Research Report on Deep Learning Methods in Prostate Cancer

Andrew Hwang, Emlyn Evans, Ernest Mo, Jiatong Li, Sasha Jenner & Tom Linstrom

Abstract

Background: Contemporary approaches to prostate cancer diagnosis are fallible and inefficient, often requiring multiple specialists to confirm a single diagnosis. Computer-automated decision-making systems can reduce the rate of misdiagnosis and increase reproducibility. In particular, convolutional neural networks (CNNs) are the leading algorithm for image recognition.

Materials and Methods: Magnetic resonance imaging (MRI) scans were collected from 204 patients suspected of having prostate cancer, alongside their true diagnosis and potential tumour location(s). Sequences included T2-weighted images, apparent diffusion coefficient (ADC) maps, and forward transfer constant (Ktrans) maps. Image cropping and augmentation were performed to enlarge the dataset and prepare it for modelling. 13 models were trained using CNNs with each categorical cross-entropy loss function weighted against a 77% bias towards non-cancerous findings in the data. Model performance was evaluated using accuracy, sensitivity, specificity and receiver operating characteristic (ROC) curves.

Results: Binary classification models were successfully trained and tested using different CNN architectures. The best model achieved an accuracy of 86.9% and an area under the ROC (AUROC) of 0.903.

Conclusion: Prostate cancer can be identified by CNNs with a higher performance than the current rate of index detection (77.6%) (Mirak, 2019). In the future, we intend to construct models for detecting the coordinates of a suspicious tumour given an MRI scan of the prostate.

Introduction

The current methods for detecting and grading prostate cancer include a digital rectal examination, a prostate specific antigen blood test, MRI imaging and/or a biopsy. For an MRI scan of the prostate, the size or density of a tumour is estimated and scored using the Prostate Imaging Reporting and Data System (PI-RADS) (Turkbey *et. al*, 2019) to determine the likelihood of metastasis. This is prone to error due to a high level of inter-observer variability between MRI contouring decisions (Steenbergen, 2015).

However, a higher level of reproducibility can be achieved by deep learning. CNNs are able to create highly automated and reliable models for prostate cancer detection. Although on a smaller scale, the present research is similar to that being achieved by the Biofocused RadioTherapy (BiRT) project, which attempts to deliver personalised treatments to prostate cancer patients.

Deep learning includes a wide range of machine learning methods based on artificial neural networks. Binary classification CNNs were employed to process MRI images of the prostate and thus make informed predictions regarding the presence of a clinically significant tumour.

CNNs consist of an input, output and several hidden layers. Convolutional layers serve to extract features from imagery, pooling layers combine these features by reducing the spatial size of the input, and the flatten layer then transforms this representation into a one-dimensional vector. The fully-connected neural network (FNN) serves to classify the input by connecting every neuron in one layer to every neuron in the next. CNNs learn by iteratively adjusting weights and biases to each neuron during training in order to minimise the difference between its predictions and the ground truth (the loss function).

Prior studies (Elahi, 2019) followed a similar method of image pre-processing. The present study differs from previous research in that CNN models are the sole investigation and the most recent version of Keras (2.3.0 released on 23rd Sep) is employed.

Materials and Methods

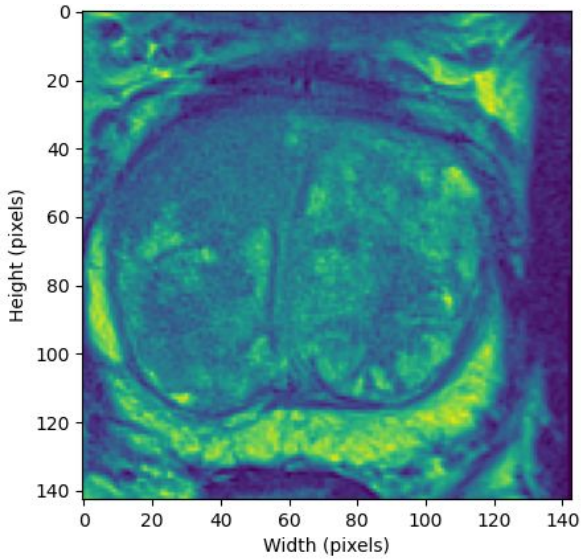
The data employed during our research includes a comma-separated values (CSV) file and three MRI parametric maps in the NIfTI-1 data format for each of the 204 patients. It was collated by the Radboud University Medical Centre (Radboudumc) in the Netherlands. The CSV file contained:

