On the Effectiveness of 3D Vision Transformers for the Prediction of Prostate Cancer Aggressiveness

Eva Pachetti[1,2][0000−0002−1321−9285], Sara Colantonio[1][0000−0003−2022−0804], and Maria Antonietta Pascali[1][0000−0001−7742−8126]

[1] "Alessandro Faedo" Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR), Pisa, Italy
{eva.pachetti, sara.colantonio, maria.antonietta.pascali}@isti.cnr.it
[2] Department of Information Engineering (DII), University of Pisa, Pisa, Italy

Abstract. Prostate cancer is the most frequent male neoplasm in European men. To date, the gold standard for determining the aggressiveness of this tumor is the biopsy, an invasive and uncomfortable procedure. Before the biopsy, physicians recommend an investigation by multiparametric Magnetic Resonance Imaging, which may serve the radiologist to gather an initial assessment of the tumor. The study presented in this work aims to investigate the role of Vision Transformers in predicting prostate cancer aggressiveness based only on imaging data. We designed a 3D Vision Transformer able to process volumetric scans, and we optimized it on the ProstateX-2 challenge dataset by training it from scratch. As a term of comparison, we also designed a 3D Convolutional Neural Network and optimized it in a similar fashion. The results obtained by our preliminary investigations show that Vision Transformers, even without extensive optimization and customization, can ensure an improved performance with respect to Convolutional Neural Networks and might be comparable with other more fine-tuned solutions.

Keywords: Vision Transformers · Prostate Cancer · ProstateX-2

1 Introduction

According to the World Health Organization, prostate cancer (PCa) is the most common tumor among European men [25]. For PCa patients, a biopsy followed by a microscopic examination of the collected specimen is, at the moment, the gold standard for diagnosis. Usually, before resorting to biopsy, the patient undergoes a multiparametric magnetic resonance imaging (mpMRI) examination. mpMRI investigations typically involve the acquisition of axial T2-weighted (T2w) images, used to investigate the anatomy, and diffusion-weighted images (DWI), from which the Apparent Diffusion Coefficient (ADC) maps are derived. By comparing T2w images and ADC maps, radiologists make an early qualitative diagnosis according to the Prostate Imaging Reporting and Data System

(PI-RADS) [21] guidelines. The PI-RADS score assigns a numerical value between 1 and 5 to the suspected lesion, which is an index of probability that the lesion constitutes an aggressive prostate neoplasm. The higher the PI-RADS score, the greater the likelihood that the suspected nodule is malignant. Typically if PI-RADS $\geqslant 3$, the patient undergoes a biopsy. At this point, the tumor's aggressiveness is assessed by examining the biopsy specimen, and a grade known as the Gleason Score (GS) is associated with the lesion. If GS $\geqslant$ 3+4, the tumor is considered clinically significant [1]. In particular, for patients with lesions having GS $>$ 3+4, treatment is foreseen; in all other cases, the patient usually undergoes active surveillance [16].

However, this early diagnosis is affected by inter-operator variability since most depend on the radiologist's experience and the acquisition protocol used. For this reason, the patient may be over-diagnosed if the biopsy reveals a tumor that is not clinically significant [23]. Because of all these reasons, there is now an increasing need for an automated tool that can diagnose PCa in a non-invasive, robust, and reliable manner. Several studies to date are focusing on building machine learning models that exploit the potential of deep learning for the automatic classification of PCa lesions from mpMRI images. Most of the works attempt to classify clinically significant from non-significant PCa (i.e., GS $\leqslant$ 3+3 vs. GS $\geqslant$ 3+4) [10, 13, 14, 20]. Only a few studies have addressed the issue of PCa aggressiveness, i.e., Low-Grade (LG) (GS $\leqslant$ 3+4) vs. High-Grade (HG) (GS $\geqslant$ 4+3) lesions. In [27], the authors exploit a 2D Convolutional Neural Network (CNN) trained on sagittal T2w images, axial T2w images, and ADC maps according to a transfer-learning approach, getting an AUROC of 0.869. In [2], CNNs with Attention Gates trained on T2w images yield 0.875 AUROC.

