Deep Learning Features Can Improve Radiomics-Based Prostate Cancer Aggressiveness Prediction

Nuno M. Rodrigues[1,2*], José Guilherme de Almeida[2], Ana Rodrigues[2, 3], Leonardo Vanneschi[4], Celso Matos[2], Maria V. Lisitskaya[5], Aycan Uysal[6], Sara Silva[1], Nickolas Papanikolaou[2]

[1] LASIGE, Department of Informatics, Faculty of Sciences, University of Lisbon, Lisbon, Portugal

[2] Champalimaud Foundation, Centre for the Unknown, Lisbon, Portugal

[3] Faculty of Medicine, University of Porto, Porto, Portugal

[4] NOVA Information Management School (NOVA IMS), Campus de Campolide, Universidade Nova de Lisboa, 1070-312 Lisboa, Portugal

[5] Cand. of Sci. (Med.), Radiologist at Radiology Department with CT and MRI, Medical Research and Educational Center, Lomonosov Moscow State University, Moscow, Russia

[6] Gulhane Medical School, University of Health Sciences, Ankara, Turkey

* Corresponding author: nmrodrigues@fc.ul.pt

This is the author accepted version of the paper published in the ASCO journal JCO Clinical Cancer Informatics:

Rodrigues, N. M., Almeida, J. G. D., Rodrigues, A., Vanneschi, L., Matos, C., Lisitskaya, M., Uysal, A., Silva, S., & Papanikolaou, N. (2024). Deep Learning Features Can Improve Radiomics-Based Prostate Cancer Aggressiveness Prediction. JCO Clinical Cancer Informatics, 8, Article e2300180. https://doi.org/10.1200/CCI.23.00180

Deep Learning Features Can Improve Radiomics-Based Prostate Cancer Aggressiveness Prediction

Nuno M. Rodrigues[a,b], José Guilherme de Almeida[b], Ana Rodrigues[b,c], Leonardo Vanneschi[d], Celso Matos[b], Maria V. Lisitskaya[e], Aycan Uysal[f], Sara Silva[a**], Nickolas Papanikolaou[b**]

[a]LASIGE, Department of Informatics, Faculty of Sciences, University of Lisbon, Lisbon, Portugal
[b]Champalimaud Foundation, Centre for the Unknown, Lisbon, Portugal
[c]Faculty of Medicine, University of Porto, Porto, Portugal
[d]NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa
[e]Cand. of Sci. (Med.), Radiologist at Radiology Department with CT and MRI, Medical Research and Educational Center, Lomonosov Moscow State University, Moscow, Russia
[f]Gulhane Medical School, University of Health Sciences, Ankara, Turkey

[**]Contributed equally

Corresponding Author:

Nuno M. Rodrigues

Champalimaud Foundation

Avenida Brasilia, 1400-038

Lisbon, Portugal

+351 21 048 0200

nmrodrigues@fc.ul.pt

Acknowledgments

This work was also partially supported by FCT, Portugal, through funding of the LASIGE Research Unit (UIDB/00408/2020 (https://doi.org/10.54499/UIDB/00408/2020), UIDP/00408/2020 (https://doi.org/10.54499/UIDP/00408/2020)), and Nuno Rodrigues PhD Grant 10.54499/2021.05322.BD (https://doi.org/10.54499/2021.05322.BD). Ana Rodrigues and José Guilherme de Almeida were supported by the European Union H2020: ProCAncer-I project (EU grant 952159). This work was supported by national funds through FCT (Fundação para a Ciéncia e a Tecnologia), under the project - UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS (https://doi.org/10.54499/UIDB/04152/2020).

Context Summary

Key Objectives: Can deep features learned from tasks which are complementary to the radiomics pipeline improve the predictive performance of radiomics-based models for prostate cancer grading?