- Patient ID
- Finding number
- 3-D coordinate of the finding
- Whether or not it is evidence of clinically significant prostate cancer (Ground truth)

Each entry records the finding of a suspicious lesion in the prostate, of which there are 326. The three MRI imaging sequences recorded for each patient were T2-weighted images (highlights differences in the T2 relaxation time of tissues), ADC maps (measures the magnitude of the diffusion of water molecules within the tissue), and Ktrans maps.

First, all MRI images, located at the coordinates of their respective finding, were scaled to a common size then cropped to 143 pixels in height and width around the centre in order to remove all other extraneous information except the prostate gland.

Figure 1. The T2-weighted MRI cropped image for patient 5.



Three cropped MRI images for each finding were combined to form an image with three channels (in the order T2-weighted, Ktrans, followed by ADC) just as an RGB image is

composed of a red, green, and then blue channel. Ground truth labels were categorised into two columns, with the first identifying whether the image was not indicative of prostate cancer and the second corresponding to the reverse.

Next, in order to increase the number of images available for training and testing, the data pool was augmented 11-fold from 326 to 3586 images. For our particular dataset, the augmented images were set to rotate up to 20 degrees from their original position, flip horizontally and zoom up to 50% inwards.

Following this, the data pool was randomly split into training and testing datasets in a 6:1 ratio. Next, 13 models were constructed using Keras' Sequential model which consists of a linear stack of layers. Many varieties were tested, however, in general the models consisted of two to three convolution layers using the rectified linear unit (ReLU) or hyperbolic tangent activation functions with successive max-pooling in between. Then the output was flattened into a one-dimensional feature vector followed by two to three FNN layers. The final FNN layer consisted of just two neurons for binary classification using the softmax activation function to convert a vector of real numbers into a probability distribution.

Each model's loss was defined by a categorical cross-entropy function, optimised using the Adadelta or Adam gradient descent algorithms. Training was then achieved by fitting each model to the partitioned training dataset for up to 150 epochs, usually with a batch size of 32 images. Due to the heavy skewing of the dataset to images without cancer, the loss function is weighted such that the model is forced to focus more on the under-represented class.

After training, we evaluated each of the 13 models on the testing dataset based on a variety of performance metrics. The simplest of these was accuracy, which measures the proportion of the model's predictions that parallel the ground truth. However, more interesting and useful to evaluation, sensitivity, or the true positive (TP) rate, measures the proportion of positives that were correctly identified as such. While specificity, or the true negative (TN) rate, measures the proportion of negatives correctly identified. In particular, a TP prediction occurs when the model predicts true and is correct, while a FN prediction occurs when the model predicts false and is incorrect. Similarly, we employed a binary confusion matrix for each model to visualise these values.

Figure 2. The binary confusion matrix for model 8 with threshold 0.5 tested on 513 images. The sensitivity is 0.653 and specificity 0.871 in this case.

	Predicted: True	**Predicted: False**
Actual: True	81	43
Actual: False	50	339

Furthermore, the ROC curve graphs sensitivity against the complement of specificity for different threshold values, while the AUROC measures how capable the model is at distinguishing between classes and is represented by a value between 0 and 1.

Results

Thirteen models were created with different architectures and hyperparameters.

Figure 3. A table consisting of the performance of each model number ranked by **AUROC** from largest to smallest. The best performing model is highlighted in green; the worst in orange.

