Decision fusion of 3D convolutional neural networks to triage patients with suspected prostate cancer using volumetric biparametric MRI

Mehta, Pritesh, Antonelli, Michela, Ahmed, Hashim, Emberton, Mark, Punwani, Shonit, et al.

Pritesh Mehta, Michela Antonelli, Hashim Ahmed, Mark Emberton, Shonit Punwani, Sebastien Ourselin, "Decision fusion of 3D convolutional neural networks to triage patients with suspected prostate cancer using volumetric biparametric MRI," Proc. SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis, 1131433 (16 March 2020); doi: 10.1117/12.2547682

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Decision fusion of 3D convolutional neural networks to triage patients with suspected prostate cancer using volumetric biparametric MRI

Pritesh Mehta[a], Michela Antonelli[b], Hashim Ahmed[c], Mark Emberton[d], Shonit Punwani[e], and Sébastien Ourselin[b]

[a]Medical Physics and Biomedical Engineering, University College London, London, UK
[b]School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK
[c]Department of Surgery and Cancer, Imperial College London, London, UK
[d]Division of Surgery and Interventional Science, University College London, London, UK
[e]Centre for Medical Imaging, University College London, London, UK

ABSTRACT

In this work, we present a computer-aided diagnosis system that uses deep learning and decision fusion to classify patients into one of three classes: "Likely Prostate Cancer", "Equivocal" and "Likely not Prostate Cancer". We impose the group "Equivocal" to reduce misclassifications by allowing for uncertainty, akin to prostate imaging reporting systems used by radiologists. We trained 3D convolutional neural networks to perform two binary patient-level classification tasks: classification of patients with/without prostate cancer and classification of patients with/without clinically significant prostate cancer. Networks were trained separately using volumetric T2-weighted images and apparent diffusion coefficient maps for both tasks. The probabilistic outputs of the resulting four trained networks were combined using majority voting followed by the max operator to classify patients into one of the three classes mentioned above. All networks were trained using patient-level labels only, which is a key advantage of our system since voxel-level tumour annotation is often unavailable due to the time and effort required of a radiologist. Our system was evaluated by retrospective analysis on a previously collected trial dataset. At a higher sensitivity setting, our system achieved 0.97 sensitivity and 0.31 specificity compared to an experienced radiologist who achieved 0.99 sensitivity and 0.12 specificity. At a lower sensitivity setting, our system achieved 0.78 sensitivity and 0.77 specificity compared to 0.76 sensitivity and 0.77 specificity for the experienced radiologist. We envision our system acting as a second reader in pre-biopsy screening applications.

Keywords: prostate cancer, multiparametric MRI, biparametric MRI, computer-aided diagnosis, convolutional neural network, decision fusion

1. INTRODUCTION

Prostate cancer (PCa) is the second most frequently diagnosed cancer in men worldwide and the fifth leading cause of cancer death in men.[1] It is estimated that there were 1.3 million new diagnoses of PCa in 2018 and 359,000 deaths.[1] When diagnosed at its earliest stage, all men with PCa will survive their disease for five years or more, compared with less than a third of men when diagnosed at the latest stage.[2]

Multiparametric magnetic resonance imaging (mpMRI) is increasingly being incorporated into the diagnostic pathway to enable non-invasive cancer detection, targeted biopsy and targeted treatment planning.[3] Whilst imaging protocols vary across centres, the most commonly collected sequences, and those recommended by the revised Prostate Imaging Reporting and Data System[4] (PI-RADS v2) and Likert assessment system,[5] are T2-weighted imaging (T2WI), diffusion-weighted imaging (DWI) and dynamic contrast-enhanced imaging (DCEI). Whilst many differences exist between PI-RADS v2 and Likert, both agree that T2WI and DWI are the dominant sequences for PCa diagnosis, while DCEI plays and ancillary supporting role. A recent systematic review and meta-analysis of twenty studies compared the diagnostic accuracy of mpMRI with biparametric MRI (bpMRI),[6]