Knowledge Generated: Our findings show that deep features can be helpful, leading to significant improvements in classifying the aggressiveness of prostate cancer. Additionally, we determine the best feature extraction techniques from segmentation and reconstruction models, reducing the burden of developing future models combining radiomics and deep features.

Relevance (Warner): Radiomic features may provide added benefit to conventional classification techniques, as shown by the authors. Use of such features may lead to risk-based clinical decisions and the consequences of such decisions are best evaluated through prospectively conducted studies.

Abstract

Purpose

Emerging evidence suggests that the use of Artificial Intelligence can assist in the timely detection and optimization of therapeutic approach in patients with prostate cancer. The conventional perspective on Radiomics encompassing segmentation and the extraction of radiomic features, considers it as an independent and sequential process. However, it is not necessary to adhere to this viewpoint. In this paper, we show that besides generating masks from which radiomic features can be extracted, prostate segmentation and reconstruction models provide valuable information in their feature space, which can improve the quality of radiomic signatures models for disease aggressiveness classification.

Material and Methods

We perform 2244 experiments with deep learning features extracted from thirteen different models trained using different anatomical zones, and characterize how modelling decisions such as deep feature aggregation and dimensionality reduction impact performance.

Results

While models using deep features from full gland and radiomic features consistently lead to improved disease aggressiveness prediction performance, others are detrimental. Our results suggest that the use of deep features can be beneficial, but an appropriate and comprehensive assessment is necessary to ensure that their inclusion does not harm predictive performance.

Conclusion

The study findings reveal that incorporating deep features derived from autoencoder models trained to reconstruct the full prostate gland (both zonal models show worse performance than radiomics only models), combined with radiomic features, often leads to a statistically significant increase in model performance for disease aggressiveness classification. Additionally, the results also demonstrate that the choice of feature selection is key to achieving good performance, with PCA and PCA+Relief being the best approaches, and that there is no clear difference between the three proposed latent representation extraction techniques.

1 Introduction

Prostate cancer (PCa) is one of the most frequent cancer types in the world, and the second most common cancer in males, affecting approximately one in every eight men. If detected early, it has a high survival rate, however, diagnosis is substandard, requiring a biopsy to be performed in order to classify its level of aggressiveness.

One research field that has been growing in popularity over the past years is radiomics, which transforms medical images into high-dimension mineable data, allowing for the extraction of large amounts of quantitative imaging features, as opposed to the qualitative features extracted by radiologists. These features can be used to train a wide variety of machine learning models for a vast selection of clinical scenarios. Several studies have been conducted using bi-parametric Magnetic Resonance Imaging (bpMRI) radiomics for PCa disease aggressiveness classification[1–9].

Besides radiomics, deep learning models that use the whole MRI scans have also been applied to both MRI and histopathology images for PCa aggressiveness classification, as shown in the literature[10–14]. These models deal with some of the weaknesses of radiomics, such as learning their own features from the data, instead of relying on handcrafted radiomic features that are dependent on the quality of the segmentations of our volume of interest. Relying on the segmentations introduces several possible concerns, such as inter- and intra-reader variability[15,16] when the masks are manually curated by expert radiologists, or the overall quality of the masks generated by deep learning models, which varies depending on the prostate zone, MRI sequence type and several other factors. Another downside of using deep learning models is that they require a considerably larger

amount of data, particularly if working on 3D MRI volumes, and are notably sensitive to the potential class imbalance.

In this work, we aim to establish a technical bridge between radiomics and deep learning. We perform a thorough study on the usage of deep learning features, otherwise known as latent representations, learned by deep learning segmentation models, together with radiomic features, for PCa aggressiveness classification. Based on two prior studies[2,17], we use radiomic features extracted from the prostate gland, as it was shown they are more stable and resilient to overfitting when compared to radiomic features extracted from the lesions. Consequentially, we couple them with deep learning features extracted from both segmentation models trained for the whole gland, and peripheral (PZ) and transition (TZ) zones, and auto-encoder models trained to reconstruct the original volume.

