Multi-Scale Hybrid Transformer Networks: Application to Prostate Disease Classification

Ainkaran Santhirasekaram[1,2,3], Karen Pinto[3], Mathias Winkler[3], Eric Aboagye[1], Ben Glocker[2], and Andrea Rockall[1,3]

[1] Department of Surgery and Cancer, Imperial College London, London, UK
[2] Department of Computing, Imperial College London, London, UK
a.santhirasekaram19@imperial.ac.uk
[3] Imperial College Healthcare Trust, London, UK

Abstract. Automated disease classification could significantly improve the accuracy of prostate cancer diagnosis on MRI, which is a difficult task even for trained experts. Convolutional neural networks (CNNs) have shown some promising results for disease classification on multi-parametric MRI. However, CNNs struggle to extract robust global features about the anatomy which may provide important contextual information for further improving classification accuracy. Here, we propose a novel multi-scale hybrid CNN/transformer architecture with the ability of better contextualising local features at different scales. In our application, we found this to significantly improve performance compared to using CNNs. Classification accuracy is even further improved with a stacked ensemble yielding promising results for binary classification of prostate lesions into clinically significant or non-significant.

Keywords: Prostate Cancer · Convolutional Neural Network · Transformer.

1 Introduction

Multi-parametric MRI differentiates non-significant from significant cancers with high accuracy [6]. The scoring system for prostate cancer classification, however, is still somewhat subjective with a reported false positive rate of 30-40 percent [4]. Therefore there is great interest and clinical need for more objective, automated classification methods for prostate lesions with the goal to improve diagnostic accuracy [7]. There have been various approaches for automated methods based on hand-crafted, quantitative features (radiomics) such as textural and statistical measures that are extracted from regions of interest and used in a machine learning classifier [16]. Textural features have already shown to be beneficial to identify significant prostate cancer [13,19]. Convolutional neural networks are now the most popular approach for automating prostate disease classification on MRI with some promising results [2,7,12,17,21,23]. Due to weight-sharing, the resulting translational invariance and the relatively small receptive field of shallow CNNs, the extracted features tend to capture mostly local information

[1]. One has to build much deeper CNNs, combined with down-sampling and multi-scale processing to extract more global information. This problem is relevant to the classification of prostate cancer where global anatomical information and the relationship between different features is important. More recently, the vision transformer has shown competitive performance with convolutional neural networks on natural image classification tasks when pre-trained on large-scale datasets such as ImageNet [20]. However, while transformers can extract better global features by leveraging the power of self-attention for modelling long-range dependencies within the data, the direct tokenisation of patches from an image makes it more difficult to extract more fine low level features while also ignoring locality unlike in deep CNNs. Yet, we know local pixels are highly correlated and this lack of inductive bias means Visual Transformer need a large scale dataset to compete with deep CNN's. It has also been shown that self-attention in the initial layers of a model can learn similar features to convolutions [8]. This shows that the inductive biases imposed by CNNs is appropriate and helpful for feature extraction. Therefore, the use of transformer-only models such as the vision transformer on smaller medical imaging data-sets seems limited and an intriguing approach would be to combine the best of both worlds, giving rise to multi-scale hybrid CNN/transformer networks.

1.1 Contribution

We devise a deep learning approach capable of contextualising local features of prostate lesions through a novel hybrid CNN/transformer architecture with the aim to improve classification of prostate lesion into clinically significant and non-significant. The first stage of our architecture extracts features in a shallow multi-resolution pathway CNN. Each scale extracts CNN based features ranging from more fine grained textural features in the high resolution pathway to more coarse global information from the larger receptive field in the low resolution pathway. Instead of extending the depth of the CNN and using fully connected layers to combine the features within and across different scales, we leverage the powerful self-attention mechanism of the transformer to do this. We specifically use a transformer architecture which takes as input the feature maps from the CNN pathways which we hypothesise will learn better contextualised features to build a richer representation of the input lesion. Our model demonstrates excellent classification accuracy and outperforms a CNN only based approach. We finally further improve performance through a stacked ensemble of our model to outperform other baseline models in the ProstateX challenge [15].