Address all correspondence to Pritesh Mehta, Email: pritesh.mehta.17@ucl.ac.uk

Medical Imaging 2020: Computer-Aided Diagnosis, edited by Horst K. Hahn,
Maciej A. Mazurowski, Proc. of SPIE Vol. 11314, 1131433 · © 2020 SPIE
CCC code: 1605-7422/20/$21 · doi: 10.1117/12.2547682

where bpMRI does not include DCEI. It concluded that the diagnostic accuracy of bpMRI is similar to that of mpMRI. This, paired with the costs and risks of contrast agent administration, prompts our use of bpMRI in this work. Reading mpMRI/bpMRI requires considerable expertise, can be prone to inter/intra-observer variability and is a time-consuming task.[3] Computer-aided diagnosis (CAD) systems can enable a more consistent assessment and reduce reading time.

The typical workflow of those CAD systems that do not employ deep learning is described in a review article by Wang et al.[3] The preliminary step is often image intensity normalisation, particularly in the case of MR images, whose intensities may lack fixed meaning due to scanner-dependent variations. In the case of DWI and DCEI, preprocessing may also involve the extraction of quantitative parameters such as the apparent diffusion coefficient (ADC) for DWI or Tofts Model[7] parameters such as influx mass transfer rate of gadolinium (K^{trans}) and reflux rate of gadolinium (k^{ep}) for DCEI. Registration may feature to correct for patient movement during acquisition and differences in voxel resolution and gland segmentation may be performed to confine analysis to the prostate. The next stages involve candidate lesion generation, feature extraction and classification by a machine learning classifier. Candidate lesions can either be hand contoured by radiologists[8–10] or obtained by post-processing probability maps generated by a voxel classification stage.[11–13] Deep learning based CAD systems differ in that the most discriminative features are learned directly from a training set of images rather than selected and extracted from a candidate lesion. The typical tasks performed using deep learning tend to be voxel classification,[14–16] lesion-centred patch classification[17–19] and slice classification.[20–22]

In this work, we perform classification at the patient-level. A CAD system is proposed that uses deep learning and decision fusion to classify patients into three classes: “Likely PCa”, “Equivocal” and “Likely nPCa”, where nPCa indicates a patient absent of PCa. Our system employs 3D residual convolutional neural networks (rCNNs) to perform two different patient-level classification tasks: differentiating patients with PCa from those without PCa (benign or normal gland) and differentiating patients with clinically significant PCa (CSPCa) from those without CSPCa. CSPCa is characterised by the presence of Gleason Score (GS) equal to or greater than 7 during histological analysis. For both tasks, we trained rCNNs independently using volumetric T2WI and ADC maps as input. The probabilities output by the four trained rCNNs are combined using a two-level scheme. In the first level, the majority voting operator is separately applied to the outputs of the two rCNNs trained on the binary PCa classification task and to the outputs of the two rCNNs trained on the binary CSPCa classification task. In the second level, we combine the outputs of the majority voting using the max operator to produce a unified classification into one of the three classes mentioned earlier. Our system was evaluated by retrospective analysis on a previously collected trial dataset.[23]

Allowing for uncertainty akin to prostate imaging reporting systems used by radiologists, through the “Equivocal” class, is a key advantage of our system. In our case, patients are allocated to this class when contradictory evidence is presented. A further advantage is that our system is trained using patient-level labels only. The ability to train using patient-level labels is advantageous as voxel-level annotation is often unavailable due to the time and effort required of a radiologist.

2. MATERIALS

Our Institutional Review Board approved the study and waived the requirement for individual consent for retrospective analysis of prospectively acquired patient data collected as part of clinical trials/routine care (R&D No: 12/0195, 16 July 2012).