We evaluate several pipelines using four different machine learning classifiers that are commonly used with radiomics data, three different feature selection and dimensionality reduction techniques, as well as three distinct ways to extract the deep features from the deep learning models. The obtained results reveal that models that used a combination of deep features and radiomic features often outperform models that only use radiomic features. Additionally, the results also show that the deep features learned by the auto-encoder transformer models are the ones that produce the best results when coupled with the radiomic features.

2 Material and Methods

2.1 Data

The ProstateX dataset (SPIE-AAPM-NCI PROSTATEx challenge) is a collection of prostate MRI volumes that include T2W, DWI and ADC modalities. These volumes were obtained by the Prostate MR Reference Center — Radboud University Medical Centre (Radboudumc) in the Netherlands, using two Siemens 3T MR scanners (MAGNETOM Trio and Skyra). A full description of the image acquisition can be found on the projects page.. For pre-processing, the images were bias field corrected using N4ITK[18], standardized (such that $\mu = 0$ and $\sigma = 1$ for intensity values), resampled to a 0.5mm × 0.5mm × 3.0mm spacing, and resized to 256 × 256 × 32 voxels when processed by segmentation/reconstruction models.

The ground truth manual segmentations of the prostate gland were performed independently by two radiologists (M.L., 10 years of experience, and A.U., radiology resident) on T2W and DWI sequences separately (153 volumes total for each sequence), while the transition and peripheral zone ground truth masks were obtained from the public dataset repository. The ground truth for the presence of clinically significant prostate cancer (aggressive vs non-aggressive) was based on the ISUP score of the index lesion as this shows a strong association with clinical outcome and survival[19]. If the index lesion had an ISUP ≥ 2 - Gleason score 3+4 - then the sample would be labelled as aggressive. We note here that including ISUP = 2 in the aggressive category is conservative — indeed, while a large number of ISUP = 2 are non-aggressive, others with cribriform patterns have been shown to constitute more aggressive forms of PCa (Hollemans et al.

2019). Lastly, the data was partitioned patient-wise, following an 80/20% split, into a training (n=116) and a hold-out test set (n=29). During training, a stratified 5-fold cross-validation strategy was applied.

2.2 Radiomic features

Radiomic features were extracted from the whole gland segmentation using the package Pyradiomics (version 3.0). The extraction of the features was performed in a 2D fashion due to the *z* axis spacing differing from both the *x* and *y* axis spacing. Additionally, since in the T2W sequences, the *x* and *y* axis spacing differed within and between patients, they were resampled to the most common value of 0.5. The width of the bin was chosen to yield between 30 and 130 bins. The resulting bin width for T2W sequences was 20, for DWI it was 5, and for ADC it was 70. All image filters and feature classes were enabled, resulting in a total of 3111 features extracted (1037 from each MRI modality). In the feature extraction of the ADC map, the mask drawn on the DWI was used.

2.3 Segmentation and reconstruction

Based on prior work[20], we trained eight different Unet-based[21] segmentation models, for prostate gland, peripheral and transition zone segmentation, for both full volume data, and volumes cropped around the prostate gland by means of a YoLo-v5[22] object detection model. We decided to use different models incorporating distinct mechanisms (*e.g.,* attention blocks) to study whether this alters the usefulness of the produced representations: Unet[21]; Unet++[23]; Attention Unet (AUnet)[24] ; Dense Attention Unet

(DUnet); Dense-2 Unet (D2Unet)[25]; Dense-2 Attention Unet (D2AUnet); Recurrent Residual Unet (R2Unet)[26] ; Recurrent Residual Attention Unet (R2AUnet).

In addition to the supervised segmentation models, we also explored unsupervised approaches by training thirteen auto-encoder models. This approach was of particular interest as unlabeled data outnumbers labeled data. Thus, positive results would imply we could potentially leverage far more data in future studies. The following is a list of all the unsupervised models that have been used: All the previous models that were used for segmentation; Unetr[27] , both regular and large; Unetr + deep attention[28], both regular and large; Swin-Unet[29] .

