647, 2021.

[21] 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.

[22] Loïc Duron, Daniel Balvay, Saskia Vande Perre, Afef Bouchouicha, Julien Savatovsky, Jean-Claude Sadik, Isabelle Thomassin-Naggara, Laure Fournier, and Augustin Lecler. Gray-level discretization impacts reproducible mri radiomics texture features. *PLOS ONE*, 14(3):1–14, 03 2019.

[23] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, Peter A Humphrey, and Grading Committee. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. *Am. J. Surg. Pathol.*, 40(2):244–252, February 2016.

[24] William Falcon and The PyTorch Lightning team. PyTorch lightning, March 2019.

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

[26] Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv*, 2022.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[28] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv*, November 2021.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv*, December 2015.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank adaptation of large language models. *arXiv*, June 2021.

[33] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation networks. *arXiv*, September 2017.

[34] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[35] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[36] Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digit Med*, 6(1):74, April 2023.

[37] Lauren M Hurwitz, Ilir Agalliu, Demetrius Albanes, Kathryn Hughes Barry, Sonja I Berndt, Qiuyin Cai, Chu Chen, Iona Cheng, Jeanine M Genkinger, Graham G Giles, Jiaqi Huang, Corinne E Joshu, Tim J Key, Synnove Knutsen, Stella Koutros, Hilde Langseth, Sherly X Li, Robert J MacInnis, Sarah C Markt, Kathryn L Penney, Aurora Perez-Cornago, Thomas E Rohan, Stephanie A Smith-Warner, Meir J Stampfer, Konrad H Stopsack, Catherine M Tangen, Ruth C Travis, Stephanie J Weinstein, Wu Lang PhD, Eric J Jacobs, Lorelei A Mucci, Elizabeth A Platz, Michael B Cook, and Prostate Cancer Cohort Consortium (PC3) Working Group. Recommended definitions of aggressive prostate cancer for etiologic epidemiologic research. *J. Natl. Cancer Inst.*, 113(6):727–734, June 2021.

[38] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv*, February 2015.

[39] 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*, 18(2):203–211, Dec 2020.

[40] 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.

[41] 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.

[42] Ana Jimenez-Pastor, Rafael Lopez-Gonzalez, Belén Fos-Guarinos, Fabio Garcia-Castro, Mark Wittenberg, Asunción Torregrosa-Andrés, Luis Marti-Bonmati, Margarita Garcia-Fontes, Pablo Duarte, Juan Pablo Gambini, et al. Automated prostate multi-regional segmentation in magnetic resonance using fully convolutional neural networks. *European Radiology*, pages 1–10, 2023.

[43] 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.

[44] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *ML Workshop*, 1992.

[45] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv*, April 2023.

[46] Miron B. Kursa and Witold R. Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010.

[47] 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.

[48] Vincenzo Lagani, Giorgos Athineou, Alessio Farcomeni, Michail Tsagris, and Ioannis Tsamardinos. Feature selection with the r package mxm: Discovering statistically-equivalent feature subsets. *arXiv preprint arXiv:1611.03227*, 2016.

[49] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327, February 2020.

[50] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Computer-aided detection of prostate cancer in mri. *IEEE Transactions on Medical Imaging*, 33(5):1083–1092, 2014.

[51] 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.

[52] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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

[54] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv*, April 2023.

[55] 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.

[56] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks*, pages 52–59. Springer, 2011.

[57] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: An experimental study. *Med. Image Anal.*, 89:102918, October 2023.

[58] Natalie McGauran, Beate Wieseler, Julia Kreis, Yvonne-Beatrice Schüler, Heike Kölsch, and Thomas Kaiser. Reporting bias in medical research - a narrative review. *Trials*, 11:37, April 2010.

[59] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

[60] 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.

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

[62] 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.

[63] 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.

[64] Ramakrishnan Muthukrishnan and R Rohini. Lasso: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE international conference on advances in computer applications (ICACA)*, pages 18–20. IEEE, 2016.

[65] Hai-Long Nguyen, Yew-Kwong Woon, Wee-Keong Ng, and Li Wan. Heterogeneous ensemble for feature drifts in data streams. In Pang-Ning Tan, Sanjay Chawla, Chin Kuan Ho, and James Bailey, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–12, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[66] 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.

