A comparative assessment of Feed-Forward and Convolutional Neural Networks for the classification of prostate lesions

Sabrina Marnell[1[0000-0003-2788-6182]], Patrick Riley[1[0000-0002-4917-5827],
Ivan Olier[1[0000-0002-5679-7501]], Marc Rea[2[0000-0001-9982-1703]],
Sandra Ortega-Martorell[1[0000-0001-9927-3209]]

[1] Department of Applied Mathematics, Liverpool John Moores University, Liverpool, UK, L3 3AF
{S.Marnell@2016.; P.J.Riley@2014.; I.A.OlierCaparroso@; S.OrtegaMartorell@}ljmu.ac.uk
[2] Clatterbridge Cancer Centre NHS Foundation Trust, Wirral, UK, CH63 4JY
m.rea@nhs.net

Abstract. Prostate cancer is the most common cancer in men in the UK. An accurate diagnosis at the earliest stage possible is critical in its treatment. Multiparametric Magnetic Resonance Imaging is gaining popularity in prostate cancer diagnosis, it can be used to actively monitor low-risk patients, and it is convenient due to its non-invasive nature. However, it requires specialist knowledge to review the abundance of available data, which has motivated the use of machine learning techniques to speed up the analysis of these many and complex images. This paper focuses on assessing the capabilities of two neural network approaches to accurately discriminate between three tissue types: significant prostate cancer lesions, non-significant lesions, and healthy tissue. For this, we used data from a previous SPIE ProstateX challenge that included significant and non-significant lesions, and we extended the dataset to include healthy prostate tissue due to clinical interest. Feed-Forward and Convolutional Neural Networks have been used, and their performances were evaluated using 80/20 training/test splits. Several combinations of the data were tested under different conditions and summarised results are presented. Using all available imaging data, a Convolutional Neural Network three-class classifier comparing prostate lesions and healthy tissue attains an Area Under the Curve of 0.892.

Keywords: Feed-Forward Neural Networks, Convolutional Neural Networks, SPIE ProstateX, mpMRI, prostate cancer.

1 Introduction

The tools for the diagnosis of prostate cancer (PCa) are seeing change in recent times. In the UK, trials for diagnosis by Magnetic Resonance Imaging (MRI) scans, as a non-invasive test, are proving to provide benefit in various areas that the current first line of diagnosis (a PSA blood test and biopsy) lack [1]. In addition, active surveillance using multiparametric MRI (mpMRI) has gained popularity as an acceptable management option for low-risk prostate cancer patients, as it can delay or prevent unnecessary interventions - thereby reducing morbidity associated with overtreatment [2]. However, the volumetric analysis of mpMRI scans remains challenging as it requires specialist knowledge and is time consuming. This has motivated the use of machine learning techniques to assist with the analysis of these many and complex images, with the aim of increasing accuracy (especially relevant in places without access to specialist knowledge) and speed up the process (also reducing costs).

This study conducts a comparative analysis with both Feed-Forward Neural Network (FFNN) [3] and Convolutional Neural Network (CNN) [4] architectures as methods for utilising machine learning methodologies for the classification of prostate cancers and healthy tissue on mpMRI scans. It extends on work conducted by the authors previously [5], where the healthy prostate class was added to the SPIE ProstateX challenge dataset [6]. Using advice from collaborating clinicians, the contralateral of the lesion location was taken as healthy prostate tissue, extending from the two classes available in the dataset – clinically significant lesions and non-significant lesions. The latter, "non-significant" prostate lesions do not always require treatment as they hold a lower Gleason score [7].

Previously, various methods were applied to the original SPIE ProstateX challenge problem for classification against the two lesion classes – including transfer learning [8], SVM [9] and convolutional neural networks [10], but not including the healthy class, which is of clinical interest. Only the authors' previous work included the healthy class, in which SVM was used for binary classification and a voting ensemble system was implemented for the diagnosis of individual cases [5]. Hence, the natural next step was to test more sophisticated approaches, which led us to the use of neural networks, and perform a comparative assessment of both FFNNs and CNNs to model classification of prostate lesions against healthy tissue using SPIE ProstateX mpMRI data.

The structure of the rest of the paper is as follows: the Data section details a description of the SPIE ProstateX dataset used in the study, with the Classification Methods section describing the setup of the FFNN and CNN applications. The Results and Discussion section provides insights into the comparative assessment of the two machine learning algorithms to model the diagnosis of prostate cancer through mpMRI.

