D5.2

Pre-processing tools

Related Work Package	**WP5 – Development of the Master models**
Related Task	Task 5.2. - Development of pre-processing pipelines on the platform
Lead Beneficiary	FORTH
Contributing Beneficiaries	CNR, CF, QTIM
Document version	**v.**1.0
Deliverable Type	**Report**
Distribution level	**Public**
Contractual Date of Delivery	31/07/2022
Actual Date of Delivery	31/07/2022

Authors	Kostas Marias [FORTH], Sara Colantonio [CNR], José Almeida [CF], Katerina Nikiforaki [FORTH], Eugenia Mylona [FORTH], Eleftherios Trivizakis [FORTH], George Manikis [FORTH], Dimitris Zaridis [FORTH], Elisavet Stamoulou [FORTH], Katerina Dovrou [FORTH], Charalampos Kalatzopoulos [FORTH], Maria Antonietta [CNR], Eva Pachetti [CNR], Gianluca Carloni [CNR], Claudia Caudai [CNR], Danila Germanese [CNR], Nikolaos Tachos [FORTH].
Contributors	-
Reviewers	Kalpathy-Cramer, Jayashree [QTIM], Ana Jiménez Pastor [QUIBIM], M. Tsiknakis [FORTH],

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

Version history

Version	Description	Date completed
0.1	Preparation of deliverable ToC [FORTH]	29/04/2022
0.2	Section 1 updates [FORTH]	10/05/2022
0.3	Section 2 updates [FORTH]	29/05/2022
0.4	Section 4,5 updates [FORTH]	17/06/2022
0.5	Section 6 updates [CNR]	29/06/2022
0.6	Section 7,8 updates [FORTH]	06/07/2022
0.7	Section 3 updates [CF]	20/07/2022
0.8	Deliverable refinements and internal review [FORTH, CNR, CF]	22/07/2022
0.9	Deliverable review from consortium reviewers	25/07/2022
1.0	Final version for submission [FORTH]	30/07/2022

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Disclaimer

Executive summary.

ProCAncer-I aims to develop novel AI models in order to add precision in the management of prostate cancer patients by providing diagnostic tools able to reduce overdiagnosis and overtreatment. To this end, a unique dataset from the project's clinical sites is being assembled making sure that image diversity and variability is properly addressed, which would enhance models' generalizability. That said, most current multi-centric studies employ harmonization methods either in centralized or federated architectures since it is still impossible to control or predict all sources of variation. It should be stressed that there is currently no gold standard to harmonize imaging data in a multi-centric setting towards an optimal radiomics pipeline. Instead, there are research trends that have shown promising results based on which the ProCAncer-I consortium has implemented a number of pre-processing functionalities that can be used to experiment and optimize the ProCAncer-I AI modeling tasks (Deep Learning and Radiomics based). The key achievements and findings of the work presented in this deliverable are summarized below:

Regarding bias field signal correction, an image filter has been developed to correct the bias field caused by magnetic field inhomogeneities during the acquisition process. Such inhomogeneities can alter the textural descriptors in the radiomics extraction process hampering the planned modeling efforts. The developed tool incorporates a process to compute the optimal filter parameters which may be beneficial to promote standardized settings. The presented method is assessed in two public datasets while future work will investigate the added value of the method in optimizing the planned AI models using the ProstateNet dataset.

Motion-related artifacts and spatial inconsistencies e.g. due to patient motion can have a detrimental effect when executing radiomics pipelines in multi-parametric MRI (mpMRI). For this reason, a non-rigid registration tool is incorporated in the ProstateNet platform (available in the environment of AI Vault) in order to align images and avoid errors in feature computation due to geometrical inconsistencies. Initial results favor the use of rigid registration only for the needs of the project.

A well-known methodology to reduce image value variability emanating from different imaging systems and protocols is image enhancement. A Regional Adaptive Contrast Limited Adaptive Histogram Equalization (RACLAHE) tool was implemented and integrated in the ProstateNet platform. Moreover, the tool was compared and tested on various U-Net based architectures, which were developed for prostate regions segmentation demonstrating initial encouraging results. Future work within the project will focus on careful assessment of the added value of this histogram equalization technique on the multi-centric modeling tasks.

When collecting a large number of images from diverse clinical sites there is always the need to address the problem of noise which is also present when high-resolution examinations are obtained with a fast acquisition protocol. In the context of D5.2 a novel Deep Learning noise reduction tool (deep learning based model) was developed and assessed in public datasets. Since

such datasets are relatively small and well-curated the added value of this preprocessing step will be assessed in the fully-fledged ProstateNet repository.

The last important family of tools focus on normalization either before or after the extraction of radiomics features from a given MRI dataset. In terms of the former, four different normalization approaches (fat-based, muscle-based, double tissue non-linear, and single tissue non-linear) are presented, that are able to reduce image appearance variability across vendors. In parallel, a radiomics normalization pipeline based on the well-known ComBat method is also incorporated in the ProstateNet platform, and initial assessment on the PI-CAI dataset demonstrates that variability in texture measures across vendors is significantly reduced. Last, radiomic feature stability was assessed on the PI-CAI dataset indicating that the use of N4 filter for bias field correction can be safely applied to the MR images prior to radiomic feature extraction while noise reduction and, particularly, image enhancement filters should be treated with caution when developing radiomic-based models. Further, extensive experiments using single or combinations of the developed pre-processing tools are planned within the wider AI modeling tasks in WP5 and WP6.

Table of Contents

List of Abbreviations

Abbreviation	Explanation
ADAM	ADAptive Moment estimation
ADC	Apparent Diffusion Coefficient
AGCCPF	Adaptive Gamma Correction with Color Preserving Framework
AGCWD	Adaptive Gamma Correction with Weighting Distribution
AHE	Adaptive Histogram Equalization
ASD	Average Surface Distance
AUC	Area Under Curve
BA	Balanced Accuracy
BBHE	Bi-Histogram Equalization
BPDHE	Brightness Preserving Dynamic Histogram Equalization
CCC	Concordance Correlation Coefficient
CDF	Cumulative Distribution Function
CLAHE	Contrast Limited Adaptive Histogram Equalization
ComBat	Combine Batches
CrAE	Convolutional Autoencoder with residual Connections
CT	Computed Tomography
DL	Deep Learning
DnCNN	Denoising Convolutional Neural Network
DrCNN	Denoising Convolutional Neural Network with residual connections
DS	Dice Score index
ERC	EndoRectal Coil
FC-CLAHE	Fuzzy Clipped Contrast Limited Adaptive Histogram Equalization
FWHM	Full Width at Half Maximum
HBV	High b-value
HD	Hausdorff Distance
HE	Histogram Equalization
ICC	Intraclass Correlation Coefficient
IoU	Intersection over Union
MMBEBHE	Minimum Mean Brightness Error Bi-Histogram Equalization
mpMRI	multi-parametric MRI
MR	Magnetic Resonance
MSE	Mean Squared Error
NDR	Normalized Dynamic Range
NMI	Normalized Mean Intensity
PDF	Probability Density Function
PSNR	Peak Signal-to-Noise Ratio
RACLAHE	Region Adaptive Contrast Limited Adaptive Histogram Equalization

REI	Rand Error Index
RIDN	Real Image Denoising Network
RLBHE	Range Limited Bi-Histogram Equalization
RMSHE	Recursive Mean-Separate Histogram Equalization
SDI	Structural Differences Index
SNR	Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
STD	Standard deviation
T2W	T_2-weighted
TCIA	The Cancer Imaging Archive
TE	Time to Echo
WG	Whole Gland

List of Tables

List of Figures

1. Introduction

Deliverable D5.2 entitled "Pre-processing tools" describes the work performed by the participating technical partners throughout the lifetime of the T5.2 "Development of pre-processing pipelines on the platform". Specifically, as described in the DoA, the ProstateNet platform will offer tools and preprocessing pipelines to the AI model developers and to the platform users related to: common distortions correction such as bias field signal; motion-related artifacts and spatial inconsistencies correction; image enhancement and noise reduction techniques for improving the fidelity of MR images and the feature detection process; image normalisation and radiomics harmonization.

Accurate quantitative characterisation of medical images through machine learning has opened new horizons in diagnostic decision support systems. The role of radiomic studies as a supportive tool has emerged through a large number of studies, in which high throughput extraction of quantitative features enhances confidence in medical decisions. However, image-based descriptors usually suffer from variability related to the specific vendor and imaging conditions (e.g. field strength). It is therefore mandatory to decrease uncertainty, as a combined metric of systematic error or bias and random measurement error. The latter is defined by a number of terms such as precision, reliability or repeatability that yield specific differences, although are sometimes used interchangeably. As opposed to accuracy, these measurements regard closeness of measured values acquired under identical or near identical conditions. For instance, repeatability can be perceived as the ability to differentiate among subjects given the uncertainty of measurement errors related to a certain metric. In more strict terms, it can be defined as the ratio of between subject measurements to the total variance based on the observed measurement. Reproducibility describes the closeness of measurements under different conditions, such as different locations, operators, measuring systems or replicate measurements on the same or similar subjects. The bottom line is that decreasing the random measurement error is an essential part for the integration of quantitative metrics as clinically meaningful values extending beyond research interest.

Radiomics analyses entail high-throughput feature extraction (e.g., volume, shape, texture, etc.), from medical images coupled with sophisticated machine learning techniques in order to develop powerful prognostic and diagnostic models. Its non-invasive nature and the promising results reported in the literature have created an increased momentum towards developing models for tumor detection, segmentation, classification as well as treatment response prediction. In ProCAncer-I, significant effort has been put to develop powerful precision oncology models that will improve disease management by reducing over and underdiagnosis. In more detail, the multicenter radiomics has been proven to be sensitive to different scanner models, acquisition protocols and reconstruction settings, the methodologies used for delineation (or segmentation) of the region (or volume) of interest and the feature extraction, deriving imaging features of

increased variability which may hamper robustness and generalizability of the models[1]. The ProCAncer-I models will be able to generalize across vendors and imaging protocols based on a large number of images across multiple centers.

Recent literature has stressed the importance of utilizing harmonization techniques to compensate for multicenter effects. In more detail, multi-centric radiomics can be sensitive to image acquisition and reconstruction variability, to the delineation (or segmentation) of the region of interest and to the feature extraction methodology. Based on the literature findings, this deliverable focuses on the two most common harmonization pre-processing strategies. The first one focuses on the image domain and more specifically both in removing quality detrimental factors such as bias-field or noise and in reducing inherent differences in the intensity profile across vendors and protocols. The second strategy proposes methodologies to reduce feature variability related to different scanners or imaging conditions by appropriately manipulating their values posteriori in order to enable more robust statistical/machine learning analysis[2].

In this deliverable, these strategies led to the development of a number of pre-processing tools designed to harmonize imaging data or radiomics measurements from different clinical sites. Inhomogeneity correction, noise filtering, and intensity normalization were considered as the main preprocessing steps needed in order to minimize intensity related variations that could potentially affect the performance of radiomics studies. As an example, concerning MR images in particular, one of the initial steps in the chain of actions is the correction for the bias field which is a low frequency variation in the acquired signal. This non-uniformity is the result of a number of contributing factors, such as poor radiofrequency coil design, gradient eddy currents, local variations in flip angle and subject-scanner interactions. Higher main magnetic field scanners are more severely affected and the local intensity gradients can be seen by the naked eye. Starting from bias field correction, a number of pre-processing functionalities have been developed and included in the ProCAncer-I platform in order to deal with motion artifacts, noise reduction, image and radiomics feature normalization.

1.1 Motivation

Radiomics aim to objectively characterize clinically relevant information even beyond human perception. That said, sources of variation exist in each step of the image acquisition and radiomics analysis workflow potentially hampering the generalization and reproducibility of radiomics pipelines. While it is impossible to control or predict all sources of variation, currently many research efforts focus on image quality enhancement and harmonization strategies in order to improve both the performance and the trustworthiness of radiomics towards translation

[1] Ibrahim A, Refaee T, Leijenaar RTH, Primakov S, Hustinx R,Mottaghy FM, Woodruff HC, Maidment ADA, Lambin P. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. PLoS One. 2021 May 7;16(5):e0251147.

[2] R Da-Ano, D Visvikis, and M Hatt, “Harmonization strategies for multicenter radiomics investigations”, Physics in Medicine & Biology, Volume 65, Number 24,2020

of the developed radiomics signatures from the bench to the bedside[3]. These arguments comprise the central motivation for the work in D5.2 focusing on important pre-processing actions that target both the augmentation of image quality e.g. by removing bias field effects, as well as harmonization methodologies to reduce the image appearance inhomogeneity across vendors and imaging protocols.

It should be stressed that there is currently no gold standard to harmonize imaging data in a multi-centric setting towards an optimal radiomics pipeline. Instead, there are research trends[4] that have shown promising results based on which we implemented a number of optional pre-processing functionalities that can be used to experiment and optimize the ProCAncer-I radiomics AI modeling tasks.

1.2 Document structure

In this deliverable, every chapter describes a pre-processing pipeline developed by the ProCAncer-I consortium. The general format of each chapter is: i) a brief state-of-the-art of the process (State-of-the-art methods), ii) a fully detailed description of the proposed method, the selected parameters and the executed experiments by the ProCAncer-I partners (The ProCAncer-I tool), iii) how the experimental evaluation was conducted (Experimental evaluation) and iv) a brief description of the proposed pipeline (Pipeline execution and integration to the platform). The outline of the deliverable is given below.

Section 2 Bias Field Signal Correction: Introduces a filter dedicated to counter the bias field caused by the magnetic inhomogeneities during the acquisition process. Here, the retrospective version of the filter is analyzed, along with the process to tune its parameters and evaluate its results. A brief discussion how each parameter affects the filter is given.

Section 3 Motion-related Artifacts and Spatial Inconsistencies Correction: Explains how to tackle the artifacts generated by the organ motion or involuntary patient motion during the acquisition.

Section 4 Image Enhancement: Describes the process in how to improve the visual quality of the image by modifying the intensity values of individual pixels, especially in the case where intensities are overlapping. Regional Adaptive Contrast Limited Adaptive Histogram Equalization (RACLAHE) is introduced and tested on various U-Net based architectures, which were developed for prostate regions segmentation.

Section 5 Noise Reduction: Explains how to improve the diagnostic capabilities of an image, by reducing the noise signal. It addresses the issue by using convolutional neural networks which are trained to denoise the images by retaining the important characteristics of it. The four developed networks are: i) convolutional autoencoder with residual connections (CrAE), ii)

[3] Binsheng Zhao, "Understanding Sources of Variation to Improve the Reproducibility of Radiomics", Front. Oncol., Sec. Cancer Imaging and Image-directed Interventions, 2021

[4] S. Gitto et al., "Diffusion-weighted MRI radiomics of spine bone tumors: feature stability and machine learning-based classification performance," Radiol. Med., no. 0123456789, 2022, doi: 10.1007/s11547-022-01468-7.

denoising convolutional network (DnCNN), iii) denoising convolutional network with residual connections (DrCNN), and iv) a real image denoising network (RIDNet).

Section 6 Image Normalization: Describes the issues that arise from MRIs, especially in the case of heterogeneous dataset (center, vendor) and why normalization is important. ProCancer-I, based on the "White-Stripping" method used in brain images, develops four different approaches: i) fat-based, ii) muscle-based, iii) single tissue piece-wise, and iv) double tissue piece normalization.

Section 7 Radiomics Normalization: Focuses mainly on a heterogeneous dataset (center, vendor), that affect the performance and accuracy of radiomics. Here, "Combine Batches" (ComBat), a specialized normalization process for the aforementioned issues, is introduced.

Section 8 Radiomic Feature Stability Assessment after Image Preprocessing: The section presents how the radiomic feature extraction is affected by the pre-process pipeline. After the radiomics extraction from the raw T_2-weighted images, a comparison is performed against three of the pre-processing steps: i) N4 bias field corrections, ii) RACLAHE, and iii) noise reduction. To quantify the radiomics stability inter-class correlation coefficient (ICC) and the concordance correlation coefficient (CCC) are used.

1.3 Relation to the DoA

The following table describes the activities performed by the Task 5.2 participating partners during the lifecyle of the latter reported in the current document addressing the description of the project DoA.

Task 5.2 Description	Relation of D5.2 to the DoA description
Medical image acquisition processes inherently introduce different types of random noise complicating the computation of quantitative biomarkers or deep learning analysis. At the same time inherent grey-scale variations across vendors and protocols leads to variability in radiomics features which may hamper ML model development. The proposed preprocessing will include: a) common distortions correction such as bias field signal, b) motion-related artifacts and spatial inconsistencies correction, c) image enhancement and noise reduction techniques for improving the fidelity of MR images and the feature detection process, and, d) radiomics normalisation. The latter will focus on repeatability and reproducibility of	In the deliverable D5.2, pre-processing tools have been developed to deal with the MRI's intensity variation from different vendors and medical centers. As it is stated in the description, the D5.2 provides: a) N4 bias field correction (Section 2), b) elastix to handle motion-related artifacts and spatial inconsistencies (Section 3), c) RACLAHE for image enhancement (Section 4) and four deep learning models for denoising (CrAE, DnCNN, DrCNN, RIDNet) (Section 5) and d) ComBat for radiomics normalization (Section 7). The deliverable also includes image normalization techniques (fat-based, muscle-based, single/double tissue piece-wise) (Section 6). On section 8, the stability of radiomics feature extraction has been tested on a large MRI

radiomics features focusing on spatial stability (ROI dependence), temporal stability using a small test –retest dataset and definition of vendor invariant radiomic features also considering novel Gray-level invariant texture analysis paradigms. Considering the multi-institutional focus of the ProCAncer-I platform standardization on the examined data will be performed by applying uniform geometric properties on the imaging data and reducing the computational complexity by removing redundant data.	prostate dataset of 1500 patients, which consist of two vendors (Siemens, Philips) and from four medical centers (Radboud University Medical Center, Ziekenhuis Groep Twente, University Medical Center Groningen, Norwegian University of Science and Technology)[5]. The results show that N4 bias field correction will not affect the radiomics stability, however, extra caution is advised when image enhancement or noise reduction is performed.

[5] A. Saha et al., "Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge (Study Protocol)", publication date: June 19, 2022.

2. Bias Field Signal Correction

Multicentric studies are a sine qua non condition for a comprehensive evaluation of a disease and are also necessary to achieve high volume data for reliable AI applications. Although image scanning conditions can be set to comply with strict specifications, many other factors induce variability across the data. MR images are frequently affected by a low frequency variation in the acquired signal known as bias field corruption, resulting in intensity inhomogeneities. This non-uniformity is generated due to different contributing factors, such as poor radiofrequency coils, gradient eddy currents, local variations in flip angle and subject-scanner interactions. Although there are dedicated procedures for maintaining MR quality assurance, bias field corruption in a patient image cannot be predicted or addressed by such procedures as it appears randomly at different locations among patients and at different acquisitions even subsequent acquisitions of the same individuals. Moreover, each MR pulse sequence is affected to a different degree from this corruption. What remains constant is the pattern of a low frequency variation, most severely affecting higher magnetic field strength acquisitions. Therefore, several methods have been proposed for bias field correction in order to reduce the intensity inhomogeneities in the medical images and provide high quality images for AI applications and diagnostic purposes.

2.1 State-of-the-art methods

Bias correction methods are broadly categorized into prospective and retrospective methods[6]. The former eliminates the bias field caused by the hardware devices by calibrating and improving the acquisition process. The retrospective approaches reduce the bias field arising from the properties of the object in the scanner and are more general methods in their concept. The retrospective approaches for bias field correction are divided into 4 categories:

1. filtering methods;
2. surface fitting-based methods;
3. intensity-based methods and
4. histogram-based methods
5.

In the category of histogram-based methods, there is a well-known intensity inhomogeneity correction method, the N4ITK algorithm. The N4ITK filter[7] is considered as the state-of-the-art method for bias field correction and has been very widely used as a preprocessing step in recent

[6] S. Song, Y. Zheng, and Y. He, "A review of Methods for Bias Correction in Medical Images," *Biomed. Eng. Rev.*, vol. 1, no. 1, Sep. 2017, doi: 10.18103/BME.V3I1.1550.
[7] N. J. Tustison *et al.*, "N4ITK: Improved N3 bias correction," *IEEE Trans. Med. Imaging*, vol. 29, no. 6, pp. 1310–1320, Jun. 2010, doi: 10.1109/TMI.2010.2046908.

bibliography in classification[8], segmentation[9,10,11,12,13] and radiomic studies[14,15]. The N4 filter was produced by simpleITK as an improvement of the N3 bias correction[16] and is available as a python module, called N4ITK. The method seeks the multiplicative field that smoothens the image histogram and aims at high-frequency maximization of the image histogram. The N4 bias filter is an iterative process that uses an improved B-spline fitting routine, allowing for the use of multiple resolution levels.

However, the evaluation of the performance of the bias field correction methods is not a straightforward task. Since there is no ground truth or a reference standard for real patient data, most published works focus on simulated data where the bias field is introduced by a known function and is corrected for with the proposed methods[17,18]. A phantom method is also inappropriate in order to assess the bias field prior to the acquisition of real patient data as the bias field pattern strongly depends on the imaging subject/object and its specific anatomical and physical characteristics. Recent studies use the reproducibility of radiomic features as an indicator to assess the performance of the bias field correction algorithm[19,20,21]. Furthermore, some published works focusing on brain tissue use the segmentation performance as a metric to evaluate the performance of the algorithm[22,23,24]. However, there are no published works for the

[8] S. H. Hsu *et al.*, "Quantitative characterizations of ultrashort echo (UTE) images for supporting air-bone separation in the head," *Phys. Med. Biol.*, vol. 60, no. 7, pp. 2869–2880, 2015, doi: 10.1088/0031-9155/60/7/2869.
[9] L. Fang and X. Wang, "Brain tumor segmentation based on the dual-path network of multi-modal MRI images," *Pattern Recognit.*, vol. 124, 2022, doi: 10.1016/j.patcog.2021.108434.
[10] F. Ullah *et al.*, "Brain mr image enhancement for tumor segmentation using 3d u-net," *Sensors*, vol. 21, no. 22, pp. 1–14, 2021, doi: 10.3390/s21227528.
[11] M. Wang, J. Yang, Y. Chen, and H. Wang, "The multimodal brain tumor image segmentation based on convolutional neural networks," *2017 2nd IEEE Int. Conf. Comput. Intell. Appl. ICCIA 2017*, vol. 2017-Janua, pp. 336–339, 2017, doi: 10.1109/CIAPP.2017.8167234.
[12] S. Saman and S. J. Narayanan, "Active contour model driven by optimized energy functionals for MR brain tumor segmentation with intensity inhomogeneity correction," *Multimed. Tools Appl.*, vol. 80, no. 14, pp. 21925–21954, 2021, doi: 10.1007/s11042-021-10738-x.
[13] A. A. Nguyen *et al.*, "Post-Processing Bias Field Inhomogeneity Correction for Assessing Background Parenchymal Enhancement on Breast MRI as a Quantitative Marker of Treatment Response," *Tomography*, vol. 8, no. 2, pp. 891–904, 2022, doi: https://doi.org/10.3390/tomography8020072.
[14] S. Gitto *et al.*, "Diffusion-weighted MRI radiomics of spine bone tumors: feature stability and machine learning-based classification performance," *Radiol. Med.*, no. 0123456789, 2022, doi: 10.1007/s11547-022-01468-7.
[15] E. Chang *et al.*, "Comparison of radiomic feature aggregation methods for patients with multiple tumors," *Sci. Rep.*, vol. 11, no. 1, pp. 1–7, 2021, doi: 10.1038/s41598-021-89114-6.
[16] J. Sled, A. Zijdenbos, and A. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans Med Imaging*, vol. 17, no. 1, pp. 87–97, 1998, doi: 10.1109/42.668698. PMID: 9617910.
[17] J. B. Arnold *et al.*, "Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects," *Neuroimage*, vol. 13, no. 5, pp. 931–943, 2001, doi: 10.1006/nimg.2001.0756.
[18] Y. C. Zin, W. Zheng, M. W. L. Chee, and V. Zagorodnov, "Evaluation of performance metrics for bias field correction in MR brain images," *J. Magn. Reson. Imaging*, vol. 29, no. 6, pp. 1271–1279, 2009, doi: 10.1002/jmri.21768.
[19] Y. Li, S. Ammari, C. Balleyguier, N. Lassau, and E. Chouzenoux, "Impact of preprocessing and harmonization methods on the removal of scanner effects in brain mri radiomic features," *Cancers (Basel).*, vol. 13, no. 12, pp. 1–22, 2021, doi: 10.3390/cancers13123000.
[20] H. Moradmand, S. M. R. Aghamiri, and R. Ghaderi, "Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma," *J. Appl. Clin. Med. Phys.*, vol. 21, no. 1, pp. 179–190, 2020, doi: 10.1002/acm2.12795.
[21] M. Bologna, V. Corino, and L. Mainardi, "Technical Note: Virtual phantom analyses for preprocessing evaluation and detection of a robust feature set for MRI-radiomics of the brain," *Med. Phys.*, vol. 46, no. 11, pp. 5116–5123, 2019, doi: 10.1002/mp.13834.
[22] W. Zheng, M. W. L. Chee, and V. Zagorodnov, "Improvement of brain segmentation accuracy by optimizing non-uniformity correction using N3," *Neuroimage*, vol. 48, no. 1, pp. 73–83, 2009, doi: 10.1016/j.neuroimage.2009.06.039.
[23] L. Liao, T. Lin, and B. Li, "MRI brain image segmentation and bias field correction based on fast spatially constrained kernel clustering approach," *Pattern Recognit. Lett.*, vol. 29, no. 10, pp. 1580–1588, 2008, doi: 10.1016/j.patrec.2008.03.012.
[24] Y. Chen, B. Zhao, J. Zhang, and Y. Zheng, "Automatic segmentation for brain MR images via a convex optimized segmentation and bias field correction coupled model," *Magn. Reson. Imaging*, vol. 32, no. 7, pp. 941–955, 2014, doi: 10.1016/j.mri.2014.05.003.

pelvic area. The main challenge of this area compared to the brain is the increased heterogeneity among patients.

The N4ITK filter has several parameters that their values should be defined. Most studies use the default values of the parameters of the N4ITK filter and there are no studies in the literature to provide a comprehensive analysis on the effect of the different values of the parameters on the acquired images. Martin et al.[25] applied the N4ITK algorithm on breast phantoms and evaluated certain values of specific filter's parameters. The number of 50 iterations, a fitting level of 5 and the use of a full mask were identified as optimal configuration for the bias field correction by assessing the segmentation results and measuring the coefficient of variation in the mean intensity of specific regions. The impact of the mask's shape and the spline distance on the performance of N3 was assessed on brain MR images acquired on 3T scanners in a paper published prior to the N4 by Boyes et al.[26]. In this work, the performance of the N3 filter was assessed by measuring the variation of the normalized white matter intensity and the variation of the normalized image difference. The masks that enclose more precisely the brain tissue and smaller values of spline distances resulted in better performance of the N3. Zheng et al.[27] confirmed the effectiveness of using smaller spline smoothing distances in the N3 filter in 3T brain MR images.

2.2 The ProCAncer-I tool

In ProCAncer-I, systematic research was performed on the effect of certain parameters of the N4 filter on the acquired results in order to establish a trend – in the frame of a very large patient cohort, where individualized settings are not feasible. Certain default values of N4 parameters for bias field correction in prostate images are proposed and, in addition, the user is offered the ability to define the most appropriate settings for a specific task. Furthermore, the optimal configuration of the N4ITK filter can be defined automatically either for a single image or a batch of images by the algorithm.

More specifically two functionalities are offered to the user (Figure 1). The first functionality is to apply the N4 bias field correction method to one or more images. For this task, the values of the parameters of the N4 filter should be defined. The user has the ability to choose either the default values of the N4 parameters that have been proposed as optimum through our exploratory analysis (Functionality 1a) or to set other preferred values for some or all the N4 parameters (Functionality 1b). The second functionality offered to the user is the automatic identification of the optimal configuration of the N4 filter by the system's algorithm. The proposed pipeline can be applied to a specific image to identify the optimal set of parameters values for this specific image. Furthermore, the whole pipeline can be applied to a batch of images in order to identify automatically the most appropriate configurations for a specific cohort. However, this

[25] M. J. Saint Martin *et al.*, "A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study," *Magn. Reson. Mater. Physics, Biol. Med.*, vol. 34, no. 3, pp. 355–366, 2021, doi: 10.1007/s10334-020-00892-y.
[26] R. G. Boyes *et al.*, "Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils," *Neuroimage*, vol. 39, no. 4, pp. 1752–1762, 2008, doi: 10.1016/j.neuroimage.2007.10.026.Intensity.
[27] W. Zheng, M. W. L. Chee, and V. Zagorodnov, "Improvement of brain segmentation accuracy by optimizing non-uniformity correction using N3," *Neuroimage*, vol. 48, no. 1, pp. 73–83, 2009, doi: 10.1016/j.neuroimage.2009.06.039.

functionality is time-consuming due to the experimentation on various N4 configurations. This set of values, which are derived from functionality 2, are offered as an extra option to the user, in the event that the user wants to subsequently use them to apply the N4 bias field method to the images. For functionalities 1b and 2, the user has also the ability to use custom or automatically derived prostate masks. The last capabilities, which are the ability to apply the N4 bias field correction (Functionality 1) after identifying the optimal configuration (Functionality 2) and the use of custom or automatically extracted masks, are offered to the user only when using the N4 bias field correction module in interactive mode (see details about the interactive mode in Section 2.4.). For instance, if the user selects the functionality 2 by passing the corresponding argument from the command line (not using the interactive mode), then only the optimal configuration is extracted. In general, the interactive mode of this module provides more flexibility and options to the user.

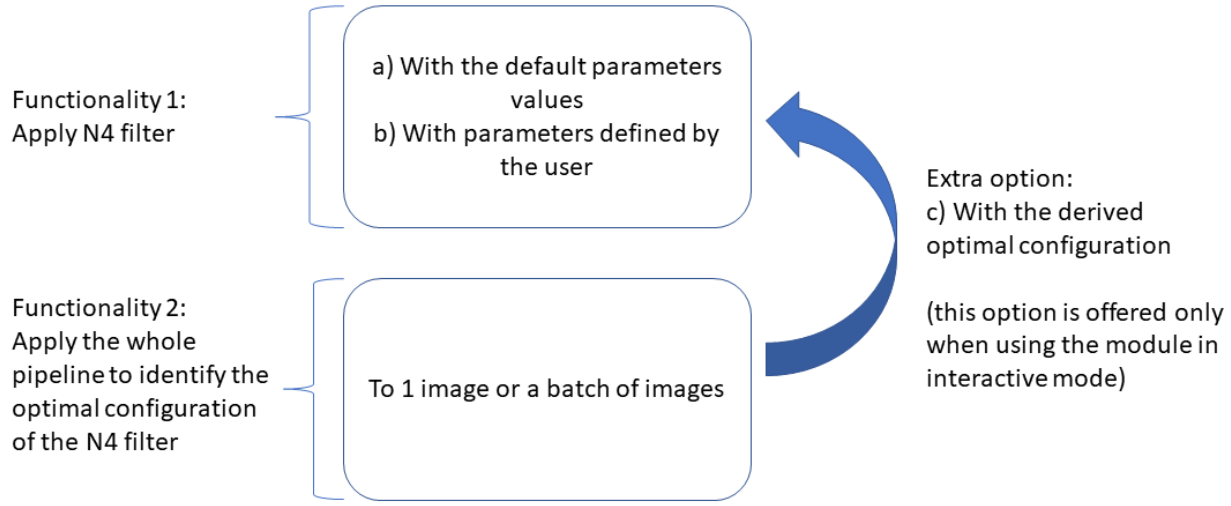


Figure 1 User's options of ProCAncer-I tool for bias field correction

In the systematic research, the effect of five parameters of the N4 filter is examined, which are:

i) the convergence threshold, which is the stopping criterion of the iterative process;
ii) the shrink factor, which defines how much the original image will be downsampled before estimating the bias field;
iii) the fitting level, which defines the number of levels that will be used to determine the resolution of the B-spline grid, where the previous mesh grid resolution is doubled and thus the spline distance gets smaller at each level;
iv) the number of iterations, which refers to the number of maximum iterations at each level and
v) the use of mask

The voxels that correspond to the mask region are used to estimate the bias field, whereas the non-zero voxels of the image are used when a mask is not provided. Larger values of convergence

threshold and shrink factor and lower values of fitting level and number of iterations result in lower computational complexity and execution time of the algorithm.