2.4 Deep learning feature space

The deep learning feature space, also known as latent space, is an abstract multi-dimensional vector space that encodes compressed meaningful features of the data. In convolutional neural network-based models with encoder-decoder architectures (Fig. 1), this space usually corresponds to the bottleneck (the inner most part of the network), where the image input data is compressed to a feature vector representation. This feature vector, which contains only the most salient and informative aspects of the input data, is then used to reconstruct the original image through a symmetrical decoding process.

Using the models devised for the prostate gland and zone segmentation, in conjunction with the unsupervised models, we devised distinct approaches to extract the learned deep features (Fig. 1). The first approach entails averaging all the representations into a unified representation before flattening it. This method generates 200 features for the full sequence data and 129 features for the cropped sequence data. The second and third

approaches consist of applying a pooling operation, either max-pool or average-pool, to the representations, condensing them into a singular value, followed by concatenating all values. These operations remove all spatial dependencies, making the extracted feature's location invariant. Both generate 1024 features for both full and cropped sequences.

2.5 Dimensionality reduction and feature selection

Radiomics data is, by nature, of very high dimensionality. Since we are adding additional feature vectors and have a small amount of data, dimensionality reduction is crucial. In this work, we try three different combinations of feature selection/dimensionality reduction mechanisms: Relief; PCA; Relief→PCA. For Relief[30], we performed initial tests reducing the total number of features to 100, 50 or 10. In the end, results between 50 and 10 were identical and better than with 100, so we decided to use the 10 features. For PCA, we select the number of components that achieved a cumulative variance value of 95% on a per-dataset basis. As for the combination of PCA and Relief, the selected approach was to apply Relief to reduce the number of components to 10 if more than 10 were selected. Full description of the parameters used in Appendix A (Data Supplement). We used the PCA algorithm from Scikit-Learn[31] and the Relief algorithm from skrebate[32].

2.6 Training pipelines

We chose a variety of commonly used machine-learning algorithms in medical imaging[33] : Logistic Regression, Decision Trees, Random Forests[34] and AdaBoost[35] with shallow decision trees. First, we defined the baselines as the results produced by each machine

learning algorithm using only the radiomic features, for the samples of the prostate gland and prostate zones, with no feature selection nor dimensionality reduction, totaling 12 baselines. Then, we evaluated each combination of a zone (gland, peripheral, transition), full and cropped data, for all latent representation types of all the segmentation and reconstruction models, using PCA, Relief and PCA + Relief for a total of 2244 results. Figure 2 provides a schema of the training pipelines. For each pipeline, we performed both nested[36] and regular cross-validation, doing a grid search hyper-parameter optimization on both(Table SI1). To avoid overfitting, we used shallow trees for tree-based algorithms. To reduce the computational costs, we used a number of trees large enough for our dataset as this guarantees that the ensemble converges to its asymptotic generalization error[34,37,38]. Training pipeline is available at https://github.com/CCIG-Champalimaud/Deep-Features-PCa.

2.7 Statistical nalysis

Statistical analysis was conducted using R version 4.1.2. Two-way comparisons were performed using Wilcoxon rank-sum tests and statistical significance was defined using α = 0.05 We focus on the F2-score, a weighted version of the harmonic mean between recall and precision which places a heavier weight on recall ($F_2 = \frac{5}{\frac{4}{precision}+\frac{1}{recall}}$). Nonetheless, we also present the F1-score ($F_1 = \frac{2}{\frac{1}{precision}+\frac{1}{recall}}$), the precision, the recall, the accuracy and the AUROC for all models for the sake of completeness.