2 Data

The SPIE ProstateX challenge training data was attained for this study and was extended for clinical use. The data was distributed in DICOM format. Table 1 provides class label information for the dataset used for this study, with a more extensive description available in [5]. The Gleason score (GS) determines the aggressiveness of PCa. At least one lesion was found in every patient. The lesion significance level was stored in the metadata for the respective DICOM files; however, the Gleason Score was not provided.

Various parameters of MRI were provided; Apparent Diffusion Coefficient (ADC), K^{trans} and T2-weighted imaging are used in this paper due to their link in detecting clinical significance [11]. ADC is a measure of the magnitude of diffusion (of water molecules) within tissue and is calculated from diffusion weighted imaging. K^{trans}, a type of perfusion imaging, represents a measure of capillary permeability, calculated from dynamic contrast-enhanced imaging. T2-weighted imaging is a form of spin-echo pulse sequencing, showing fatty tissue and fluid brightly.

The contralateral was taken from 54 patients as described in [5]. This dataset was the training dataset of the challenge, however in this paper it has been used as the sole dataset – used for training and testing; at the time of writing, the challenge dataset test labels have not been released. Different planes were used whilst scanning the patients to create a 3D view around the region of interest (ROI): coronal, sagittal and transverse.

Table 1 – Class label information on the extended SPIE ProstateX dataset used in this study.

Class	Gleason Score	Available (N = 384)	Under-sampled data	Over-sampled data
Clinically Significant	≥7	76	76	228
Non-Significant	≤6	251	75	251
Healthy tissue	N/A	54	54	216

Both under-sampling and over-sampling techniques have been applied to the dataset and compared. Under-sampling is the practice of randomly deleting observations from the larger class, ensuring a good comparative ratio. Synthetic Minority Oversampling Technique (SMOTE) has been utilised for over-sampling, which synthetically manufactures observations of the unbalanced class which share a likeness with the said class; using the k-Nearest Neighbours technique. These methods were tested on the data as provided. Table 1 shows the class sizes for each sampling method, as well as the original class sizes.

3 Classification methods

For creating the FFNN and CNN classifiers, we followed the steps detailed below:

a) Pre-processing, region of interest extraction, vectorisation and standardisation: Various rules were required for pre-processing this large data set; an extensive description, including a description of the contralateral for healthy prostate tissue, is available in [5]. For this study, the FFNNs utilised a 5*5mm centred patch extracted at 1 px/mm around the ROI, while and the CNNs utilised 25*25mm centred patch extracted at 1 px/mm around the ROI, with the input in image format i.e. [25 25 1]. All data was standardised by subtracting the mean and dividing by the standard deviation, ensuring that each dimension was approximately normal.

b) Lesion classification using FFNNs: The data for FFNNs were inputted as flattened vectors. Both binary and three-class classifiers were tackled in this study.

The network architecture for this model FFNN consisted of input layers, hidden layers and output layers, utilising both dense and dropout layers. The dropout layer is used for regularisation. *ReLU* and *Softmax* activation functions were utilised, the loss function was *sparse categorical cross-entropy* and the *Adam* optimiser was utilised.

c) Lesion classification using CNNs: CNNs were used for both binary and three-class classification. The CNN architecture in this work used an array of different layers – two-dimensional convolutional layers, pooling layers, dropout layers and a dense layer. The CNN employed *ReLU* and *Softmax* activation functions, *categorical cross-entropy* loss function and a *stochastic gradient descent* optimizer.

d) Hyperparameter selection and finetuning: Hyperparameter selection has been utilised upon a large selection of combinations of parameters for model training.

e) Table **2** lists the values for hyperparameter selection.

f) Validation: For all models, an 80/20 out-of-bag sampling method for training/test has been utilised.

Table 2 – Hyper-parameter tuning values performed (six for FFNN and twelve for CNN).