Assessment of PCa aggressiveness is a challenging task for several reasons. First of all, the lesion occupies very few pixels within the image. In addition, it may occur in different areas of the prostate; therefore, the network must be able to identify it among other tissues before classifying it. For this reason, many works are now focusing on building an end-to-end model, which first detects the lesion and then classifies it [15, 24, 26].

Recently, Vision transformers (ViTs) have gained popularity in Computer Vision, exceeding the performance of CNNs in almost all tasks: classification [7], object detection [3] and segmentation [19]. They have seen an increase in their application also in medical imaging [12]. Classic ViTs require large amounts of data to be trained. Because of this, usually transfer learning approach is exploited. In this work, we wanted to verify ViTs' effectiveness in addressing a challenging task as the prediction of PCa aggressiveness without any pre-training steps but by training them from scratch on PCa 3D volumes.

In the following sections, we describe our experiments with 3D ViTs and basic 3D CNNs applied to a freely available dataset (i.e., ProstateX-2 [9]). Firstly, we introduce the dataset used and how this was prepared for training the deep learning models. Afterward, we give a description of the 3D ViT architecture used and of the training pipeline. We do the same for the 3D CNN models that we exploited to compare and evaluate the performance ensured by 3D ViTs.

Therein, we report the results and compare our work to one belonging to the state-of-the-art addressing the same task. Finally, according to the results, we establish the potential effectiveness of 3D ViTs in determining PCa severity.

2 3D Vision Transformer and 3D CNN Development for Prostate Cancer Classification

The work aims to develop a 3D ViT model for assessing PCa aggressiveness based on axial volumetric T2w imaging data. Starting from the ViT model proposed in [7], we modified the architecture by reducing the number of parameters to train the model from scratch on the ProstateX-2 challenge dataset [9]. We also designed a 3D CNN and trained it from scratch on the same dataset as a reference model against which we compared our 3D ViT.

2.1 Dataset composition

The dataset for the ProstateX-2 challenge [9] was acquired at the Radboud University Medical Centre (Radboudumc) in the Prostate MR Reference Center. The dataset contains 112 lesions from 99 patients. GS is provided for each lesion to be used as ground truth. Each study was performed through mpMRI, of which we exploited only axial T2w acquisitions since according to [2], they provide better results in the application of deep learning models for the assessment of PCa severity. In terms of aggressiveness, the dataset is composed as follows: 77 LG (69%) and 35 HG (31%). As for the location of the lesion, the dataset is organized as follows: 50 peripheral (PZ) (44%), 47 anterior fibromuscular stroma (AS) (43%), and 15 transition (TZ) (13%).

2.2 Data preparation

To provide the model with only the most meaningful information, we selected only a subset of slices for each MRI scan, thus reducing the size of the 3D volume processed by the deep learning models. Based on the supplementary information provided with the dataset, we first selected the slice that contains the lesion. Hence, starting from that slice, we selected two slices above and below for a total of five slices per lesion. This approach ensured us to consider the slices that contain the lesion or that are strictly around it. Next, we harmonised the pixels dynamic from 0-2^{16} to 0-2^{8}, and we converted each image type from uint16 to uint8. This operation did not affect the image quality since the uint16 range is barely exploited. Indeed, the maximum value assumed by the pixels in all acquisitions was 800. This procedure ensures that each image had the same range of pixel values.

Since not all the patients had equal image sizes, to make the procedure reproducible to further processing, we rescaled all the images to the most common and largest ones in the dataset (i.e., 384x384). This approach limited the number of patients that required resampling and avoided losing information due to down-sampling.

Assuming that the prostate is placed within the center of the image, we center-cropped each slice to let the model focus only on the prostate gland. The final size of each image was 128x128. Through a visual inspection, we verified that this size was appropriate to include the prostate glands of all sizes in the crop's field of view and yet, at the same time, remove most of the tissue that does not belong to the gland. Eventually, for each lesion, we obtained a volume of size 128x128x5.