3 Results and Discussion

Here, a comprehensive analysis of the efficacy of utilizing deep features for disease aggressiveness classification is presented, followed by a detailed evaluation of the contributions of individual components, namely, the origin of the deep features, feature selection, and latent representation extraction techniques. Additionally, deep features extracted from models trained with cropped data did not produce any meaningful results, and thus are not presented in the analysis. We note here that a compromise between clarity and extension was attempted; as such, specific results may be more or less discussed, depending on whether they were relevant for the topic at hand.

Deep features can improve performance

We will start by comparing the overall performance of radiomics-only models to that of hybrid models composed of radiomic features and deep features (referred to as deep feature models).

Figure 3 shows the average cross-validation F2-score of all models using full gland data. Several different conclusions can be extracted from these results. First and most importantly, it is shown that regardless of the feature selection type, there exist multiple deep feature models that outperform their radiomics counterparts. In particular, weak shallow learners such as decision trees and AdaBoost were capable of leveraging this additional information in most scenarios, often greatly outperforming the results they obtained when using radiomics-only data. Logistic regression was able to leverage the small search space produced by applying Relief to the entire dataset, however, the obtained

results are still subpar when compared to the weak learners using the other combinations of feature selection. One unexpected finding was the overall bad performance of the random forest models. They were the worst-performing models in any given scenario, immediately overfitting regardless of the hyperparameter configurations, which led to poor generalization, as shown in the figure.

Figure 4 shows the results of the more conservative nested cross-validation results (we show the complete picture of F2-scores in Appendix B Figure SI1 (Data Supplement) and the results for other metrics in Figures SI2, SI3, SI4, SI5 and SI6). Here we can further strengthen our previous findings. We can observe that several deep feature models are still statistically significantly better than their radiomics counterparts. All these better models are weak learners, and despite not being statistically better than their radiomics-only counterparts, linear regression models using Relief still outperform them.

Comparing the distribution of cross-validation and hold-out performances of different models (Figure 5), we can observe that there is a high degree of correlation between them. This indicates that the produced models are robust and able to generalize on unseen data, and not just good in a controlled cross-validation environment.

Overall, the obtained results show that hybrid models can indeed improve the performance of the models for disease aggressiveness classification.

Deep Features from the entire gland are the most informative

Regarding the results of the peripheral zone, Figure 6 shows the same analysis that was done previously for the full gland. Only two models show statistically significantly better

performance when compared to their radiomics-only counterparts. Despite proving not to be statistically better, there are still several deep feature models that perform better than the radiomics-only models regardless of feature selection type. For the transition zone, the deep feature models did not yield results which were as satisfactory, producing no models which were better than those produced using only radiomics (Fig. 7). Figures SI7, SI8, SI9 show the global rank distribution of all solutions for all three zones, further highlighting these differences.

The results show that utilizing zonal information produced inferior results compared to using full gland data. Our findings align with our expectation: training models with full gland data leads to a more robust and informative latent space.

Role of feature selection

As demonstrated earlier, feature selection plays a key role regarding the overall quality of the solutions. Observing Figure 3, deep feature models are highly dependent on the feature selection type, being able to leverage both PCA and PCA+Relief. The more conservative nested cross-validation results (Figure 4) further confirm this: all the deep feature models that are significantly better than their radiomics-only counterparts use one of these feature selection techniques. This is expected given that the dataset is rather limited, and tremendously increasing its the dimensionality by including hundreds of new features would result in data sparsity and overfitting (the commonly called curse of dimensonality) We can also see that radiomics-only models are hindered by most feature selection types, being resilient only to Relief. This was an unexpected result, in particular regarding the underperformance of PCA, as it should be able to remove the redundant information and

keep only meaningful features. We suspect that this might have happened because there was not enough variation in the most meaningful radiomic features, resulting in their loss.