The examined values of the N4 parameters are:

i) with and without mask;
ii) the convergence threshold of 0.01, 0.001 and 0.0001;
iii) the shrink factor of 2 and 3;
iv) the fitting level of 3 and 4; and
v) the number of iterations of 5, 10, 25 and 50.

Each parameter is handled individually while keeping the rest of the parameters unchanged, in order to gain an overview of the effect of each parameter.

The Full Width at Half Maximum (FWHM) of the periprostatic fat distribution is proposed as a quantitative metric for assessing the performance and identifying the optimal configuration of the N4 bias filter. The periprostatic fat tissue was chosen as preference tissue due to its beneficial position around the prostate and its robust magnetic properties among individuals yielding high signal intensity. The fat distribution is desired to be as narrow as possible in order to be considered homogeneous tissue having similar intensity values. To this end, the narrowest fat distribution, and subsequently the minimum FWHM, indicates the reduction of the inhomogeneities that account for fat distribution broadening.

The proposed pipeline for the calculation of the FWHM and the identification of the optimal N4 configuration for a cohort is presented in Figure 2. More specifically, 96 different configurations of the N4 filter are applied to the MR images. The mask of each image is derived automatically by the proposed deep learning model[28] that extracts a cubic adaptive box around the estimated position of the whole prostate gland. The original and the filtered images are cropped automatically to include only the periprostatic region, which consists mainly of fat and muscle tissue, by removing the heterogeneous prostate gland and the area distant from the prostate. The k-means clustering algorithm with number of K equal to 2 is performed to the masked image to produce a cluster with the low intensity values and a cluster with the high intensity values that correspond to the fat distribution. The part of the histogram of the masked image that corresponds to the intensity values that belong to the cluster with the high intensity values is identified and the FWHM of this distribution is calculated for the original and the filtered images of each patient. The filter configuration that results in the minimum FWHM for each patient is considered as optimum. Furthermore, the number of patients and the corresponding percentage that result in each optimal configuration are calculated in order to identify the optimal settings for all the dataset. The settings that are identified as optimum to the largest percentage of patients are recommended to be used.

[28] D. Zaridis, E. Mylona, N. Tachos, K. Marias, M. Tsiknakis, and D. I. Fotiadis, "A Deep Learning-based cropping technique to improve segmentation of prostate's peripheral zone," *BIBE 2021 - 21st IEEE Int. Conf. Bioinforma. Bioeng. Proc.*, 2021, doi: 10.1109/BIBE52308.2021.9635576.

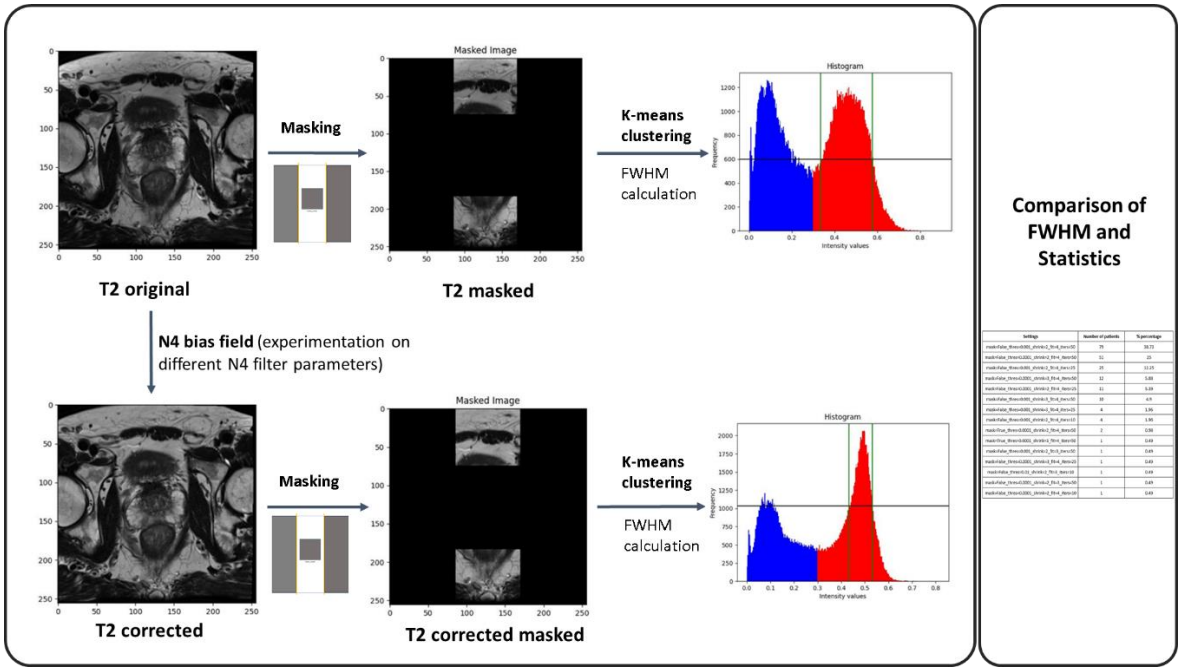


Figure 2. Workflow of proposed methodology for assessing the N4 parameters. The pipeline is performed to each patient and statistics are derived from the calculated FWHM values from all patients.

2.3 Experimental evaluation

As a preparatory step, a corn oil phantom was used in order to provide homogeneous fat mimicking material covering a very wide proportion of the field of view with appropriate signal intensities. The corn oil phantom was acquired on a SonataVision Siemens scanner and scanned using a 1.5T magnet and 3 different echo time (TE) values, which are 60, 80 and 120ms, resulting in 3 scanning settings. The main analysis was performed on two different MRI datasets, the Prostate-Diagnosis and the ProstateX dataset, with T2-weighted prostate images scanned with 1.5T and 3T, respectively, to investigate the impact of N4 filter to both magnetic field strengths, widely used in clinical practice. The MRI prostate images of Prostate-Diagnosis dataset were acquired on a 1.5T Philips Achieva scanner using combined surface and endorectal coil and can be downloaded from PROSTATE-DIAGNOSIS - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki (accessed on 3 February 2022)[29]. The MRI prostate images of ProstateX dataset were acquired on two different types of Siemens 3T scanners (the MAGNETOM Trio and Skyra) using surface coil and can be downloaded from SPIE-AAPM-NCI PROSTATEx Challenges (PROSTATEx) - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging

[29] B. N. Bloch, A. Jain, and C. C. Jaffe, "Data From PROSTATE-DIAGNOSIS. The Cancer Imaging Archive.," 2015, doi: http://doi.org/10.7937/K9/TCIA.2015.FOQEUJVT.

Archive Wiki[30,31,32]. The analysis was also applied to an external dataset of 30 image series acquired on 1.5T GE scanner using combined surface and endorectal coil, which was provided by ProCAncer-I clinical partner, FPO. This dataset consists of 15 positive and 15 negative cases of prostate cancer.

Phantom

The effect of the shrink factor and the use of mask was not examined in the corn oil phantom as it is a homogeneous material consisting of only 2 slices. Thus, 24 different configurations of the N4 filter were applied to the corn oil phantom. The FWHM of the original and all the filtered images of the phantom was calculated. The minimum FWHM was achieved using threshold = 0.01, fitting level = 4 and number of iterations equal to either 5 or 10 or 25 or 50. Thus, increasing the number of iterations does not improve the results as the algorithm converges early to the optimal estimation of the bias field. The relative difference between the FWHM of the original and the minimum FWHM of the filtered phantom showed a decrease greater than 80% in the value of the FWHM (Table 1), demonstrating and quantifying the effectiveness of the N4 filter to the image. This decrease in the value of FWHM can be observed by the narrower distribution of image intensities in the histogram of N4 filtered than the original phantom in Figure 3. The original unfiltered image and the corresponding N4 filtered image with the optimal setting of corn oil phantom with TE = 60 are depicted in Figure 4, as well as the estimated bias field map. The homogeneous material of the phantom becomes brighter after filtering the image with the N4 filter. The bias field map shows the variations of the bias field across the corn oil phantom.

Table 1. Selection of the optimal filter settings for the corn oil phantoms based on the difference of FWHM before and after N4 filtering.

Cornoil_TE	Optimal Filter Configuration	FWHM original	FWHM filtered	Relative Difference
cornoil60	thres=0.01_fit=4 _iters=5,10,25,50	41.89	5.92	-85.87%
cornoil80	thres=0.01_fit=4 _iters=5,10,25,50	37.06	5.26	-85.81%
cornoil120	thres=0.01_fit=4 _iters=5,10,25,50	27.1	4.52	-83.32%

[30] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, "'ProstateX Challenge data', The Cancer Imaging Archive," 2017, doi: 10.7937/K9TCIA.2017.MURS5CL.
[31] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, "Computer-aided detection of prostate cancer in MRI," *IEEE Trans. Med. Imaging*, vol. 33, no. 5, pp. 1083–1092, 2014, doi: 10.1109/TMI.2014.2303821.
[32] K. Clark *et al.*, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013, doi: 10.1007/s10278-013-9622-7.

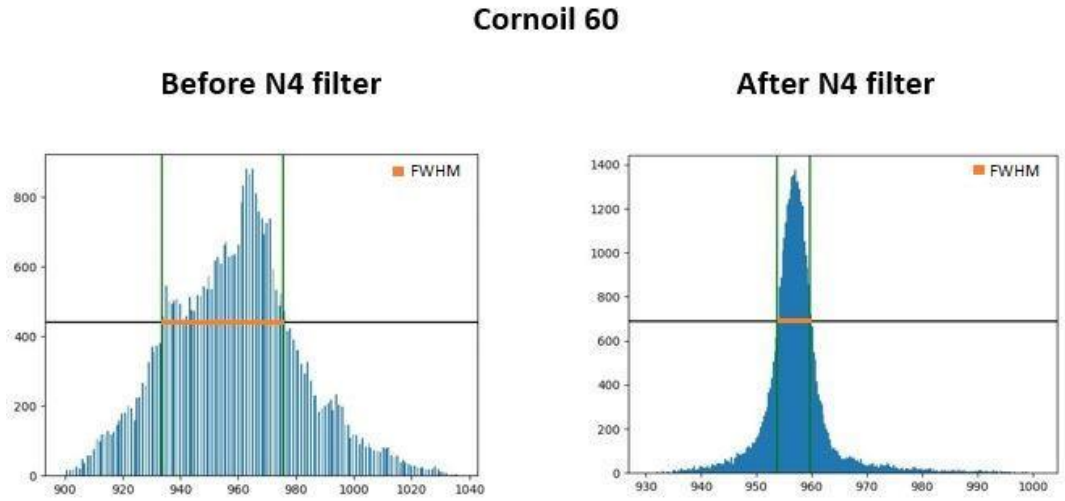


Figure 3. Histograms of the corn oil phantom with TE = 60, before and after N4 filtering.

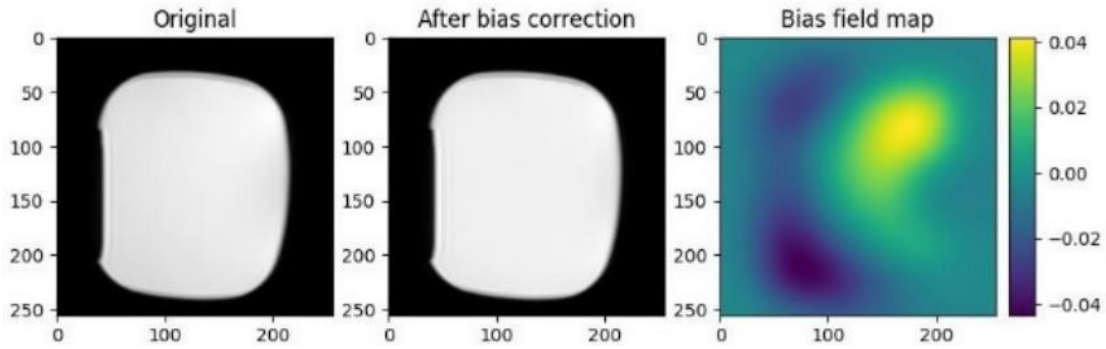


Figure 4. Original image of corn oil phantom with TE 60, N4 filtered image after bias field correction and the corresponding bias field map.

Prostate images

The experimentation of the different N4 filter parameters on the heterogeneous environment of the prostate was performed on the two datasets of MR images scanned with 1.5T and 3T magnetic field strength. The proposed pipeline was implemented and the 4 configurations that were identified as optimum to the largest percentages of patients are depicted in Table 2 and Table 3 for the 1.5T MR images and the 3T MR images, respectively. In the dataset with the MR images scanned with 1.5T, the largest percentage of patients that resulted in the same optimal configuration are only 25% and the following 3 percentages have similar values. On the contrary, the two largest percentages of patients in the dataset with the scanner of 3T magnetic field strength indicated a clearer trend in the optimal N4 configuration. More precisely, the combination of the use of the convergence threshold of 0.001, shrink factor of 2, fitting level of 4, number of iterations of 50 and without mask resulted in the minimum FWHM for the most patients in the dataset with the 3T MR images. This configuration was also in the top 4 settings in the dataset of 1.5T MR images.

Table 2. Percentage of occurrences of the derived optimal settings for the PROSTATE-DIAGNOSIS dataset with 1.5T magnetic field strength.

Settings	Number of patients	% percentage
mask=False_thres=0.001_shrink=3_fit=4_iters=50	23	25.84
mask=False_thres=0.0001_shrink=2_fit=4_iters=50	16	17.98
mask=False_thres=0.0001_shrink=3_fit=4_iters=50	14	15.73
mask=False_thres=0.001_shrink=2_fit=4_iters=50	13	14.61

Table 3. Percentage of occurrences of the derived optimal settings for the PROSTATEx dataset with 3T magnetic field strength.

Settings	Number of patients	% percentage
mask=False_thres=0.001_shrink=2_fit=4_iters=50	79	38.73
mask=False_thres=0.0001_shrink=2_fit=4_iters=50	51	25.00
mask=False_thres=0.001_shrink=2_fit=4_iters=25	25	12.25
mask=False_thres=0.0001_shrink=3_fit=4_iters=50	12	5.88

To validate the results, the relative difference between the minimum FWHM of the optimal setting and the FWHM achieved with a specific setting was calculated for both datasets. For instance, the relative difference between the minimum FWHM of the optimal setting and the FWHM achieved with the setting “mask=False, threshold=0.001, shrink factor=2, fitting level=4, iterations=50” was calculated for all the patients of each dataset for whom this was not the optimal setting. This difference was also calculated for the other settings of Table 2 and Table 3, as well as for other configurations to evaluate the effect of the N4 parameters. The results are shown in Figure 5 and Figure 6.

As depicted in Figure 5, the relative differences in the values of FWHM are smaller for the top 4 settings (Table 2) than the other examined settings. This confirms that these 4 configurations have similar impact on the cohort of the 1.5T MR images. In these 4 settings, the median value is lower than 10%, indicating that the value of FWHM obtained from these settings is less than 10% greater than the minimum FWHM of each patient. The distribution of the values of the relative difference achieved by the “mask=False, threshold=0.001, shrink factor=2, fitting level=4, iterations=50” and the “mask=False, threshold=0.0001, shrink factor=2, fitting level=4, iterations=50” settings has slightly smaller values than the rest two settings.

In the dataset with the MR images scanned with 3T, the settings with the 2 largest percentages confirmed their superior performance due to the small difference in the value of FWHM from the minimum FWHM (Figure 6). The setting with the highest frequency (38.73% of the patients), which is the “mask=False, threshold=0.001, shrink factor=2, fitting level=4, iterations=50”, showed great performance for the rest 61.27% of the patients, as the value of FWHM achieved with this setting was close to the minimum FWHM. The relative difference in their values was

less than 10% for the largest proportion of patients. The second most frequent setting, which is the configuration with the same parameters' values except the threshold value, which was equal to 0.0001, resulted also in small differences in the value of FWHM.

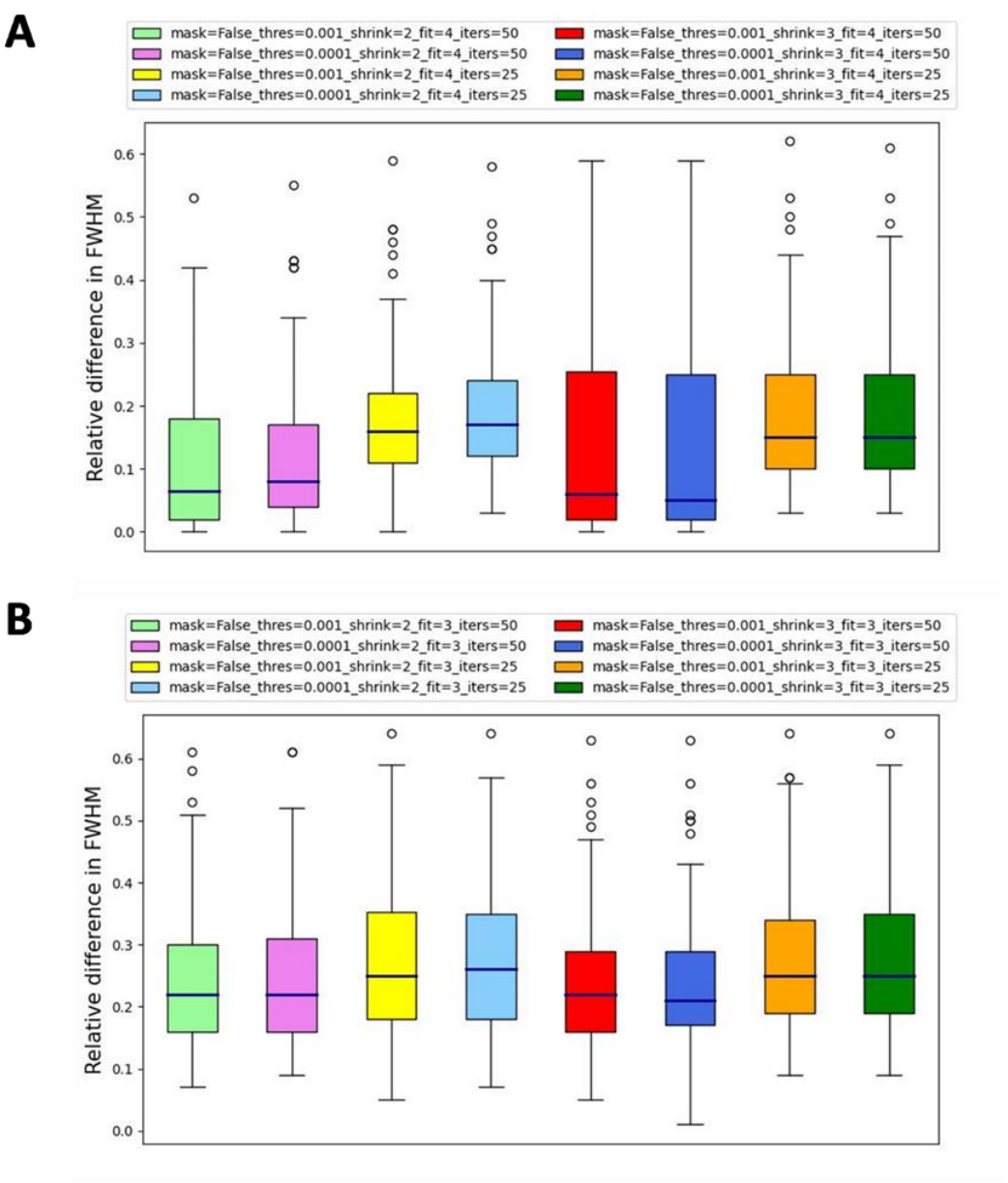


Figure 5. ***A.*** *Boxplots showing the relative difference between the FWHM of the optimal setting and the FWHM of various specific settings with fixed fitting level = 4* ***B.*** *Boxplots showing the relative difference between the FWHM of the optimal setting and the FWHM of various specific settings with fixed fitting level = 3. These results are obtained from the patients of PROSTATE-DIAGNOSIS dataset with 1.5T magnetic field strength. For each box, the relative difference is calculated only for the patients whose optimal setting is not the setting defined in the legend.*

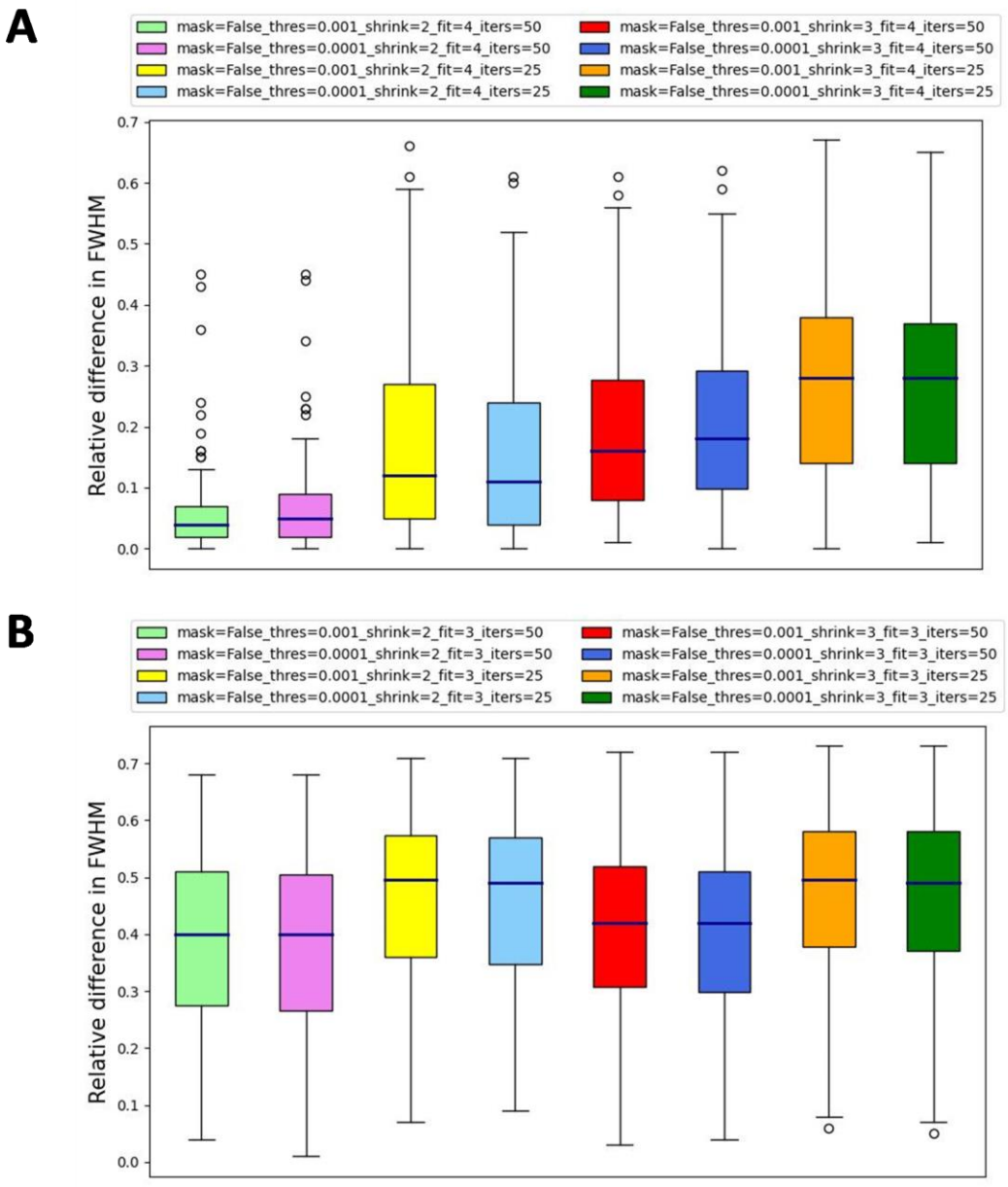


Figure 6. ***A.*** *Boxplots showing the relative difference between the FWHM of the optimal setting and the FWHM of various specific settings with fixed fitting level = 4.* ***B.*** *Boxplots showing the relative difference between the FWHM of the optimal setting and the FWHM of various specific settings with fixed fitting level = 3. These results are obtained from the patients of PROSTATEx dataset with 3T magnetic field strength. For each box, the relative difference is calculated only for the patients whose optimal setting is not the setting defined in the legend.*

Therefore, the configuration of convergence threshold of 0.001, shrink factor of 2, fitting level of 4, number of iterations of 50 and without mask (i.e. using the default mask) has great performance in both 1.5T and 3T MRI datasets. Even for the patients of this dataset that did not reach this optimal setting, the difference between their minimum FWHM and the FWHM achieved from this setting is small. Hence, the values of this configuration are recommended as

default values for the parameters of the N4 filter, when the N4ITK filter is applied for bias field correction in prostate images.

External Validation

The 30 images of the external dataset acquired with combined surface and endorectal coil are severely affected by the variable sensitivity of the receiver coil resulting in extremely high signal in the coil-body interface and relatively low signal intensity in structures far from the prostatic region. Due to this steep gradient in the receive coil sensitivity, the deep learning algorithm, which was used for the automatic extraction of the cubic box around the position of the prostate gland, was not able to produce the cubic box. To this end, a certain threshold was used in order to cut off the extremely high intensity values around the position of the coil. The threshold was set equal to the 11.5% of the maximum intensity value of the image. An experienced medical physicist evaluated the effect of this cut-off in order to ensure that only the areas i) outside the prostate gland and ii) appearing with abnormally enhanced signal in the artifact affected area are excluded. Hence, the voxels with intensity values larger than the 11.5% of the maximum intensity value of the image, were excluded. In Figure 7, the original slices of an image acquired with an endorectal coil, as well as the slices of the same image after applying thresholding are depicted. The intensities of the whole image are better represented after thresholding. Thus, the image after thresholding was given as input to the deep learning algorithm in order to extract the cubic box around the prostate gland. In this case, the deep learning algorithm was able to produce the cubic box and thus the whole pipeline for the identification of the optimal parameters could then be applied to the images similarly to the images acquired by surface coil only. Hence, the thresholding method efficiently mitigates the problem of the inability of the deep learning algorithm to produce the cubic box for the original image.

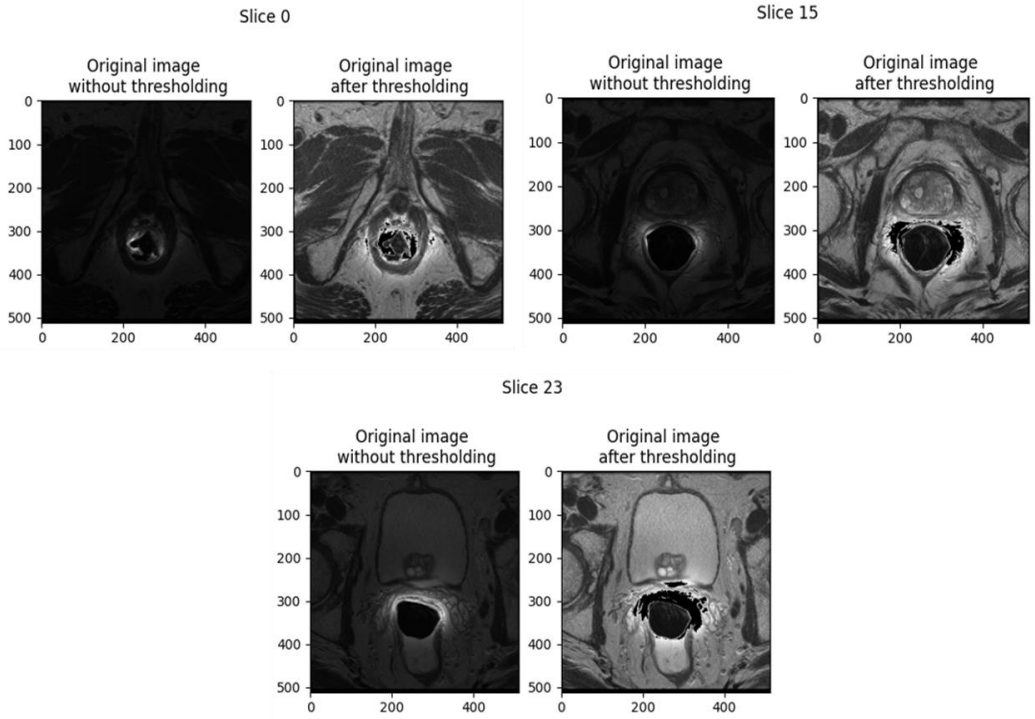


Figure 7. Slices from an image acquired with endorectal coil, with and without applying thresholding.

Therefore, the pipeline for the identification of the optimal configuration of the N4 filter was applied to the dataset of 30 images after applying the thresholding. The four configurations that were identified as optimal for the largest proportion of patients are depicted in Table 4. These four configurations are included in the top four configurations that had been extracted also for the ProstateX and the PROSTATE-DIAGNOSIS dataset. The configuration with a convergence threshold of 0.001, a shrink factor of 2, a fitting level of 4, 50 iterations and without the use of a mask resulted in the minimum FWHM for the most patients of this dataset. This optimal configuration is the same configuration which had been extracted as optimum to the ProstateX and the PROSTATE-DIAGNOSIS dataset, confirming the thresholding efficiency for bias field correction in pelvic MR images acquired with a combination of surface and endorectal coil and thus suffering from receive coil spatial sensitivity variation.

Table 4. Percentage of occurrences of the derived optimal settings for the dataset scanned on an endorectal coil.

Settings	Number of patients	% percentage
mask=False_thres=0.001_shrink=2_fit=4_iters=50	7	23.33
mask=False_thres=0.001_shrink=3_fit=4_iters=50	6	20
mask=False_thres=0.001_shrink=2_fit=4_iters=25	3	10
mask=False_thres=0.0001_shrink=2_fit=4_iters=50	3	10

Parameters effect

Different values of the parameters of the N4ITK filter were examined in order to understand their effect on the performance of the filter in prostate images. A small value of convergence threshold, equal to at least 0.001, is required to effectively reduce the intensity inhomogeneities in real prostate images. A smaller value of threshold equal to 0.0001 can also be used; however, it increases the computational complexity and the required execution time of the N4ITK filter. Regarding the number of iterations, 50 iterations are required when the N4 filter is applied in the heterogeneous MR prostate images. The use of 25 iterations, which lead to faster execution of the N4 filter, resulted in substantial difference in the value of FWHM (Figure 5 and Figure 6), as the value of FWHM was significantly larger than the minimum FWHM, especially in the 3T MRI images. A shrink factor of 2 leads to better performance of the N4ITK method. The use of a shrink factor equal to 3 resulted in considerable increase in the value of FWHM in the 3T MR images (Figure 6). However, in the 1.5T MR images the use of shrink factor of 3 did not result in substantial difference in the value of FWHM, as the median value remained at the same level or even smaller. Thus, a shrink factor of 3 can be used only for lower magnetic field strength (i.e. 1.5T) with lower bias field corruption, in cases that the decrease of the computational complexity

is necessary. The use of a fitting level of 4 results in better performance of the N4ITK algorithm. Higher magnetic field strengths result in higher bias field corruption. Hence, the spline density requirements are more demanding for higher field strengths and thus the use of smaller spline distances are necessary. This explains the fact that the decrease in the performance of the N4ITK with the use of a fitting level of 3 is more obvious in the 3T MR images (Figure 6B) than in the 1.5T MR images (Figure 5B). The greater the difference from the minimum FWHM, the poorer the bias field correction.

2.4. Pipeline execution and integration to the ProstateNet platform

The developed pipeline as described in the previous section has been packaged in a container image and integrated in the ProstateNet platform and provided as a tool in the market place of the AI Vault environment of ProCAncer-I infrastructure. Specifically, the steps for the execution of the N4 filter pipeline are depicted in the following figure.

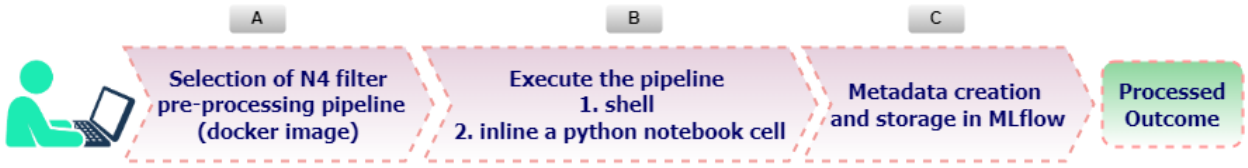


Figure 8.The user selects the N4 filter preprocessing pipeline (A) to be performed on the specified input and executes the pipeline in a shell script / python notebook (B). The metadata from the process are created and stored in the MLflow and the results are extracted in the destination folder specified by the user.