Since the dataset was unbalanced, we applied, to the training dataset only, three data augmentation techniques: vertical flip, horizontal flip, and rotation. Since the training set was composed of 54 LG and 27 HG volumes, we chose 9 HG images randomly with a fixed seed, and, for each one, we added three augmented versions to the set. In the end, the training set was composed of 54 LG and 54 HG images.

Eventually, we applied a mean normalization by calculating the mean value of the pixels across all the volumes within the training set only and subtracting it from all the slices in the training, validation, and test sets.

2.3 3D ViT Architectures

The ViT model used in this work stemmed from the one introduced in [7]. Since this model was designed to be trained on 2D images, we modified its structure so that it could work on 3D volumes, by processing 3D patches. All the three architectures described in the original work [7] were designed to be pre-trained on large datasets and then fine-tuned on smaller datasets. As we were working on 3D data, we avoided transfer learning and train the 3D ViT from scratch. Considering the limited size of the ProstateX-2 dataset, we then rescaled the original architecture to significantly reduce the number of parameters to be set.

We determined the most suitable architecture with a grid search on 18 different configurations (see Table 1), designed by varying the following parameters: Multi Layer Perceptron (MLP) size (d), hidden size (D), number of layers (L), and number of attention heads (k). In all configurations, we used a patch size (p) of 16 in one dimension (i.e., the 3D size of the patch was 16x16x5). This value seemed reasonable to allow the ViT processing enough information for each patch. In addition, some preliminary tests using $p = 8$ showed significantly worse results. We chose L and k values with the purpose of reducing the number of parameters w.r.t the architecture proposed in [7]. After, we derived D value by exploiting the relation (1):

$$D = p^2 c/k \tag{1}$$

where c is the number of channels in the image. Finally, we calculated d value according to (2):

$$d = p^2 cn \tag{2}$$

where n is the number of patches. We also tested the d value used in the ViT-Base architecture described in [7], which is equal to 3072.

Table 1. The values considered in the grid-search.

Patch size	d	L	D	k	N configuration
16	2048	4	64	4	1
			32	8	2
			16	16	3
		6	64	4	4
			32	8	5
			16	16	6
		8	64	4	7
			32	8	8
			16	16	9
	3072	4	64	4	10
			32	8	11
			16	16	12
		6	64	4	13
			32	8	14
			16	16	15
		8	64	4	16
			32	8	17
			16	16	18

2.4 3D ViTs Training

Training, validation, and test of the models were coded in Python by employing the following modules: pytorch (v. cuda-1.10.0) [17], keras (v. 2.7.0) [4], tensorflow (v. 2.7.0) [6], numpy (v.1 .20.3) [8], scikit-learn (v. 0.24.2) [18], pydicom (v. 2.1.2) [11] and pillow (v. 9.0.1) [5].

Since the goal of this work was a preliminary investigation of the effectiveness of ViTs in PCa aggressiveness, we did not perform a comprehensive hyperparameter optimization; instead, we focused mainly on optimizing the architectural features of ViTs via the grid search described above. The hyperparameters' values used are: Learning rate = 1e-4, Weight decay = 1e-2, Number of steps = 1000, Batch size = 4, Warmup steps = 1000, Optimization algorithm = Adam, Loss function = Binary Cross Entropy.

To make each training run reproducible, we exploited the reproducibility flags provided by pytorch [17], numpy [8], and random [22] libraries, choosing a seed equal to 42. The detailed code is reported in Listing 1.1

We split the entire dataset into two: 90 lesions (80%) were used for the grid search and the final training of the best-performing architecture; 22 lesions (20%) were kept for the final test of the best-performing architecture. We ensured a strict patient separation by this split. This means that all the lesions of the same patient were contained only in one of the two splits to avoid any data leakage. In addition, we stratified w.r.t the aggressiveness label (2/3 LG and 1/3 HG) and the lesion location (2/5 PZ, 2/5 AS, and 1/5 TZ).

We used the 90-lesion sub-set to carry out the grid search. This sub-set was further split into two sub-sets: 90% used for training and 10% used for validation. As a result, the validation set comprised 9 lesions (4 PZ [3 LG 1 HG] + 4 AS [3 LG + 1 HG] + 1 TZ HG).