	Hyper-parameter	Tested values	Hyper-parameter	Tested values
FFNN	Dense units 1st layer	25, 50, 100, 150, 200, 250	Dropout units 4th layer	0.05, 0.1
	Dropout rate 2nd layer	0.05, 0.1, 0.2	Epochs	25, 50, 100
	Dense units 3rd layer	50, 100, 150	Batch size	32, 64
CNN	Filters 1st layer	16, 32, 64	Filters 5th layer	32, 64
	Kernel size 1st layer	2, 3	Kernel size 5th layer	2, 3
	Filters 2nd layer	16, 32, 64	Dropout rate 6th layer	0.05, 0.1, 0.2, 0.3
	Kernel size 2nd layer	2, 3	Dense units 7th layer	50, 100, 200
	Dropout rate 3rd layer	0.1, 0.2	Dropout rate 8th layer	0.1, 0.2, 0.3
	Filters 4th layer	32, 64	Epochs	15, 25, 50
	Kernel size 4th layer	2, 3	Batch size	32, 64

The total number of models developed with the FFNN architecture was 864, and the total of models with CNN was 994. This allowed us to identify the best set of hyper-parameter values for each architecture and the data at hand. Classification was conducted against the clinical significance label denoted to the prostate lesion, or healthy tissue. Binary classifiers have been tested as well as the three-class classification problem, with the Area Under the Curve (AUC) reported for each test. A McNemar test was used to determine whether the results between the two architectures were statistically significant.

4 Results and Discussion

Various comparative experiments of FFNNs and CNNs were performed. Tests for the binary classifiers are averaged over single modality tests – for example, for the Significant vs. Non-Significant classifier, the classification was conducted for T2-weighted only, for ADC only and for K^{trans} only, and averaged. Results are summarised in Table 3.

Table 3 - Summarised results for the comparative assessment of the FFNN and CNN application to the extended SPIE ProstateX dataset. The binary classifiers were performed on each of the mpMRI scan parameters alone and are then averaged for each separate classifier. S: Significant; N: Non-Significant; H: Healthy.

	Under-sampled data		Over-sampled data	
	FFNN	CNN	FFNN	CNN
Single mpMRI, binary: S vs. N	0.703 ± 0.025	0.619 ± 0.059	0.831 ± 0.023	0.837 ± 0.047
Single mpMRI, binary: S vs. H	0.871 ± 0.033	0.583 ± 0.074	0.954 ± 0.011	0.905 ± 0.054
Single mpMRI, binary: N vs. H	0.645 ± 0.091	0.720 ± 0.048	0.873 ± 0.020	0.960 ± 0.031
All mpMRI, three-classes: S vs. N vs. H	0.628	0.648	0.824	0.892

The first thing to notice from the results is that always the over-sampling strategy outperformed the under-sampling, regardless of the neural network architecture used. This is not surprising as during the under-sampling process we are bound to lose what can turn out to be valuable information from the cases left out, whilst the over-sampling process would benefit from keeping those cases in the dataset. This is even more the case since the size of the dataset is not very large; hence, the reduction of a number of observations may limit the generalisation capabilities of the models since the data used would not be properly representing the population.

Focusing the attention from this point onwards on the results obtained with the over-sampled data for the binary classifiers, using a single mpMRI, we can see that both neural network architectures have produced more competitive results than the previous ones in [5] using SVM, where the AUC for Significant vs. Non-significant was 0.72 (in this study, FFNN: 0.83, CNN: 0.84), Significant vs. Healthy was 0.87 (in this study, FFNN: 0.95, CNN: 0.91), and Non-significant vs. Healthy was 0.71 (in this study, FFNN: 0.87, CNN: 0.96). We compared the obtained results with the ones in [5] since the dataset used is the same. One of the reasons the results are so much improved in the current study can be explained by the use of the SMOTE over-sampling (which was not the case in the previous study), leading to a better informed dataset of which both neural network methods made the most of.

Previous works that looked at the discrimination of the Clinically Significant lesions from the Non-significant ones, reported AUCs of 0.83 in [8], 0.811 in [9] and 0.84 in

[10]. This compares with AUCs of approximately 0.83 (when using FFNN) and 0.84 (when using CNN) in our results for the same discrimination problem, hence we can conclude from here that we have matched the results from these previous study, whilst adding extra value from the inclusion of the Healthy class. In the case of the comparison of the other two binary classifiers (i.e. Significant vs. Healthy and Non-Significant vs. Healthy) with the rest of the literature was not possible since the Healthy class was introduced to this dataset following the request of clinical collaborators.