The N4 bias field correction module requires as input data the images in NifTI format (*.nii, *.nii.gz). Any examination shape of the image is accepted. If the corresponding masks of the prostate gland are given as input to the module, the masks should also be in NifTI format (*.nii, *.nii.gz). There are two basic functionalities of this module: i) apply the N4 filter and ii) identify the optimal configuration of the N4 filter. In the former, the output is the N4 filtered images, which are exported in the same file format (*.nii, *.nii.gz) and shape as the input images. In the latter, the output is a text or an excel file with the optimal configuration/configurations of the N4 filter for 1 image or a batch of images, respectively. The module requires the use of at least 8GB of RAM in order to process a number of images that indicatively have the size of a full image series as produced in medical routine practice (>30). The basic instructions for executing this module are described in Figure 9 and Figure 10 by analyzing each step of Figure 8. In the run command, the user should specify at least one of the available arguments that correspond to a functionality (i.e. -i, -N, -P) in order to execute an available task. The user can specify which functionality will be performed and provide the values of the parameters of the N4 filter from the command line or during the execution of the module in the interactive mode. The priority of the argument -i (i.e. interactive mode) is higher than the other arguments. For instance, if the user declares -i -N, then the module will run in interactive mode (argument -i) rather than start applying the N4 filter (ignores argument -N). Furthermore, if the user declares the argument -N, but he/she does not provide values for the arguments that correspond to the parameters of the

N4 filter (e.g. -i 0.01), then the module will use the default values of the parameters of the N4 filter that are extracted from our analysis.

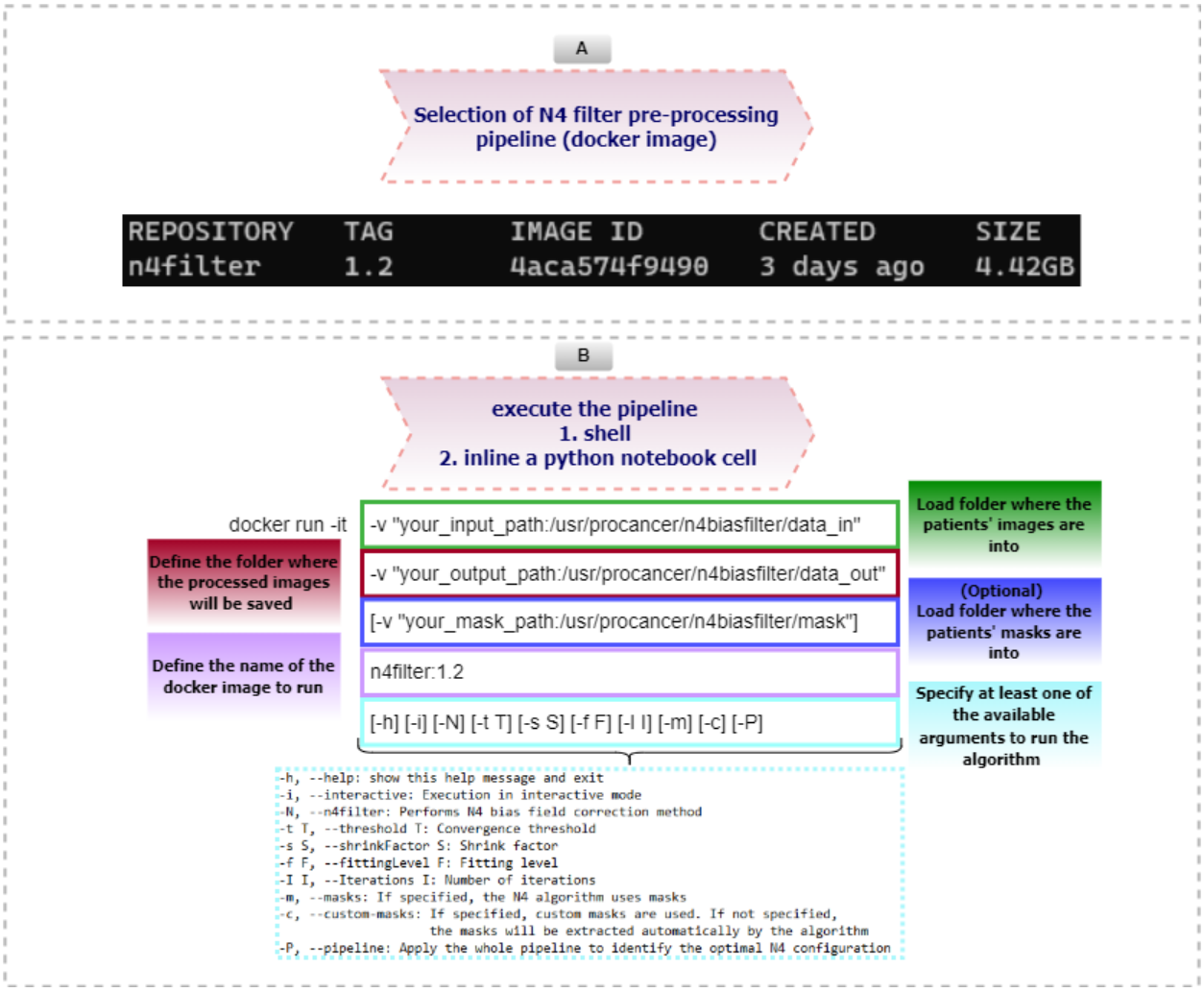


Figure 9. Description of the running command of the N4 filter pipeline.

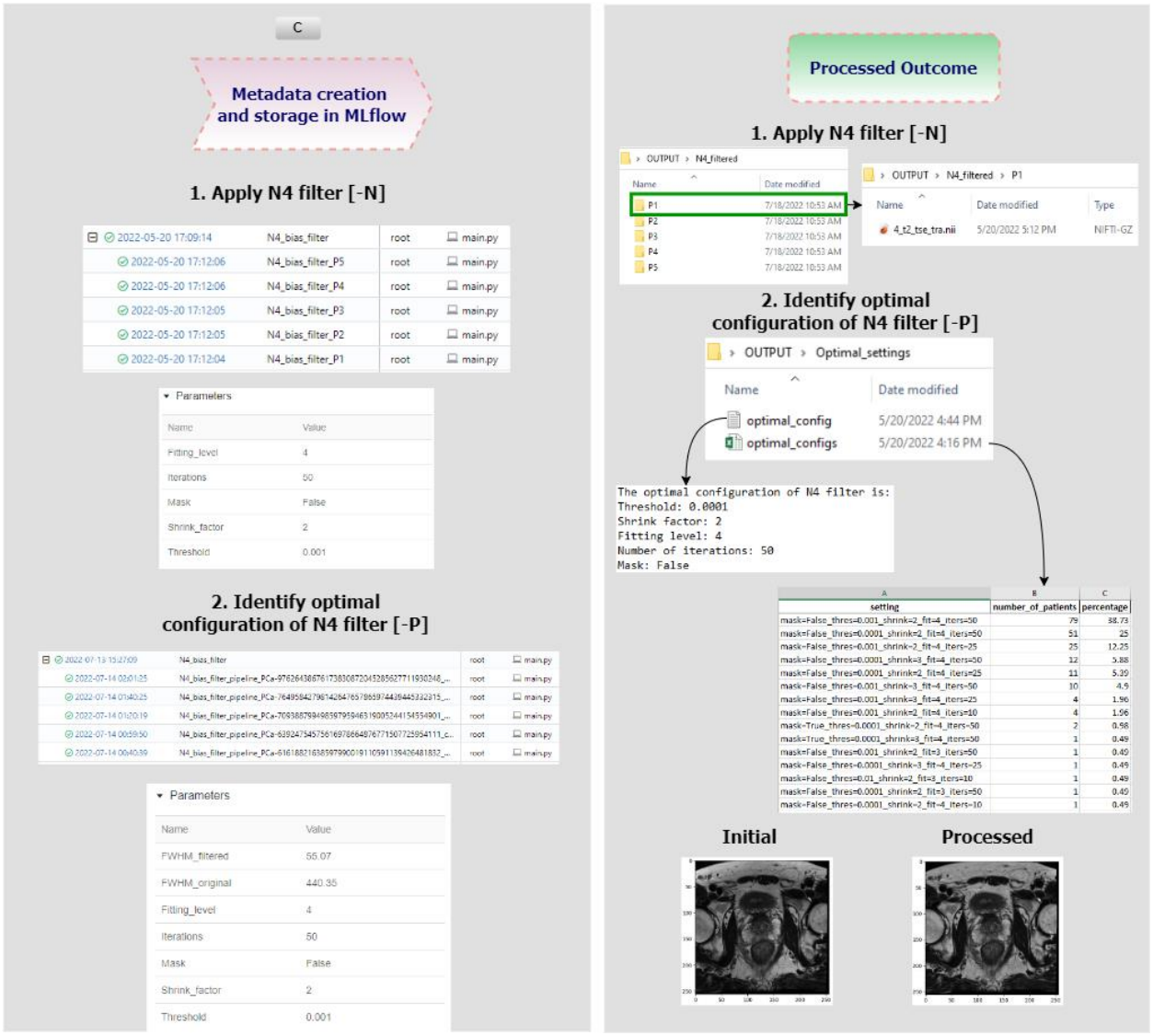


Figure 10. Description of the metadata and the processed outcome of the N4 filter pipeline.

3. Motion-related Artifacts and Spatial Inconsistencies Correction

3.1 State-of-the-art methods

Motion related artifacts can affect mpMRI of the prostate in a twofold manner: Firstly in a single volume 3D series patient motion can negatively affect image quality by increasing image blurring and ghost artifacts from a subtle degree to the degree of lacking a diagnostically useful image. Secondly motion in between different series comprising a dynamic or multi-series acquisition results in spatial inconsistencies among images of the same z-position, which in turn can be related to errors in further image exploitation and quantification, but also it hinders successful image fusion that is often used in clinical practise for diagnosis, annotation or segmentation.

3.2 The ProCAncer-I tool

Concerning the first case, in the frame of ProCancer-I involuntary incoherent patient motion causing image blurring and ghost artifacts will be addressed by implementing a basic image quality control that aims to identify and exclude images that do not meet certain criteria. A very basic metric to define motion corrupted images is a signal to noise ratio, where the two regions of interest that define the ratio are taken from the whole prostate gland and an area inside the field of view that does not contain anatomy (air) respectively. The latter area will be used to define the standard deviation of noise in the phase encoding direction in order to give an indication of ghost artifacts

Concerning patient motion between different acquisitions rendering consecutive series misalignment, a motion correction algorithm is provided within the curation toolkit to correct for inter-volume mismatches. This algorithm is based on relatively simple registration techniques – particularly, we:

1. Determine the transform that best aligns the apparent diffusion coefficient acquisition (ADC) to the T2-weighted acquisition (T2W);
2. Apply this transform to both the ADC and high b-value acquisitions (HBV) (this is feasible as both the ADC and HBV acquisitions share the same or a very similar coordinate system).

By aligning ADC, rather than HBV, to T2W we avoid possible registration artifacts caused by the low signal-to-noise ratio in HBV images as shown in Figure 11, comparing T2W, ADC and HBV. Identifying identical anatomical structures is much simpler when comparing T2W and ADC images than T2W and HBV images. This figure also highlights how the regions of most intensity in HBV are not attributable to a clinically relevant region, and are instead artefactual.

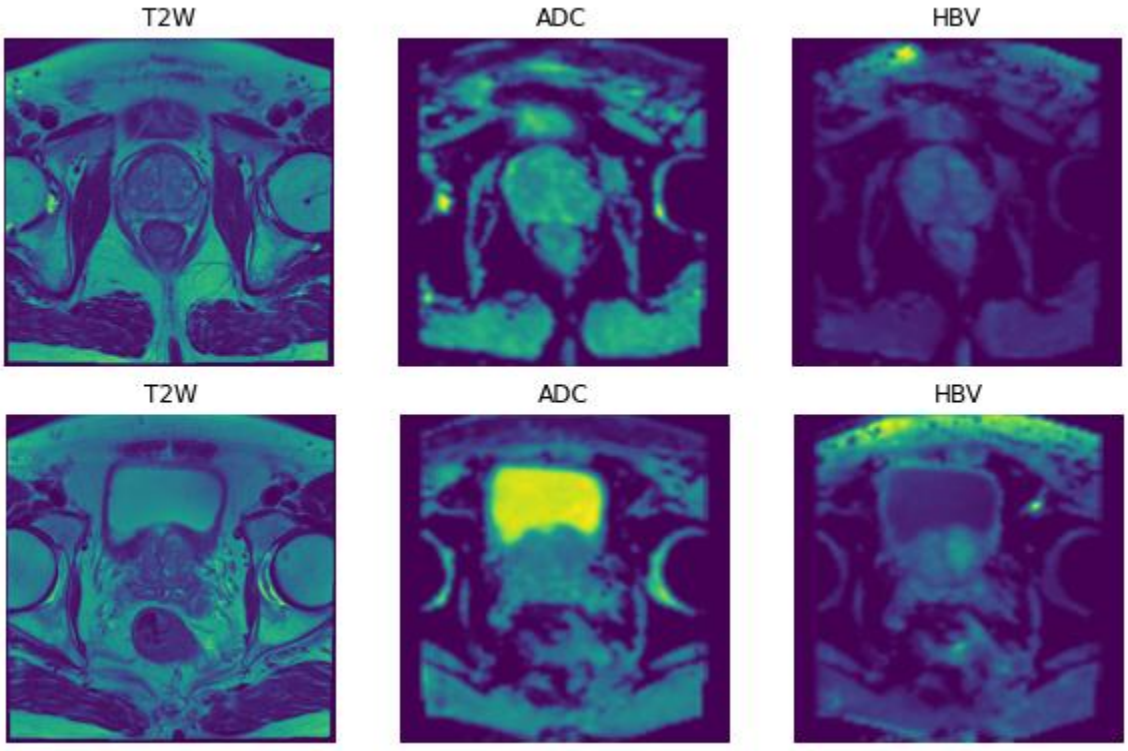


Figure 11. Slices extracted from an MRI study with 3 imaging modalities for the same individual, highlighting the differences between them. While T2W and ADC images share some similarities in terms of the expected voxel intensity of each zone, HBV has little similarity with either ADC or HBV based on this (nonetheless, it still is possible to identify the same structures in the HBV image). The three columns correspond to T2W, ADC and HBV images, respectively, whereas the rows correspond to the 13th and 17th slices, respectively.

3.3 Experimental evaluation

Many registration softwares are available[33,34]. In the context of the Task 5.2 activities we picked elastix[35], an image registration software featuring some desirable advantages over other alternatives[36] and implemented as a part of the Insight Toolkit (ITK) and SimpleITK [37,38,39] for the Python programming language[40]. Like many other solutions, elastix optimizes a given metric such that its optimal value should correspond to the largest overlap between fixed (T2W) and moving (ADC) images. For ProCAncer-I, we focused on relatively simple but effective registration methods:

[33] Zitová, B. & Flusser, J. Image registration methods: a survey. Image Vis. Comput. 21, 977–1000 (2003).
[34] Brown, L. G. A survey of image registration techniques. ACM Comput. Surv. 24, 325–376 (1992).
[35] Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. W. elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29, 196–205 (2010).
[36] Keszei, A. P., Berkels, B. & Deserno, T. M. Survey of Non-Rigid Registration Tools in Medicine. J. Digit. Imaging 30, 102–116 (2017).
[37] Avants, B. B. et al. The Insight ToolKit image registration framework. Front. Neuroinform. 8, 44 (2014).
[38] McCormick, M., Liu, X., Jomier, J., Marion, C. & Ibanez, L. ITK: enabling reproducible research and open science. Front. Neuroinform. 8, 13 (2014).
[39] Marstal, Berendsen & Staring. SimpleElastix: A user-friendly, multi-lingual library for medical image registration. Proc. IEEE.
[40] Van Rossum & Drake. The python language reference. Python software foundation.

- Translation registration (3 degrees of freedom) – the moving image can be shifted along the x, y and z axes;
- Rigid body registration (6 degrees of freedom) – the moving image can be shifted along and rotated about the x, y and z axes;
- Affine registration (12 degrees of freedom) – the moving image can be shifted along the x, y and z axis (3 parameters), and an affine matrix (9 parameters) is inferred, allowing for the image to be anisotropically scaled, rotated about the x, y and axes, and sheared.

While more complex techniques are available, based, for example, on non-linear registration using splines [41] or deep-learning [42], the focus of ProCAncer-I is on prostate cancer lesions, which can represent a volume of less than 0.1% of the original image. Considering such small volumes of interest, such complex methods can introduce a range of artifacts whose detection and curation would be laborious and counterproductive to our aim of automated prostate cancer prediction.

To empirically assess the best combination of methods, we devised an experiment using the ProstateX dataset [43], for which prostate gland masks were annotated by a radiologist with 10 years of experience for T2W and ADC/HBV images separately. In total, 178 sets of T2W, ADC and HBV and their respective prostate gland masks were used. Then, we performed the following registration protocol:

1. Resample ADC/HBV images to a voxel spacing identical to that of the T2W image;
2. Infer the ADC-to-T2W transform using either (the → sign implies the sequential application of these registrations):
    a. Image resampling (no registration, just resample the moving image to match the origin, direction and spacing of the fixed image). Otherwise referred to as image resampling;
    b. Translation registration. Otherwise referred to as translation (registration);
    c. Translation → rigid body registration. Otherwise referred to as rigid body (registration);
    d. Translation → rigid body → affine registration. Otherwise referred to as affine (registration);
3. Apply the ADC-to-T2W to the ADC and HBV images and to the prostate gland mask annotated using the ADC/HBV images;
4. Measure the degree of overlap between T2W and the ADC/HBV prostate gland mask using the Jaccard index (otherwise known as the intersection over the union or IoU).

[41] Szeliski, R. & Coughlan, J. Spline-based image registration. Int. J. Comput. Vis. 22, 199–218 (1997).
[42] Haskins, G., Kruger, U. & Yan, P. Deep learning in medical image registration: a survey. Mach. Vis. Appl. 31, 8 (2020).
[43] PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images

The validity of this assessment depends on whether or not these images had already been aligned or curated to some extent. For this reason, it may be hard to assess how either registration would perform under the presence of deviations. To address this, we run a second set of experiments, where the origin and direction of the ADC/HBV is perturbed; particularly, for the direction perturbation, we sum a random 3-dimensional quantity to the 3-dimensional origin vector, while for the origin perturbation we sum a 9-dimensional quantity to the 9-dimensional vector representing the 3x3 direction cosine matrix and consequently altering it to ensure its orthogonality. The origin perturbation in millimeters O is generated by sampling from an anisotropic uniform distribution such that each element of this distribution $O_i \sim U(-30,30)$, while the direction perturbation D is generated by sampling from a 9-dimensional uniform distribution such that each element $D_i \sim U(-0.05,0.05)$. We apply both perturbations before step 1. of the experimental protocol described above.

In any of the registrations, we optimize the Mattes mutual information criterion [44], where a set of n voxels values (in our case $n = 4096$) calculated through linear interpolation from both images is randomly selected for the calculation of the mutual information between both images after discretizing the voxel values into 30 different bins. All registrations are performed at 4 resolutions using image pyramids and optimized using the adaptive stochastic gradient descent algorithm as implemented in elastix as this provides out-of-the-box good solutions. The translation, rigid body and affine registrations are optimized for a maximum of 1,000, 1,000 and 500 iterations, respectively. Voxel interpolation is performed using b-splines of the 3rd order in the case of ADC/HBV images and using nearest neighbor interpolation when transforming the mask.

Regarding the results of these analyses, we start by focusing on the first experiment (no perturbation) with the top panels of Figure 12 and Figure 13, showing the IoU and IoU improvement (difference between IoU for registration and IoU for resampling). For reference, the presented test-statistics (t) and p-values (p) refer to one-sample t-tests of the IoU improvement where the null hypothesis is "the IoU improvement is 0" unless otherwise noted. In effect, we can attest that under such cases – where images may already be relatively well aligned – little is achieved by the registration (Table 1), with no evidence for change in the cases of translation and rigid body registration (p=0.8 and p=0.6, respectively), and a statistically significant decrease in IoU for the affine registration case (t=-4.8, p=$3*10^{-6}$). This behavior from the affine registration, while unexpected, is explainable – deformations in the ADC image associated with rectal air may cause spurious anisotropic scalings as seen in Figure 14.

[44] Rahunathan, Stredney & Schmalbrock. Image registration using rigid registration and maximization of mutual information. 13th Annu. Med. Meets.

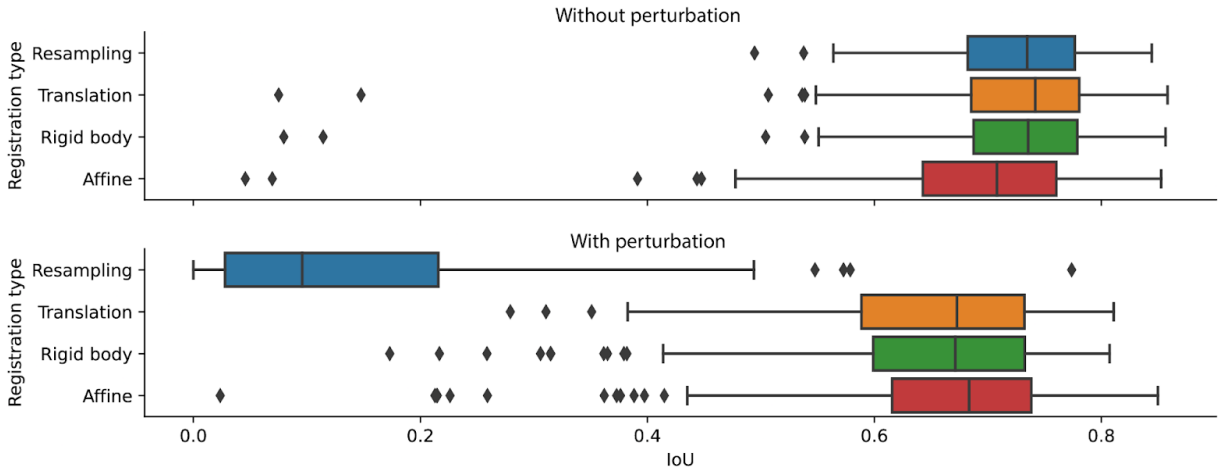


Figure 12. IoU (Jaccard index) distribution calculated between each mask-registered mask pair. The top panel represents the results without perturbation while the bottom panel represents the results with origin and direction perturbations (n=178 MRI studies).

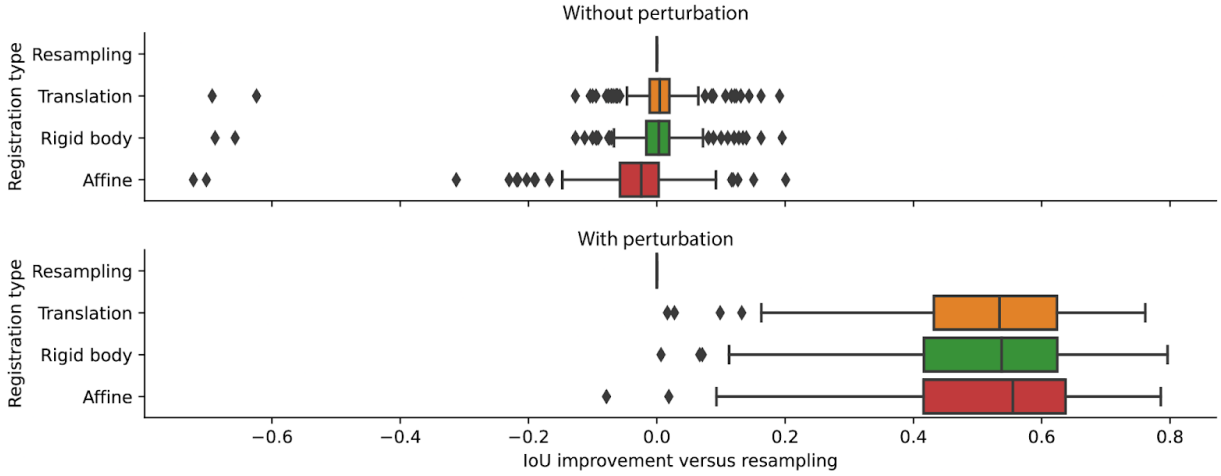


Figure 13. IoU improvement compared with image resampling distribution calculated between each mask-registered mask pair. The top panel represents the results without perturbation while the bottom panel represents the results with origin and direction perturbations (n=178 MRI studies).

Table 5. Intersection over union (IoU) between the prostate gland across different MRI modalities for three different registration protocols and its improvement when compared with image resampling.

Registration	Without perturbation		With perturbation	
	IoU	Improvement	IoU	Improvement
Resampling	73.5%	0%	9.5%	0%
Translation	74.2%	0.4% (p=0.8)	67.3%	53.4% ($p=6*10^{-100}$)

Rigid body	73.5%	0.3% (p=0.6)	67.1%	53.7% ($p=3*10^{-96}$)
Affine	70.8%	-2.4% ($p=3*10^{-6}$)	68.3%	55.5% ($p=3*10^{-94}$)

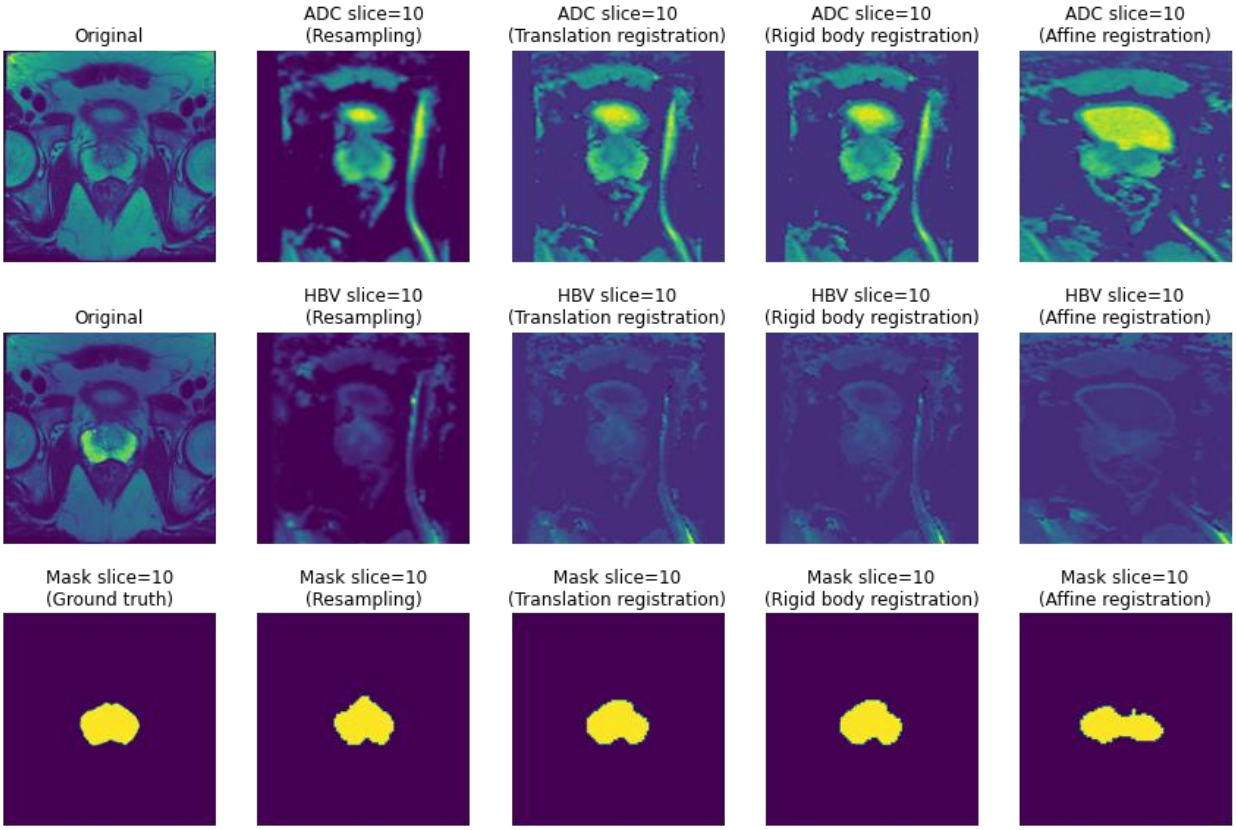


Figure 14. Visualization of the effect of the different and sequentially applied registration protocols on the ADC/HBV images and their masks for the first experiment. The first column represents the fixed image (Original), the overlap of the fixed image with its mask (2nd row) and the fixed image prostate gland mask, while the remaining columns represent different registrations/image resampling. For the affine registration (last column), it is possible to observe an anisotropic and incorrect scaling of the prostate gland caused by the affine registration.

From the first experiment one would be inclined to believe that registration is redundant or, in the worst case scenario, deleterious. However, if we turn our attention to the second experiment, where small perturbations are present, one can see how a simple resampling is highly inaccurate, with translation, rigid body and affine registration leading to considerable and statistically significant improvements (t=45.7 and $p=6*10^{-100}$, t=43.4 and $p=3*10^{-96}$, t=42.1 and $p=3*10^{-94}$, respectively; Table 1). In this last case it is also worth noting that no particular registration yields different results from other registrations – for two-sample t-tests the IoU of translation with rigid body, translation with affine and rigid-body with affine we see that no comparison is statistically significant (p=0.6, p=0.96 and p=0.6, respectively).

Finally, we show a final example of this image registration process. In Figure 15 (where 0 is ADC and 1 is HBV), we can observe an example of an origin and direction perturbation to the moving images and their consequent realignment – one can see that the misaligned images (Moving 0 and Moving 1 in Figure 5) are correctly aligned in almost all slices (Moved 0 and Moved 1).

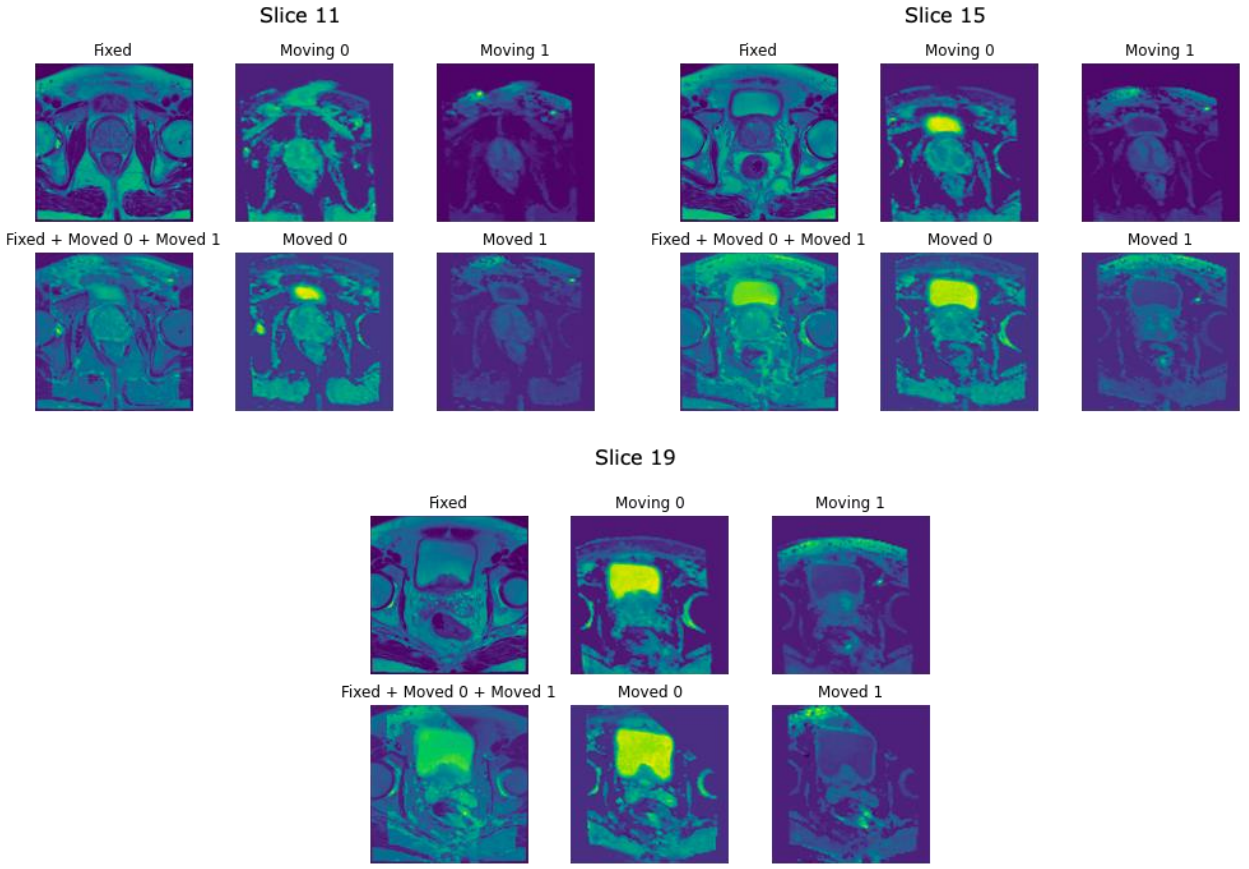


Figure 15. Visualization of fixed, moving and moved slices for our translation → rigid-body registration protocol for an MRI study (the 11th, 15th and 19th slices are shown). The Fixed + Moved 0 + Moved 1 images represent the sum of the scaled Fixed, Moved 0 and Moved 1 images for each of the represented slices.