For each ViT configuration, we evaluated the following metrics: specificity, sensitivity, accuracy, AUROC, and F2-score. The training was performed according to an ad-hoc early-stopping criterion defined as follows.

Early-stopping criterion. On the validation set, we measure both the specificity and the sensitivity at each epoch. If both metrics are greater than 0.6, we save the model at that epoch. In the subsequent epochs, if the specificity and sensitivity condition still occurs, as well as an increase in AUROC, the best model is updated. If this condition never occurs, we save the model that has the higher AUROC. When possible, this criterion ensures that the model can distinguish between both classes more accurately.

At the end of the grid search, we chose the best configuration based on the performance on the validation set, and we re-trained it with a 5-fold cross-validation (CV) to obtain more statistically reliable results. Namely, the training set was divided into five equally distributed folds, of which, in turn, one was used as a validation set. This way, we minimized possible splitting bias. Moreover, also, in this case, we stratified w.r.t classes and lesion zones.

The five models were finally evaluated on the same test set (i.e., the 22 lesions mentioned above). We reported performance as mean and standard deviation across each training run.

Listing 1.1. Reproducibility code.

```
import numpy as np
import random
import torch

def set_reproducibility(seed=42):
    random.seed(seed)
    np.random.seed(seed)
    torch.manual_seed(seed)
    torch.cuda.manual_seed_all(seed)
```

2.5 CNNs Architectures

As a comparison, we designed 3D CNN and trained it by following the same approach used to train the 3D ViTs.

The 3D CNN model consisted of three convolutional blocks (the composition of each block is described in Table 2) and four fully connected layers. We performed an architecture optimization of this model as done for ViT's architecture. A total of five configurations was considered. In each configuration, we varied the size of the max-pooling kernel within the three convolutional blocks. As detailed in the Pytorch library [17] documentation, a kernel consists of (kD, kH, kW),

so we investigated five different combinations of the placement and number of kernels acting only on the plane ((1,2,2)), and kernels acting also on the third dimension ((2,2,2)). We provide a complete description of the different configurations in the Table 3.

Table 2. The composition of 3D-CNN convolutional blocks. In the first block, k=7, whilst k=3 in the other two blocks.

Convolutional block
3D Convolutional layer (kernel kxk)
3D Max Pooling layer
Batch Normalization layer
3D Convolutional layer (kernel 1x1)

Table 3. The composition of the five alternative configurations of the 3D-CNN. MP = Max-Pooling.

N configuration	MP kernel size
1	(1,2,2) (1,2,2) (2,2,2)
2	(1,2,2) (2,2,2) (2,2,2)
3	(2,2,2) (1,2,2) (1,2,2)
4	(1,2,2) (2,2,2) (1,2,2)
5	(2,2,2) (2,2,2) (1,2,2)

To train each 3D-CNN's configuration, we exploited the same dataset partitioning used for 3D ViTs. To make the results comparable, we again evaluated the performance of each configuration by training the model with the fixed splitting of the dataset. Regarding the early-stopping criterion, we established that if validation loss did not decrease for more than five consecutive epochs, training was stopped.

We then re-trained the best configuration by applying the 5-fold CV, and we evaluated all five models on the test set, reporting the mean and standard deviation results. The training hyperparameters were set as follows: Learning rate = 1e-4, Epochs = 20, Batch size = 4, Optimization algorithm = Adam, Loss Function = Cross Entropy.

3 Results

3.1 3D ViT Results

The following parameters led to the best-performing 3D ViT: $p = 16$, $d = 2048$, $L = 6$, $D = 32$ and $k = 8$. This corresponds to the configuration number five

in Table 1. An overview of the model is depicted in Figure 1. On the 5-fold CV training this model provided 0,775 AUROC and 0,523 F2-score. In particular, the best split w.r.t the AUROC metric yielded 0,927 AUROC and 0,735 F2-score. We reported complete results for all the five CV models in Table 4.