2 Methods

Our proposed model (Fig. 1) has two stages. Stage 1 consists of 3 parallel pathways, each with a different resolution input. There is no weight-sharing between parallel pathways, so each learns discriminative features for a specific resolution. We use $5 \times 5 \times 3$ convolutional filters in the first layer followed by 2D max pooling in order to account for the anisotropic nature of the input patches. This is followed by residual block layers using 3 × 3 × 3 convolutions and 3D max pooling. We also used grouped convolutions of size 4, which we found to have better performance than single-grouped convolutions during cross-validation for hyper-parameter tuning [22]. The final 128 features maps for each resolution pathway are concatenated to form a stack of 384 feature maps of size 8 × 8 × 4 for stage 1. Each feature map is flattened into a one-dimensional vector forming the input for stage 2 of our model.

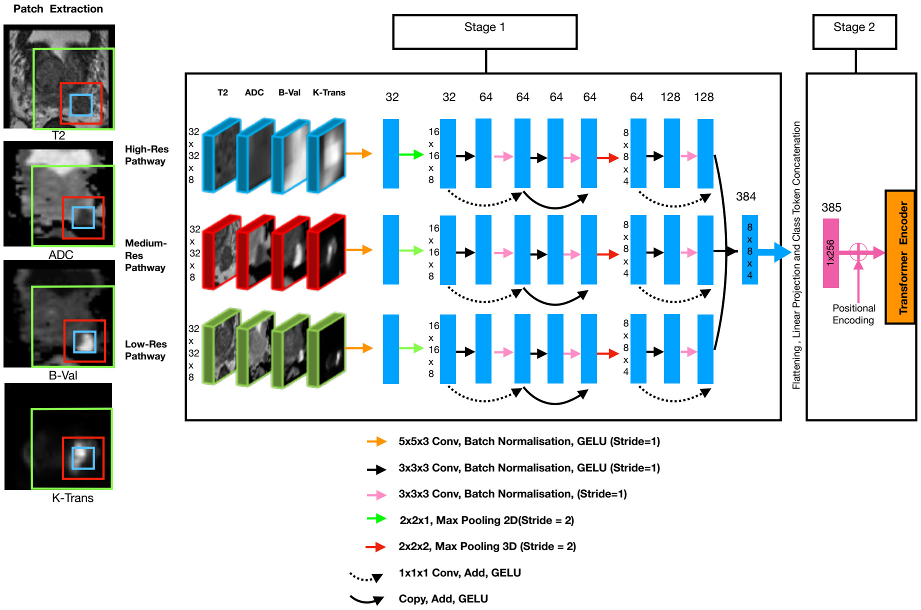

Fig. 1. Proposed model for prostate disease classification. The first stage is the feature extractor for 3 resolution pathways. The second stage is the transformer encoder.

Stage 2 of the model (Fig. 1) involves a linear transformation of the flattened feature maps to an embedding space of dimension 256. A random vector of size 1 × 256 denoted as the classification token is concatenated to the embedding matrix to learn an image level representation of the feature maps through self-attention. The embedding matrix is of size 385 × 256. We encode position p using a combination of sine and cosine waves as used in the vision transformer [9]. Each positional encoding i is a vector of $1 \times d$, where d is the embedding dimension with each element in the vector denoted with j. The formulation is described in equation 1 below.

$$p_{i,j} = \left\{ \begin{array}{ll} \sin(\frac{i}{1000^{\frac{j}{d}}}, & \text{if j is even.} \\ \cos(\frac{i}{1000^{\frac{j-1}{d}}}, & \text{if j is odd.} \end{array} \right\} \tag{1}$$

We sum the positional encodings to the embeddings to form new embeddings as input into the transformer encoder visualised in Fig. 2. The first part of the encoder consists of layer normalisation followed by a linear transformation of the embedding to Query(Q), Key(K) and value(V) matrices. The Q, K and V matrices are logically split by the number of heads (h) to be of dimension $385 \times 256/h$. Multi-headed self attention is calculated using scaled dot product attention as:

$$head_i = softmax(\frac{Q \times K^T}{\sqrt{256/h}}) \times V \tag{2}$$

$$Q = M \in \mathcal{R}^{385 \times 256/h}, K = M \in \mathcal{R}^{385 \times 256/h}, V = M \in \mathcal{R}^{385 \times 256/h}$$

$$MultiHead(Q, K, V) = Concat(Head_1, Head_2 ... Head_h) \times W_0 \tag{3}$$

$$MultiHead(Q, K, V) = M \in \mathcal{R}^{256 \times 385}, W_0 = M \in \mathcal{R}^{256 \times 256}$$