From these analyses we conclude that translation or rigid body registration lead to no significant decrease in the expected anatomical overlap when different modalities are already aligned and greatly improves this overlap in cases where there are slight perturbations to this alignment. On the other hand, while affine registration leads to similar improvements given the presence of perturbations, it leads to a decrease in the expected anatomical overlap when images are already reasonably well aligned. Taking this into account, ProCAncerI will use rigid body registration prior to downstream analysis as it can lead to improved alignment with no expected decrease in anatomical overlap.

3.4 Pipeline execution and integration to the ProstateNet platform

The registration protocol presented in this section and illustrated in Figure 16 requires input data to be in either NifTI (*.nii or *.nii.gz) or DICOM formats (folder containing several *.dcm files).

Images should all be three dimensional and image shape is accepted. Given that the registration protocol is relatively simple, it is possible to execute the registration in relatively little time (between 20 seconds and one minute). This method (register.py) is available in a Bitbucket repository[45], where instructions are available for its execution. Briefly, the user should specify the path to the T2W fixed image (with the --fixed_image_path argument) and the paths to the ADC and HBV moving images (with the --moving_image_paths argument; the first moving image will be used to infer the transform which will be applied to all moving images). Given that this script assumes that images use the affine registration, the option "--no_affine" should be added to skip the affine registration step and perform translation and rigid body registrations. If the user wants to use DICOM images as input and output, the --input_is_dicom and --dicom_output flags should be used, respectively. Other options, such as registering masks should be specified with the --labels argument, which should note the index of the moving image corresponding to a mask.

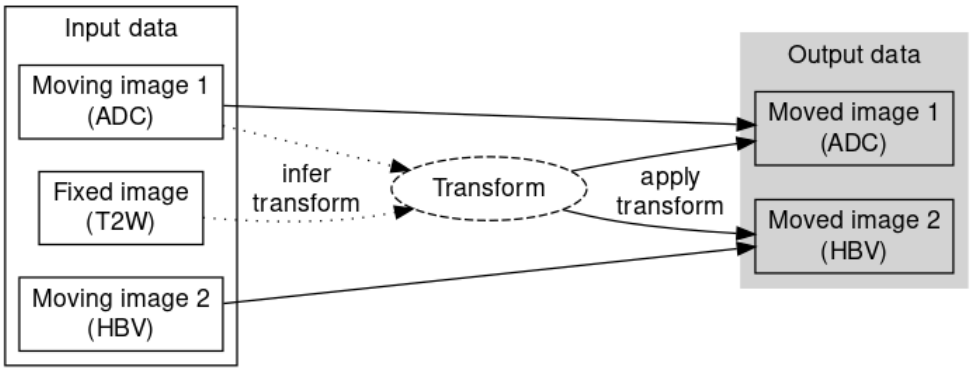


Figure 16. Illustrative example of the pipeline. The ADC to T2W coordinate system transform is inferred using both T2W and ADC as the fixed and moving images, respectively, and the transform is then applied to both ADC and HBV sequences.

[45] https://bitbucket.org/josegcpa/mri-registration

4. Image Enhancement

The field of medical image preprocessing encompasses a large set of techniques that aim to overcome issues arising from degraded images, due to the presence of motion, noise or artifacts, as in the case for MR images. An indispensable part of image preprocessing is image enhancement which aims to improve the visual quality of the image by modifying the intensity values of individual pixels so that anatomical structures can more easily be recognised by humans and machines. This is achieved by means of adequate gray-scale transformations[46] aiming to disentangle intensity distributions arising from adjacent regions with similar gray level intensities[47]. Therefore, by sharpening the boundaries between different tissues[48],contrast enhancement has emerged as a powerful method for improving the accuracy of DL segmentation models[49].

4.1 State-of-the-art methods

Image enhancement is a process to emphasize the desired information of an image, to obtain better results for a user (e.g., diagnosis) or machine (e.g., automated segmentation). Generally, it can be divided into two groups, i) image filters (e.g., Gaussian, median, Wiener filters) and ii) histogram manipulation.

Concerning the histogram manipulation methods, their main target is to adjust the image contrast, especially in the case of overlapping intensities where the desired features (e.g., biomarkers) are indistinguishable. Starting with the most basic method, histogram equalization (HE)[50] is performed to distribute the intensities of the histogram by taking into account the frequencies of the gray level values, but since it is applied to the histogram of the whole image, it may cause over-enhancement and produce unrealistic images. Thus, the adaptive histogram equalization (AHE)[51] was evolved to perform histogram manipulation on multiple regions of the image. Even though AHE can deal with heterogeneous regions, it may introduce artifacts in homogeneous regions. To overcome the generated issues from the previous aforementioned methods, contrast limited adaptive histogram equalization (CLAHE)[52] was developed to achieve the enhancement of the local regions with respect to retain the quality of the input image. The CLAHE divides the initial image into smaller squared patches. Afterwards, the local maxima and the histogram for each patch are extracted, to perform histogram modification for each patch. These steps will generate a contrast enhanced image with altered dynamic range of the input,

[46] L. Xiong, H. Li, and L. Xu, "An enhancement method for color retinal images based on image formation model," Computer Methods and Programs in Biomedicine, vol. 143, pp. 137–150, May 2017, doi: 10.1016/J.CMPB.2017.02.026.
[47] R. B. Paranjape, "Fundamental Enhancement Techniques," Handbook of Medical Image Processing and Analysis, pp. 3–18, Jan. 2009, doi: 10.1016/B978-012373904-9.50008-8.
[48] K. D. Toennies, "Guide to Medical Image Analysis," 2017, doi: 10.1007/978-1-4471-7320-5.
[49] L. Rundo et al., "A novel framework for MR image segmentation and quantification by using MedGA," Computer Methods and Programs in Biomedicine, vol. 176, pp. 159–172, Jul. 2019, doi: 10.1016/J.CMPB.2019.04.016.
[50] C. Moore, "Medical image processing: the characterization of display changes using histogram entropy," Image and Vision Computing, vol. 4, no. 4, pp. 197–202, Nov. 1986, doi: 10.1016/0262-8856(86)90046-6.
[51] S. M. Pizer et al., "Adaptive histogram equalization and its variations," Computer Vision, Graphics, and Image Processing, vol. 39, no. 3, pp. 355–368, Sep. 1987, doi: 10.1016/S0734-189X(87)80186-X.
[52] A. M. Reza, "Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement," Journal of VLSI signal processing systems for signal, image and video technology 2004 38:1, vol. 38, no. 1, pp. 35–44, Nov. 2004, doi: 10.1023/B:VLSI.0000028532.53893.82.

without amplifying the locally residual noise. Studies show that the CLAHE algorithm is superior to both HE and AHE, along with other contrast enhanced methods[53,54,55,56,57,58].

Another alternative method to deal with the inconveniences from HE is to retain the original brightness of the image. Wang et al.[59] developed the brightness preserving bi-histogram equalization (BBHE). BBHE divides the image histogram in two. From the input image, the mean brightness value is extracted to compute the threshold of this division. HE is performed for both histograms and generating an image with mean brightness between the input mean and the middle gray level. Chen and Ramli[60] suggest an improved BBHE, named minimum mean brightness error bi-histogram equalization (MMBEBHE). In order to retain initial brightness values of the input image, this method searches for all plausible intensity thresholds and for each threshold HE is applied. From the generated images, the one with the lowest difference from the initial mean brightness is selected. The same authors suggest the recursive mean-separate histogram equalization (RMSHE), follows the same routine as BBHE, however RMSHE repeats the process for a number of times. In this case, the user defines the number of divisions of each sub-histogram produced by the previous step. For a large number of divisions, the final generated image will converge to the brightness level of the original one. Haidi et al.[61] developed the brightness preserving dynamic histogram equalization (BPDHE), a pipeline of four stages. First, gaussian filtering is applied, followed on histogram partitioning in regions with respect to the local maxima. Afterwards, it dynamically sets the intensity range of each region and finally it applies HE in each one. The Range Limited Bi-Histogram Equalization[62] (RLBHE) uses the Otsu filter to divide the histogram into two sub-histograms. To retain the original brightness of the input image, the upper and lower boundaries are extracted, then, the sub-histograms are equalized. According to the author, the enhanced images have a similar brightness level to the input, do not contain “undesirable” artifacts and unrealistic effects from the process.A modified

[53] Maison, T. Lestari, and A. Luthfi, “Retinal Blood Vessel Segmentation using Gaussian Filter,” Journal of Physics: Conference Series, vol. 1376, no. 1, Nov. 2019, doi: 10.1088/1742-6596/1376/1/012023.
[54] G. F. C. Campos, S. M. Mastelini, G. J. Aguiar, R. G. Mantovani, L. F. de Melo, and S. Barbon, “Machine learning hyperparameter selection for Contrast Limited Adaptive Histogram Equalization,” Eurasip Journal on Image and Video Processing, vol. 2019, no. 1, pp. 1–18, Dec. 2019, doi: 10.1186/S13640-019-0445-4/FIGURES/11
[55] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. Muller, “Contrast-limited adaptive histogram equalization: Speed and effectiveness,” Proceedings of the First Conference on Visualization in Biomedical Computing, pp. 337–345, 1990, doi: 10.1109/VBC.1990.109340.
[56] I. A. M. Ikhsan, A. Hussain, M. A. Zulkifley, N. M. Tahir, and A. Mustapha, “An analysis of x-ray image enhancement methods for vertebral bone segmentation,” Proceedings - 2014 IEEE 10th International Colloquium on Signal Processing and Its Applications, CSPA 2014, pp. 208–211, 2014, doi: 10.1109/CSPA.2014.6805749.
[57] S. F. Mat Radzi et al., “Impact of Image Contrast Enhancement on Stability of Radiomics Feature Quantification on a 2D Mammogram Radiograph,” IEEE Access, vol. 8, pp. 127720–127731, 2020, doi: 10.1109/ACCESS.2020.3008927.
[58] I. A. M. Ikhsan, A. Hussain, M. A. Zulkifley, N. M. Tahir, and A. Mustapha, “An analysis of x-ray image enhancement methods for vertebral bone segmentation,” Proceedings - 2014 IEEE 10th International Colloquium on Signal Processing and Its Applications, CSPA 2014, pp. 208–211, 2014, doi: 10.1109/CSPA.2014.6805749.
[59] Y. Wang, Q. Chen, and B. Zhang, “Image enhancement based on equal area dualistic sub-image histogram equalization method,” IEEE Transactions on Consumer Electronics, vol. 45, no. 1, pp. 68–75, 1999, doi: 10.1109/30.754419.
[60] S. der Chen and A. R. Ramli, “Preserving brightness in histogram equalization based contrast enhancement techniques,” Digital Signal Processing: A Review Journal, vol. 14, no. 5, pp. 413–428, 2004, doi: 10.1016/J.DSP.2004.04.001.
[61] H. Ibrahim and N. S. P. Kong, “Brightness preserving dynamic histogram equalization for image contrast enhancement,” IEEE Transactions on Consumer Electronics, vol. 53, no. 4, pp. 1752–1758, Nov. 2007, doi: 10.1109/TCE.2007.4429280.
[62] Zuo, C., Chen, Q., & Sui, X. (2013). Range Limited Bi-Histogram Equalization for image contrast enhancement. Optik, 124(5), 425–431. https://doi.org/10.1016/J.IJLEO.2011.12.057

CLAHE was developed from Sheeba et al.[63], called fuzzy clipped CLAHE (FC-CLAHE). The process enhances the local contrast by preserving the original brightness of the image. FC-CLAHE was tested on mammogram images. The authors set an optimal clip value to enhance masses and micro-calcifications with respect to the surrounding tissues. By introducing the fuzzy rules, they are increasing the adaptability of the clip to automate the procedure and deal with the variability between images effectively.

The Adaptive Gamma Correction with Weighting Distribution[64] (AGCWD) makes use of the power law transformation $T = I_{max}\,(I\,/\,I_{max})\,\gamma$, where I is the pixel's intensity, I_{max} the maximum intensity of the input image and γ the adaptive parameter to calibrate for contrast. enhancement. In contrast to power law transformation, the AGCWD sets automatically the γ. The process is to extract the weighted probability density function $(PDF_w(I))$ from the histogram, followed by the weighted cumulative distribution function $(CDF_w(I))$. Finally, the optimal $\gamma = 1 - CDF_w(I)$. Gupta et al.[65] propose an improved version of AGCWD, the gamma correction adaptive gamma correction with color preserving framework (AGCCPF) . In a similar manner, AGCPF estimates the probability density function (PDF), cumulative density function (CDF) and gamma correction and then retains the original information by applying a color-preserving framework. The main fix of this process is to smooth the CDF curve transition from lower to higher intensities, thus avoiding the over-enhancement.

4.2 The ProCAncer-I tool for image enhancement

ProCAncer-I proposes an original image enhancement technique based on the CLAHE method, named Region Adaptive CLAHE (RACLAHE), to serve as a model-invariant preprocessing method for improving prostate and prostatic zone segmentation. The aim was to provide a universal pipeline that will enhance models performance regardless of the choice of model for the segmentation task.

Conventionally, the CLAHE algorithm[66] is applied globally on the entire frame of the image. The algorithm utilizes the histogram equalization method[67] in a close neighborhood around a central pixel. Although the histogram equalization is applied frame-wised, the CLAHE algorithm is applied patch-wise enhancing further the contrast of the sub-regions within the frame. The proposed method RACLAHE utilizes the CLAHE algorithm along with the steps described below to transform selected features to be more interpretable for the model. The pipeline is visually presented in

[63] S. Jenifer, S. Parasuraman, and A. Kadirvelu, "Contrast enhancement and brightness preserving of digital mammograms using fuzzy clipped contrast-limited adaptive histogram equalization algorithm," Applied Soft Computing, vol. 42, pp. 167–177, May 2016, doi: 10.1016/J.ASOC.2016.01.039.
[64] S. Huang, F. Cheng and Y. Chiu, "Efficient Contrast Enhancement Using Adaptive Gamma Correction With Weighting Distribution," in IEEE Transactions on Image Processing, vol. 22, no. 3, pp. 1032-1041, March 2013, doi: 10.1109/TIP.2012.2226047
[65] Gupta, B., & Tiwari, M. (2016). Minimum mean brightness error contrast enhancement of color images using adaptive gamma correction with color preserving framework. Optik, 127(4), 1671–1676. https://doi.org/10.1016/J.IJLEO.2015.10.068
[66] Reza, A.M. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology 38, 35–44 (2004). https://doi.org/10.1023/B:VLSI.0000028532.53893.82
[67] W. Zhihong and X. Xiaohong, "Study on Histogram Equalization," 2011 2nd International Symposium on Intelligence Information Processing and Trusted Computing, 2011, pp. 177-179, doi: 10.1109/IPTC.2011.52.

Figure 17 and Figure 18. The algorithm that describes the RACLAHE operation is described in the following section.

Let Z be a space where the intensity features values of each frame lie in:

$$FM_Z \in Z, 0 \leq FM_Z \leq Frame\ width * Frame\ height.$$

Each frame is passed from a DL U-Net like structure[9,44] that proposes a reduced size area that includes the prostate gland. Specifically, the initial space Z is reduced into a subspace $Q \subset Z$ and features $FM_Q \in Q$ are selected by reducing the dimensionality of Z space in Q space. The latter operation is described by:

$$Q \simeq 0.25\ Z \pm 0.12\ Z \quad \text{[Eq.1]}$$

Frame is then separated into two subframes, features FM_Q, $FM_Z - FM_Q$ and those areas are the proposed area that contains the whole gland and the remaining area respectively while this process is presented in Figure 17(b). The CLAHE algorithm is then applied on the features FM_Q (proposed area). Specifically, FM_Qpixel intensity features are divided into $8X8$ patches and the number of those patches in each FM_Q is approximately 196. Then, the probability of the occurrence of each pixel's unique intensity value $P^{patch}(i_{FM_Q})$ is given by the following equation.

$$P^{patch}\left(i_{FM_Q}\right) = \frac{Num\left(i_{FM_Q}\right)}{TotNum}, 0 \leq i_{FM_Q} \leq LD^{patch} \quad \text{[Eq.2]}$$

where $Num(i_{FM_Q})$ is the number of occurrences of pixel intensity i_{FM_Q} within the patch, $TotNum$ is the total number of pixels the patch has, LD^{patch} is the range of values, inside each patch, the intensity could be assigned. Consequently, the cumulative distribution for each patch is calculated by the following equation:

$$CDF^{patch}(i_{FM_Q}) = \sum_{k=0}^{i_{FMQ}} P^{patch}(k = i_{FM_Q}) \quad \text{[Eq.3]}$$

The histogram equalized patch is obtained by the following.

$$EqHist^{patch} = round(LD^{patch} - 1) \times CDF^{patch}\left(i_{FM_Q}\right) \quad \text{[Eq.4]}$$

The enhanced area of Figure 17 is constructed from the aggregation of the histogram equalized patches and it is obtained by the following equation.

$${FM_Q}^{trans} = Enhanced\ Area = \sum_{t=0}^{Patches} EqHist^t \quad \text{[Eq.5]}$$

Where with $Patches$ we denote the total number of patches within FM_Q while ${FM_Q}^{trans}$indicates the enhanced area. Finally, the RACLAHE resulte image is given by the following equation as shown in Figure 18.

$$RACLAHE = {FM_Q}^{trans} + FM_Z - FM_Q, \quad \text{[Eq.6]}$$

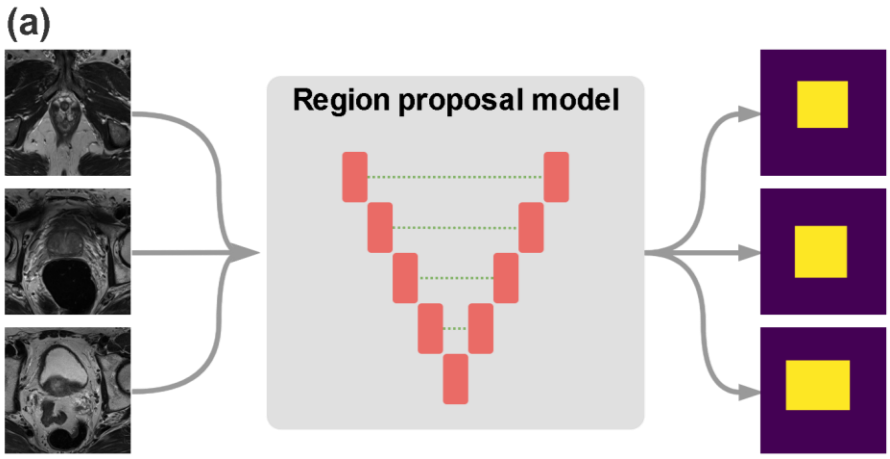


Figure 17. The RACLAHE algorithm. From the initial $256\ X\ 256$ *frame an area of* $\{134 \pm 15\}\ X\ \{134 \pm 15\}$ *pixels are selected that contain the region of interest (A) in a reduced dimensional space which simplify the complexity of the problem.*

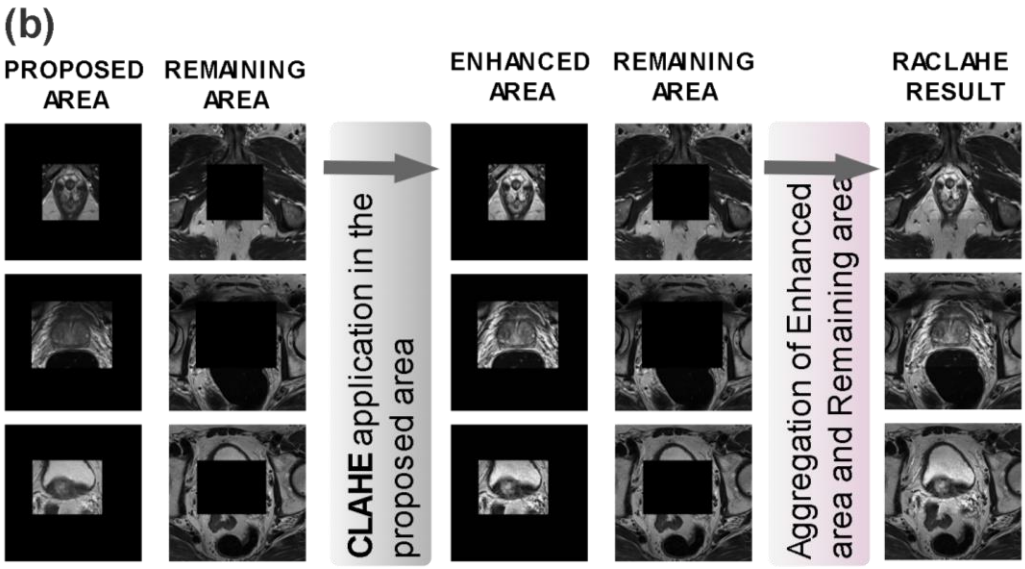


Figure 18. The RACLAHE algorithm. Image separation into the area shown in Fig.1a and the rest of the image, while CLAHE is applied in the proposed area. RACLAHE result is the aggregation of the CLAHE enhanced area and the remaining area.

To provide trustworthiness regarding the RACLAHE preprocessing method in comparison with other methods we tried to quantify how each model is important for the task features diverge from the ground truth binary masks' density map. Specifically, the pipeline is shown in Figure 19.

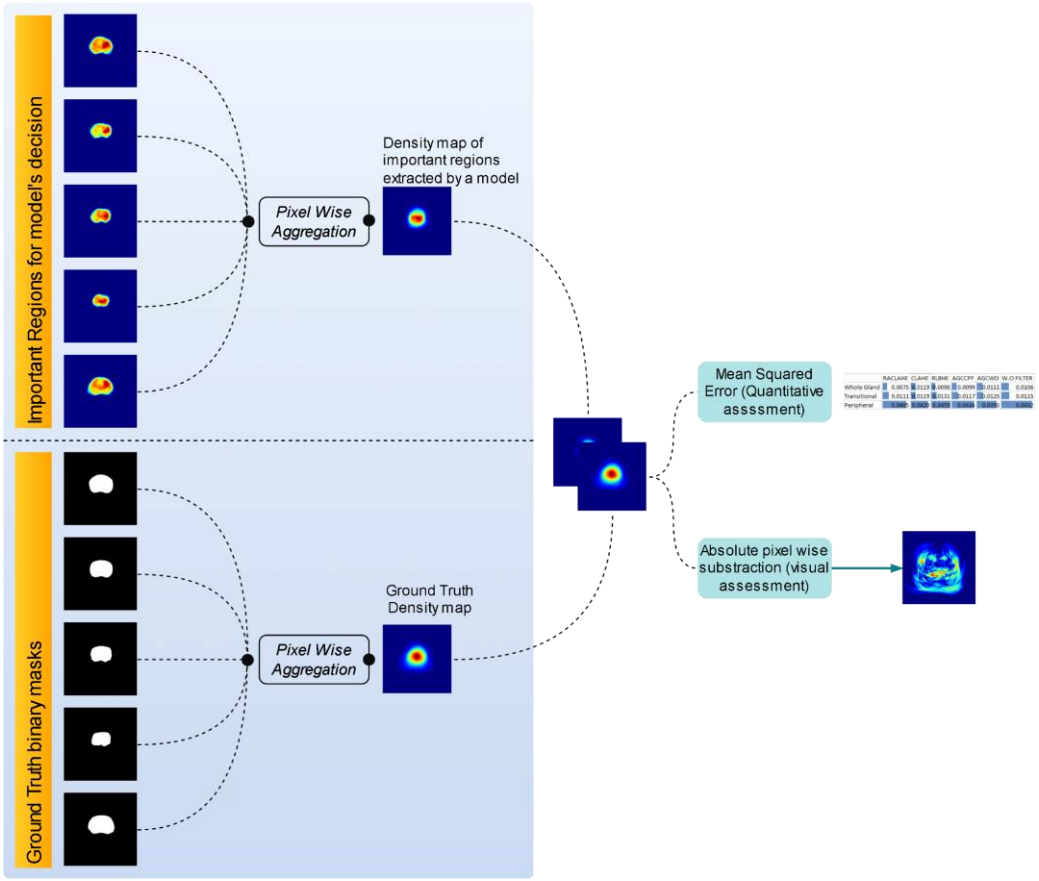


Figure 19. The explainability assessment pipeline. Density maps for GT binary masks and Feature maps are extracted via a pixel wise aggregation. Mean squared error and absolute pixel wise subtraction are performed on the density maps for quantitative and visual inspection.

4.3 Experimental evaluation

The RACLAHE is compared against four popular histogram-based image enhancement methods: i) AGCWD, ii) AGCCPF, iii) RLBHE, and iv) CLAHE. To assess how it affects the performance of the Deep Learning models the i) U-Net, ii) U-Net++, iii) U-Net3+, iv) ResU-Net, and v) USE-NET were exploited. The evaluation metrics were: i) Sensitivity, ii) Balanced Accuracy (BA), iii) Dice Score index (DS), iv) Hausdorff Distance (HD), v) Average Surface Distance (ASD) and vi) Rand Error Index (REI). The tested regions were: i) the whole gland (WG), ii) the peripheral and iii) the transitional prostatic zones.

Whole gland

Although, for WG segmentation the most robust networks tend to perform best without any image preprocessing (i.e., U-Net++, Unet3+, USE-NET), the proposed RACLAHE algorithm was able to improve sensitivity and BA in most cases. AGCWD was efficient in improving Unet and Unet++ but degraded Unet3+. With AGCCPF, only the performance of Unet was improved, achieving results similar to RACLAHE but degraded slightly Unet++, Unet3+ and USE-NET. The CLAHE algorithm was the best preprocessing method for ResUnet model, marginally

outperforming RACLAHE, but degraded other networks such as Unet3+. The RLBHE had the poorest performance compared to other methods with particularly high variability. It is worth noting that the USE-NET model was the best performing network and remained invariant to image preprocessing. Even without any preprocessing, USE-NET has better scores for WG segmentation than all other models (i.e., AUC= 0.88± 0.12). Results are presented in Figure 20.

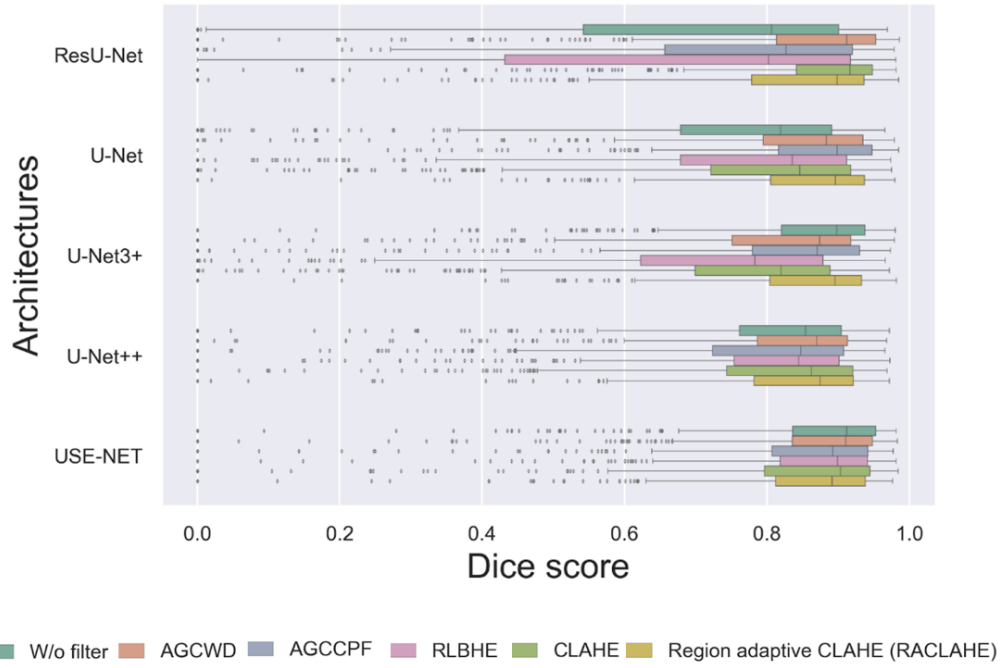


Figure 20. Boxplot of WG segmentation performance for each model and preprocessing techniques.

Peripheral zone

In general, the RACLAHE algorithm consistently improved the performance of the majority of DL models, as it is shown in Figure 21. The only exception was the ResU-net, for which AGCWD and AGCCPF achieved superior performance. Again, for PZ segmentation, models' performance was degraded when the RLBHE was used. The CLAHE algorithm also degraded models' performance, except for USE-NET. The best performance was achieved with the ResU-net model (DS=0.75 ± 0.17 for AGCCPF).

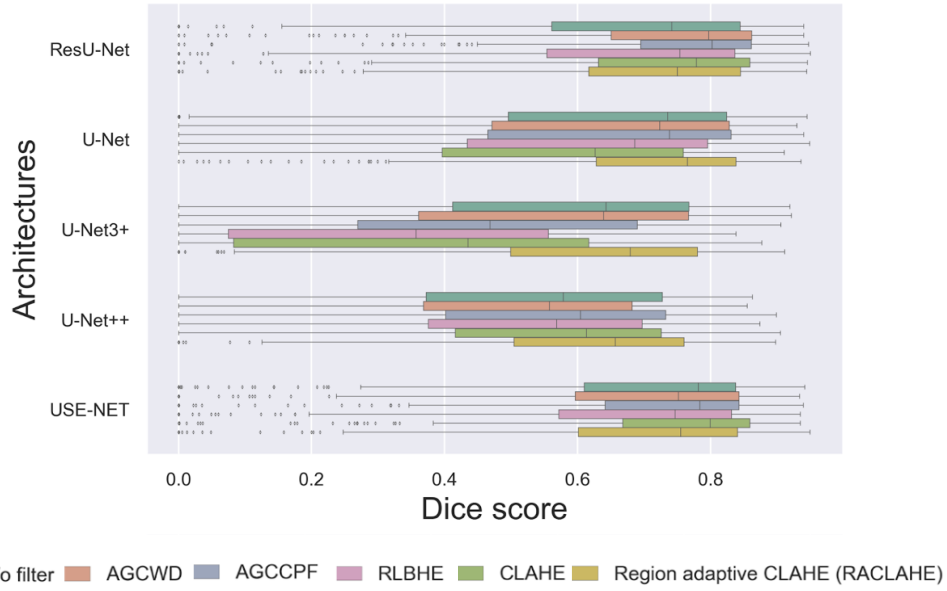


Figure 21. Boxplot of peripheral zone segmentation performance for each model and preprocessing techniques.