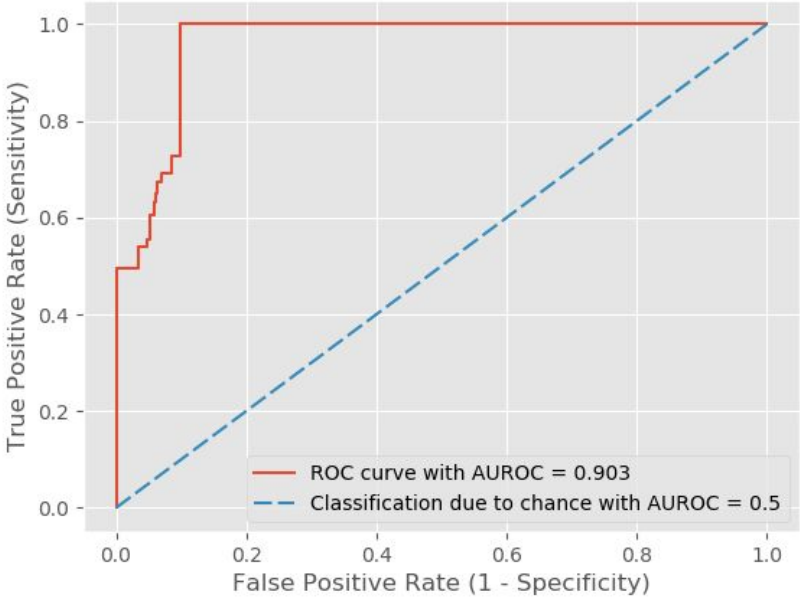
Model Number	**Accuracy (%)**	**Loss**	**Specificity at Threshold 0.5**	**Sensitivity at Threshold 0.5**	**AUROC**
5	86.9	0.855	0.939	0.632	0.903
8	81.9	0.844	0.871	0.653	0.861
13	57.5	0.881	0.49	0.883	0.774
6	54.4	0.704	0.489	0.737	0.65
1	31.2	1.543	0.079	0.977	0.644

4	27.9	1.854	0.095	0.946	0.624
11	76.8	0.605	1	0	0.578
9	76.4	0.659	1	0	0.526
2	77	3.707	1	0	0.525
10	75.4	0.658	1	0	0.512
3	80.7	3.111	1	0	0.508
12	76	0.68	1	0	0.502
7	17	13.375	0	1	0.5

In terms of the specificity and sensitivity metrics, ideally both should be close to 1 but specificity in particular should be high such that the number of FNs predicted is minimised. FNs occur when a patient has prostate cancer but the model predicts otherwise and is the worst-case scenario.

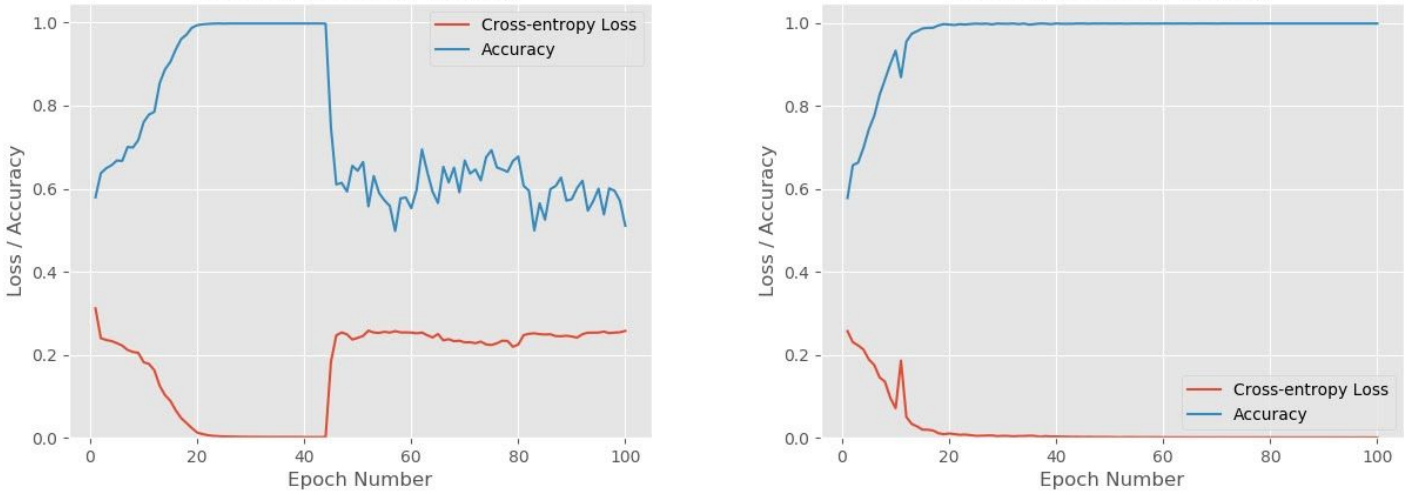
The optimal ROC curve should follow the periphery of the graph and was plotted for each model with model 5 displaying a very high AUROC of 0.903. This means that it is very good at distinguishing between images containing prostate cancer and those without.