The different heads are concatenated and undergo linear transformation (equation 3). The next stage is a multi-layer perceptron (MLP) with two fully connected layers (Fig. 2). The transformer encoder network is repeated L(layers) times. Residual connections are incorporated to aid with gradient flow. We use Gaussian error linear unit (GELU) activation for both stages [11] and a dropout rate of 0.2 after every dense layer except for linear transformation of Q, K, V in stage 2 of the model. During 5-fold cross validation we observe the optimum number of heads, layers and MLP hidden dimension to be 8, 8 and 1024 respectively. Finally, the learnt classification token vector is consumed by an MLP classifier which consists of two fully connected layers with a hidden dimension of 1024 (Fig. 2). Stage 1 and 2 of the model are trained end-to-end.

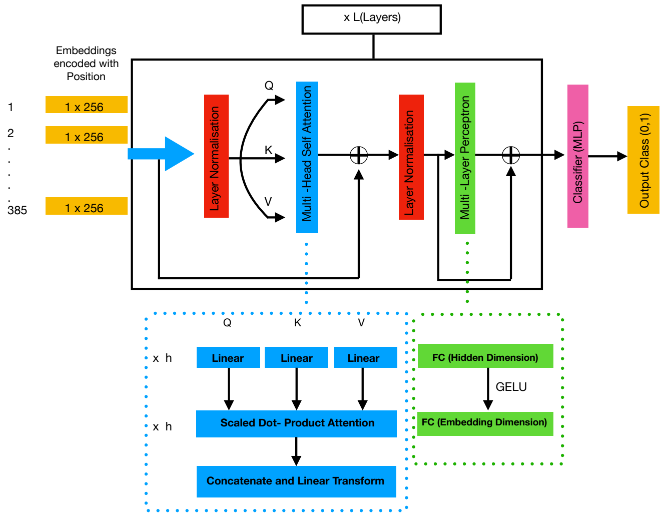

Fig. 2. Transformer encoder architecture with L layers and h heads for multi headed self attention. An MLP is used as the classifier after L layers of the Transformer encoder

2.1 Model Comparison

CNNs: We compare with a number of models using convolutions only. Firstly, we train the three resolution pathways in stage 1 of our model separately followed by 2 fully connected layers of size 4096 and 512 with dropout (0.5) to form 3 separately trained CNN models without masking. We then train a new high and medium resolution CNN with tumour masking followed by a new low resolution CNN with whole prostate masking as its goal is to extract prostate anatomical as well as lesion information. After evaluating the effect of masking on classification performance, we trained two multi-resolution CNNs (high and medium resolution vs all 3 resolutions). We replace the second stage with 2 fully connected layers of size 8192 and 1024 with dropout (0.5) for both multi-resolution CNNs.

Model Ensemble: To improve classification accuracy of our proposed model, We trained an ensemble of our multi-Scale hybrid transformer by varying the number of epochs during training (14-16 epochs), number of layers (6-9 Layers) and the MLP hidden dimension (1024 and 1280) to produce 24 trained models for each training fold in 5-fold cross-validation. We employ a stacked ensemble method and use the class output probability from each model as input to train a logistic regression model with L2 penalty ($\lambda = 1.0$) to predict the class outputs.

Radiomics: A total of 2724 features were extracted using TexLab2.0; a Radiomics analysis software [3]. The tumour segmentations were used for masking. Various feature selection methods were explored during 5 fold cross-validation. The best feature selection process firstly involved removing highly correlated features(correlation coefficient > 0.95) using Pearson Correlation [5]. Secondly, uni-variate feature selection with Analysis of variance (ANOVA) analysis was performed to select the best 20 features [18]. Least absolute shrinkage and selection operator (LASSO) logistic regression was then employed which identified 6 features useful for model building. 18 different classifiers were evaluated for classification. Logistic regression with L2 penalty ($\lambda = 1.0$) optimised during cross-validation demonstrated the best classification accuracy.

Other Baslines: We trained only on the ProstateX challenge training dataset and therefore validate our final stacked ensemble model on the ProstateX challenge test set to compare to other baseline models in the literature trained and tested on the same dataset [15]. A challenge entry returns a single AUC-ROC value and position on the leaderboard.