Transitional zone

As it is shown in Figure 22, the proposed RACLAHE algorithm was the only consistent preprocessing method for improving the performance of all the five networks. The best results were obtained for RACLAHE with USE-NET (DS=0.81 ± 0.16)

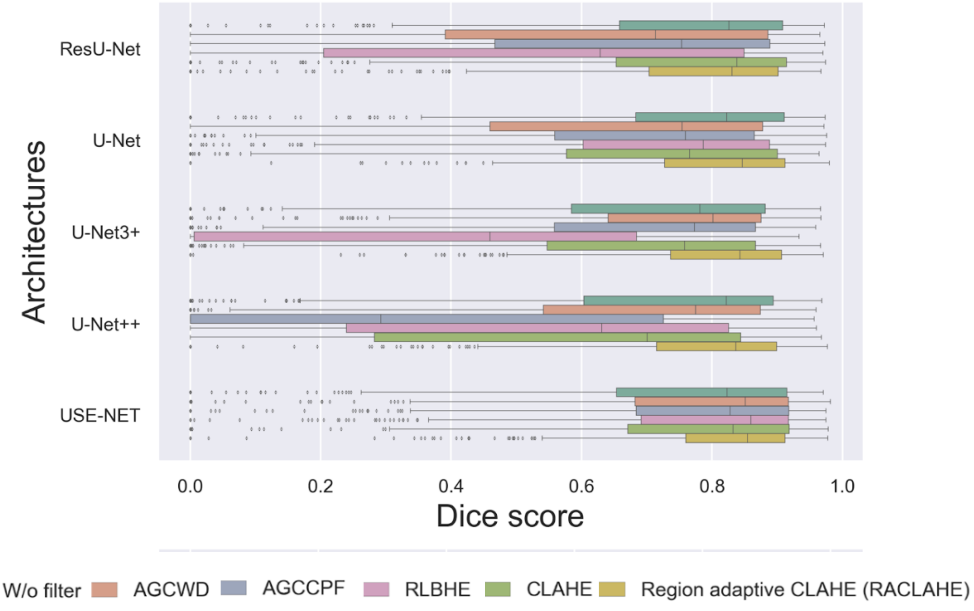
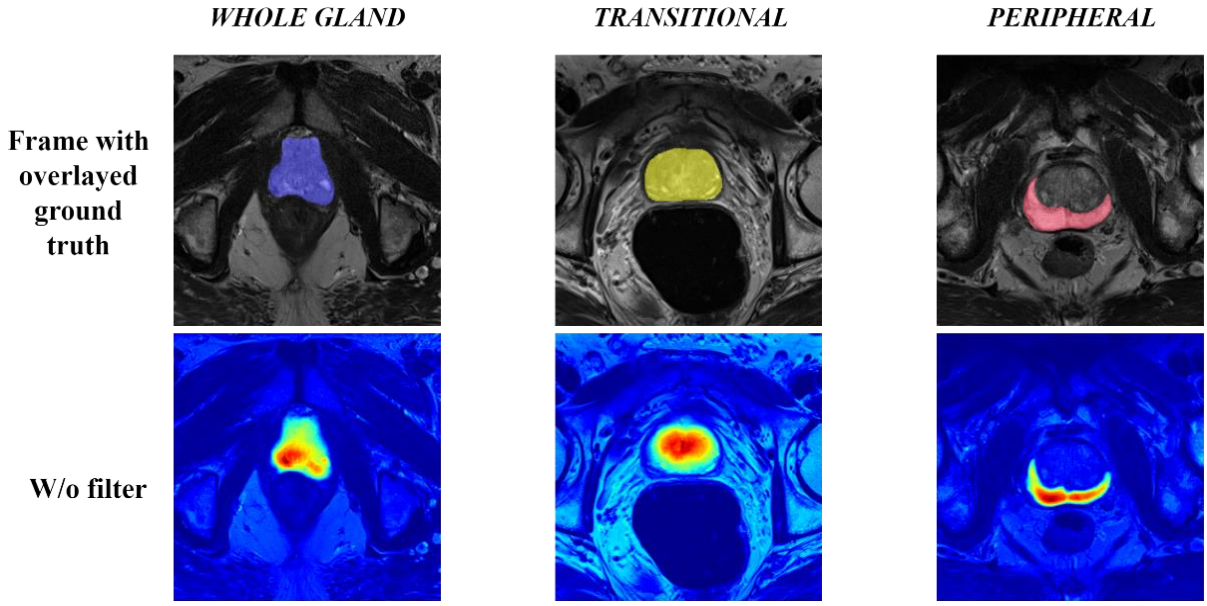


Figure 22. Boxplot of transitional zone segmentation performance for each model and preprocessing techniques.

RACLAHE filter

Figure 23 presents the feature importance the USE-Net model observes under the effect of each filter, applied in Whole gland, transitional and peripheral zone. It appears that RACLAHE filter estimates better the boundaries by giving them more importance compared to the other filters. In general, RACLAHE filter assists further the model to extract robust features and to reduce the uncertainty of the model. Red areas indicate that the model is certain that those pixels belong to the object of interest while yellow areas point that the model is ambiguous if those clusters of pixels are into the object of interest. Blue areas indicate that those pixels have no meaning for the model's final decision. Grad-Cam[68] technique was used to extract the feature importance maps.



[68] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision, 128(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

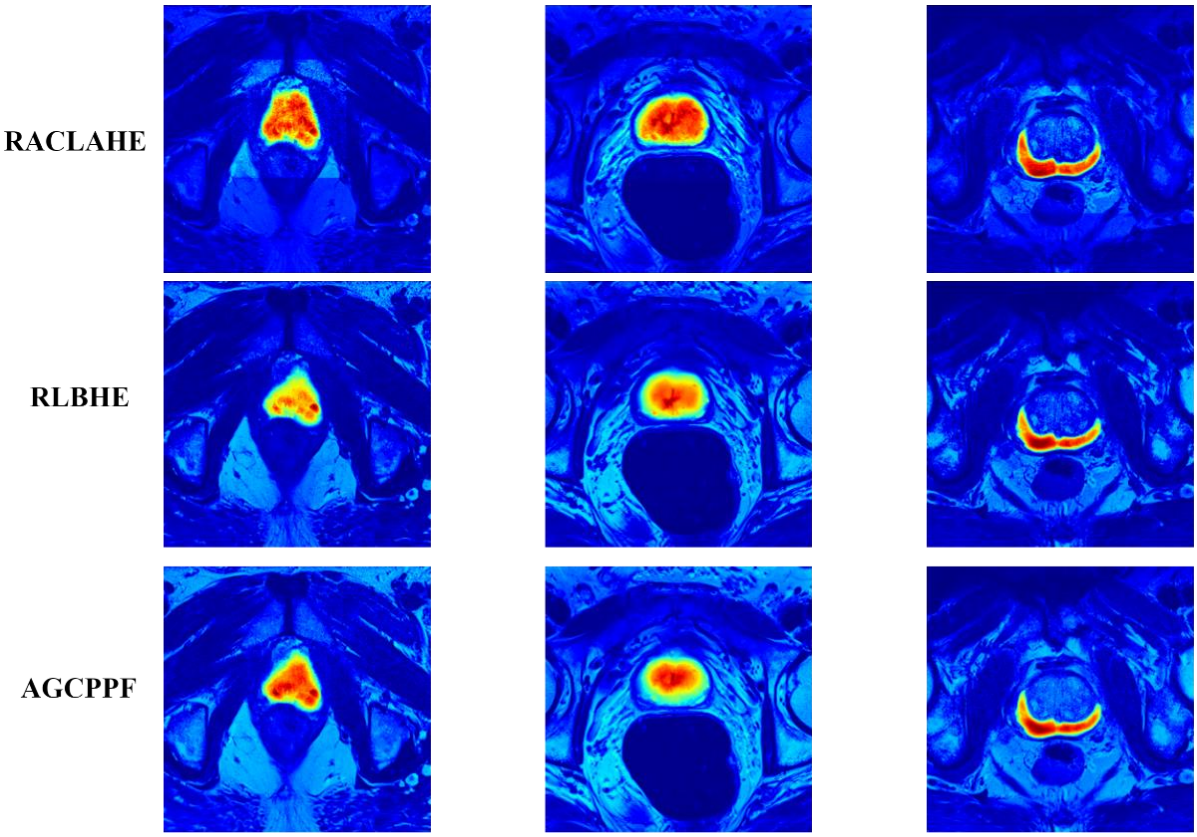


Figure 23. Weight heatmap for USE-Net model and for the used filters. Columns are the prostatic zones while rows are the evaluated filters.

The absolute difference of the GT density maps and the important feature maps a model observes from Figure 23 are presented in Figure 24.

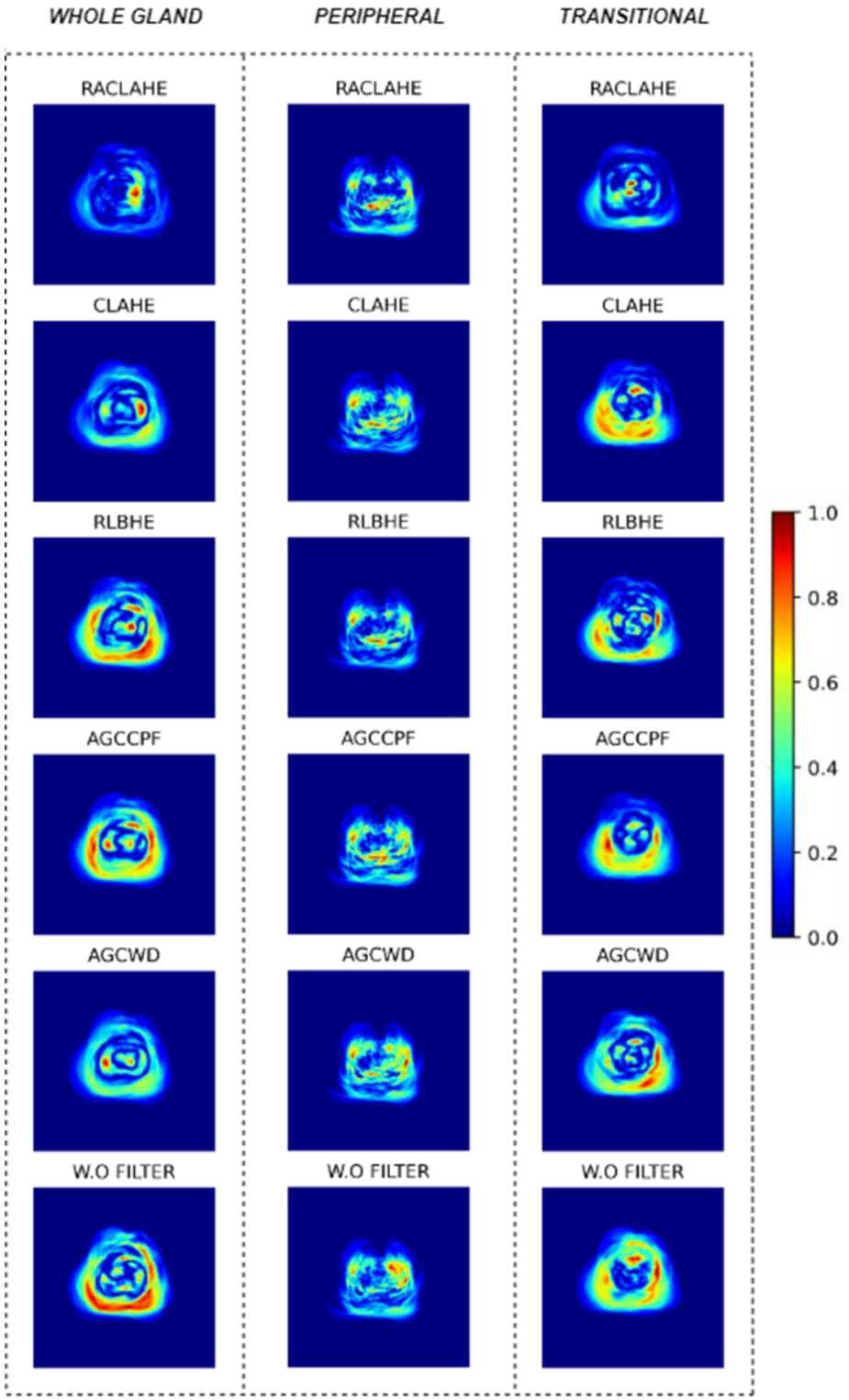


Figure 24. The visual assessment after the absolute pixel wise subtraction of GT density map and Feature map for each filter applied on USE-NET network.

Moreover, another external validation has been performed regarding the RACLAHE algorithm, on the PI-CAI dataset, where for the U-Net model RACLAHE outperformed non-processed pipeline while for the USE-NET model non-processed pipeline outperformed RACLAHE. The testing performed on 2 models which were carefully selected based on their performance and generalizability.

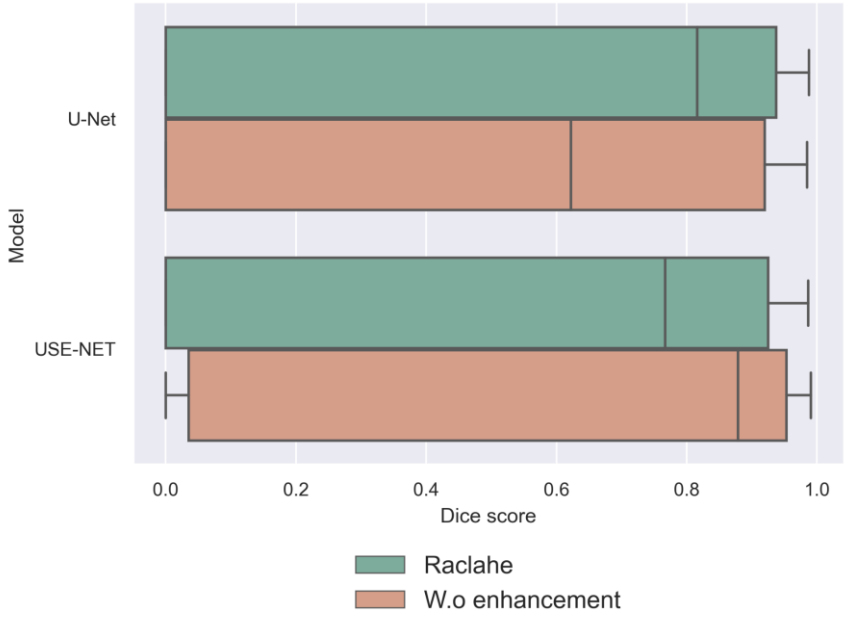


Figure 25. Boxplot of whole gland segmentation performance for 2 models for RACLAHE and without preprocessing.

4.4 Pipeline execution and integration to the ProstateNet platform

The steps for the execution of the RACLAHE pipeline are depicted in the following figure.

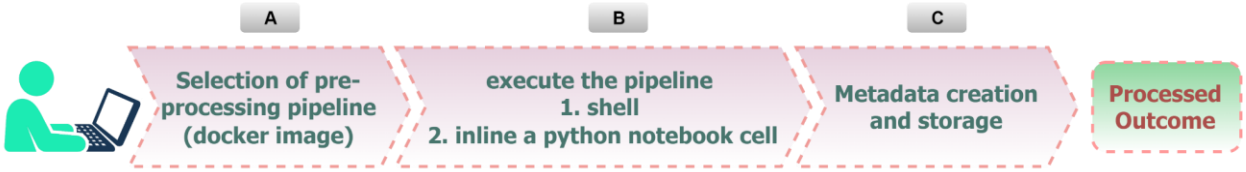


Figure 26. The user selects the desired preprocessing pipeline (A) to perform on the specified input, executes the pipeline in a shell script / python notebook (B). the metadata from the process are stored in the Mlflow, while the result is extracted in the desired folder specified by the user.

Figure 27 analyzes each step of the workflow diagram shown in Figure 26, where multiple diagrams illustrate how the pipeline should operate inside the AI Vault ProstateNet environment.

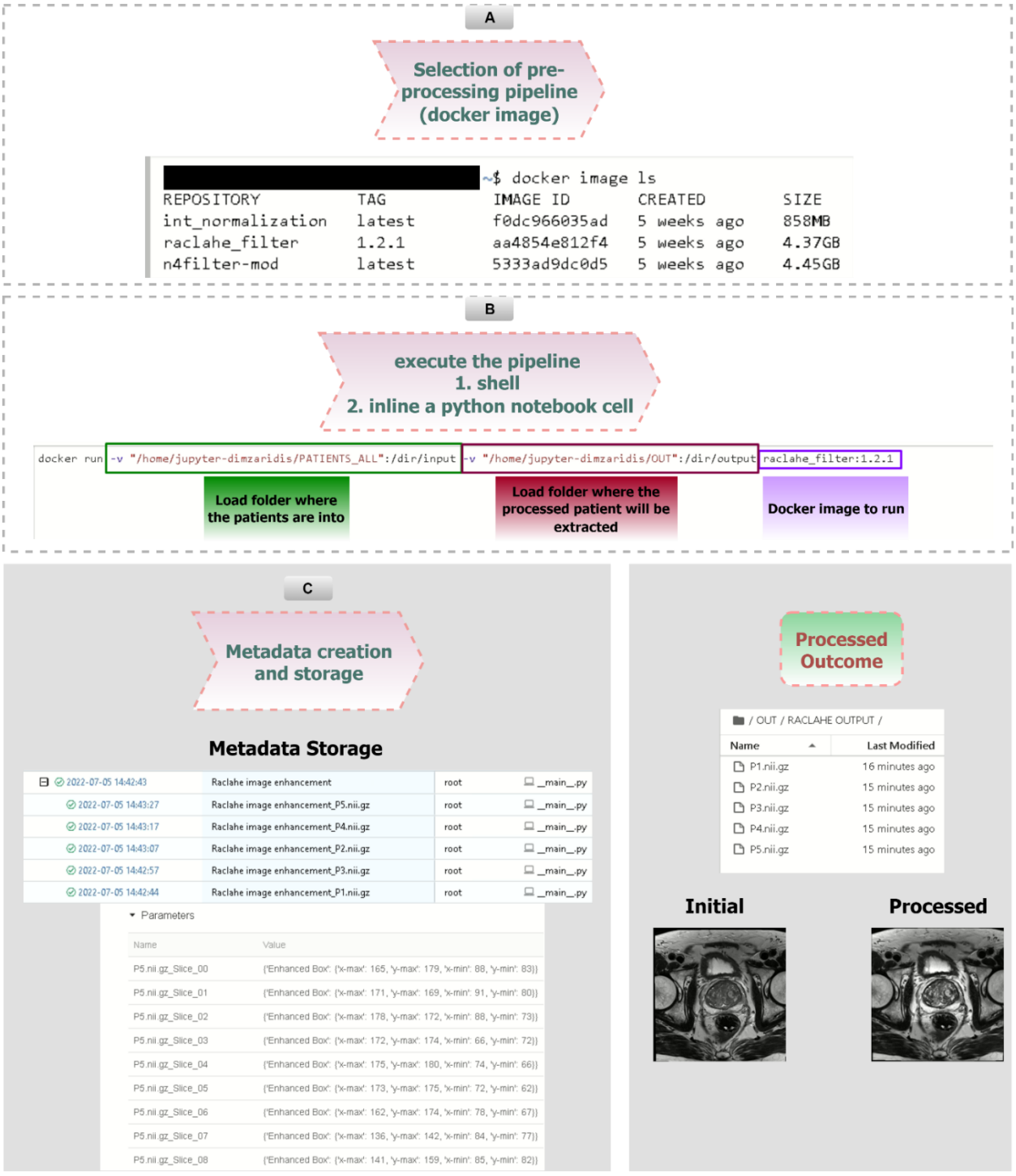


Figure 27. Execution for each step of the RACLAHE pipeline

5. The ProCAncer-I tool for noise reduction

Noise reduction, also referred to in image processing as denoising, is a key preprocessing step in imaging, which is frequently a necessity for accurate analysis. The noisy information must be minimized without compromising the underlying texture component during denoising.

High-resolution MR images with high SNR allow imaging with more detailed anatomical features, boosting diagnostic ability and aiding earlier detection of a variety of illnesses. On the other hand, the signal-to-noise ratio (SNR) is reduced when high-resolution examinations are obtained with a fast acquisition protocol. There are various other ways of achieving higher quality with high SNR in MR scans, including increasing the number of acquired slices, utilizing a strong magnetic field, and adjusting the acquisition bandwidth. However, some of those methods for increasing the scan quality also increase the acquisition time, thereby introducing a significant limitation in the clinical setting. Therefore, several denoising methods can be employed to further enhance scan quality in low SNR examinations. In particular, deep learning (DL) techniques for noise reduction have gained popularity in recent studies[69] because of their texture reconstruction and edge-preserving properties.

Endorectal coil (ERC) MR imaging is used in the prostate because of the distinct advantage it has over surface array coils in terms of image quality that is free from artifacts. This is a critical technique that provides high-resolution scans, which are necessary for optimal examination of the prostate anatomy and other surrounding regions of interest, as well as for detecting cancerous tissue. However, using ERC has some key drawbacks, such as cost and painful or inconvenient application. This might lead to patients refusing or rejecting future MRI exams, resulting in a late diagnosis or inefficient monitoring of the condition. Deep learning reconstruction can not only improve noisy ERC images but also enhance diagnosis without the need for an "invasive" ERC, therefore changing the way prostate cancer scans are currently acquired.

A key advantage of using DL-based models over traditional image processing denoising methods such as average or median filtering is that the parameters of the deep model are optimized during training, while with traditional denoising, a predefined method is used and cannot be tuned for the examined set of data. Unlike traditional methods, deep learning denoising preserves the edges and granular details in texture[70]. This is not the case with traditional methods, which corrupt the original signal and make the image blurry.

5.1 State-of-the-art studies for denoising

Despite the popularity of DL in many image processing and analysis applications, only a handful of studies have been published regarding denoising in medical imaging. Employing these types of commercial or open-source DL methods in computed tomography (CT) for several anatomical

[69] Nakamura, Yuko, et al. "Possibility of deep learning in medical imaging focusing improvement of computed tomography image quality." *Journal of Computer Assisted Tomography* 44.2 (2020): 161-167.

[70] Geng, M., Meng, X., Yu, J., Zhu, L., Jin, L., Jiang, Z., ... & Lu, Y. (2021). Content-noise complementary learning for medical image denoising. *IEEE Transactions on Medical Imaging*, 41(2), 407-419.

areas, including abdominal scans for renal cancer[71], liver[72], lung[73], and pelvis[74], led to significant image quality enhancements. Kidoh et al.[75] successfully applied three types of DL denoising architectures for brain MRI. The fully-convolutional architectures advanced denoising by featuring novel layers such as soft shrinkage, in which a custom activation function is trained to adapt to the noise levels, and discrete cosine transform layers, where the input examinations are processed in the frequency domain. Wang et al.[76] implemented a diverse set of denoising models with prostate MR data from different acquisition protocols. In particular, the author found that the DL-based method for reconstructing non-ERC images performed the best out of the models that were tested. This could indicate that a simple protocol with DL denoising could be used in the clinic.

5.2 The ProCAncer-I tool for noise reduction

Dataset: The T2-weighted images of the ProstateX dataset were used to train and evaluate the DL denoising models. The scans were produced using two Siemens 3T MRI scanners, the MAGNETOM Skyra and Trio. T2-weighted images with a resolution of roughly 0.5 mm in plane and a slice thickness of 3.6 mm were obtained using a turbo spin echo procedure. There are available two patient cohorts: a) 203 patients with their clinical data, gland and lesion annotations; and b) 143 with only the multi-parametric MRI available. In the context of this task, all the 346 scans were used since only the imaging data without annotations was required for the convergence of the denoising models.

Data stratification: The examined dataset of 326 was split into three different sets on a patient-basis. The training set consisted of 276 patients, and it was used for fitting the deep learning models. A validation set of 25 patients was used for tuning the hyperparameters of the deep learning models, stopping the training in models' convergence (early-stopping) and assessing the status of overfitting during training. Finally, an unseen testing set of 25 patient scans was utilized for the model evaluation protocol, providing a fair and robust assessment of the denoising effectiveness. The models were trained and evaluated on a slice-basis, yielding approximately: a) 5500 unique slices for the training set, b) 450 slices for the validation set, and c) 470 slices for the testing set.

Preprocessing activities: A protocol for selecting the highest quality of MRI examinations was established to ensure that the best slices are used for model convergence and evaluation. An experienced radiophysicist evaluated all the 346 T2-weighted scans of the ProstateX dataset. As

[71] Akagi M, Nakamura Y, Higaki T, et al. Deep learning reconstruction improves image quality of abdominal ultra-high-resolution CT. Eur Radiol. 2019;29:6163–6171.
[72] Hur BY, Lee JM, Joo I, et al. Liver computed tomography with low tube voltage and model-based iterative reconstruction algorithm for hepatic vessel evaluation in living liver donor candidates. J Comput Assist Tomogr. 2014;38:367–375
[73] Kakinuma R, Moriyama N, Muramatsu Y, et al. Ultra-high-resolution computed tomography of the lung: image quality of a prototype scanner. *PLoS One*. 2015;10:e0137165.
[74] Tian SF, Liu AL, Liu JH, et al. Potential value of the PixelShine deep learning algorithm for increasing quality of 70 kVp+ASiR-V reconstruction pelvic arterial phase CT images. *Jpn J Radiol*. 2019;37:186–190.
[75] Kidoh, Masafumi, et al. "Deep learning based noise reduction for brain MR imaging: tests on phantoms and healthy volunteers." *Magnetic Resonance in Medical Sciences* 19.3 (2020): 195.
[76] Wang, Xinzeng, et al. "Novel deep learning-based noise reduction technique for prostate magnetic resonance imaging." *Abdominal Radiology* 46.7 (2021): 3378-3386.

a result, 20 scans (approximately 6% of the dataset) were rejected due to severe noise, motion, and other types of artifacts. Additionally, to mitigate the variation in spacing across the MRI examinations, an aspect ratio preserving reshaping with zero-padding and interpolation was applied to the original scans. This resulted in a pixel array of 384 by 384 pixels across all the slices. Prior to the analysis, the pixel intensities of the MRI slices were normalized (unity-based normalization). A Gaussian noise pattern was assumed for generating the synthetic noisy slices. Six noisy images for each real slice were generated with noise thresholds spanning from 4% to 14%. Therefore, approximately 38500 noisy slices were used for convergence and evaluation of the examined deep denoising models. A few noisy slices are depicted in Figure 28.

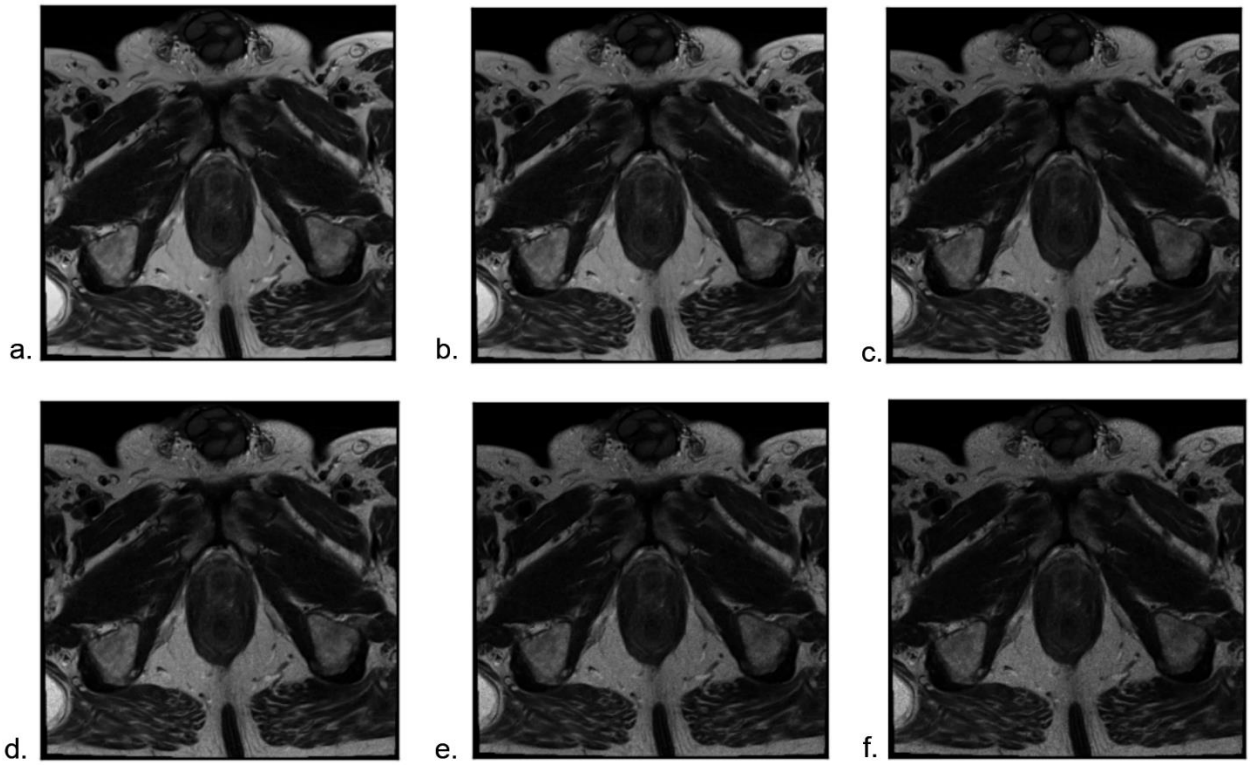


Figure 28. An original (a) prostate T2w slice with different levels of noise (4-12%, b-f) applied.

Data augmentation: In the context of DL analysis, this step is essential for increasing the number of samples that are used during model fitting and to minimize overfitting of the deep models. Aside from increasing the training sample count, data augmentation results in translation, perspective invariance, and artificially introduced variety in the examined dataset, which strengthens the generalization power of the deep models. Four types of off-line transformations were performed: 1) pixel flipping from right to left, 2) pixel flipping from top to bottom, 3) 90-degree image rotation, and 4) 270-degree image rotation. The final training was comprised of approximately 132000 slices. A sample of augmented data is presented in Figure 29.

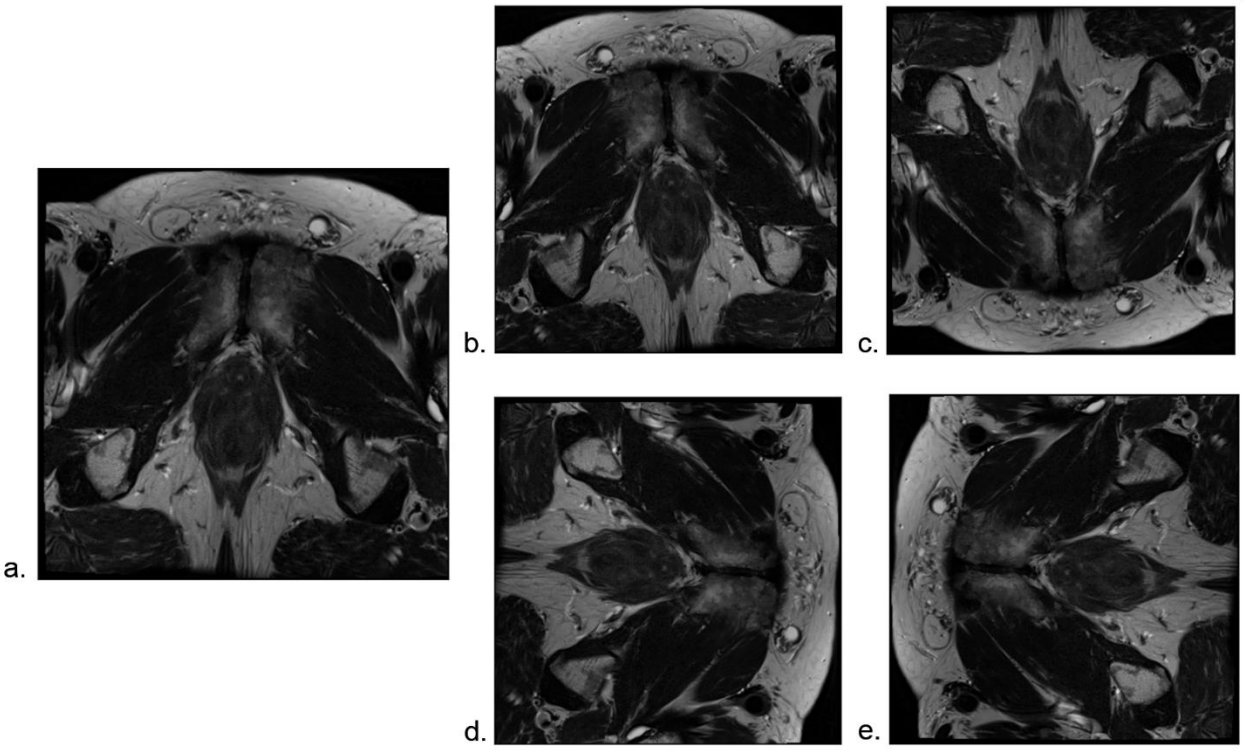


Figure 29. Data augmentation applied to a slice of the training cohort. This includes flipping the original image (a) from right to left (b) and top to bottom (c), rotating 90o (d) and 270o (e).