Full details of the trial have been previously reported.[24] Men were included in the trial if they had undergone previous transrectal ultrasound (TRUS) biopsy and were suitable for further characterisation using transperineal template prostate mapping (TTPM) biopsy. Men with a previous history of treatment were excluded. 330 men enrolled in the trial, and following 81 withdrawals, 249 men completed mpMRI, TTPM biopsy and targeted biopsy.

MpMRI was acquired using a 3 Tesla magnetic field scanner (Achieva, Philips Healthcare) and a pelvic-phased array coil. Sequences collected included T2WI, DWI with high b-value (2000), ADC map computed from DWI at multiple b-values (0, 150, 500, 1000) and DCEI with gadolinium. A detailed description of acquisition parameters can be found in the protocol publication for the trial.[24] All mpMRI studies were reported on by a radiologist with over 10 years of experience in reading prostate mpMRI. A 5-point scoring system (see table 1) was used to score at the lesion level and patient level. Three distinct definitions of cancer aggressiveness were scored against during the trial. We consider the definition in which the presence of low grade GS 3+3 and higher GS, warrant scores 4 or 5.

Table 1: 5-point MRI scoring system utilised to indicate presence of disease.[24]

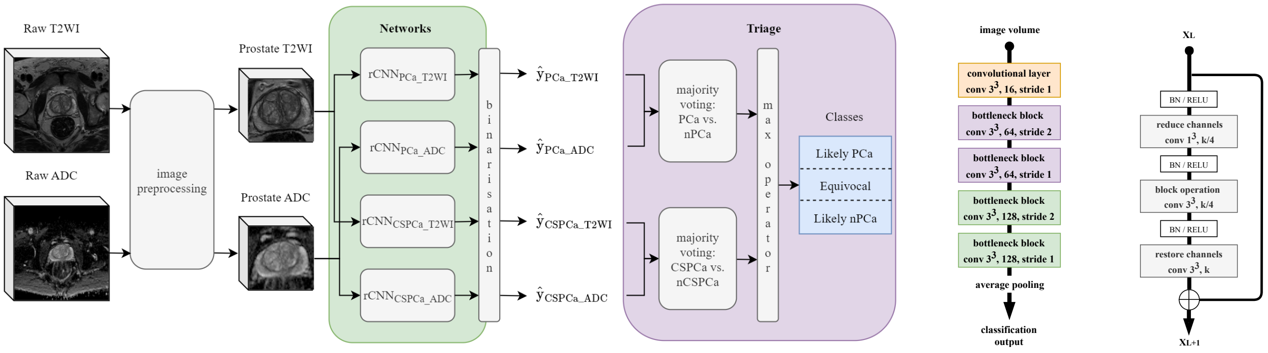
1	Highly likely benign
2	Likely benign
3	Equivocal
4	Likely malignant
5	Highly likely malignant

Ultrasound-guided TTPM and targeted biopsy acted as the reference standard. The whole gland was sampled through a brachytherapy template-grid placed on the perineum using a 5-mm sampling frame. Focal index lesions underwent cognitive MRI-targeted biopsies at the time of TTPM. One of two expert uropathologists each with over 20 years experience analysed all biopsy cores blinded to the MRI results and negative biopsies were double-reported for quality control. Of the 249 patients who completed the trial, 71 patients DWI was distorted by magnetic susceptibility artefacts caused by air in the rectum. These patients are excluded from evaluation in this work. Histological analysis found (post patient exclusion shown in brackets): 34(27) patients with no cancer, 61(38) patients with max GS 3+3, 114(83) patients with max GS 3+4, 34(25) patients with max GS 4+3 and 6(5) patients with max GS greater than 8.

In addition to the data described above, a publicly available dataset released for the PROSTATEx challenge, was used to augment the data available to train our CAD system,[11] but not evaluated upon due to the lack of available radiologist interpretation to compare to.

3. METHOD

Given bpMRI of a patient, the goal is to classify that patient into one of three classes ("Likely PCa", "Equivocal", "Likely nPCa"). In Figure 1a, we present the framework of the proposed CAD system. Each component of the system is described in the subsections to follow.