Architectures and extraction techniques

There is an interesting finding regarding the types of models from where the deep features are extracted, as well as the extraction techniques. The vast majority of deep feature models (Figure 4) that show significantly better results compared to their radiomics-only counterparts use deep features extracted from autoencoder models. This indicates that the features these models learn when performing the reconstruction of the original MRI image contain valuable information that can then be used for additional tasks, in this case specifically the task of disease aggressiveness classification. This finding is very relevant since training reconstruction models is far easier in the sense that it only requires the MRI sequences themselves, and does not require manually labeled regions or volumes, which is a very time-consuming task for radiologists.

Additionally, when evaluating the different feature extraction techniques, we can observe that out of the proposed three techniques, there is no clear best. All of these techniques are used by the significantly best-performing models to a different degree depending on the type of feature selection that is used. In Figure 4, we can observe that, for both PCA and PCA+Relief, max-poll and mean are the most used techniques, while the average pool is not very present despite appearing on the top-ranked model. For the peripheral zone (Figure 6), three out of the four statistically significant best models used max-pool, with the remaining one using average-pool. So, while max-pool seems to appear more often in the

top-performing solutions, there is no solid basis to say it is, in fact, better than the remaining alternatives.

On alternative risk-stratification models

Other valuable and informative risk assessment models are available and could have been used to make our analysis more informative. CAPRA, which is associated with survival[39], and NCCN, which also considers metastisation and provides concrete treatment recommendations[40]. However, both require PSA, age and T stage; given that PROSTATEx provides no information on these variables, we opted for the simpler, ISUP-based stratification of PCa aggressiveness.

4 Conclusion

We perform an extensive analysis regarding the potential of including deep learning features, also known as deep features, in radiomic signatures models for disease aggressiveness prediction. We evaluate multiple different deep features taken from both segmentation and reconstruction models for the full prostate gland, and peripheral and transition zones. Furthermore, we assess three different latent representation extraction techniques and three different feature selection types across four distinct machine learning classifiers. The study findings reveal that incorporating deep features derived from autoencoder models trained to reconstruct the full prostate gland, combined with radiomic features, often leads to a statistically significant increase in model performance for disease aggressiveness classification. Additionally, the results also demonstrate that the choice of

feature selection is key to achieving good performance and that there is no clear difference between the three proposed latent representation extraction techniques.

Limitations

The main caveat of this study comes from the limited variability of the data. As mentioned previously, the ProstateX data is a single institution dataset obtained through two different Siemens scanner models - MAGNETON Trio and Skyra - that use a magnet strength of 3T. Considering this, there is no indication if the results would be generalizable to other common scanner manufacturers, such as Philips, GE or Toshiba, or even other Siemens models, considering that the MAGNETON series has models which go from 0.35T to 7T. A good future direction would be to study this approach in large multi-institution and multi-vendor data, such as that of the ProCancer-I project[i].

Figure Legends

Figure 1: *Illustration of a simplified schema for the UNet-based models employed for feature extraction, along with the proposed feature extraction techniques.*

Figure 2: *Representation of the various proposed training pipelines and corresponding components.*

Figure 3: *Average cross-validation F2 scores for all models, stratified by feature selection type (colour). Points represent the median F2 and horizontal lines represent the interval spanning 95% of the data points centred around the median. Each shape represents whether the model was trained on radiomics only (circles) or if it has been trained using deep features (triangles).*

Figure 4: *Comparison of score distribution for the models trained using deep features taken from the entire prostate gland. For each feature selection type, both the radiomics baseline and all models whose F2 is greater than the F2 of the 10th ranked model, following a rank test, are shown. Models that show a statistically significant difference compared to the radiomics-only models (p-value*0.05*) are marked with *. For each model the latent representation extraction technique is indicated in parenthesis.*

Figure 5: *Correlation plots showing the distribution of cross-validation and hold-out performances of all models, stratified by prostate zone.*

Figure 6: *Comparison of score distribution for the models trained using deep features taken from the peripheral zone. For each feature selection type, both the radiomics baseline and all models whose F2 is greater than the F2 of the 10th ranked model, following a rank test, are shown. Models that show a statistically significant difference compared to the radiomics-only models (p-value*0.05*) are marked with *.For each model the latent representation extraction technique is indicated in parenthesis.*