Figure 4. The ROC curve for model 5 in red. The blue dotted line represents the ROC curve when classification is done at random.



Throughout, training loss and accuracy was similarly charted for each model with some trajectories converging to a value while others taking completely unexpected paths.

Figure 5. The trajectory of accuracy (blue) and cross-entropy loss (red) throughout training. Left: model 10, accuracy drastically decreases as loss increase after epoch 45. Right: model 8, accuracy tends to 100% and loss to 0 as training progresses.



Discussion

Although the results indicate reasonably high performance for certain models, there are a number of factors impacting the validity of these findings and the integrity of the models.

Firstly, the lack of original data available for training and testing the models has limited the scope of each CNN to recognising only a small subset (namely 326 images) of the total population of prostate MRI imagery. However, training data on the order of 1000 images per class is necessary according to (Quinlan, 1986) to properly train a deep learning model. Although image augmentation has been used as a secondary solution, it is not sufficient in order to accurately train a model for deployment, since the process creates only superficial alterations of the initial dataset.

In addition, the data was biased by 77% towards non-cancerous images. As such, six models favoured a false prediction, even after weighting the loss function against this imbalance.

Future studies could benefit from a larger dataset with more balanced findings, or through implementing the Python Smote library.

Figure 6. The binary confusion matrix for model 12 with threshold 0.5 tested on 513 images. The sensitivity is 0 and specificity 1 in this case. The bias towards a false prediction still results in a 76% accuracy on the testing data.

	Predicted: True	**Predicted: False**
Actual: True	0	123
Actual: False	0	390

Prior research recognised similar issues within their models and suggested that there needs to be greater public accessibility to MRI imaging data before significant improvements can be made (Elahi *et al.,* 2019).

Concerning the experimental process itself, a more systematic approach to improving models could be utilised in the future. In particular, hyperparameter optimisation algorithms may be used to improve the research testing framework.

The novelty of our research emerges from the final model's accuracy. In comparison to a recent review of misdiagnosis rates of prostate cancer, our best model was 75.2% better than human diagnosis (Mirak, 2019). This is suggestive of the significant ability for deep learning to transform the contemporary methods for detecting prostate cancer.

For future research, we are interested in building models to predict the precise location of suspicious lesion(s) on an MRI scan of the prostate. Also, improvements could be made to the accuracy of the ground truth data, which still relies on human labelling which is subject to systematic error.