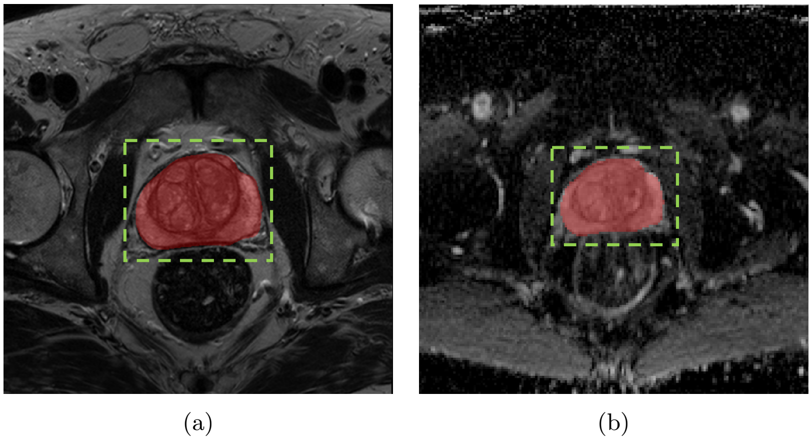

Raw T2WI
Raw ADC
image preprocessing
Prostate T2WI
Prostate ADC
Networks
$\mathrm{rCNN}_{\mathrm{PCa\_T2WI}}$
$\mathrm{rCNN}_{\mathrm{PCa\_ADC}}$
$\mathrm{rCNN}_{\mathrm{CSPCa\_T2WI}}$
$\mathrm{rCNN}_{\mathrm{CSPCa\_ADC}}$
binarisation
$\hat{y}_{\mathrm{PCa\_T2WI}}$
$\hat{y}_{\mathrm{PCa\_ADC}}$
$\hat{y}_{\mathrm{CSPCa\_T2WI}}$
$\hat{y}_{\mathrm{CSPCa\_ADC}}$
Triage
majority voting: PCa vs. nPCa
majority voting: CSPCa vs. nCSPCa
max operator
Classes
Likely PCa
Equivocal
Likely nPCa
image volume
convolutional layer conv 3^3, 16, stride 1
bottleneck block conv 3^3, 64, stride 2
bottleneck block conv 3^3, 64, stride 1
bottleneck block conv 3^3, 128, stride 2
bottleneck block conv 3^3, 128, stride 1
average pooling
classification output
XL
BN / RELU
reduce channels conv 1^3, k/4
BN / RELU
block operation conv 3^3, k/4
BN / RELU
restore channels conv 3^3, k
XL+1

(a) (b) (c)

Figure 1: (a) Proposed framework of our CAD system, where nPCa indicates normal prostate or prostate with benign condition and nCSPCa indicates normal prostate or prostate with benign condition or prostate with indolent PCa. (b) 3D rCNN architecture used to perform patient-level classification tasks. (c) A bottleneck block, where $k = \#$kernels.

3.1 Image preprocessing

A histogram-based standardisation method[25] was applied to each patient's T2WI to give image intensities tissue meaning, absent on acquisition due to scanner and patient dependent variations. We trained a CNN (HighRes3DNet[26]) to segment the prostate from background structures in all T2WI using a network implementation available on NiftyNet,[27] an open-source platform for deep learning in medical imaging. The binary segmentation mask output was used to crop the prostate on T2WI. The cropping operation creates a simpler classification task unsullied by background information. ADC maps were cropped by transforming the segmentation from T2WI space into DWI space using affine registration[28] followed by a cubic b-spline non-rigid registration.[29]


(a)

(b)

Figure 2: (a) Segmentation of T2WI generated using CNN (HighRes3DNet), with cropping area shown by green dashed border. (b) Segmentation of ADC map in DWI space, generated using registration derived transformation of T2WI segmentation, with cropping area shown by green dashed border.