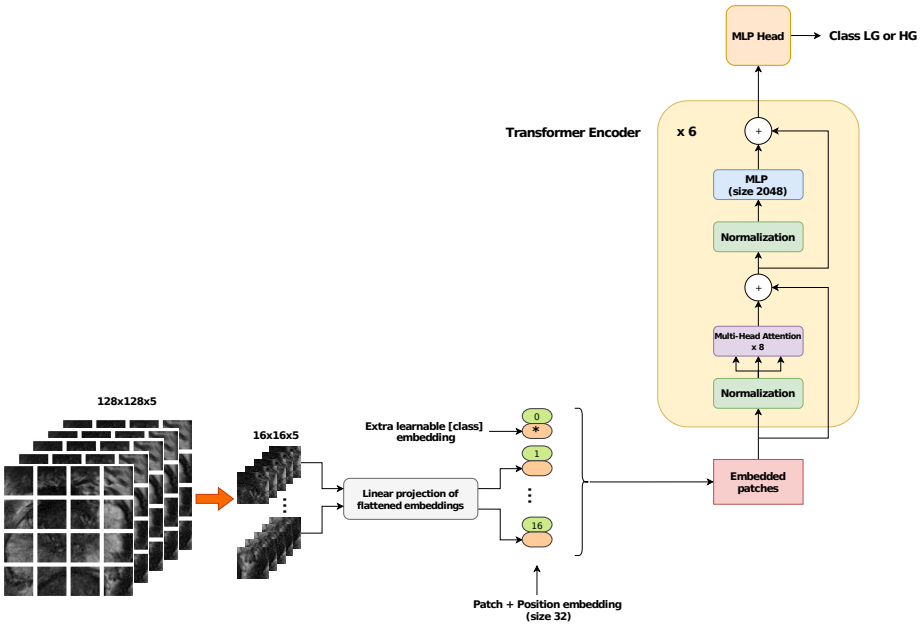

Fig. 1. Our best-performing ViT architecture.

3.2 CNN Results

The best 3D-CNN configuration resulted as the number four of those shown in Table 3. By applying the 5-fold CV on the test set, this model yielded 0.585 mean AUROC and 0.215 mean F2-score. The best split w.r.t the AUROC metric provided 0.635 AUROC and 0.3125 F2-score. We reported all the results for the five CNN models in Table 5.

4 Discussion and Conclusions

This study aimed to evaluate the effectiveness of 3D ViTs in assessing the aggressiveness of PCa, as this deep learning model is emerging as a fresh gold standard in several computer vision tasks. As a starting point, we exploited the architecture proposed in [7], and we modified it to preocess 3D patches and to significantly reduce the number of parameters so that we could train it from scratch

Table 4. Best 3D ViT configuration results in the 5-fold CV on the test set.

Cross-validation Fold	Specificity	Sensitivity	Accuracy	AUC	F2-score
1	0,688	0,667	0,682	0,74	0,606
2	0,875	0,167	0,682	0,698	0,185
3	0,75	0,333	0,636	0,708	0,333
4	0,75	0,833	0,773	0,802	0,758
5	0,688	0,833	0,727	0,927	0,735
Mean (SD)	0,750 (0,076)	0,567 (0,303)	0,700 (0,052)	0,775 (0,094)	0,523 (0,254)

Table 5. Best CNN configuration results in the 5-fold CV on the test set.

Cross-validation Fold	Specificity	Sensitivity	Accuracy	AUC	F2-score
1	1.0	0.167	0.773	0.583	0.2
2	0.813	0.167	0.636	0.604	0.179
3	0.938	0.167	0.727	0.552	0.192
4	0.625	0.333	0.545	0.635	0.313
5	0.9375	0.167	0.727	0.552	0.192
Mean (SD)	0,8625 (0,145)	0,2 (0,068)	0,682 (0,215)	0.585 (0,089)	0.215 (0,05)

using a small amount of data, such as the freely available ProstateX-2 challenge dataset [9]. With a grid search on the architectural features of the newly defined 3D ViT model, we selected the best-performing architecture and evaluated it via a CV approach. A 3D CNN model was designed and trained from scratch to have a basic reference model against which to compare our 3D ViT. It is worth noting that, to the best of our knowledge, this is the first work that trains a 3D CNN on volumetric scans to predict PCa aggressiveness. Three-dimensional CNN models have been previously exploited only to distinguish clinically significant from non-significant lesions [10, 13]. The results of our comparison showed that, when trained with the same training pipeline, 3D ViT outperforms the 3D CNN. Although both models exploited volumetric information, the 3D CNNs likely suffered more from the lack of data. Whilst, despite the small amount of data and without any specific structural optimization, the best-performing 3D ViT provided quite good results, reaching an AUROC of 0.927 on the test set in the best dataset partitioning.