3 Experiments and Results

3.1 Dataset

The PROSTATEx challenge dataset which was acquired on two, 3 Tesla scanners [15]. For the purpose of this study, the T2 weighted axial, diffusion weighted imaging (b-800), apparent diffusion coefficient (ADC) maps and K-trans images were used. The dataset consists of 330 pre-selected lesions [15]. Clinically significant lesions are classed as Gleason grade group 2 and above. Non-biopsied lesions were considered non-significant as ground truth and histology results were used as ground truth for biopsied lesions. The dataset is imbalanced with only 23 percent of lesions labelled as significant.

Segmentation of each lesion and prostate was performed by a Radiologist using ITKsnap [24]. The lesion segmentation was performed on the T2 axial, ADC, b-800 and K-trans. All patients had a lesion visible on at least one sequence which were segmented manually. The sequence with the maximum lesion volume (from sequences with the lesion visible) is mapped to the sequences where a lesion is not visible. Whole prostate segmentation was performed on the T2 axial only.

3.2 Pre-processing and Augmentation

We resample to form 3 sets of images: high resolution ($0.5mm \times 0.5mm \times 1.5mm$), medium resolution ($1.0mm \times 1.0mm \times 3.0mm$) and low resolution ($2.0mm \times 2.0mm \times 4.0mm$). Cubic B-spline interpolation is used for resampling of the MR images. Nearest-neighbour is used for resampling of the mask. The high resolution images were used for radiomics input. The ADC, b800 and K-trans images were then registered with the T2 axial images using affine transformation.

We then mask out irrelevant background areas that are further away from the boundary of the prostate as only nearby extra-prostatic regions are assumed to be valuable for tumour classification. We do this by using the boundaries of the whole prostate segmentation mask to form an initial bounding box for each slice around the prostate. The bounding box is then extended to include more extra-prostatic region by adding an optimal length to the width and height of the bounding box defined as 10 divided by the resolution(mm) in the axial plane.

Patches were then extracted for each resolution. For, high and medium resolution images, $32 \times 32 \times 8$ patches were extracted centred on the lesion. In the low resolution images, the goal is to capture as much of the prostate in the region of interest (ROI). Therefore we limit the area outside of the prostate, by centering a $32 \times 32 \times 8$ patch on a point equidistant between the lesion and whole prostate centre. All patches are then processed with and without masking as described in section 2.1. Finally, we re-scaled the intensities between 0 and 1 for normalisation.

We augment the significant class to handle class imbalance through vertical or horizontal flipping followed by random rotations between -90 and 90 degrees and random translation of 0 to 10mm.

3.3 Training

Despite accounting for class imbalance using augmentation. This would not completely account for natural variations of significant tumour appearance. Therefore we use a weighted binary cross entropy loss function (equation 4) with the weighting parameters fine tuned during cross validation. This was applied to all models.

$$WeightedCrossEntropy = -0.6log(p) - 0.4log(1-p) \tag{4}$$

All model were implemented using PyTorch. Experiments were run on three NVIDIA Geforce RTX 2080 GPUs. Stratified (using label) 5 fold cross-validation is used for model training/ validation and to optimise the number of epochs, batch size, learning rate, weight decay, loss function weightings, dropout rate, the transformer encoder hidden dimension and number of CNN/transformer encoder layers. The model weights are initialised with Kaiming initialisation [10]. The CNN based models use Adam optimisation with a base learning rate of 0.0001 [14] and weight decay of 0.001 for all models. The hybrid model uses Adam optimisation with weight decay(0.001) and cosine annealing (base learning rate: 0.001) as a learning rate scheduler. We use a batch size of 40 for training. The single and two resolution CNNs were trained for 10 epochs. The multi-resolution CNN and our multi-Scale hybrid transformer were trained for 12 and 15 epochs, respectively.

For testing, our final proposed stacked ensemble of multi-scale hybrid transformers is trained on the entire dataset and submitted for external validation on the ProstateX test set.

3.4 Results

Table 1. Mean and standard error on 5-fold cross-validation for metrics comparing all trained models. Best result for each metric is highlighted in bold.