3.2 Networks

Each 3D rCNN is composed of a convolutional layer followed by four bottleneck blocks.[30] A network diagram is shown in Figure 1b. Bottleneck blocks reduce the computational load of 3D convolutional layers by performing a channel reduction and restoration operation either side of the core convolution operation, as shown in figure 1c. Preactivation[31] (batch normalisation and rectified linear unit activation prior to weight layer computation) is used, as shown in figure 1c, to ease optimisation and regularise the networks. The number of blocks/layers and all other hyperparameters were determined through hyperparameter search using a 5-fold cross validation. Hyperparameters selected using validation sets were used to evaluate over inference sets. For rCNN training, we used the cross-entropy loss, Adam optimisation,[32] learning rate equal to 0.0001 and a batch size of 8. We employed flip and random deformation augmentations during training to balance classes, reduce overfitting and increase generalisability. We implemented the network architecture using NiftyNet.

3.3 Triage

We converted the continuous valued network output of each rCNN into a binary categorical variable by selecting a specificity for each network, from which a cut-off score between classes was derived. The binarised rCNN outputs are combined using a two-level scheme as shown in figure 1a. Let $m_1, \ldots, m_k$ be k available modalities, with k equal to some positive even-valued integer. The total number of trained rCNNs is equal to $2k$:

$$\begin{aligned} \text{rCNN}_{\text{PCa\_M}}, &\quad M = m_1, ..., m_k, \\ \text{rCNN}_{\text{CSPCa\_M}}, &\quad M = m_1, ..., m_k, \end{aligned}$$

where each $\text{rCNN}_{\text{PCa\_M}}$ is trained on task 1 (PCa vs. nPCa) and each $\text{rCNN}_{\text{CSPCa\_M}}$ is trained on task 2 (CSPCa vs. nCSPCa). Each $\text{rCNN}_{\text{PCa\_M}}$ gives a probabilistic output, which is transformed into a binary variable $\hat{y}_{PCa\_M}$. Similarly, each $\text{rCNN}_{\text{CSPCa\_M}}$ gives a probabilistic output, which is transformed into a binary variable $\hat{y}_{CSPCa\_M}$. The set of $\{\hat{y}_{PCa\_M} | M = m_1, \ldots, m_k\}$ take value 1 representing the class PCa or value 0 representing the class nPCa. The set of $\{\hat{y}_{CSPCa\_M}|M = m_1, \ldots, m_k\}$ take value 1 representing the class CSPCa or value 0 representing the class nCSPCa. Majority voting (MV) is applied for each task separately to combine the classifications of all k imaging modalities:

$$c_{PCa} = MV(\{\hat{y}_{PCa\_M}|M = m_1, \ldots, m_k\}) = \begin{cases} \text{Likely PCa}, & \sum_k \hat{y}_{PCa\_m_k} > k/2, \\ \text{Equivocal}, & \sum_k \hat{y}_{PCa\_m_k} = k/2, \\ \text{Likely nPCa}, & \sum_k \hat{y}_{PCa\_m_k} < k/2. \end{cases}$$

$$c_{CSPCa} = MV(\{\hat{y}_{CSPCa\_M}|M = m_1, \ldots, m_k\}) = \begin{cases} \text{Likely CSPCa}, & \sum_k \hat{y}_{CSPCa\_m_k} > k/2, \\ \text{Equivocal}, & \sum_k \hat{y}_{CSPCa\_m_k} = k/2, \\ \text{Likely nCSPCa}, & \sum_k \hat{y}_{CSPCa\_m_k} < k/2. \end{cases}$$