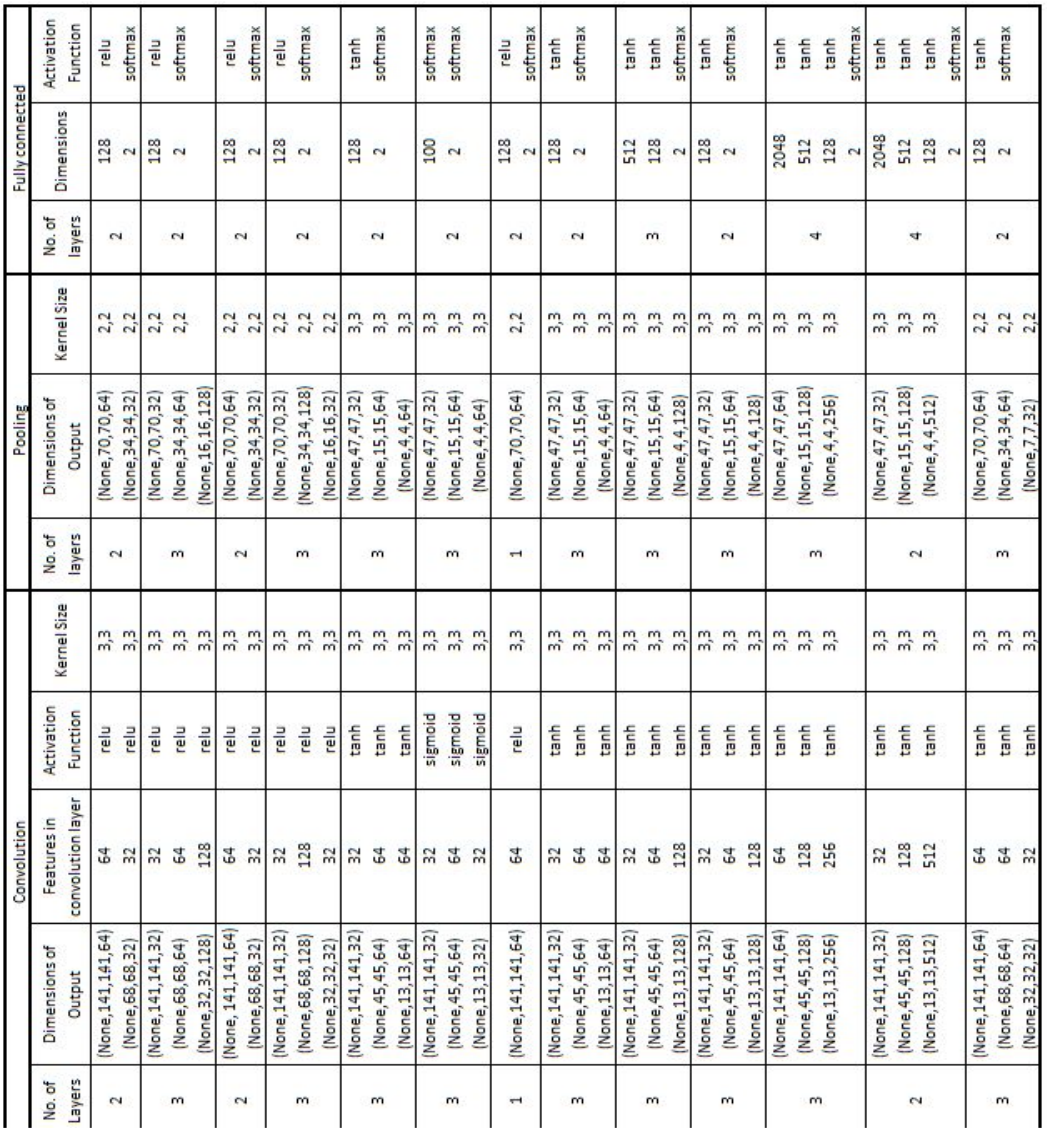
Conclusion

This study investigated the use of CNNs for the automation of prostate cancer diagnosis. Thirteen different CNN architectures were constructed and tested, and the best model achieved an AUROC of 0.903. Future work will involve the automated detection of the location of suspicious lesion(s) from MRI scans of the prostate.

Appendix

Figure 7. A tabulated representation of the hyperparameters and model structure for each of the 13 models.

Figure 8. A summary of model 5's structure.