	Accuracy	**Specificity**	**Precision**	**Recall**
High Res CNN(no mask)	0.841 ± 0.009	0.811 ± 0.025	0.826 ± 0.020	0.866 ± 0.015
High Res CNN(mask)	0.840 ± 0.009	0.813 ± 0.015	0.828 ± 0.056	0.868 ± 0.022
Medium Res CNN(no mask)	0.810 ± 0.001	0.794 ± 0.019	0.802 ± 0.004	0.866 ± 0.043
Medium Res CNN(mask)	0.818 ± 0.043	0.800 ± 0.019	0.803 ± 0.035	0.857 ± 0.045
Low Res CNN(no mask)	0.771 ± 0.037	0.751 ± 0.039	0.739 ± 0.070	0.788 ± 0.053
Low Res CNN(mask)	0.764 ± 0.012	0.742 ± 0.058	0.752 ± 0.028	0.782 ± 0.043
Two Res CNN(no mask)	0.875 ± 0.009	0.854 ± 0.021	0.860 ± 0.006	0.888 ± 0.019
Three Res CNN(no mask)	0.883 ± 0.008	0.865 ± 0.028	0.868 ± 0.009	0.895 ± 0.022
Radiomics	0.775 ± 0.053	0.751 ± 0.055	0.753 ± 0.048	0.799 ± 0.040
Our Model (no mask)	0.900 ± 0.018	0.899 ± 0.024	0.879 ± 0.014	0.918 ± 0.017
Our Model Ensemble	**0.944 ± 0.013**	**0.927 ± 0.023**	**0.933 ± 0.009**	**0.959 ± 0.022**

We observe slightly improved performance from not using a mask for all three single resolution CNN models (Table 1). This suggests that the prostate area outside the volume of the tumour and nearby area outside the prostate itself provides useful information for classification. We therefore did not use whole prostate and tumour masks for the multi-resolution CNNs and our model.

We also find that increasing the input resolution of the CNN improves classification performance (Table 1). This is most likely due to class imbalance as the ROI increases which is more pronounced after masking. This is also likely due to the loss of fine features important for classification at coarser resolutions. However, we find the two resolution CNN demonstrates improved performance in evaluation metrics which is further enhanced with the three resolution CNN (Table 1). This demonstrates the importance of combining local and global features to learn better contextual information of the tumour lesion using the lower resolution pathways to provide more informative anatomical localisation of the tumour region. Our model outperforms all CNN only models in all metrics (Table 1). This shows our model improves CNN performance by harnessing self-attention in the transformer to extract better contextual features by learning important relationships between the features maps extracted in each CNN pathway. We also demonstrate significantly better performance of using multi-scale CNNs and our model compared to the radiomics approach (Table 1) highlighting the benefit of feature learning. A stacked ensemble of our model leads to overall best performance on all evaluation metrics (Table 1) and achieves an average AUC of 0.95 during 5- fold cross-validation (Fig. 3). Our model ensemble achieves an AUC of 0.94 and 3rd place on the leader-board for the ProstateX challenge test set.

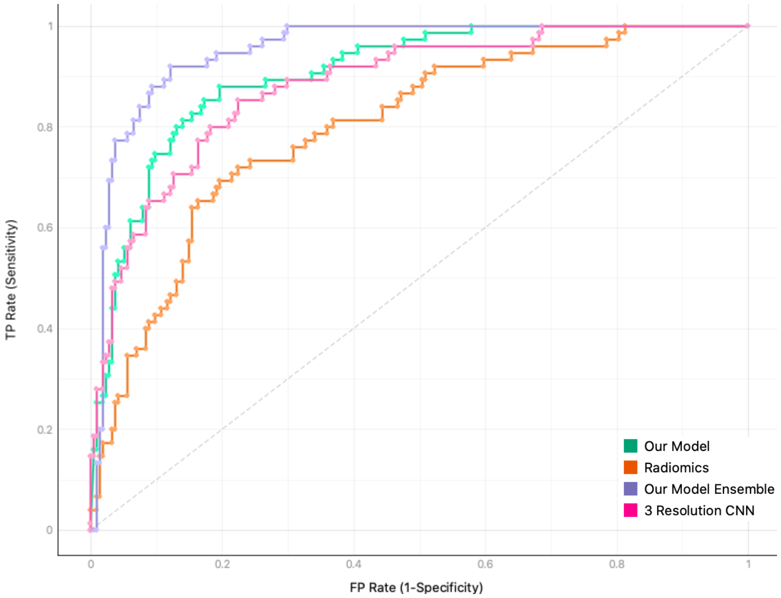

Fig. 3. ROC curves from the merged predictions of each fold comparing our model and model ensemble to radiomics and the three resolution CNN.