DL architectures for denoising: Four (4) fully-convolutional architectures were examined: a) a convolutional autoencoder with residual connections (CrAE), b) a denoising convolutional network (DnCNN[77]), c) a denoising convolutional network with residual connections (DrCNN), and d) a real image denoising network (RIDNet[78]). The fully-convolutional models were trained with a supervised learning strategy employing pairs of images; the high-quality ground truth image and the slice with synthetic noise. In most studies, mean squared error (MSE) is used as a loss function, despite the fact that this type of metric does not capture the statistical distribution of image texture. During hyperparameter optimization, the structural similarity index measure[79] (SSIM) was identified as a better method to formulate the denoising task. SSIM encapsulates three key factors for comparing the aforementioned pair of slices: a) luminance (captures the pixel distortions for brighter regions), b) contrast (captures the pixel distortions of regions with high diversity), and c) structure (a sliding window calculates the statistical local dependencies of texture regions). Therefore, the proposed loss integrates these three factors, and it is formulated as an index that captures the structural differences (SDI) between two images. The adaptive moment estimation (ADAM) optimizer was used to minimize the proposed SDI loss between the

[77] Zhang, Kai, et al. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising." IEEE transactions on image processing 26.7 (2017): 3142-3155.
[78] Anwar, Saeed, and Nick Barnes. "Real image denoising with feature attention." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
[79] Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity." *IEEE transactions on image processing* 13.4 (2004): 600-612.

ground truth and the noisy slice. An L1 penalty was applied to the kernels of each layer, constraining the trainable weights of the model from taking outlier values and consequently preventing the model from learning noisy representations that can lead to overfitting. Residual connections were incorporated into the model's architecture to prevent the vanishing gradients[80] effect of the very deep convolutional networks. The integration of the soft-shrinkage activation function[81] was a key integration in the proposed fully-convolutional architecture because it allowed the network to learn thresholds that were proportional to the noise power levels of the examined dataset.

Hyperparameter optimization: This process was very important to the success of the denoising task because hyperparameters are the least reported information in the published studies, and their value is dependent on the dataset that is used. To find the best model hyperparameters, the performance of different configurations on new unseen data (validation set) was evaluated. These hyperparameters are comprised of learnable elements of the architecture (number of layers, modules, kernels, and neurons) as well as other fundamental factors such as the learning rate, optimizer, activation functions, kernel initializers, and regularization penalties. To minimize model overtraining, obtain the most optimal model, and prevent redundant training iterations, early-stopping was implemented with a threshold of 20 epochs after minimizing the validation loss function. Furthermore, comparing the learning curves for loss can reveal information about the fitting state of the deep model. Therefore, to assess the denoising performance and model generalization ability, the learning curves were examined by juxtapositioning the minimum distance between the training and validation loss curves.

5.3 Experimental evaluation

Evaluation metrics: The evaluation of the proposed methodology was conducted exclusively on the unseen testing set as: a) qualitative score by expert radiophysicists or clinicians, and b) quantitative evaluation using the juxtaposition of noisy versus denoised images with metrics such as SSIM, and PSNR, as depicted in Figure 30.

Internal validation (ProstateX): The objective of this task was to identify the best performing model architecture out of several deep learning architectures for denoising in terms of image quality improvements. A modified version of the DrCNN achieved the best image quality performance in terms of PSNR and SSIM on the unseen testing set. The findings suggest that denoised scans from the examined deep model have higher image quality than the synthetically noisy images. DL denoising delivered significant noise reduction with no visible blurring or loss of image quality. Additionally, the edges were preserved, and in many cases, enhanced, the original texture distribution was partially restored, and pixel intensities were closer to the values of the original scan. In particular, the delta between the PSNR of the noisy scan and the denoised scan shows significant improvements of up to 22% and up to 20% for SSIM compared to the ground

[80] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[81] Isogawa, Kenzo, et al. "Deep shrinkage convolutional neural network for adaptive noise reduction." *IEEE Signal Processing Letters* 25.2 (2017): 224-228.

truth image. The quantitative analysis of the image quality improvements for different noise thresholds is depicted in Figure 30. This advancement in prostate MRI denoising, as indicated by the experienced radiophysicist, might be employed in conjunction with ERC or independently with conventional T2-weighted sequences reconstructed by the DL model.

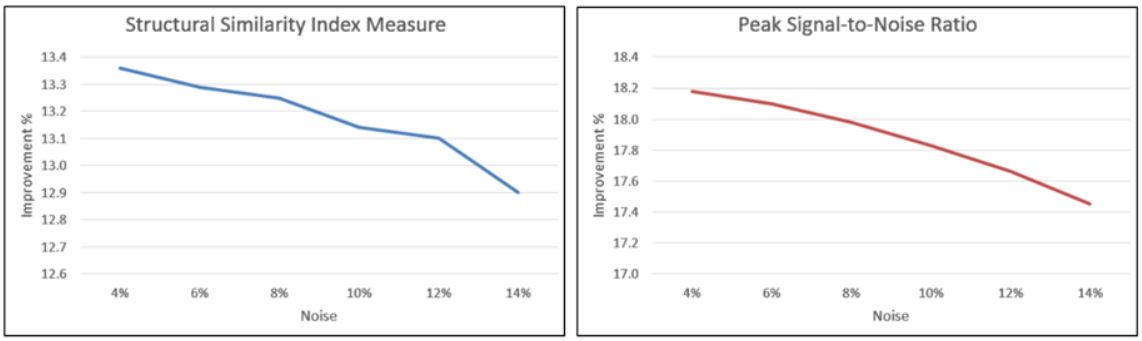


Figure 30. The improvement in image quality of the denoised (custom DrCNN) versus the noisy image in different noise thresholds.

External validation (PI-CAI): The DL denoising module was tested on the previously unseen PI-CAI dataset to evaluate potential undesirable impacts of the DL model, radiomics stability (section 8), and quantify the texture differences. In particular, no artifacts or other distortions were observed after denoising the PI-CAI dataset. Since synthetic noise was not introduced in the dataset, a SSIM-based metric that estimates the differences between denoised and the original examinations was calculated. A difference mean of 0.02±0.04 (4e-4-0.35) was obtained, which is to be expected because in such a dataset that has been curated for image analysis challenges, only a few examinations were found to be noisy (less than 30 with more than 0.1 difference).

5.4 Pipeline execution and integration to the ProstateNet platform

Deployment: The denoising module requires as input data the following meta-image formats: nifti (*.nia, *.nii, *.nii.gz, *.hdr, *.img, *.img.gz), nrrd (*.nrrd, *.nhdr), and meta-images (*.mha, *.mhd) with the original MRI intensities. A normalization on an examination-basis is performed prior to the DL denoising. Any examination shape is accepted, since the DL model is fully-convolutional and can be adapted to the input shape. A universal image orientation (RAI) is enforced across the input meta-images to ensure that the same orientation is applied throughout the denoising of the examinations. The denoised images are exported while preserving the original meta-data, orientation, spacing, MRI intensities, and filenames. The basic instructions for executing this module are provided in Figure 31.

Fine-tuning deep denoising model: To further advance the generalization ability and robustness of the first iteration of the denoising model, fine-tuning with the ProCAncer-I data will be applied to the already trained deep model. This will allow the denoising module to capture the subtle textural and intensity-based differences across multiple scanners and the diverse set of acquisition parameters. The curated and high-quality dataset of this project will further enhance the fine-tuning process by boosting the texture reconstruction performance of the model and by improving the image quality of the denoised examinations.

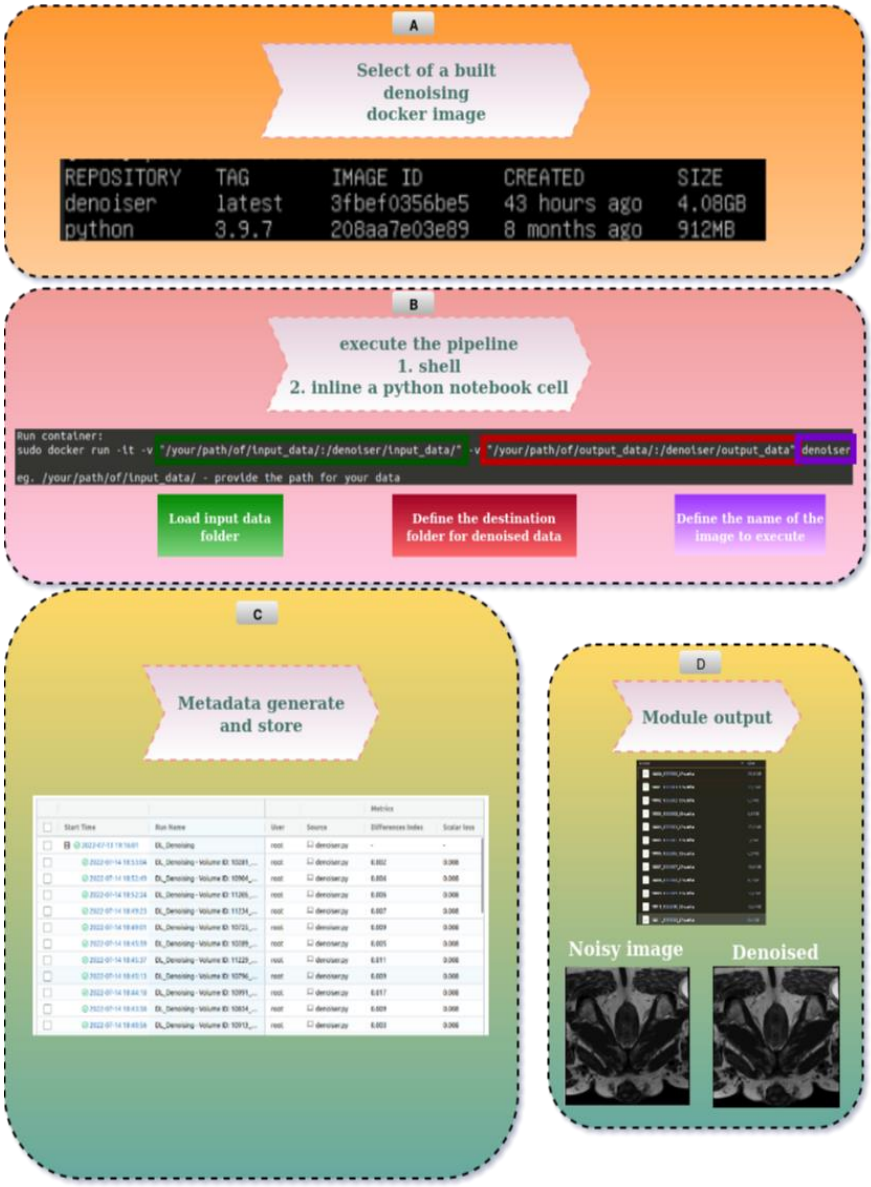


Figure 31. Model execution in the ProstateNet platform: Select the pre-built docker image, execute the corresponding shell command, monitor the progress of the task in the MLFlow frontend interface, and find the denoised data in the defined output folder.

6. Image normalization and harmonization methods

The deviation in intensities of MRI scans is a central bottleneck for quantitative MRI analysis, in terms of comparison and reproducibility: the nature of MR signal measured in arbitrary units hinders direct quantitative analysis of MR images from conventional T1 or T2 contrasts without any prior (post-acquisition) processing. In more details, intensity values are not only highly dependent on acquisition parameters, but also on the subject and body region being scanned; this is especially true in large-scale multicenter studies, involving a significant number of subjects scanned with different scanner types and scanning parameters as in ProCAncer-I. Moreover, when considering radiomics analyses, several works have proved that the extracted features are sensitive to acquisition/reconstruction parameters, pre- and post-processing methods. This issue is worsened in the case of MR Imaging data due to the lack of physical interpretability and arbitrary nature of MRI intensity values.

As a consequence, in the last decades, the significant number of publications focusing on image normalization and data harmonization increased a lot. We refer to normalization methods as to the family of techniques aiming at bringing all images to a common signal intensity space, when among data, a certain variability is expected (e.g. also within a single subject at consecutive MRI scans); on the other hand, we refer to harmonization methods as to the family of techniques (such as machine learning, image/signal processing) aiming at integrating multicenter datasets (reducing the so-called center effect, or cohort bias), and reducing their non-biological heterogeneity[82]. The harmonization can be performed ex-ante on the MRI data or ex-post on the extracted radiomics features. Both normalization and harmonization focus only on intra-modality datasets.

When considering the ex-ante approach, the harmonization usually relies on intensity normalization of MRI data, aiming at the correction of scanner-dependent variations. The goal is twofold: (a) to make intensities have similar distributions for same tissue types within and across patients and (b) make intensities have a common interpretation across locations within the same tissue type. This way, it is possible to steer appropriate and consistent radiomic features or deep learning analyses for diagnostic or prognostic decision support.

In the frame of ProCAncer-I, hosted images come from 3 different vendors, two different magnetic field strengths and, moreover, images are acquired either with surface coils or combination of surface and endorectal coil acquisitions; therefore, data normalization and harmonization either at a signal intensity level or at a feature level is an important stage of the pre- and post-processing pipeline.

In the following sections, we firstly analyse the most common techniques at the state of the art and then report on their application to prostate cancer data. We also introduce an ad-hoc harmonization method devised within ProCAncer-I.

[82] Yang Nan at al., Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions, Information Fusion, Volume 82, 2022, Pages 99-122, doi:10.1016/j.inffus.2022.01.001

6.1 State-of-the-art studies for normalization and harmonization methods

The following normalization methods (from a to d) have been considered and are available to enable the user to normalize images, or to harmonize among datasets, as explained at the end of this subsection (e).

a) Mean normalization

This is one of if not the simplest normalization method. It consists in rescaling the image intensity values by subtracting the mean value of the whole scan.

b) Gaussian normalization

This is one of the simplest methods. It rescales the intensity values I by the standard deviation SD of the image data in the whole scan:

$$INEW = I/SD$$

The method assumes that each scan has the same intensity distribution, and the rescaling is done based on this assumption.

c) Z-Score normalization

This method, referred to as Z-score, is also known as zero mean unit variance. It rescales and shifts the intensities by:

$$INEW=(I-\mu)/SD$$

where I is the intensity, μ is the mean intensity and SD is the standard deviation of the whole scan. The method assumes that each scan has the same intensity distribution, to perform rescaling and shifting. All means are rescaled to zero, but the means do not automatically correspond to the same tissue type.

d) Histogram normalization

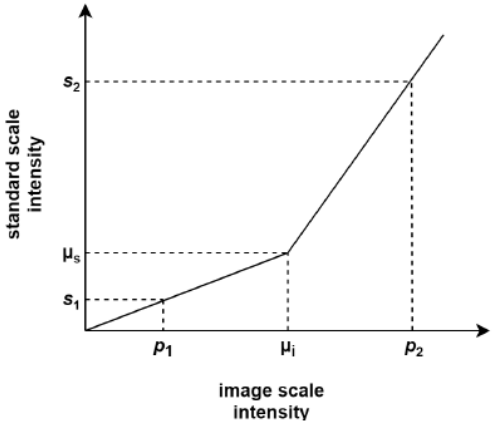
Histogram normalization is one of the first consistently successful normalization methods. It was developed by Nyul and Udupa in 1999 [83][1] and further refined in a succeeding work [84][2]. Although relatively old, the method is still used today in various studies, one of the most recent ones on prostate cancer [85].

[83] L. Nyul e J. Udupa, On standardizing the mr image intensity scale., Magnetic Resonance in Medicine, vol. 2, n. 6, pp. 1072-1081, 1999
[84] L. Nyul, J. Udupa e X. Zhang, New variants of a method of mri scale standardization, IEEE transactions on medical imaging, vol. 19, n. 2, pp. 143-150, 2000
[85] V. Giannini, S. Mazzetti et al. A Fully Automatic Artificial Intelligence System Able to Detect and Characterize Prostate Cancer Using Multiparametric MRI: Multicenter and Multi-Scanner Validation Frontiers in Oncology, 11, 2021, 10.3389/fonc.2021.718155

The method consists in matching the intensity histogram of the images to a standard histogram that is determined by a training process. This is done in a piece-wise linear fashion that ensures that certain landmark points (such as the mean or median intensity) are mapped onto the same value on the standard scale.

The first step of histogram normalization is identifying the images' foregrounds to isolate pixels of interest from the background. This can be done by thresholding the image using the average pixel intensity of all the slices of the scan. The ensuing normalization training procedure is then done only on the foreground of the images. The mode of the foreground is taken as the median landmark. Two other landmarks can be chosen as the 0th and 99.8th percentiles of the entire intensity histogram, although the 2nd and 98th percentiles are also a possible choice.

During the transformation step these three landmarks are derived from the intensity histogram and piece-wise linear mapped to the standard scale landmarks. The standard scale landmarks are determined during the training step with the use of multiple scans. This is done for a specific body region and protocol, usually with a subset of images from the study cohort.



$$x' = \begin{cases} \mu_s + \frac{s_1 - \mu_s}{p_{1i} - \mu_i}(x - \mu_i), & \text{if } x \leq \mu_i \\ \mu_s + \frac{s_2 - \mu_s}{p_{2i} - \mu_i}(x - \mu_i), & \text{if } \mu_i < x \end{cases}$$

Figure 32. Illustration of the piece-wise linear mapping used in the different normalization methods. The landmark intensities (μ_i) for the different images on the original image scale are all mapped to the same value, μ_s on the standard scale. Different linear functions map the intensities less than the landmark value and greater than the landmark value.

Standard scale landmarks s1 and s2 are chosen in such a way that the mapping is one-to-one, and that the histogram is not compressed. The mode of the foreground is mapped on this new standard scale and the average remapped mode of the training scans is taken as the standard scale median landmark [86].

[86] M. Jansen, Evaluation of intensity normalization methods for MR images., University Medical Center, Master Thesis, Utrecht, 2015

In the following Section (6.2.3 and 6.2.4) two normalization methods based on the Nyul-Udupa algorithm will be described. The piece-wise linear mapping is applied looking at the distributions of one or two specific tissues, i.e.periprostatic muscle and the fat.

(e) From normalization to harmonization: an example

Both the state-of-the-art normalization methods described above, and the four different normalization approaches described in Section 6.2, based on "white stripping" and on the piece-wise normalization, can be used to harmonize data coming from different datasets.

In a few words, any normalization method, applied on a large scale, as required to extract knowledge from a large dataset, is a way to perform an ex-ante harmonization on the MRI data.

Recently, Crombé et al.[87] compared standard normalization, Z-score normalization, standardization per signal intensities of healthy tissue, histogram matching (with diverse reference histograms) and ComBat harmonization methods for enhancing the MFS (metastatic-relapse-free survival) predictive models on T2W MRI images of sarcoma patients. In that study, intensity histogram matching, which is the alignment of all intensity histograms to a *reference intensity histogram*, performed better with an AUC of 0.823 in an unsupervised analysis. There could be more than one reasonable choice of the reference intensity histogram, e.g., one patient histogram (randomly chosen), the average intensity histogram of the whole normalized MRI dataset, or the average among the intensity histograms of each datasets are the most used.

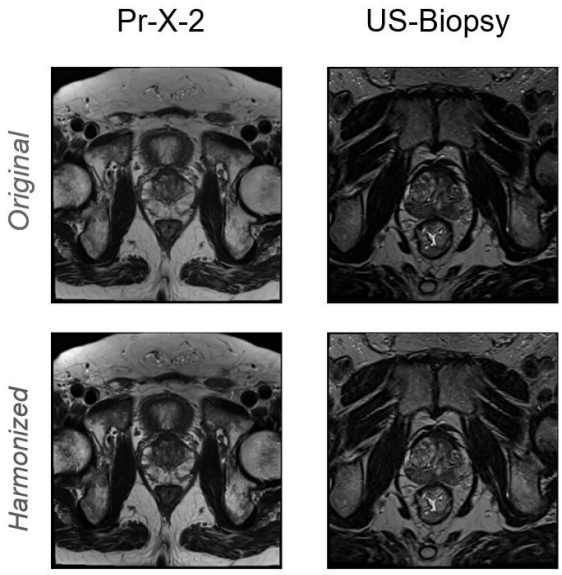


Figure 33. The original and the harmonized (by Nyul histogram matching) images of the preliminary test are shown: one sample image for the ProstateX-2 and one for the Prostate-MRI-US-Biopsy.

Concerning the harmonization of prostate cancer images, we performed a test using two well known prostate cancer datasets: the histogram normalization algorithm (using 10 landmark

[87] Crombé A., Kind M., Fadli D., Le Loarer F., Italiano A., Buy X., Saut O. Intensity harmonization techniques influence radiomics features and radiomics-based predictions in sarcoma patients. Sci. Rep. 2020;10:15496. doi: 10.1038/s41598-020-72535-0

points with a 10 percentile-step) was applied to all T2W images of ProstateX-2 and Prostate-MRI-US-Biopsy[88]. An example of original and harmonized images for both datasets is provided in Figure 33, while Figure 34 shows, qualitatively, that intensity histograms get closer after harmonization.

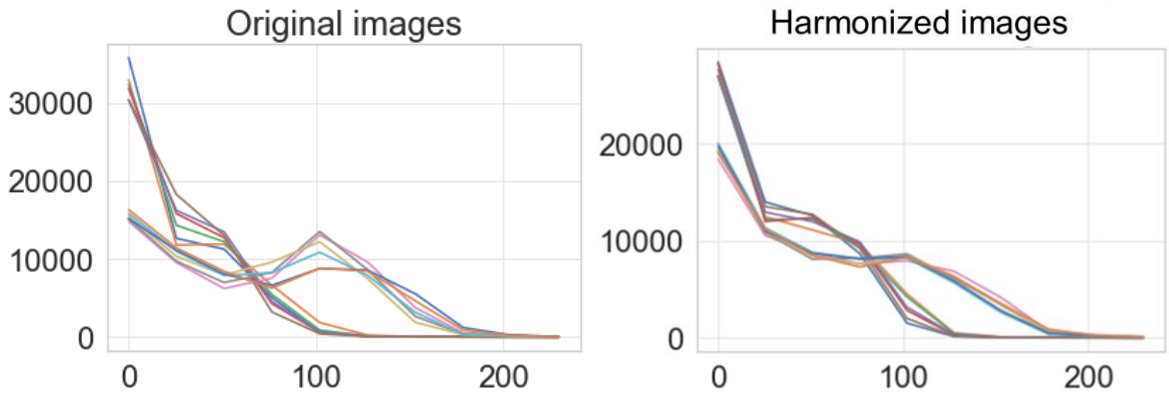


Figure 34. Visual comparison of how the histograms of 12 images (one randomly chosen subject per dataset; 6 images extracted per subject from the T2w sequence) change after the harmonization performed through histogram matching. Even though it is possible to easily distinguish the two sequences, the two groups of histograms are much closer after the harmonization.

Of course, further investigation is required: (i) to choose how to compute the best reference histogram, and the parameters of the histogram matching (e.g. landmarks); and (ii) to assess the impact of such harmonization in the specific setting of ProCAncer-I (e.g. assessing the stability of radiomic features, or the accuracy performance of DL models), exploiting at best the ProCAncer-I platform functionalities (e.g. filters for patient stratification, and model training frameworks).

6.2 ProCAncer-I normalization method

As mentioned before, the notion of normalization is any transformation of assigning a certain value to the initial value of a voxel based on a well-defined rationale. The initial value is subject and visits specific intensity density. The final pixel value can be described/conceptualized as a mixture of densities independent of subjects and/or visits. In order to constitute a robust and reliable method for this correspondence, Shinohara et al [89] proposed a set of 7 statistical principles that have to be respected. In detail, those principles are:

[88] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=68550661

[89] Shinohara RT, Sweeney EM, Goldsmith J, Shiee N, Mateen FJ, Calabresi PA, Jarso S, Pham DL, Reich DS, Crainiceanu CM; Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing; Alzheimer's Disease Neuroimaging Initiative. Statistical normalization techniques for magnetic resonance imaging. Neuroimage Clin. 2014 Aug 15;6:9-19. doi: 10.1016/j.nicl.2014.08.008. Erratum in: Neuroimage Clin. 2015;7:848. PMID: 25379412; PMCID: PMC4215426.

1. To have a common interpretation across locations within the same tissue type. This principle depends on the definition of the normalization and is crucial for the population-level interpretability of statistical inference from the image intensities.
2. To be replicable. This may be assessed using simulations, or the analysis of data containing replicates.
3. To preserve the rank of intensities. This implies that the result of normalization should be as close to one another as possible to the initial distribution and ordering of its clusters.
4. To have similar distributions for the same tissues of interest within and across patients
5. To be not influenced by biological abnormality or population heterogeneity
6. To be minimally sensitive to noise and artifacts
7. do not result in loss of information associated with pathology or other phenomena.

Principles 5-7 require validation studies in large biologically heterogeneous populations with varying levels of noise and artifacts, as is the case of the ProstateNet repository. In their pioneer work, Shinohara et al set the foundation for robust biologically motivated signal intensity normalization and state that no single normalization method may be available to satisfy all 7 SPIN criteria simultaneously, but the prioritization and the relevance of its criterion is related to the specific purpose of the post-processing.

Having in mind those principles, in the frame of ProCAncer-I a pelvis specific method was proposed in accordance with the concept of “White-Stripping” used to normalize brain images as a pilot robust and biologically motivated technique. In detail, four different approaches were tested regarding their performance for homogenizing the signal intensity space, namely fat-based normalization, muscle-based normalization, single tissue piece-wise normalization and double tissue piece-wise normalization.

The first two methods, the fat-based and the muscle-based normalization technique, are similar to the white stripe normalization technique which was developed for the normalization of the brain images. The single tissue piece-wise normalization and the double tissue piece-wise normalization methods were developed based on the concept of the histogram normalization method of Nyul and Udupa and the white stripe technique.

The common basis of all four proposed methods is the use of an extended area of segmentation assigned to one or two tissue types which are the most abundant in the examined region. The advantage over Nuyl based methods is that the histogram landmarks have biological reference and are tissue type specific. The other advantage of those methods over tissue based signal intensity normalization is the large extent of the reference areas, meaning that apart from an excluded area the whole volume of pixels assigned as fat or muscle participate in the histogram definition. The reason for excluding a certain area in the central part of the image is to exclude the prostate tissue that contains heterogeneous pixel intensities among patients, that can be both in the fat or muscle distributions without the feasibility to automatically identify them and exclude them. Moreover, the prostate is the tissue where lesions are expected to be found, thus

tissue that has unpredictable signal intensities beyond normal. However, the segmentation method allows for a very extensive part of the image to be used as reference.

Therefore, a segmentation method is proposed to automatically segment the muscle and the fat tissue in MR pelvic images and was used in the 4 proposed normalization techniques. More specifically, the N4 bias field correction method was applied to the images to produce images free from bias field artifacts. The configuration of the N4 filter that was identified as optimum from our analysis in Section 2, was used for the bias field correction of the prostate images. Each image was subsequently cropped by removing the 20% of the columns in the middle of the image to automatically remove the heterogeneous prostate gland and simultaneously maintain the largest area of the fat and the muscle tissue. The K-means algorithm was applied to the cropped image, setting K equal to 2, in order to identify the 2 clusters of the low intensity values (i.e. muscle tissue approximation) and the high intensity values (i.e. fat tissue approximation). Especially for the segmentation of the muscle tissue, the 12^{th} percentile of the distribution is calculated in order to remove the 12% of the lower values that correspond to background pixels (representing air) and the vessels. An example of an N4 filtered image after removing the 20% in the middle and the corresponding fat and muscle segmentations (with red color) are presented in Figure 35.

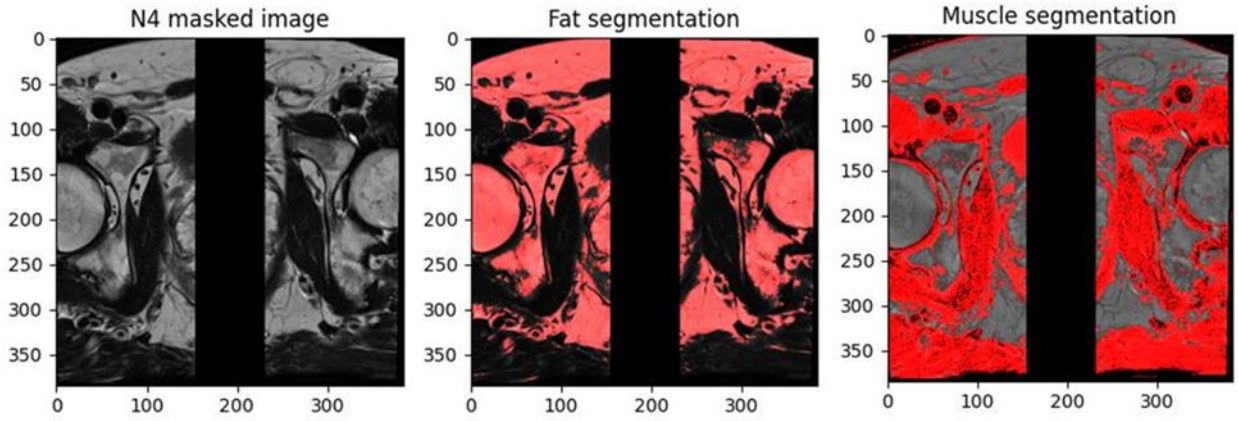


Figure 35. From left to right, the N4 filtered image after cropping the 20% of the image in the middle, the fat segmentation (with red color) and the muscle segmentation (with red color) are presented.

6.2.1 Fat-based normalization

The fat tissue, which was automatically segmented, was used as reference tissue in order to normalize the whole image according to statistics derived from the fat's distribution. More specifically, the fat signal intensity is expected to be the highest among the other abundant tissues of the pelvic region in T2W imaging of the prostate. Thus, in a histogram image obtained from the whole periprostatic area, fat tissue is the image component that forms the distribution at the high values of the spectrum, unless a fat suppressed sequence is used.

The proposed method, called fat-based normalization technique, is a biologically motivated normalization technique, as it is based on the intensity values of the fat tissue. More precisely, in the fat-based normalization method, the image intensity values are transformed according to the following equation:

$$I_{fat-normalized}(x) = \frac{I(x) - \mu_{fat}}{\sigma_{fat}}$$

where μ_{fat} is the mean intensity value of the voxels that correspond to the fat tissue and σ_{fat} is the standard deviation of the voxels that correspond to the fat tissue.

6.2.2 Muscle-based normalization

As opposed to the fat tissue in the whole image histogram, muscle tissue is expected to occupy the low signal distribution of the spectrum. Any other distribution between the fat and muscle components cannot be excluded in the case that no part of the image is selectively excluded. A challenge for muscle segmentation lies in the cut-off value at the low signal intensity side, i.e., the correct cut-off to exclude areas depicting air (noise) and vessels, that both present very low signal intensities. The 12th percentile was used as the cut-off to exclude these areas. The effect of this percentile was assessed by an experienced radiophysicist evaluating the results obtained by different percentiles.

After segmenting the muscle tissue, the muscle-based normalization technique can be applied for the normalization of the image, similarly to the fat-based normalization technique. More specifically, in the muscle-based normalization technique, the image intensity values are transformed according to the following equation:

$$I_{muscle-normalized}(x) = \frac{I(x) - \mu_{muscle}}{\sigma_{muscle}}$$

where μ_{muscle} is the mean intensity value of the voxels that correspond to the muscle tissue and σ_{muscle} is the standard deviation of the voxels that correspond to the muscle tissue.

6.2.3 Single tissue piece-wise normalization

Another approach was to use the distribution of a single tissue, either periprostatic muscle or fat tissue, for learning and extracting landmarks for signal intensity normalization. The metrics derived from the distribution of a single tissue can serve as the upper and lower landmarks, respectively, for restricting the signal intensity spectrum effectively. In this approach, we utilized the piece-wise histogram normalization algorithm proposed by Nyul and Udupa[90] with the only difference of using the muscle or the fat tissue distribution to identify specific landmarks and learn the standard histogram. More specifically, the muscle or the fat tissue were automatically segmented using the aforementioned pipeline. The intensity values of the voxels that correspond to the approximations of the fat or the muscle tissue were given as input to the piece-wise

[90] L. Nyul e J. Udupa, On standardizing the mr image intensity scale., Magnetic Resonance in Medicine, vol. 2, n. 6, pp. 1072-1081, 1999

histogram normalization algorithm of Nyul and Udupa in order to create the standard scale based on the values of a reference single tissue. To this end, we implemented the single tissue piece-wise normalization method, which is either fat-based or muscle-based. The areas of the image that correspond to the fat and the muscle approximations are depicted with white color in Figure 36.