After being reassured that the results for the Significant vs. Non-Significant problem are in line with the literature, we can focus the attention on the comparison between FFNN and CNN. Starting with the binary classifiers: In the case of the Significant vs. Non-Significant classifier, the McNemar test resulted in a p-value > 0.05 (any p-value > 0.05 is deemed as not statistically significant), hence we conclude that both FFNN and CNN models are equivalent. In the case of the other two binary classifiers, FFNN was better for the Significant vs. Healthy classifier, whilst CNN was better for the Non-significant vs. Healthy. Hence, in the case of these binary classification problems, we can only advice that both architectures should be considered, as each of them will have something to offer. However, if only one architecture was going to be used, and more weight was given to the accurate identification of the clinically significant class using binary classifiers, we would recommend the use of FFNN, since it produced a better outcome in the separation of this class from healthy tissue.

When looking at the three-class classifier that utilises all three mpMRI data available to the study, the CNN performed the best, with an AUC of 0.89. A McNemar test showed that between the FFNN and the CNN over all three classes, the differences in predictions observed are statistically significant (p-value < 0.05). This corroborates with [11], which denotes there is a relationship in clinical significance between the utilised MRI scan parameters in this study – T2-weighted, ADC and K^{trans}. Therefore, we would recommend the use of CNN in the scenario where a three-class classifier is implemented using these MR images for the simultaneous separation of the Significant, Non-significant, and Healthy classes.

5 Conclusions and further work

This study looked at a comparative assessment of FFNNs and CNNs, both in the context of binary and a three-class classifier for the separation of prostate lesions against healthy tissue. The inclusion of the healthy class to a publicly available dataset was motivated by the interest of clinical collaborators. The use of both FFNN and CNN architectures proved successful for both binary and three-class classifiers, leading to clinical impact.

Future work will look at interpreting the CNN features through sensitivity analysis. As opposed to FFNNs, CNN features are by nature sparse. This would provide insights into tissue relevance discrimination.

Acknowledgements. This work has been funded by the LJMU Scholarship Fund.

References

1. Prostate cancer screening scan hope - BBC News
2. A, J.Y., Sidana, A., Choyke, P.L., Wood, B.J., Pinto, P.A., Türkbey, İ.B.: Multiparametric Magnetic Resonance Imaging for Active Surveillance of Prostate Cancer. Balkan Med. J. 34, 388–396 (2017). doi:10.4274/balkanmedj.2017.0708
3. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133 (1943). doi:10.1007/BF02478259
4. Srivastava, N., Hinton, G., Krizhevsky, A., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. (2014)
5. Riley, P., Olier, I., Rea, M., Lisboa, P., Ortega-Martorell, S.: A Voting Ensemble Method to Assist the Diagnosis of Prostate Cancer Using Multiparametric MRI. In: Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization. pp. 294–303. Springer, Cham (2020)
6. Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H.: Computer-Aided Detection of Prostate Cancer in MRI. IEEE Trans. Med. Imaging. 33, 1083–1092 (2014). doi:10.1109/TMI.2014.2303821
7. Gallagher, J.: Prostate cancer treatment "not always needed" - BBC News / Health, https://www.bbc.co.uk/news/health-37362572, (2016)
8. Chen, Q., Xu, X., Hu, S., Li, X., Zou, Q., Li, Y.: A transfer learning approach for classification of clinical significant prostate cancers from mpMRI scans. In: Armato III, S.G. and Petrick, N.A. (eds.) SPIE Medical Imaging 2017: Computer-Aided Diagnosis. p. 101344F. International Society for Optics and Photonics, Orlando (2017)
9. Kitchen, A., Seah, J.: Support vector machines for prostate lesion classification. In: Armato III, S.G. and Petrick, N.A. (eds.) SPIE Medical Imaging 2017: Computer-Aided Diagnosis. p. 1013427. International Society for Optics and Photonics, Orlando (2017)
10. Seah, J.C.Y., Tang, J.S.N., Kitchen, A.: Detection of prostate cancer on multiparametric MRI. In: Armato, S.G. and Petrick, N.A. (eds.) SPIE Medical Imaging 2017: Computer-Aided Diagnosis. p. 1013429. International Society for Optics and Photonics (2017)
11. Langer, D.L., van der Kwast, T.H., Evans, A.J., Plotkin, A., Trachtenberg, J., Wilson, B.C., Haider, M.A.: Prostate Tissue Composition and MR Measurements: Investigating the Relationships between ADC, T2, Ktrans, ve, and Corresponding Histologic Features. Radiology. 255, 485–494 (2010). doi:10.1148/radiol.10091343