As a further means of comparison with state-of-the-art methods, we compared our results with those obtained in [27], which is the only work, to the best of our knowledge, that addressed our same clinical task on the ProstateX-2 challenge dataset. For the sake of clarity, we hereby highlight the differences between our work and [27]. In [27], the CNNs were trained on 2D images cropped around the center of the lesion rather than the prostate, and the training was performed using more data. In fact, in addition to the ProstateX-2 challenge dataset, additional 132 lesions from a private dataset were used, and a transfer-learning

approach was performed. Furthermore, in [27], the dataset was split randomly, while we ensured a stratified and complete separation among patients. Results of the comparison are reported in Table 6. Overall, although the performance achieved by our model is lower, we must stress that it was obtained with less training data. Furthermore, unlike [27], we ensured a complete separation of patients between training and test sets, as well as a double stratification, w.r.t. class and the lesion's zone. This approach suppressed any bias in favor of the model's classification capabilities.

Table 6. Comparison between 3D ViT and the CNN from [27].

Model	Specificity	Sensitivity	Accuracy	AUC	F2-score
Our ViT	0,750 (0,076)	0,567 (0,303)	0,700 (0,052)	0,775 (0,094)	0,523 (0,254)
CNN from [27]	-	0.794 (0.0124)	0.738 (0.0136)	0.809 (-)	-

Our study has been conceived as a preliminary investigation and, as such, it has some limitations. Indeed, we did not apply any image enhancement steps nor any architectural optimization of the original ViT model by, for instance, including anatomical priors or employing diverse loss functions. ProstateX-2 is a challenging dataset as it contains lesions in different areas of the prostate gland. We applied the 3D ViT only to T2w scans, as these appeared more informative according to our previous research in the field [2]. Nonetheless, the contribution of ADC maps in cancer lesions located in diverse gland zones might be informative and they could enable a multimodal 3D ViT to better predict the lesion aggressiveness. Overall, as a first exploratory step, our results are encouraging and suggest that 3D ViTs, trained from scratch, might be a viable strategy for assessing PCa aggressiveness. To confirm this statement, additional studies are needed, especially on larger datasets and on datasets acquired with different protocols and from different institutions. This would be necessary to validate the robustness and generalization capabilities of the 3D ViT model. All these additional experiments will be the subject of our future works.

Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952159 (ProCAncer-I) and from the Regional Project PAR FAS Tuscany - NAVIGATOR. The funders had no role in the design of the study, collection, analysis and interpretation of data, or writing the manuscript.