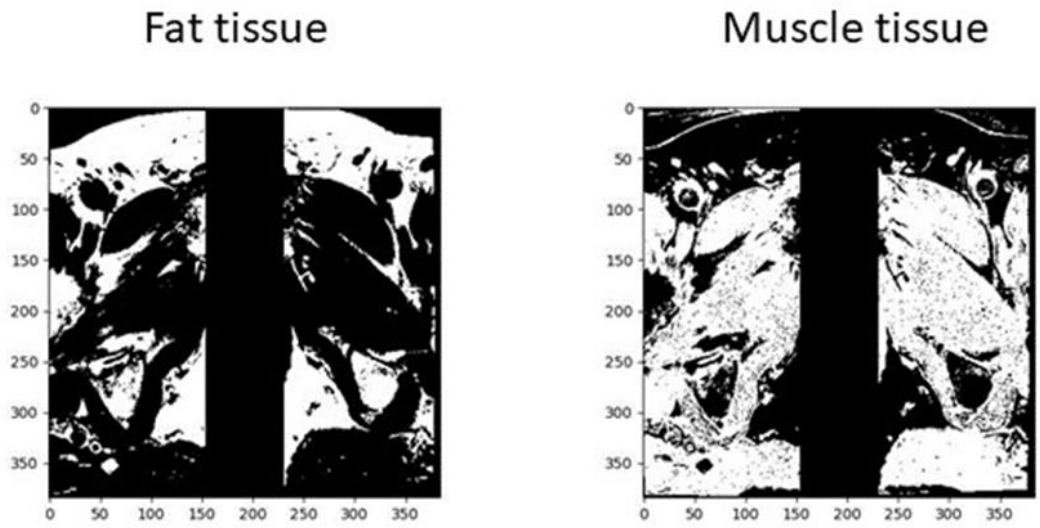


Figure 36. Binary image of the fat (left) and the muscle (right) segmentation. The intensity values that correspond to the white-colored region of the left and the right image are given as input to the piece-wise histogram normalization algorithm of Nyul and Udupa to extract the landmarks and learn the standard histogram in the fat-based and the muscle-based piece-wise normalization method, respectively.

To implement the histogram normalization algorithm of Nyul and Udupa, the values of the landmarks that describe the lower and the upper limit on the standard scale should be defined. These landmarks, called s_1 and s_2, are the minimum and the maximum value on the standard scale, respectively. According to the paper of Nyul and Udupa, the values of s_1 and s_2 should be calculated according to Theorem 2 in order to ensure that the mapping is one-to-one and the histogram is not compressed. Furthermore, intermediate landmarks were calculated in order to create the learned standard scale. The values of the intermediate landmarks were learned during the training phase using the distribution of a specific reference tissue. Hence, the values of the landmarks were extracted from one biological tissue. The lower and the upper limit in the spectrum of the standard scale are restricted by the values that describe a specific biological tissue. After completing the training phase, the standard scale, which consists of the s_1, s_2 and the intermediate landmarks, was used to map the intensity values of a test image in that scale. Hence, the aim of this method is to bring all the images to the same intensity scale based on the distribution of a specific tissue type. During the transformation phase, the corresponding percentiles were calculated in the image scale in order to be mapped with the learned landmarks on the standard scale, resulting in the output normalized image. More specifically, the minimum and the maximum percentile to consider in the image is denoted by p_{1i} (or p_1) and p_{2i} (or p_2). Thus, for each image in the test set, the p_{1i} and p_{2i} on the image scale were mapped to the s_1 and

s_2 on the standard scale, respectively, and the intermediate percentiles on the image scale were mapped to the intermediate landmarks on the standard scale.

We experimented on the number of the intermediate landmarks used in order to investigate the effect of the landmarks on the normalization. Hence, the landmarks were calculated using i) 10-percentile step (i.e. 9 intermediate landmarks); ii) 20-percentile step (i.e. 4 intermediate landmarks) and iii) only the median percentile (i.e. 1 intermediate landmark). Furthermore, we experimented on the values of the minimum and maximum percentiles using the pairs of $p_{1i} = 1^{st}$ and $p_{2i} = 99^{th}$ percentile, $p_{1i} = 2^{nd}$ and $p_{2i} = 98^{th}$ percentile and $p_{1i} = 1^{st}$ and $p_{2i} = 99.8^{th}$ percentile of the image.

6.2.4 Double tissue piece-wise normalization

The last approach was to use both segmented areas, periprostatic muscle and fat for signal intensity normalization as metrics derived from the distributions of those two tissue types. The main difference lies in the use of the distribution of two tissue types, fat and muscle tissue, for learning and extracting the standard scale. Hence, the intensity values of the voxels that correspond to the fat and the muscle tissue (Figure 37) were given as input to the piece-wise histogram normalization algorithm of Nyul and Udupa in order to create the standard scale based on the values of two tissue types. To this end, we implemented the double tissue piece-wise normalization method. Hence, the values of the landmarks of the algorithm were extracted from two specific biological tissues. The muscle tissue corresponds to the lower part of the histogram and thus this tissue type sets the lower limit in the spectrum, while the fat tissue with the high intensity values sets the upper limit in the spectrum of the standard scale. The distribution of the intensity values of the fat and muscle tissue results in a bimodal histogram as each tissue type creates one peak in the histogram.

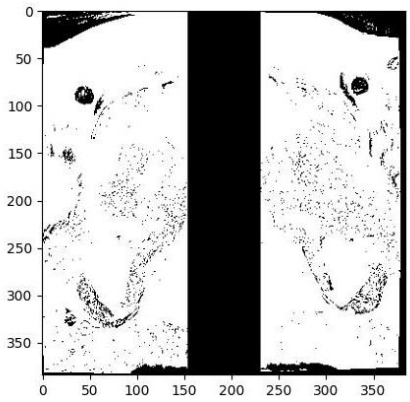


Figure 37. Binary image of the muscle and fat approximation. The intensity values that correspond to the white-colored region of the image are given as input to the piece-wise histogram normalization algorithm of Nyul and Udupa to extract the landmarks and learn the standard histogram in double tissue piece-wise normalization method.

6.3 Experimental Evaluation

To evaluate the four proposed normalization methods, histograms from the unnormalized and normalized whole or tissue-specific images were calculated. Furthermore, the standard deviation

(std) of the Normalized Mean Intensity (NMI)[91,92] is computed for the fat and the muscle tissue in the unnormalized and the normalized images within each dataset. The value of this metric should be as low as possible indicating that the variation in the intensity values of a specific tissue between patients from the same dataset has been decreased. Thus, the tissue has a more similar representation across patients. This metric was used to quantitatively assess the impact of the normalization methods on the images and is given by the following formula:

For patient p and tissue type:

$$NMI_{p,t} = \frac{\mu_{p,t}}{s_{2,p} - s_{1,p}}$$

where μ is the mean intensity, and s1 and s2 are the 2nd and 98th percentile intensity values on the image distribution.

The Prostate-Diagnosis and the ProstateX datasets were used to evaluate the performance of the 4 proposed normalization methods. Each normalization algorithm was applied independently to each dataset. For the single tissue and the double tissue piece-wise normalization methods, a training set is required to learn the standard scale. Thus, the 60% of the patients' images of each dataset was used as a training set. Furthermore, the values of s_1 and s_2 were defined according to Theorem 2[93]. Hence, the s_1 and s_2 were set equal to 1 and 5000, respectively, and 1 and 10000, respectively, for the single tissue and the double tissue piece-wise normalization methods, respectively.

The histograms from the whole images (Figure 38), fat tissue images (Figure 39) and muscle tissue images (Figure 40) draw the same conclusion. The intensity values and thus the histograms of the unnormalized images from different datasets are significantly different, covering a broad range of intensities. The histograms from different patients within the same dataset are also different as their peaks are not aligned. After applying all the normalization techniques, the histograms of either the whole or the tissue-specific images of patients from different datasets overlap, indicating that the variability has been reduced and the tissues have more similar intensity representation across patients of different datasets.

Regarding the patients within the same dataset, the shape and peaks of the histograms of their normalized images coincide and are more similar than the histograms of the unnormalized images. However, they are not perfectly aligned. Regarding the fat tissue, the shape of the histograms of the fat-based normalized and the single fat-based piece-wise normalized fat tissues within the same dataset are more similar than the muscle-based normalized and single muscle-based piece-wise normalized fat tissues (Figure 39). The same effect is observed on muscle tissue. More specifically, regarding the muscle tissue, the shape of the histograms of the muscle-based normalized and single muscle-based piece-wise normalized muscle tissues within the same dataset are more similar than the fat-based normalized and single muscle-based piece-wise

[91] L. J. Isaksson *et al.*, "Effects of MRI image normalization techniques in prostate cancer radiomics," *Phys. Medica*, vol. 71, no. August 2019, pp. 7–13, 2020, doi: 10.1016/j.ejmp.2020.02.007.

[92] L. G. Nyúl and J. K. Udupa, "On standardizing the MR image intensity scale," *Magn. Reson. Med.*, vol. 42, no. 6, pp. 1072–1081, 1999, doi: 10.1002/(SICI)1522-2594(199912)42:6<1072::AID-MRM11>3.0.CO;2-M.

[93] L. G. Nyúl and J. K. Udupa, "On standardizing the MR image intensity scale," *Magn. Reson. Med.*, vol. 42, no. 6, pp. 1072–1081, 1999, doi: 10.1002/(SICI)1522-2594(199912)42:6<1072::AID-MRM11>3.0.CO;2-M.

normalized muscle tissues (Figure 40). It should be mentioned that the double tissue piece-wise normalization method did not produce good results since no improvement is observed in the alignment of the histograms in all examined cases (i.e., whole image, fat tissue only and muscle tissue only).

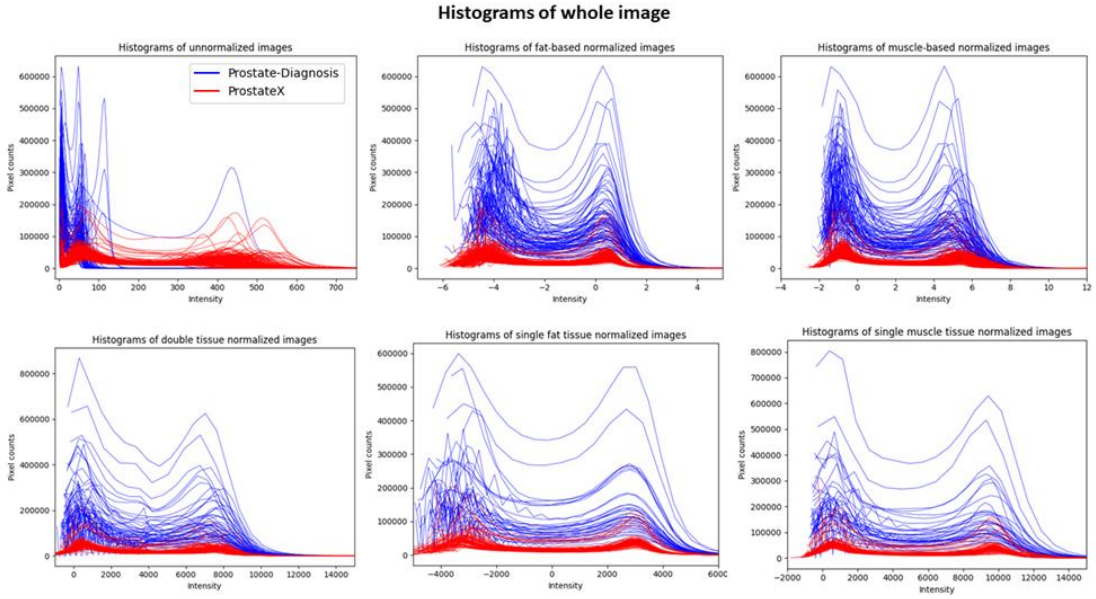


Figure 38. Histograms of whole images. In the first row, from left to right, the histograms of the unnormalized images, the fat-based normalized and the muscle-based normalized images are depicted. In the second row, from left to right, the histograms of the double tissue piece-wise normalized, the fat tissue-based piece-wise normalized and the muscle tissue-based piece-wise normalized images are illustrated.

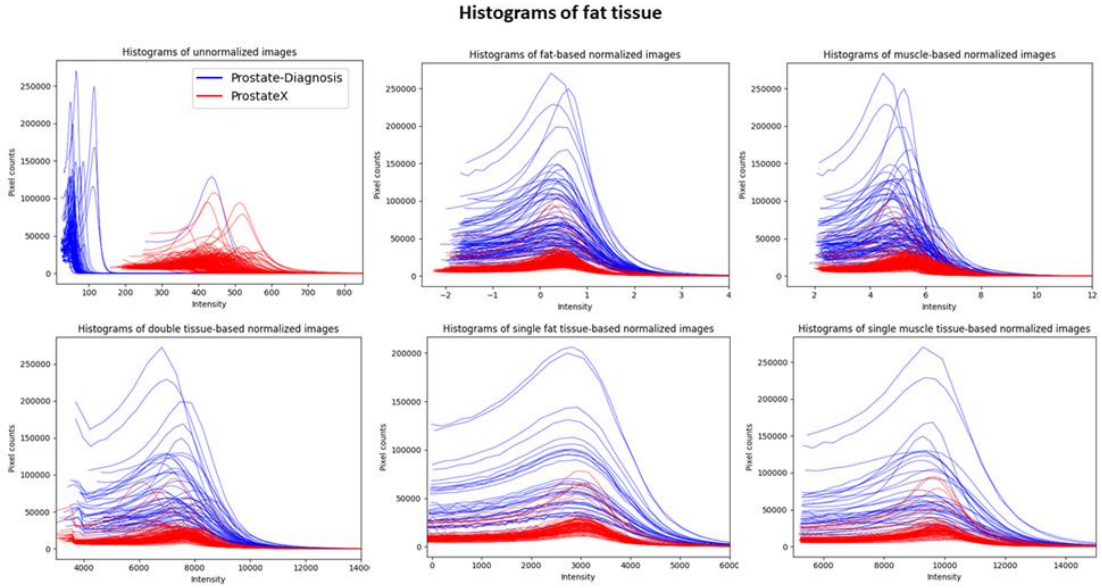


Figure 39. Histograms of fat tissue. In the first row, from left to right, the histograms of the unnormalized images, the fat-based normalized and the muscle-based normalized images are depicted. In the second row, from left to right, the histograms of the double tissue piece-wise normalized, the fat tissue-based piece-wise normalized and the muscle tissue-based piece-wise normalized images are illustrated.

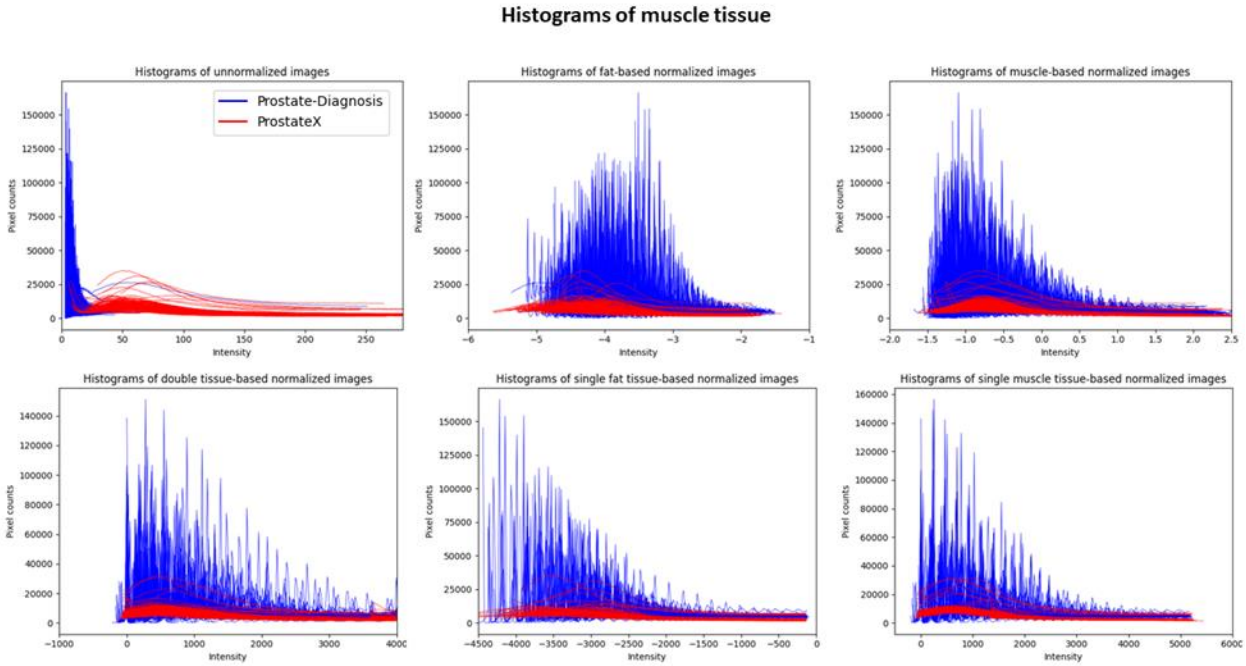


Figure 40. Histograms of muscle tissue. In the first row, from left to right, the histograms of the unnormalized images, the fat-based normalized and the muscle-based normalized images are depicted. In the second row, from left to right, the histograms of the double tissue piece-wise normalized, the fat tissue-based piece-wise normalized and the muscle tissue-based piece-wise normalized images are illustrated.

The std of NMI was calculated for both the fat and the muscle tissue in the unnormalized and the normalized images in order to quantitatively assess the performance of the proposed normalization techniques. The results are presented in Table 6.

The measurements of the std of NMI (Table 6) confirm the conclusions that had been derived from the observations of the histograms. The fat-based normalization methods (i.e., fat-based normalization and fat tissue piece-wise normalization) significantly reduce the variation of the intensity values of the fat tissue within the same dataset (lower values in std of NMI for fat tissue in both datasets). Furthermore, the muscle-based normalization methods (i.e., muscle-based normalization and muscle tissue piece-wise normalization) reduce the variation of the intensity values of the muscle tissue within the same dataset (lower values in std of NMI for muscle tissue in both datasets). However, in the fat tissue, the muscle-based normalization and the muscle tissue piece-wise normalization also reduce the variation of the intensity values compared to the unnormalized fat tissue images. This is not observed in the case of the muscle tissue and the fat-based normalization method (std of fat-based normalized muscle tissue greater than the std of the unnormalized muscle tissue images).

The fat tissue piece-wise normalization method results in lower values of std of NMI for both fat and muscle tissue in both datasets. Furthermore, it significantly reduces the value of the metric compared to the unnormalized images in all examined cases, indicating its efficiency for signal intensity normalization. The muscle-based normalization method demonstrates similar performance with the fat tissue piece-wise normalization method. The muscle tissue piece-wise

normalization method also results in lower values of std of NMI compared to the unnormalized images in all cases, but to a lesser extent in the fat tissue. The fat-based normalization method results in slightly larger values of std of NMI compared to the unnormalized images for the muscle tissue in both datasets. The double tissue piece-wise normalization method has poor performance in normalizing the fat tissue and thus it is not recommended. Therefore, the results show that the muscle-based normalization method and the fat tissue piece-wise normalization method have the best behavior. Taking the results into account, the fat-based, the muscle-based and the single tissue (either fat or muscle) piece-wise normalization methods will be integrated for signal intensity normalization in the ProCAncer-I platform.

Table 6. Intensity variation within fat and muscle tissue measured by the standard deviation of the normalized mean intensity (NMI) in the unnormalized and normalized images of Prostate-Diagnosis and ProstateX datasets. The values in the parentheses represent the rank of the methods based on the std of NMI (lower is better). For the piece-wise normalization methods, the values of the p_{1i}, p_{2i} and the intermediate landmarks-percentiles that resulted in the best performance are presented inside the square brackets.

	STD OF NMI			
	Prostate-Diagnosis (1.5T)		**ProstateX (3T)**	
Normalization method	**Fat tissue**	**Muscle tissue**	**Fat tissue**	**Muscle tissue**
Unnormalized	0.092 (5)	0.055 (5)	0.09770 (5)	0.071 (5)
Muscle-based	0.063 (3)	$2.26e^{-16}$ (1)	0.068 (3)	$1.81e^{-16}$ (1)
Fat-based	$5.84e^{-16}$ (1)	0.066 (6)	$4.32e^{-16}$ (1)	0.073 (6)
Fat tissue piece-wise [p1=2, p2= 98, 10-step percentiles]	0.00067 (2)	0.037 (4)	0.00049 (2)	0.042 (4)
Muscle tissue piece-wise [p1=2, p2=98, median percentile]	0.083 (4)	0.0085 (2)	0.09768 (4)	0.0099 (2)
Double tissue piece-wise [p1=1, p2 =99.8, median percentile]	0.16 (6)	0.024 (3)	0.14 (6)	0.030 (3)

In the muscle-based, the fat tissue piece-wise and the muscle tissue piece-wise normalization methods, the SPIN 1 and 4 are satisfied as the histograms and the std of NMI of the same tissue have low variance and thus similar interpretation across the different patients and the different datasets. The lower variance in these two imaging datasets suggests more replicable measurements (SPIN 2). Furthermore, the rank of the intensities (SPIN 3) is preserved as the fat tissue has higher values than the muscle tissue after applying the normalization methods.

The piece-wise normalization methods were implemented using various normalization schemas regarding the number of the intermediate landmarks and the values of the minimum and the maximum percentiles. The effect of the different normalization schemas was quantified by calculating the std of NMI for each different configuration. The results for the fat tissue and the muscle tissue piece-wise normalization methods are presented in Table 7 and Table 8, respectively. As shown in these tables, the use of different intermediate landmarks results in different values of the metric, while maintaining the values of the minimum and the maximum percentiles unchanged. Furthermore, the use of the same intermediate landmarks but different pairs of minimum and maximum percentiles result also in different performance. However, in the fat tissue piece-wise normalization method, the value of the std of NMI for the muscle tissue remains the same across all the different schemas in each dataset (Table 7). This is not observed neither for the fat nor for the muscle tissue in the muscle tissue piece-wise normalization method (Table 8). Hence, the effect of the percentiles is different when using different distributions for learning and extracting the landmarks. Thus, our analysis indicates that there is no golden standard regarding the number of the intermediate landmarks used and the values of the minimum and the maximum percentiles for the piece-wise normalization methods. In the framework of the ProCAncer-I platform, the schemas that resulted in the best performance (i.e. lower value of std of NMI) for each one of the two proposed piece-wise methods are used. More specifically, 10-step percentiles, p_{1i} = 2 and p_{2i} = 98 are used for the fat tissue piece-wise normalization method. Regarding the muscle tissue piece-wise normalization method, the median percentile, the p_{1i} equal to 2 and p_{2i} equal to 98 are utilized in the proposed algorithm.

Table 7. Intensity variation within fat and muscle tissue measured by the standard deviation of the normalized mean intensity (NMI) in the unnormalized and normalized images of Prostate-Diagnosis and ProstateX datasets using the fat tissue piece-wise normalization method with different landmarks.

	STD OF NMI			
	Prostate-Diagnosis (1.5T)		**ProstateX (3T)**	
Normalization method	**Fat tissue**	**Muscle tissue**	**Fat tissue**	**Muscle tissue**
Unnormalized	0.092	0.055	0.09770	0.071
p1=1, p2=99, 10-step percentiles	0.0030	0.037	0.0041	0.041

p1=1, p2=99, 20-step percentiles	0.0038	0.037	0.0047	0.041
p1=1, p2=99, median percentile	0.0058	0.037	0.0084	0.041
p1=2, p2=98, 10-step percentiles	**0.00067**	**0.037**	**0.00049**	**0.042**
p1=2, p2=98, 20-step percentiles	0.00099	0.037	0.0011	0.042
p1=2, p2=98, median percentile	0.0052	0.037	0.0067	0.042
p1=1, p2=99.8, 10-step percentiles	0.011	0.037	0.010	0.041
p1=1, p2=99.8, 20-step percentiles	0.014	0.037	0.012	0.041
p1=1, p2=99.8, median percentile	0.015	0.037	0.014	0.041

Table 8. Intensity variation within fat and muscle tissue measured by the standard deviation of the normalized mean intensity (NMI) in the unnormalized and normalized images of Prostate-Diagnosis and ProstateX datasets using the muscle tissue piece-wise normalization method with different landmarks.

	STD OF NMI			
	Prostate-Diagnosis (1.5T)		**ProstateX (3T)**	
Normalization method	**Fat tissue**	**Muscle tissue**	**Fat tissue**	**Muscle tissue**
Unnormalized	0.092	0.055	0.09770	0.071
p1=1, p2=99, 10-step percentiles	0.11	0.0027	0.14	0.00052

p1=1, p2=99, 20-step percentiles	0.10	0.0038	0.13	0.0015
p1=1, p2=99, median percentile	0.084	0.010	0.098	0.0093
p1=2, p2=98, 10-step percentiles	0.10	0.0011	0.14	0.00041
p1=2, p2=98, 20-step percentiles	0.10	0.0023	0.12	0.0016
p1=2, p2=98, median percentile	**0.083**	**0.0085**	**0.09768**	**0.0099**
p1=1, p2=99.8, 10-step percentiles	0.11	0.0028	0.14	0.00051
p1=1, p2=99.8, 20-step percentiles	0.105	0.0038	0.13	0.0013
p1=1, p2=99.8, median percentile	0.084	0.010	0.098	0.0089

6.4. Pipeline execution and integration to the ProstateNet platform

The developed methods as described in the previous section have been packaged in a docker image and integrated in the ProstateNet platform. The biologically motivated intensity normalization module contains the 3 proposed normalization techniques: i) fat-based; ii) muscle-based and iii) single tissue (fat or muscle) piece-wise normalization methods. The module requires as input data the images in NifTI format (*.nii, *.nii.gz). Any examination shape of the image is accepted. In the module, the N4 bias field correction method is applied prior to the normalization. The output is the normalized images, which are exported in the same file format (*.nii, *.nii.gz) and shape as the input images. The basic instructions for executing this module are described in Figure 41 and Figure 42. In the run command, the user should specify at least one of the available arguments that correspond to a normalization method. For instance, if the user declares -f, then the fat-based normalization will be applied. If the user wants to use the single tissue piece-wise normalization method, the tissue from which the landmarks will be extracted should be specified by declaring -f for fat tissue or -m for muscle tissue. For instance, if the user wants to use the muscle tissue piece-wise normalization method, the arguments -m -p should be specified. When the argument -p is declared and no other argument is selected, the fat tissue piece-wise normalization method is the default option.

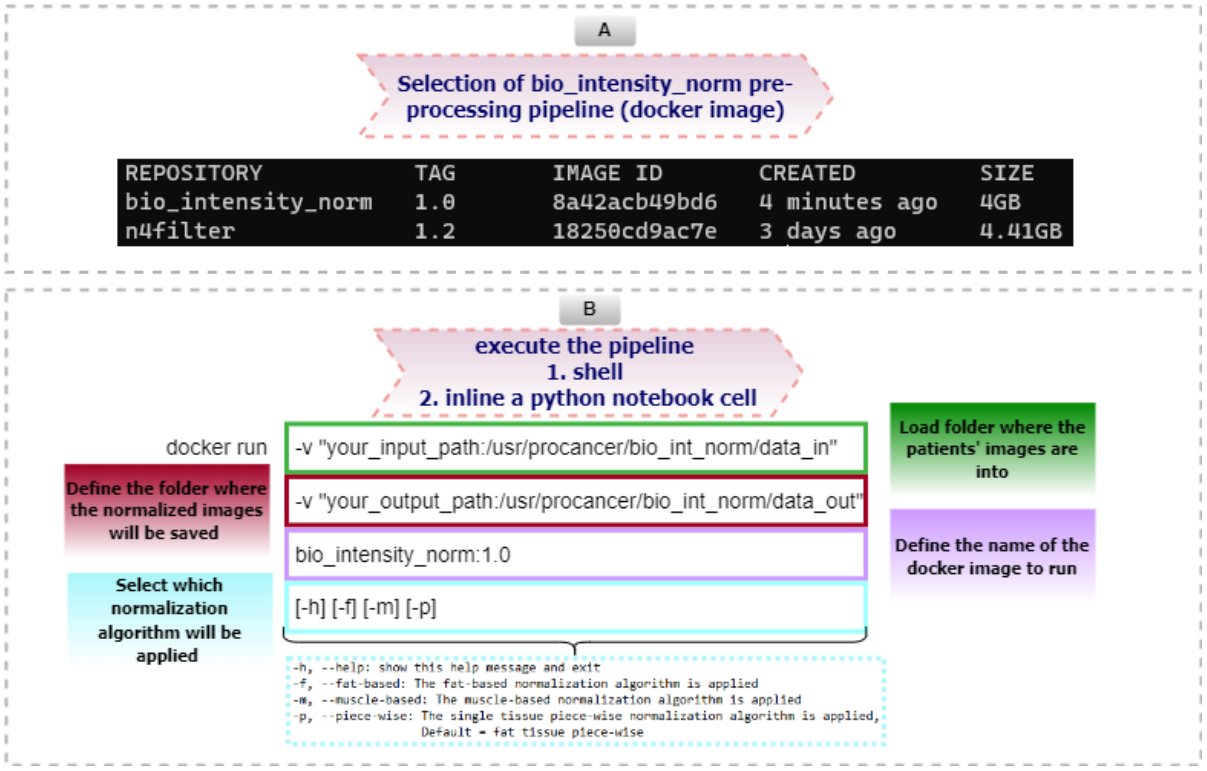


Figure 41. Description of the running command of the proposed biologically motivated normalization module.

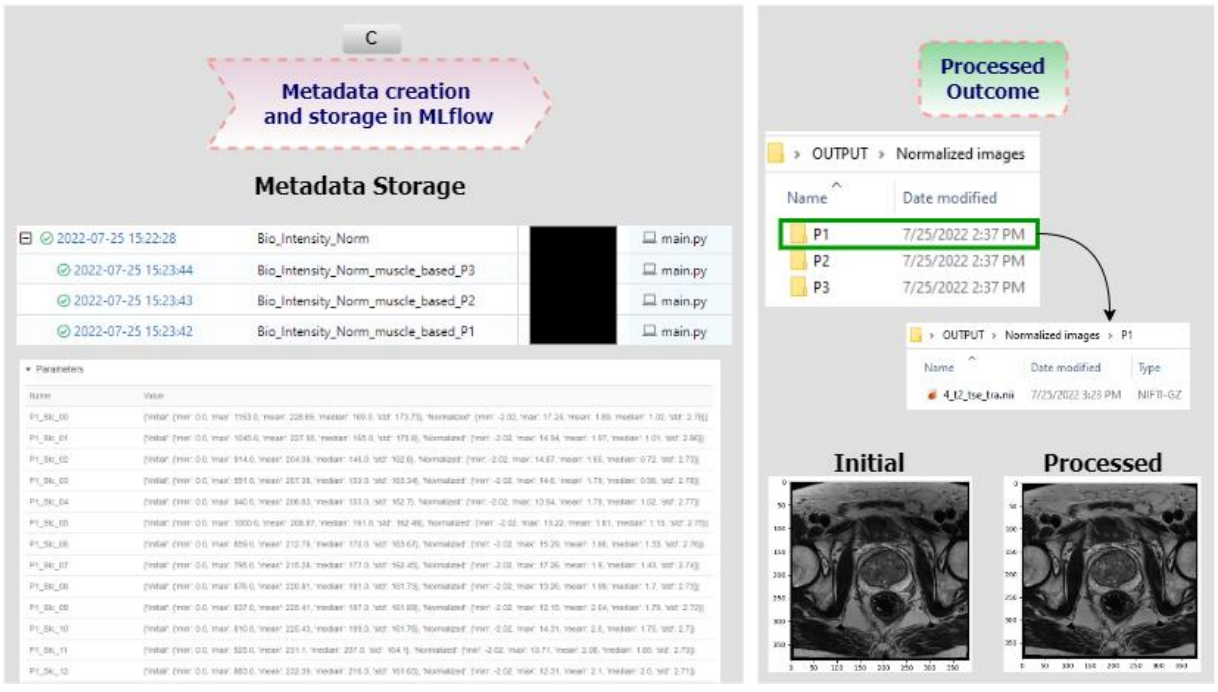


Figure 42. Description of the metadata and the processed outcome of the biologically motivated normalization module.

7. Radiomics Harmonization

Radiomics, a rapidly evolving field in medical image analysis, aims to enable digital decoding of images into quantitative features, aiding earlier and precise care in the clinical decision making [94,95]. However, when analysis is performed on data acquired from multiple centers, different scanner models, acquisition protocols and/or reconstruction settings, variability in feature extraction is unavoidable and can potentially affect radiomics generalization performance and applicability in the clinical routine. In the literature, this variability is called "center-effect" [96]. To address this issue, several computational approaches have been proposed towards making the radiomics features more stable and reproducible. This chapter focuses on the feature-based harmonization techniques developed within the ProCAncer-I project, aiming to surpass center-effect issues after prostate radiomics feature extraction [97].

7.1 State-of-the-art methods

A representative and widely used feature-based harmonization method is ComBat (Combine Batches), a post-reconstruction method to compensate for multicentric/multi-vendor effects [98].