The three categories are ranked from 1 to 3 such that the positive classes (Likely PCa, Likely CSPCa) have the highest rank and the negative classes (Likely nPCa, Likely nCSPCa), the lowest. Considering the ranks $r_1 = \text{rank}(c_{PCa})$ and $r_2 = \text{rank}(c_{CSPCa})$, we apply the max operator to r_{PCa} and r_{CSPCa} to obtain the final class C:

$$C = \begin{cases} \text{Likely PCa}, & \max(r_{PCa}, r_{CSPCa}) = 3, \\ \text{Equivocal}, & \max(r_{PCa}, r_{CSPCa}) = 2, \\ \text{Likely nPCa}, & \max(r_{PCa}, r_{CSPCa}) = 1. \end{cases}$$

4. RESULTS

The whole gland segmentation performance of HighRes3DNet was evaluated by 10-fold cross validation on a subset of 82 patients from the trial dataset, where manual annotations performed by a radiologist were available as ground truth. The dice coefficient was used as an evaluation metric over the inference sets. Transformation of the segmentation from T2WI space to DWI space, derived using registration, was evaluated by visual inspection.

Each rCNN was evaluated by 5-fold cross validation using the receiver operating characteristic (ROC) curve, area under the ROC curve (AUC) and the point on the ROC curve characterised by a specificity of 0.50 ($SN_{SP=0.50}$).

The decision fusion CAD system output was evaluated using the sensitivity and specificity at two thresholds, (i) where both the "Equivocal" group and "Likely PCa" group indicate that the patient has cancer and (ii) where the "Likely PCa" group alone indicates that the patient has cancer. We present an equivalent analysis for the reporting radiologist for comparison.

4.1 Segmentation performance

A mean dice coefficient ± 1 standard deviation over 10-folds of 0.90 ± 0.03 was obtained. Following the cross validation exercise, a single HighRes3DNet was trained to segment all remaining T2WI in the trial dataset for which no radiologist annotation was available. Through visual inspection, we observed all segmentations obtained were of sufficient quality such that no manual correction was required.

ADC map segmentations were obtained by transforming T2WI segmentations using a registration based approach as described in subsection 3.1. Visual inspection showed segmentations were of sufficient quality such that no manual correction was required.

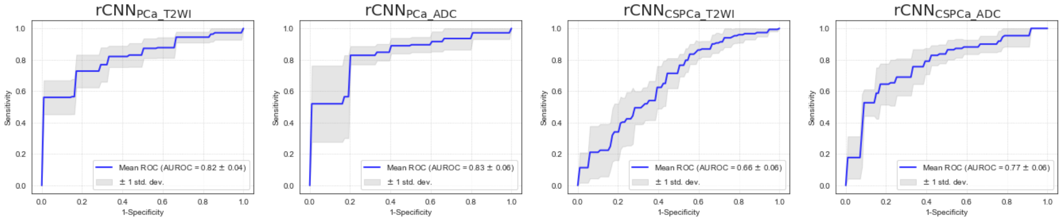
4.2 rCNN performance

Table 2 shows the evaluation metrics AUC and $SN_{SP=0.50}$ for PCa and CSPCa binary classification tasks, using T2WI and ADC maps as input. The corresponding ROC curves are shown in figure 3. The values presented in Table 2 and the ROC curves shown in Figure 3 are an average over 5-folds.

We observe a greater ability to discriminate between PCa and nPCa than between CSPCa and nCSPCa. We also observe that ADC maps allow better discrimination in both tasks.

Table 2: Classification performance of the individual rCNNs for each task/modality, averaged over 5-fold cross validation.

Task	Network	AUC	$SN_{SP=0.50}$
PCa vs. nPCa	$\text{rCNN}_{\text{PCa\_T2WI}}$	0.82 ± 0.04	0.87 ± 0.07
	$\text{rCNN}_{\text{PCa\_ADC}}$	0.83 ± 0.07	0.90 ± 0.05
CSPCa vs. nCSPCa	$\text{rCNN}_{\text{CSPCa\_T2WI}}$	0.66 ± 0.07	0.77 ± 0.08
	$\text{rCNN}_{\text{CSPCa\_ADC}}$	0.77 ± 0.07	0.86 ± 0.07