Figure 7: *Comparison of score distribution for the models trained using deep features taken from the transition zone. For each feature selection type, both the radiomics baseline and all models whose F2 is greater than the F2 of the 10th ranked model, following a rank test, are shown. Models that show a statistically significant difference compared to the radiomics-only models () are marked with *.For each model the latent representation extraction technique is indicated in parenthesis.*

References

1. Liu B, Cheng J, Guo DJ, et al. Prediction of prostate cancer aggressiveness with a combination of radiomics and machine learning-based analysis of dynamic contrast-enhanced MRI. *Clin Radiol*. 2019;74(11):1-896.

2. Rodrigues A, Santinha J, Galvão B, Matos C, Couto FM, Papanikolaou N. Prediction of Prostate Cancer Disease Aggressiveness Using Bi-Parametric Mri Radiomics. *Cancers (Basel)*. 2021;13(23).

3. Castillo T JM, Arif M, Starmans MPA, et al. Classification of Clinically Significant Prostate Cancer on Multi-Parametric MRI: A Validation Study Comparing Deep Learning and Radiomics. *Cancers (Basel)*. 2021;14(1).

4. Bertelli E, Mercatelli L, Marzi C, et al. Machine and Deep Learning Prediction Of Prostate Cancer Aggressiveness Using Multiparametric MRI. *Front Oncol*. 2021;11:802964.

5. Han C, Ma S, Liu X, et al. Radiomics Models Based on Apparent Diffusion Coefficient Maps for the Prediction of High-Grade Prostate Cancer at Radical Prostatectomy: Comparison With Preoperative Biopsy. *J Magn Reson Imaging*. 2021;54(6):1892-1901.

6. Cutaia G, Tona G La, Comelli A, et al. Radiomics and Prostate MRI: Current Role and Future Applications. *J Imaging*. 2021;7.

7. Midiri F, Vernuccio F, Purpura P, Alongi P, Bartolotta T V. Multiparametric MRI and Radiomics in Prostate Cancer: A Review of the Current Literature. *Diagnostics (Basel)*. 2021;11(10).

8. Kwon D, Reis IM, Breto AL, et al. Classification of suspicious lesions on prostate multiparametric MRI using machine learning. *J Med Imaging (Bellingham)*. 2018;5(3):34502.

9. Bernatz S, Ackermann J, Mandel PC, et al. Comparison of machine learning algorithms to predict clinically significant prostate cancer of the peripheral zone with multiparametric MRI using clinical assessment categories and radiomic features. *Eur Radiol*. 2020;30:6757-6769.

10. Nagpal K, Foote D, Tan F, et al. Development and Validation of a Deep Learning Algorithm for Gleason Grading of Prostate Cancer From Biopsy Specimens. *JAMA Oncol*. 2020;6(9):1372-1380.

11. Khosravi P, Lysandrou M, Eljalby M, et al. Biopsy-free prediction of prostate cancer aggressiveness using deep learning and radiology imaging. *medRxiv*. Published online December 19, 2019:2019.12.16.19015057. doi:10.1101/2019.12.16.19015057

12. Lawson P, Schupbach J, Fasy BT, Sheppard JW. Persistent homology for the automatic classification of prostate cancer aggressiveness in histopathology images. In: Tomaszewski JE, Ward AD, eds. *Medical Imaging 2019: Digital Pathology*. Vol 10956. SPIE; 2019:109560G. doi:10.1117/12.2513137

13. Cuocolo R, Cipullo MB, Stanzione A, et al. Machine learning applications in prostate cancer magnetic resonance imaging. *Eur Radiol Exp*. 2019;3.