It employs empirical Bayes methods to estimate the differences in features values due to a center effect. Then, these estimates are used to adjust the data. In the radiomics analysis context, ComBat performs location and scale adjustments of the features within centers in order to remove the discrepancies introduced by technical differences in the images. Specifically, it assumes that feature values can be standardized by the equation:

$$Z_{ijg} = \frac{Y_{ijg} - \widehat{a_g} - X \cdot \widehat{\beta_g}}{\widehat{\sigma_g}},$$

where Y refers to the raw radiomics feature g for sample j and center i, $\widehat{a}$ and $\widehat{\sigma}$ the features-wise mean and $\hat{\beta}$ standard deviation estimates, X is a design matrix of non-center related covariates and is the vector of regression coefficients corresponding to each covariate.

The standardized data is assumed to be normally distributed $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$, where γ and δ are the center effect parameters with Normal and Inverse Gamma prior distributions, respectively. Then, ComBat performs feature transformation based on the empirical Bayes prior estimates for γ and δ for each center. The final center effect adjusted values are given by:

$$Y_{ijg}^{ComBat} = \frac{\widehat{\sigma_g}}{\hat{\delta}_{ig}^{ComBat}} \cdot (Z_{ijg} - \hat{\gamma}_{ig}^{ComBat}) + \widehat{a_g} - X \cdot \hat{\beta}_g ,$$

[94] Gillies, Robert J., Paul E. Kinahan, and Hedvig Hricak. "Radiomics: images are more than pictures, they are data." Radiology 278.2 (2016): 563-577.
[95] Van Timmeren, Janita E., et al. "Radiomics in medical imaging "How-to" guide and critical reflection." Insights into Imaging 11.1 (2020): 1-16.
[96] Da-Ano, R., et al. "Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies." Scientific Reports 10.1 (2020): 1-12.
[97] Da-Ano, R., D. Visvikis, and M. Hatt. "Harmonization strategies for multicenter radiomics investigations." Physics in Medicine & Biology 65.24 (2020): 24TR02
[98] Ibrahim A, Refaee T, Leijenaar RTH, Primakov S, Hustinx R,Mottaghy FM, Woodruff HC, Maidment ADA, Lambin P. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. PLoS One. 2021 May 7;16(5):e0251147. doi: 10.1371/journal.pone.0251147. PMID: 33961646; PMCID: PMC8104396.

where $\hat{\delta}_{ig}^{ComBat}$ and $\hat{\gamma}_{ig}^{ComBat}$ are the Empirical Bayes estimates of $\hat{\delta}_{ig}^{2}$ and $\hat{\gamma}_{ig}$, respectively. ComBat method centers the radiomics feature data to the overall, grand mean of all samples. Consequently, the new adjusted data are shifted to an arbitrary location that no longer concurs with the location of any of the original centers. On the other hand, the modified version of ComBat called M- ComBat shifts the data to the mean and variance of the chosen reference center (r), by changing the standardizing mean and variance of the estimates $\hat{a}_{\iota}$ and $\hat{\sigma}_{\iota}$ to center-wise estimates $\widehat{a_{\iota=r,g}}$ and $\widehat{\sigma_{\iota=r,g}}$. ComBat has shown promising results towards harmonization and radiomics features reproducibility [99] [100] [101] [102] [103] and ProCancer-I will rely on existing publicly available software [104] to include both ComBat methods to the ProCAncer-I radiomics analysis platform.

7.2 The ProCAncer-I tool for Radiomics Harmonization

The ProCAncer-I feature-based harmonization tool is based on the Combine Batched method (ComBat), introduced by Jean-Philippe Fortin [105] and alternatives (i.e., M-ComBat). It consists of a training and a testing phase. Initially, radiomics features from the train set are harmonized and ComBat harmonization estimates are exported. Subsequently, the pre-trained estimates are used for harmonizing the radiomics features from the test set. The input arguments required for harmonization are the multicenter radiomics with a sufficient size and covariates that include center-related and non-center related parameters. The center-related covariates refer to the manufacturers (e.g. Philips, Siemens) or the manufacturer model (e.g., Skyra, TrioTim) that is used from each center, whereas the non-center covariates are optional parameters that refer to additional biological covariates (e.g. tumor grade, gender, etc.). In case of the M-Combat method, a center-related covariate must be also defined (e.g., the center which provides the most data in absolute numbers).

7.3 Experimental evaluation

Evaluation of the ComBine Batched harmonization performance was performed using the PI-CAI public dataset (Prostate Imaging: Cancer AI). Specifically, the harmonization method was applied on the public training dataset, comprising 1500 MRI prostate cases from two different scanners (Siemens and Philips) and seven different scanner models (Table 9). To this end, ComBat was evaluated on two different scenarios: a) choosing the scanner type as a 'center- effect', and b)

[99] Li, Yingping, et al. "Impact of Preprocessing and Harmonization Methods on the Removal of Scanner Effects in Brain MRI Radiomic Features." Cancers 13.12 (2021): 3000.
[100] Da-Ano, R., et al. "Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies." Scientific Reports 10.1 (2020): 1-12.
[101] Ibrahim A, Refaee T, Leijenaar RTH, Primakov S, Hustinx R,Mottaghy FM, Woodruff HC, Maidment ADA, Lambin P. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. PLoS One. 2021 May 7;16(5):e0251147. doi: 10.1371/journal.pone.0251147. PMID: 33961646; PMCID: PMC8104396.
[102] Fortin, J.-P.; Cullen, N.; Sheline, Y.I.; Taylor, W.D.; Aselcioglu, I.; Cook, P.A.; Adams, P.; Cooper, C.; Fava, M.; McGrath, P.J.; et al. Harmonization of Cortical Thickness Measurements across Scanners and Sites. *NeuroImage* 2018, *167*, 104–120, doi:10.1016/j.neuroimage.2017.11.024.
[103] Orlhac, F.; Lecler, A.; Savatovski, J.; Goya-Outi, J.; Nioche, C.; Charbonneau, F.; Ayache, N.; Frouin, F.; Duron, L.; Buvat, I. How Can We Combat Multicenter Variability in MR Radiomics? Validation of a Correction Procedure. *Eur. Radiol.* 2021, *31*, 2272–2280, doi:10.1007/s00330-020-07284-9.
[104] https://github.com/Jfortin1/neuroCombat, accessed on 30 November 2020, version 0.2.7
[105] https://github.com/Jfortin1/neuroCombat, accessed on 30 November 2020, version 0.2.7

choosing the scanner model. PyRadiomics was utilized to extract whole prostate gland radiomics features from T2-weighted images.

Table 9. The table shows the scanner type and scanner model variability in the PI-CAI dataset

Manufacturer	Manufacturer Model Name	Number of patients
Siemens	Skyra	1032
Siemens	Prisma	89
Siemens	Avanto	13
Siemens	TrioTim	68
Siemens	Aera	17
Philips	Ingenia	227
Philips	Achieva	52

An indicative representation of both ComBat and M-ComBat harmonization when single radiomics features were randomly selected (e.g., 1st order entropy) is given from Figure 43 and Figure 44. Boxplots and related histograms of this particular feature show entropy values extracted from the same ROIs but from different scanners and models respectively. It is obvious that after applying the harmonization methods, the histograms of the radiomic feature acquired by different scanner settings, have a much better overlap compared to the raw data, demonstrating that both scanner and model effects have been successfully removed. To test if the feature distributions are significantly different or not , the F-test ANOVA was used to compare the two distributions of the radiomics feature (in case of the different scanner types, e.g. Siemens and Philips) and between seven distributions of the same feature (in case of various scanner model types). The p-value of less than 0.05 indicates a statistically significant difference in radiomics features distributions without harmonization. On the contrary, distribution differences were found to be statistically insignificant in ComBat and M-ComBat harmonized features (p-value > 0.05).

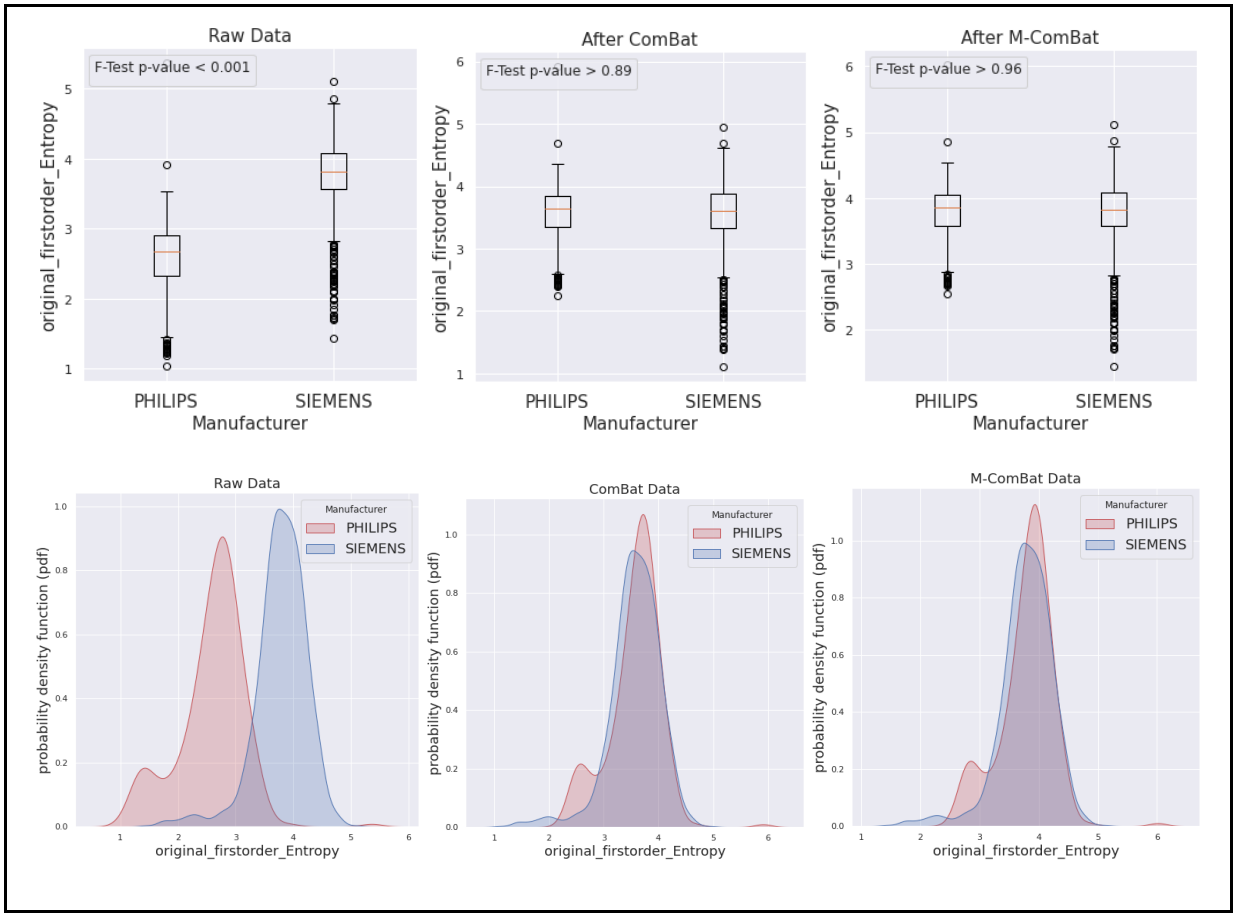


Figure 43. Box plots and probability density functions of "1st order entropy" radiomic feature before and after ComBat and M-ComBat, respectively. For the harmonization process, the scanner type was used as center-effect and the 'Siemens' manufacturer as reference batch effect for the M-ComBat. P-values are for the one-way ANOVA F-Test.

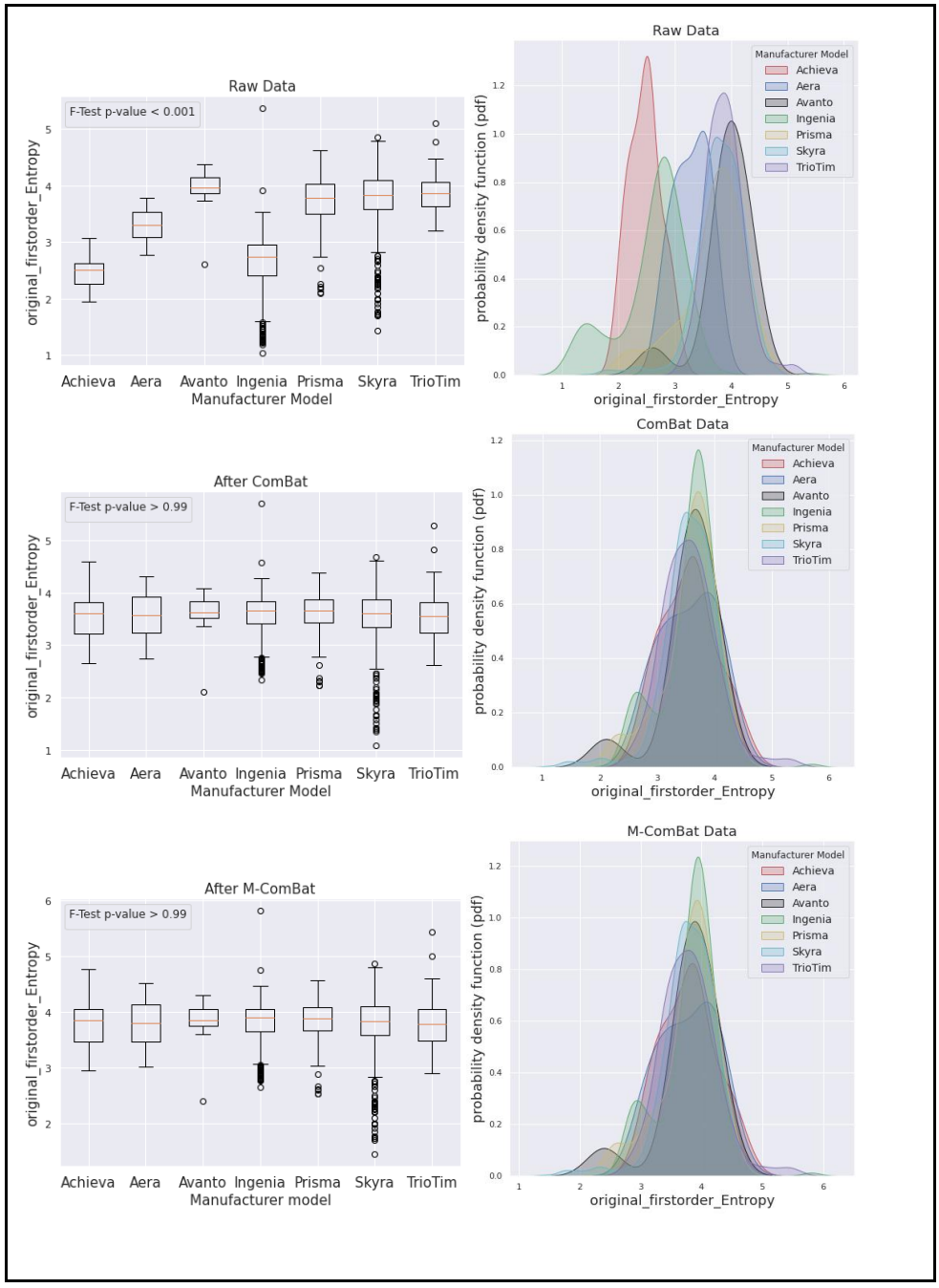


Figure 44. Box plots and probability density functions of "1st order entropy" radiomic feature before and after applying ComBat and M-ComBat method. For the harmonization process, the scanner model was used as center-effect and the 'Skyra' model as reference batch effect for the M-ComBat. P-values are for the one-way ANOVA F-Test.

To have an overall evaluation of the ability of ComBat harmonization methods, we calculated the "DiffFeatureRatio" evaluation metric. It has been introduced by Li Y et al.[106] and it is the ratio of the radiomic features with p-value < 0.05 to the overall radiomic features. For each ROI of the T2-weighted images, 100 radiomics features were extracted. The values of "DiffFeatureRatio" lie in [0,1], where 0 and 1 correspond to the perfect and worst harmonization efficiency, respectively.

$$DiffFeatureRation = \frac{Number\ of\ features\ with\ p-value\ <\ 0.05}{Number\ of\ all\ features}$$

Table 10. DiffFeatureRatio computed on the non-harmonized (raw) and harmonized (ComBat and M-ComBat) data

	DiffFeatureRatio		
Center- Effect	**Raw**	**ComBat**	**M-ComBat**
Manufacturer	0.93	0.00	0.00
Manufacturer Model	0.91	0.00	0.00

The values of the "DiffFeatureRatio", representing the ratio of features who have significantly different feature distributions among different center-effects, are shown in Table 10. Without harmonization, the raw features had a significantly different feature distribution. In contrast, the "DiffFeatureRatio" values of the harmonized data always tend to zero, which means that most radiomics features couldn't be detected to have significantly different feature distributions between the different scanner settings. In addition, the two different harmonization methods were evaluated in terms of the resulting coefficient of variation (COV) in the harmonized features, compared to the raw features. In Table 11, the raw data showed increased variability compared to the harmonized ones, while the ComBat-harmonized features showed slightly reduced variability compared to the M-ComBat-harmonized data.

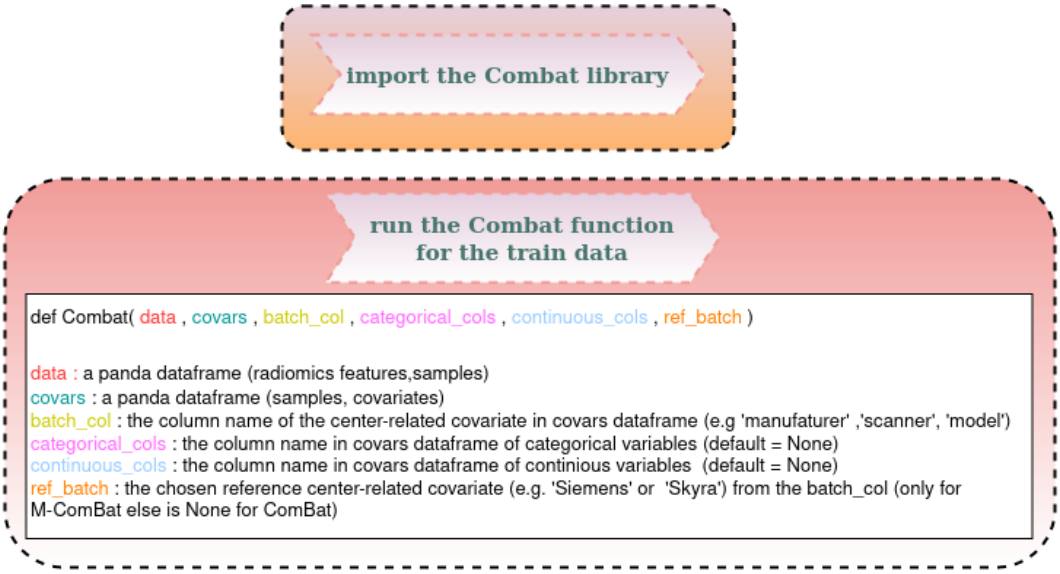
Table 11. COV computed on the raw, ComBat and M-ComBat data

	Coefficient of Variation (COV)		
Center- Effect	**Raw**	**ComBat**	**M-ComBat**
Manufacturer	0.738	0.649	0.689
Manufacturer Model	0.738	0.626	0.679

[106] Li Y, Ammari S, Balleyguier C, Lassau N, Chouzenoux E. Impact of Preprocessing and Harmonization Methods on the Removal of Scanner Effects in Brain MRI Radiomic Features. Cancers (Basel). 2021 Jun 15;13(12):3000. doi: 10.3390/cancers13123000. PMID: 34203896; PMCID: PMC8232807.

7.4 Pipeline execution and integration to the ProstateNet platform

According to Jean-Philippe Fortin's Combine Batches implementation [107], the required inputs are a) the radiomics features, b) the covariates, c) the center/batch effect, d) the categorical and continuous parameters, and e) the reference center/batch effect in case of the M-ComBat method. Both radiomics features and covariates must be in dataframe format. The first must have rows corresponding to radiomics features and columns to the samples while the second must have rows corresponding to the samples and columns to the covariates. The center/batch parameter, the categorical and the continuous parameters must be referred to the corresponding columns of the covariates' dataframe, while the reference center/batch parameter must be an element of the center/batch column. There are two basic functionalities of the harmonization module: i) apply the ComBat method in the training dataset and extract the harmonized radiomics features with the corresponding configuration of ComBat estimates (e.g., the Empirical Bayes estimates of $\hat{\delta}_{ig}^2$and $\hat{\gamma}_{ig}$) and ii) apply the ComBat method in the testing dataset using the previous configuration of ComBat pre-trained estimates. The basic instructions for executing this module are described in Figure 45. In the run command, the user should first import the public ComBat library and then specify the input arguments that correspond to a functionality in order to execute the harmonization process. The output of the functionalities consist of i) a numpy array of harmonized radiomics features with shape as the original data ii) a dictionary of the ComBat estimates used for the harmonization iii) a dictionary of the inputs needed for ComBat harmonization.



[107] https://github.com/Jfortin1/neuroCombat, accessed on 30 November 2020, version 0.2.7

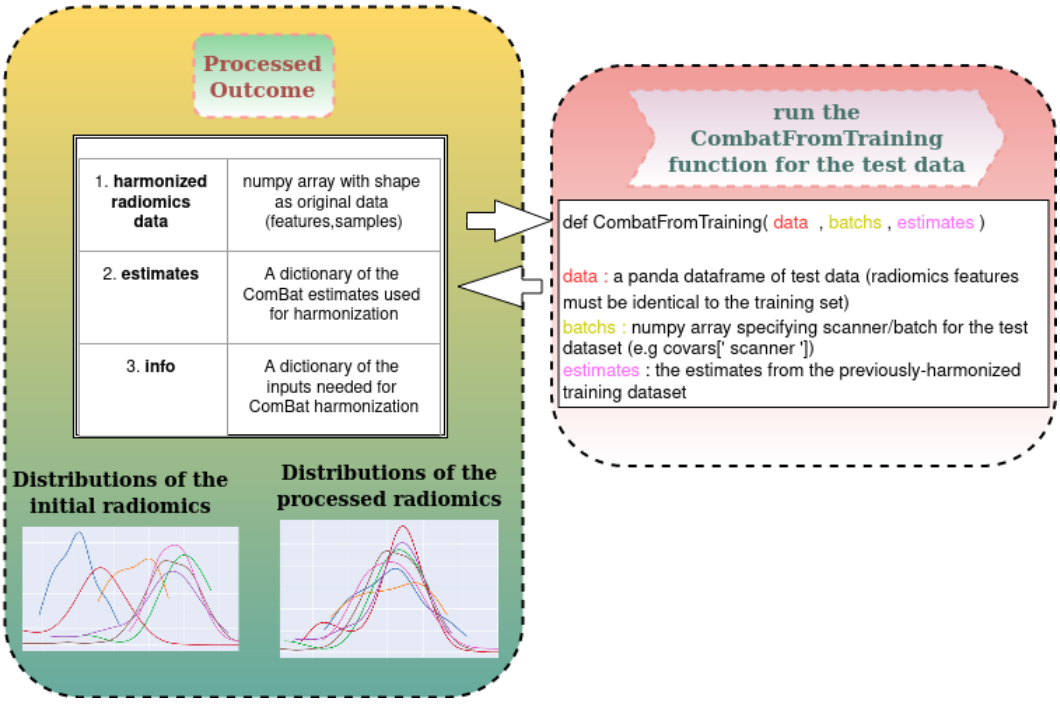


Figure 45. Execution for each step of ComBat Harmonization. Import the ComBat library, run the ComBat function, export the harmonized radiomics features with the corresponding ComBat estimates and then use the pre-training estimates for the harmonization of a new testing dataset.

8. Radiomic Feature Stability Assessment after Image Preprocessing

In regard to AI model development (specifically the radiomic-based models), image preprocessing may severely affect the reliability and reproducibility of radiomic features. Experimental results from the literature have demonstrated that MRI preprocessing can have a considerable impact on the whole radiomics analysis, and identifying a generalizable preprocessing step is crucially needed for providing clinical radiomics biomarkers[108]. In order to gain insights into the appropriateness and efficacy of the ProCAncer-I's preprocessing pipelines in developing radiomics-based AI models, a separate analysis of radiomic feature stability was performed as part of the activities of the Task 5.2.

For the analysis, the publicly available dataset from the PI-CAI challenge[109] was used, containing MRIs of 1500 prostate cancer patients. Radiomic features were extracted from three-dimensional prostate volumes on T2w MRIs using the Pyradiomics open-source Python package[110]. First, radiomic features were extracted from the original MRI images without any of the ProCancer-I's preprocessing applied to them. Subsequently, the native T2w MRIs were altered using one image preprocessing operations at a time -(i) N4 bias field correction, (ii) RACLAHE image enhancement and, (iii) noise reduction- and the same features were extracted. The effect of ProCAncer-I's preprocessing pipelines on the reproducibility of the extracted radiomic features were quantified using the intra-class correlation coefficient (ICC) and the concordance correlation coefficient (CCC). All the settings for feature extraction were defined with functions embedded within the PyRadiomics library, including pixel dimension resampling, which was set to [2,2,2] and intensity value discretization. The bin width was set to 20 which was computed based on the range of intensity values in the population.

The extracted radiomic features consist of shape features, first-order statistics features, Gray Level Co-occurrence Matrix (GLCM) features, Gray Level Run Length Matrix (GLRLM) features, Gray Level Size Zone Matrix (GLSZM) features, Neighbouring Gray Tone Difference Matrix (NGTDM) features and Gray Level Dependence Matrix (GLDM) features[111]. Apart from shape features, other texture features were also computed after applying wavelet and Laplacian transform of Gaussian (LoG) transformations to the images leading to a total of 1140 radiomic features per patient/image.

The distribution of radiomic features in excellent, good, modest and poor stability category, for each preprocessing pipeline proposed, is shown in the Table 12. Overall, after applying the N4 filter (section 2) the vast majority of radiomic features exhibit excellent stability (ICC ≥ 90). This phenomenon can be expected when MRI images do not suffer from obvious bias field effects, thus the bias-corrected images are similar to the original images. After denoising, the majority of features show relatively good stability (ICC ≥ 75). However, after applying image enhancement

[108] Moradmand, H., Aghamiri, S.M.R. and Ghaderi, R. (2020), Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. J Appl Clin Med Phys, 21: 179-190. https://doi.org/10.1002/acm2.12795
[109] https://pi-cai.grand-challenge.org/
[110] van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., Aerts, H. J. W. L. (2017). Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research, 77(21), e104–e107.
[111] https://pyradiomics.readthedocs.io/en/latest/

the vast majority of radiomic features demonstrate poor stability (ICC<50). The same pattern is observed in the distribution of the CCC, as it is shown in Figure 46.

Table 12. The absolute and relative number of radiomic features belonging to each ICC category after applying each one of the thee preprocessing pipelines. The ICC was computed between the features from the preprocessed images and the features from the original images (without preprocessing).

Preprocessing Pipeline	Radiomic Feature Family	ICC ≥ 90 Excellent Stability	75≤ICC<90 Good Stability	50≤ICC<75 Modest Stability	ICC<50 Poor Stability
N4 FILTER	Firstorder (216)	94% (204)	5% (10)	1% (2)	0
	Glcm (312)	86% (268)	12% (37)	2% (7)	0
	Gldm (182)	89% (162)	10% (18)	1% (2)	0
	Glrlm (208)	96% (200)	3% (7)	1% (1)	0
	Glszm (208)	93% (194)	5% (10)	2% (4)	0
	Shape (14)	100% (14)	0	0	0
RACLAHE	Firstorder (196)	0	6% (11)	31% (62)	63% (123)
	Glcm (240)	0	0	2% (5)	98% (235)
	Gldm (151)	0	0	7% (10)	93% (141)
	Glrlm (172)	0	0	5% (9)	95% (163)
	Glszm (167)	0	0	0	100% (167)
	Shape (14)	14% (2)	7% (1)	0	79% (11)
DENOISE	Firstorder (234)	36% (84)	35% (81)	10% (23)	20% (46)
	Glcm (312)	36% (128)	26% (81)	19% (59)	19% (60)
	Gldm (182)	39% (71)	35% (64)	20% (37)	6% (10)
	Glrlm (208)	46% (96)	33% (69)	18% (38)	3% (5)
	Glszm (208)	33% (68)	43% (89)	17% (36)	7% (15)

	Shape (14)	100% (14)	0	0	0

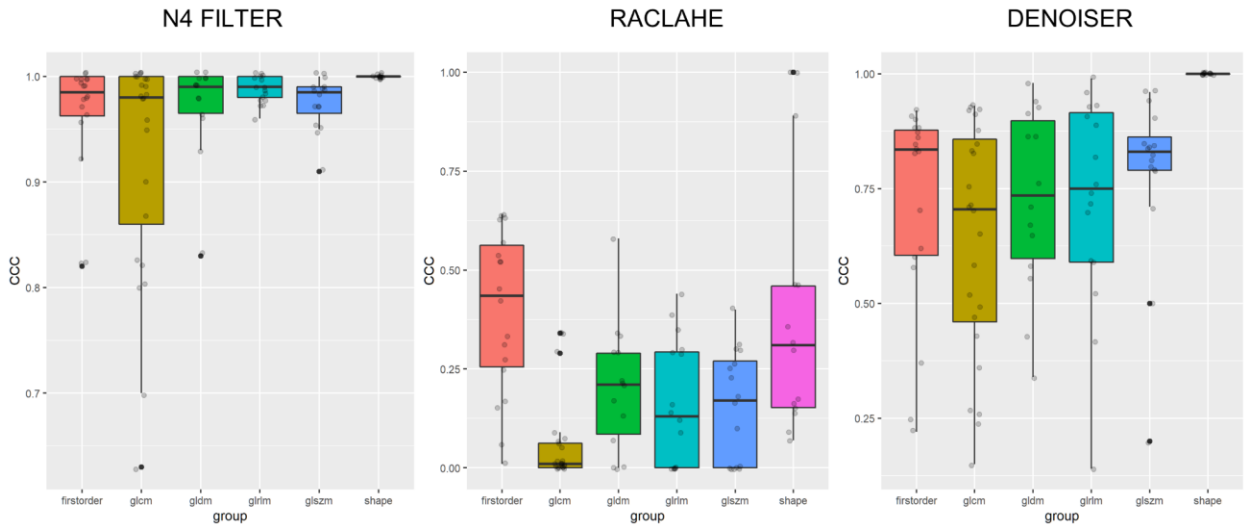


Figure 46. Boxplots of concordance correlation coefficient (CCC) for the different families of radiomic features, calculated between the original and the preprocessed images.

Overall, with respect to radiomic feature stability, the results from the PI-CAI dataset suggest that the proposed N4 filter for bias field correction can be safely applied to the MR images prior to radiomic feature extraction. This is also in line with findings from the literature demonstrating that the N4 bias field correction has no obvious impact to feature stability[112]. Nevertheless, noise reduction and, particularly, image enhancement filters should be treated with caution when developing radiomic-based models since they seem to have a greater impact on feature stability and repeatability.

[112] Li Y, Ammari S, Balleyguier C, Lassau N, Chouzenoux E. Impact of Preprocessing and Harmonization Methods on the Removal of Scanner Effects in Brain MRI Radiomic Features. Cancers. 2021; 13(12):3000.

Conclusions

In this deliverable the containerized pipelines for preprocessing the prostate MRI images that will be made available in the ProstateNet platform, are described. The proposed pipelines are carefully designed to address the most common sources of image quality degradation and eliminate their adverse effect on imaging quality, which is usually detrimental for model development. Importantly, in this deliverable we draw special attention to the emerging and challenging issue of cross-center MRI harmonization, providing pipelines to harmonize both the images and the extracted radiomic biomarkers.

The experimental evaluation of the preprocessing pipelines demonstrated that their performance is comparable -or even surpass- the performance of state-of-the-art pipelines currently available in the biomedical literature. Through an extensive external validation, we confirmed that the developed preprocessing pipelines can effectively process the prostate MRI data.

The algorithms for the execution of pipelines are available on the ProstateNet platform via container technology and user-friendly shell scripts to enable and facilitate their application for scientific purposes. In addition, the containerized form of the pipeline distribution supports its cross-platform usage.

During the lifecycle of the project, as the ProstateNet platform is populated will MRIs from different centers, all the pipelines described in this deliverable will be further refined- and potentially extended - and a de novo assessment will be performed to ensure there are no drifts in performance.

References

In the current document the references are reported as footnotes.

Annex

The current deliverable do not include any content in the ANNEX section.