4 Conclusion

We demonstrate the importance of extracting contextual information of the tumour region in regards to its anatomical location and extension. We propose a novel multi-scale hybrid CNN/transformer network with the ability to extract richer contextualised features to build stronger representations which significantly improves prostate disease classification in all evaluation metrics compared to radiomics and multi-resolution CNNs. We believe our novel transformer-based approach could be appealing for many other disease classification tasks where the contextualisation of fine-detailed local features is important. This will be explored in future work.

Acknowledgements. This work was supported and funded by Cancer Research UK (CRUK) (C309/A28804).

References

1. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET). pp. 1–6. Ieee (2017)
2. Aldoj, N., Lukas, S., Dewey, M., Penzkofer, T.: Semi-automatic classification of prostate cancer on multi-parametric mr imaging using a multi-channel 3d convolutional neural network. European radiology **30**(2), 1243–1253 (2020)
3. Arshad, M.A., Thornton, A., Lu, H., Tam, H., Wallitt, K., Rodgers, N., Scarsbrook, A., McDermott, G., Cook, G.J., Landau, D., et al.: Discovery of pre-therapy 2-deoxy-2-18 f-fluoro-d-glucose positron emission tomography-based radiomics classifiers of survival outcome in non-small-cell lung cancer patients. European journal of nuclear medicine and molecular imaging **46**(2), 455–466 (2019)
4. Bass, E., Pantovic, A., Connor, M., Gabe, R., Padhani, A., Rockall, A., Sokhi, H., Tam, H., Winkler, M., Ahmed, H.: A systematic review and meta-analysis of the diagnostic accuracy of biparametric prostate mri for prostate cancer in men at risk. Prostate Cancer and Prostatic Diseases pp. 1–16 (2020)
5. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise reduction in speech processing, pp. 1–4. Springer (2009)
6. Brizmohun Appayya, M., Adshead, J., Ahmed, H.U., Allen, C., Bainbridge, A., Barrett, T., Giganti, F., Graham, J., Haslam, P., Johnston, E.W., et al.: National implementation of multi-parametric magnetic resonance imaging for prostate cancer detection–recommendations from a uk consensus meeting. BJU international **122**(1), 13–25 (2018)
7. Castillo T, J.M., Arif, M., Niessen, W.J., Schoots, I.G., Veenland, J.F., et al.: Automated classification of significant prostate cancer on mri: A systematic review on the performance of machine learning applications. Cancers **12**(6), 1606 (2020)
8. Cordonnier, J.B., Loukas, A., Jaggi, M.: On the relationship between self-attention and convolutional layers. arXiv preprint arXiv:1911.03584 (2019)
9. 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)
10. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
11. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
12. Ishioka, J., Matsuoka, Y., Uehara, S., Yasuda, Y., Kijima, T., Yoshida, S., Yokoyama, M., Saito, K., Kihara, K., Numao, N., et al.: Computer-aided diagnosis of prostate cancer on magnetic resonance imaging using a convolutional neural network algorithm. BJU international **122**(3), 411–417 (2018)
13. Khalvati, F., Wong, A., Haider, M.A.: Automated prostate cancer detection via comprehensive multi-parametric magnetic resonance imaging texture feature models. BMC medical imaging **15**(1), 1–14 (2015)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H.: Computer-aided detection of prostate cancer in mri. IEEE transactions on medical imaging **33**(5), 1083–1092 (2014)

16. Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Fanciullo, C., Morganti, A.G., Bellomi, M.: Radiomics: the facts and the challenges of image analysis. European radiology experimental **2**(1), 1–8 (2018)
17. 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)
18. St, L., Wold, S., et al.: Analysis of variance (anova). Chemometrics and intelligent laboratory systems **6**(4), 259–272 (1989)
19. Stoyanova, R., Takhar, M., Tschudi, Y., Ford, J.C., Solórzano, G., Erho, N., Balagurunathan, Y., Punnen, S., Davicioni, E., Gillies, R.J., et al.: Prostate cancer radiomics and the promise of radiogenomics. Translational cancer research **5**(4), 432 (2016)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
21. 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 transactions on medical imaging **37**(5), 1127–1139 (2018)
22. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
23. Yang, X., Liu, C., Wang, Z., Yang, J., Le Min, H., Wang, L., Cheng, K.T.T.: Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric mri. Medical image analysis **42**, 212–227 (2017)
24. Yushkevich, P.A., Gao, Y., Gerig, G.: Itk-snap: an interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 3342–3345. IEEE (2016)