14. Duran A, Dussert G, Rouvière O, Jaouen T, Jodoin PM, Lartizien C. ProstAttention-Net: A deep attention model for prostate cancer segmentation by aggressiveness in MRI scans. *Med Image Anal*. 2022;77:102347. doi:https://doi.org/10.1016/j.media.2021.102347

15. Chen MY, Woodruff MA, Dasgupta P, Rukin NJ. Variability in accuracy of prostate cancer segmentation among radiologists, urologists, and scientists. *Cancer Med*. 2020;9(19):7172–7182. doi:10.1002/cam4.3386

16. Steenbergen P, Haustermans K, Lerut E, et al. Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation. *Radiotherapy and Oncology*. 2015;115(2):186–190. doi:10.1016/j.radonc.2015.04.012

17. Rodrigues A, Rodrigues N, Santinha J, et al. Value of handcrafted and deep radiomic features towards training robust machine learning classifiers for prediction of prostate cancer disease aggressiveness. *Sci Rep*. 2023;13(1):6206. doi:10.1038/s41598-023-33339-0

18. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: Improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29(6):1310-1320. doi:10.1109/TMI.2010.2046908

19. Egevad L, Delahunt B, Srigley JR, Samaratunga H. International Society of Urological Pathology (ISUP) grading of prostate cancer - An {ISUP} consensus on contemporary grading. *APMIS*. 2016;124(6):433-435.

20. Rodrigues NM, Silva S, Vanneschi L, Papanikolaou N. A Comparative Study of Automated Deep Learning Segmentation Models for Prostate MRI. *Cancers (Basel)*. 2023;15(5). doi:10.3390/cancers15051467

21. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv*. 2015;abs/1505.0.

22. Jocher G, Chaurasia A, Stoken A, et al. ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations. Published online August 2022. doi:10.5281/zenodo.7002879

23. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Published online 2018.

24. Oktay O, Schlemper J, Folgoc L Le, et al. Attention U-Net: Learning Where to Look for the Pancreas. *ArXiv*. 2018;abs/1804.0.

25. Aldoj N, Biavati F, Michallek F, Stober S, Dewey M. Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net. *Sci Rep*. 2020;10.

26. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. Published online 2018.

27. Hatamizadeh A, Yang D, Roth HR, Xu D. UNETR: Transformers for 3D Medical Image Segmentation. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Published online 2022:1748-1758.

28. Zhou D, Kang B, Jin X, et al. DeepViT: Towards Deeper Vision Transformer. *ArXiv*. 2021;abs/2103.1.

29. Cao H, Wang Y, Chen J, et al. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *ArXiv*. 2021;abs/2105.0.

30. Kira K, Rendell LA. A Practical Approach to Feature Selection. In: *ML Workshop*. ; 1992.

31. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*. 2011;12:2825-2830.

32. Urbanowicz RJ, Olson RS, Schmitt P, Meeker M, Moore JH. Benchmarking relief-based feature selection methods for bioinformatics data mining. *J Biomed Inform*. 2018;85:168-188. doi:https://doi.org/10.1016/j.jbi.2018.07.015

33. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. *RadioGraphics*. 2017;37(2):505-515. doi:10.1148/rg.2017160130

34. Breiman L. Random Forests. *Mach Learn*. 2004;45:5-32.

35. Freund Y, Schapire RE. A Short Introduction to Boosting. In: ; 1999.

36. Raschka S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *ArXiv*. 2018;abs/1811.1.

37. Bentéjac C, Csörgo A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev*. Published online 2020:1-31.

38. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*. 2019;9(3). doi:10.1002/widm.1301

39. Cooperberg MR, Pasta DJ, Elkin EP, et al. The University of California, San Francisco Cancer of the Prostate Risk Assessment score: a straightforward and reliable

preoperative predictor of disease recurrence after radical prostatectomy. *J Urol*. 2005;173(6):1938-1942.

40. Schaeffer EM, Srinivas S, Adra N, et al. Prostate Cancer, Version 4.2023, {NCCN} Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw*. 2023;21(10):1067-1096.

[i] https://www.procancer-i.eu/