References

1. Barentsz, J.O., Choyke, P.L., Cornud, F., Haider, M.A., Macura, K.J., Margolis, D., Shtern, F., Padhani, A.R., Tempany, C.M., Thoeny, H.C., et al.: Pi-rads prostate imaging–reporting and data system: 2015, version 2. eur urol 2016; 69: 16–40. European urology **70**(5), e137 (2016)
2. Bertelli, E., Mercatelli, L., Marzi, C., Pachetti, E., Baccini, M., Barucci, A., Colantonio, S., Gherardini, L., Lattavo, L., Pascali, M.A., et al.: Machine and deep learning prediction of prostate cancer aggressiveness using multiparametric mri. Frontiers in oncology **11**, 802964–802964 (2021)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chollet, F., et al.: Keras (2015), https://github.com/fchollet/keras
5. Clark, A.: Pillow (pil fork) documentation (2015)
6. Developers, T.: Tensorflow (2021). https://doi.org/10.5281/zenodo.5593257
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (2020). https://doi.org/10.1038/s41586-020-2649-2
9. Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H.: Prostatex challenge data. Cancer Imaging Arch **10**, K9TCIA (2017)
10. Liu, S., Zheng, H., Feng, Y., Li, W.: Prostate cancer diagnosis using deep learning with 3d multiparametric mri. In: Medical imaging 2017: computer-aided diagnosis. vol. 10134, pp. 581–584. SPIE (2017)
11. Mason, D., scaramallion, rhaxton, mrbean bremen, Suever, J., Vanessasaurus: pydicom/pydicom: pydicom 2.1.2 (2020). https://doi.org/10.5281/zenodo.4313150, https://doi.org/10.5281/zenodo.4313150
12. Matsoukas, C., Haslum, J.F., Söderberg, M., Smith, K.: Is it time to replace cnns with transformers for medical images. arXiv preprint arXiv:2108.09038 (2021)
13. Mehrtash, A., Sedghi, A., Ghafoorian, M., Taghipour, M., Tempany, C.M., Wells III, W.M., Kapur, T., Mousavi, P., Abolmaesumi, P., Fedorov, A.: Classification of clinical significance of mri prostate findings using 3d convolutional neural networks. In: Medical Imaging 2017: Computer-Aided Diagnosis. vol. 10134, p. 101342A. International Society for Optics and Photonics (2017)
14. Mehta, P., Antonelli, M., Ahmed, H.U., Emberton, M., Punwani, S., Ourselin, S.: Computer-aided diagnosis of prostate cancer using multiparametric mri and clinical features: A patient-level classification framework. Medical image analysis **73**, 102153 (2021)
15. Mehta, P., Antonelli, M., Singh, S., Grondecka, N., Johnston, E.W., Ahmed, H.U., Emberton, M., Punwani, S., Ourselin, S.: Autoprostate: Towards automated reporting of prostate mri for prostate cancer assessment using deep learning. Cancers **13**(23), 6138 (2021)

16. Mohler, J.L., Antonarakis, E.S., Armstrong, A.J., D'Amico, A.V., Davis, B.J., Dorff, T., Eastham, J.A., Enke, C.A., Farrington, T.A., Higano, C.S., et al.: Prostate cancer, version 2.2019, nccn clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network **17**(5), 479–505 (2019)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
19. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188 (2021)
20. Song, Y., Zhang, Y.D., Yan, X., Liu, H., Zhou, M., Hu, B., Yang, G.: Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric mri. Journal of Magnetic Resonance Imaging **48**(6), 1570–1577 (2018)
21. Turkbey, B., Rosenkrantz, A.B., Haider, M.A., Padhani, A.R., Villeirs, G., Macura, K.J., Tempany, C.M., Choyke, P.L., Cornud, F., Margolis, D.J., et al.: Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. European urology **76**(3), 340–351 (2019)
22. Van Rossum, G.: The Python Library Reference, release 3.8.2. Python Software Foundation (2020)
23. Vickers, A.J.: Effects of magnetic resonance imaging targeting on overdiagnosis and overtreatment of prostate cancer. European Urology **80**(5), 567–572 (2021). https://doi.org/https://doi.org/10.1016/j.eururo.2021.06.026
24. Wang, Z., Liu, C., Cheng, D., Wang, L., Yang, X., Cheng, K.T.: Automated detection of clinically significant prostate cancer in mp-mri images based on an end-to-end deep neural network. IEEE Trans on Medical Imaging **37**(5), 1127–1139 (2018)
25. World Health Organization, I.A.f.R.o.C.: (2020), https://gco.iarc.fr/today/data/factsheets/populations/908-europe-fact-sheets.pdf
26. Yoo, S., Gujrathi, I., Haider, M.A., Khalvati, F.: Prostate cancer detection using deep convolutional neural networks. Scientific reports **9**(1), 1–10 (2019)
27. Yuan, Y., Qin, W., Buyyounouski, M., Ibragimov, B., Hancock, S., Han, B., Xing, L.: Prostate cancer classification with multiparametric mri transfer learning model. Medical physics **46**(2), 756–765 (2019)