	Convolution				Pooling			Fully connected		
No. of Layers	Dimensions of Output	Features in convolution layer	Activation Function	Kernel Size	No. of layers	Dimensions of Output	Kernel Size	No. of layers	Dimensions	Activation Function
2	(None,141,141,64)	64	relu	3,3	2	(None,70,70,64)	2,2	2	128	relu
	(None,68,68,32)	32	relu	3,3		(None,34,34,32)	2,2		2	softmax
3	(None,141,141,32)	32	relu	3,3	3	(None,70,70,32)	2,2	2	128	relu
	(None,68,68,64)	64	relu	3,3		(None,34,34,64)	2,2		2	softmax
	(None,32,32,128)	128	relu	3,3		(None,16,16,128)				
2	(None, 141,141,64)	64	relu	3,3	2	(None,70,70,64)	2,2	2	128	relu
	(None,68,68,32)	32	relu	3,3		(None,34,34,32)	2,2		2	softmax
3	(None,141,141,32)	32	relu	3,3	3	(None,70,70,32)	2,2	2	128	relu
	(None,68,68,128)	128	relu	3,3		(None,34,34,128)	2,2		2	softmax
	(None,32,32,32)	32	relu	3,3		(None,16,16,32)	2,2			
3	(None,141,141,32)	32	tanh	3,3	3	(None,47,47,32)	3,3	2	128	tanh
	(None,45,45,64)	64	tanh	3,3		(None,15,15,64)	3,3		2	softmax
	(None,13,13,64)	64	tanh	3,3		(None,4,4,64)	3,3			
3	(None,141,141,32)	32	sigmoid	3,3	3	(None,47,47,32)	3,3	2	100	softmax
	(None,45,45,64)	64	sigmoid	3,3		(None,15,15,64)	3,3		2	softmax
	(None,13,13,32)	32	sigmoid	3,3		(None,4,4,64)	3,3			
1	(None,141,141,64)	64	relu	3,3	1	(None,70,70,64)	2,2	2	128	relu
									2	softmax
3	(None,141,141,32)	32	tanh	3,3	3	(None,47,47,32)	3,3	2	128	tanh
	(None,45,45,64)	64	tanh	3,3		(None,15,15,64)	3,3		2	softmax
	(None,13,13,64)	64	tanh	3,3		(None,4,4,64)	3,3			
3	(None,141,141,32)	32	tanh	3,3	3	(None,47,47,32)	3,3	3	512	tanh
	(None,45,45,64)	64	tanh	3,3		(None,15,15,64)	3,3		128	tanh
	(None,13,13,128)	128	tanh	3,3		(None,4,4,128)	3,3		2	softmax
3	(None,141,141,32)	32	tanh	3,3	3	(None,47,47,32)	3,3	2	128	tanh
	(None,45,45,64)	64	tanh	3,3		(None,15,15,64)	3,3		2	softmax
	(None,13,13,128)	128	tanh	3,3		(None,4,4,128)	3,3			
3	(None,141,141,64)	64	tanh	3,3	3	(None,47,47,64)	3,3	4	2048	tanh
	(None,45,45,128)	128	tanh	3,3		(None,15,15,128)	3,3		512	tanh
	(None,13,13,256)	256	tanh	3,3		(None,4,4,256)	3,3		128	tanh
									2	softmax
2	(None,141,141,32)	32	tanh	3,3	2	(None,47,47,32)	3,3	4	2048	tanh
	(None,45,45,128)	128	tanh	3,3		(None,15,15,128)	3,3		512	tanh
	(None,13,13,512)	512	tanh	3,3		(None,4,4,512)	3,3		128	tanh
									2	softmax
3	(None,141,141,64)	64	tanh	3,3	3	(None,70,70,64)	2,2	2	128	tanh
	(None,68,68,64)	64	tanh	3,3		(None,34,34,64)	2,2		2	softmax
	(None,32,32,32)	32	tanh	3,3		(None,7,7,32)	2,2			

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 141, 141, 32)      896
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 47, 47, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 45, 45, 64)        18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 15, 15, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 13, 13, 64)        36928
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 4, 4, 64)          0
_________________________________________________________________
flatten (Flatten)            (None, 1024)              0
_________________________________________________________________
dense (Dense)                (None, 128)               131200
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 258
=================================================================
Total params: 187,778
Trainable params: 187,778
Non-trainable params: 0
_________________________________________________________________
```

References

Lemaitre G., Marti R., Freixenet J., Vilanova P. & Rizzi A., (2015) Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review. *Computers in Biology and Medicine*, Vol 60, pp8-31

Mirak, S. A., Shakeri, S., Bajgiran, A. M., Felker, E. R., Sung, K. H., Asvadi, N. H., … Raman, S. S. (2019). Three Tesla Multiparametric Magnetic Resonance Imaging: Comparison of Performance with and without Endorectal Coil for Prostate Cancer Detection, PI-RADS™ version 2 Category and Staging with Whole Mount Histopathology Correlation. *Journal of Urology*, *201*(3), 496–502

Munir, K., Elahi, H., Ayub, A., Frezza, F., & Rizzi, A. (2019). Cancer Diagnosis Using Deep Learning: A Bibliographic Review. Cancers, 11(9), 1235

Quinlan, R. Machine Learning I, (1986) Kluwer Academic Publishers, Boston. Pp 81-106

Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Fanciullo, C., Morganti, A. G., & Bellomi, M. (2018). Radiomics: the facts and the challenges of image analysis. *European Radiology Experimental*, *2*(1)

Steenbergen, Peter et al. (2015) Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation. Radiotherapy and Oncology, Volume 115, Issue 2, 186 - 190

Turkbey B., Rosenkrantz A., Haider M., Padhani A., Villeirs G., Macura K., Tempany C., Choyke P., Coornud F., Margolis D., Thoeny H., Verma S., Barentsz J., Weinreb J (2019). Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. *European Eurology,* V76(3), pp 340-351