[67] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8026–8037. Curran Associates Inc., Red Hook, NY, USA, December 2019.

[68] 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.

[69] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems*, 29, 2016.

[70] Milos Radovic, Mohamed Ghalwash, Nenad Filipovic, and Zoran Obradovic. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, 18(1), 2017.

[71] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv*, October 2017.

[72] Indranil Ray, Geetank Raipuria, and Nitin Singhal. Rethinking ImageNet pre-training for computational histopathology. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2022:3059–3062, July 2022.

[73] Nuno M. Rodrigues, Sara Silva, Leonardo Vanneschi, and Nickolas Papanikolaou. A comparative study of automated deep learning segmentation models for prostate mri. *Cancers*, 15(5), 2023.

[74] Yaniv Romano, Matteo Sesia, and Emmanuel J Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, 33, 2020.

[75] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[76] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015.

[77] Leonardo Rundo, Changhee Han, Yudai Nagano, Jin Zhang, Ryuichiro Hataya, Carmelo Militello, Andrea Tangherloni, Marco S. Nobile, Claudio Ferretti, Daniela Besozzi, Maria Carla Gilardi, Salvatore Vitabile, Giancarlo Mauri, Hideki Nakayama, and Paolo Cazzaniga. Use-net: Incorporating squeeze-and-excitation blocks into u-net for prostate zonal segmentation of multi-institutional mri datasets. *Neurocomputing*, 365:31–43, 2019.

[78] Anindo Saha, Jasper J. Twilt, Joeran S. 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. 2022.

[79] Ivo G Schoots, Daniel F Osses, Frank-Jan H Drost, Jan F M Verbeek, Sebastiaan Remmers, Geert J L H van Leenders, Chris H Bangma, and Monique J Roobol. Reduction of MRI-targeted biopsies in men with low-risk prostate cancer on active surveillance by stratifying to PI-RADS and PSA-density, with different thresholds for significant disease. *Transl. Androl. Urol.*, 7(1):132–144, February 2018.

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

[81] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[82] 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.

[83] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. `https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer,`, 2014. Accessed: 2023-7-23.

[84] 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.

[85] Jose M Castillo T, Muhammad Arif, Wiro J Niessen, Ivo G Schoots, and Jifke F Veenland. Automated classification of significant prostate cancer on MRI: A systematic review on the performance of machine learning applications. *Cancers*, 12(6), June 2020.

[86] Saeid Asgari Taghanaki, Yefeng Zheng, S Kevin Zhou, Bogdan Georgescu, Puneet Sharma, Daguang Xu, Dorin Comaniciu, and Ghassan Hamarneh. Combo loss: Handling input and output imbalance in multi-organ segmentation. *Comput. Med. Imaging Graph.*, 75:24–33, July 2019.

[87] 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.

[88] 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.

[89] 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.

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

[91] 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.

[92] Yu Wang, Igor V. Tetko, Mark A. Hall, Eibe Frank, Axel Facius, Klaus F.X. Mayer, and Hans W. Mewes. Gene selection from microarray data for cancer classification—a machine learning approach. *Computational Biology and Chemistry*, 29(1):37–46, 2005.

[93] Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020.

[94] Isabella Sc Williams, Aoife McVey, Sachin Perera, Jonathan S O'Brien, Louise Kostos, Kenneth Chen, Shankar Siva, Arun A Azad, Declan G Murphy, Veeru Kasivisvanathan, Nathan Lawrentschuk, and Mark Frydenberg. Modern paradigms for prostate cancer detection and management. *Med. J. Aust.*, 217(8):424–433, October 2022.

[95] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv*, May 2015.

[96] 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.

[97] 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.

[98] Rongjian Zhao, Buyue Qian, Xianli Zhang, Yang Li, Rong Wei, Yang Liu, and Yinggang Pan. Rethinking dice loss for medical image segmentation. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 851–860, November 2020.

[99] Qikui Zhu, Bo Du, Baris Turkbey, Peter L Choyke, and Pingkun Yan. Deeply-supervised CNN for prostate segmentation. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 178–184, May 2017.