(a) (b) (c) (d)

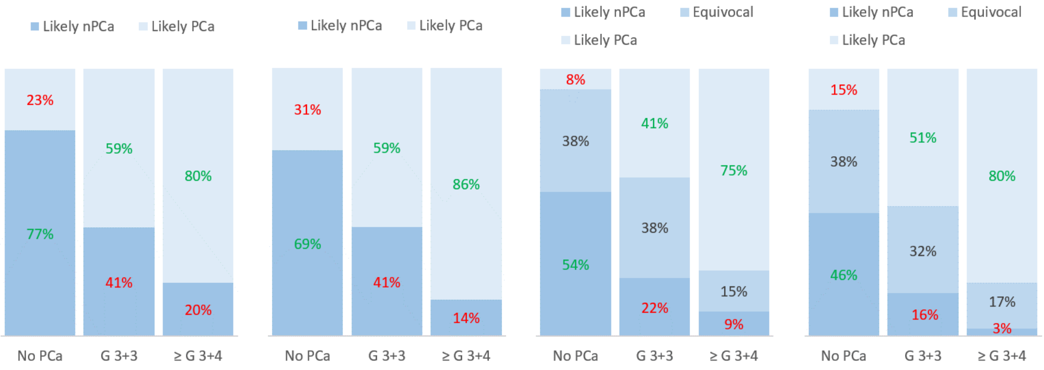
Figure 3: Classification ROC curves of the individual rCNNs for each task/modality, averaged over 5-fold cross validation.

4.3 Triage performance (decision fusion)

In subsection 4.3.1, the impact of the majority voting and max operators is shown, while in subsection 4.3.2 we present a comparison of the output of our system with the scoring of an experienced radiologist.

4.3.1 Impact of the majority voting and max operators

As shown in subsection 3.3, the majority voting operator observes whether T2WI and ADC map offer a consensus decision or not, in the latter case allocating to the "Equivocal" class. The max operator increases the sensitivity of the system to patients with clinically significant cancer. Figure 4 shows the impact of the majority voting operator and max operator.



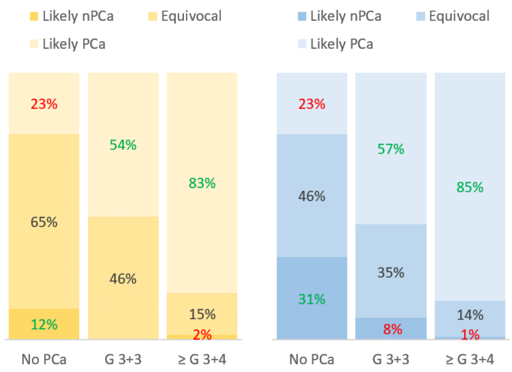
(a) (b) (c) (d)

Figure 4: (a) Classification performance of $\text{rCNN}_{\text{PCa\_T2WI}}$ only, by histological status. (b) Classification performance of $\text{rCNN}_{\text{PCa\_ADC}}$ only, by histological status. (c) Majority voting combination of $\text{rCNN}_{\text{PCa\_T2WI}}$ and $\text{rCNN}_{\text{PCa\_ADC}}$. (d) Majority voting and max operator combination of $\text{rCNN}_{\text{PCa\_T2WI}}$, $\text{rCNN}_{\text{PCa\_ADC}}$, $\text{rCNN}_{\text{CSPCa\_T2WI}}$, $\text{rCNN}_{\text{CSPCa\_ADC}}$. Red font indicates incorrect group allocation. Patients are grouped by their histological grading.

Figures 4a and 4b show the performance of $\text{rCNN}_{\text{PCa\_T2WI}}$ and $\text{rCNN}_{\text{PCa\_ADC}}$ prior to decision fusion. Figure 4c shows the result of combining the binarised outputs of $\text{rCNN}_{\text{PCa\_T2WI}}$ and $\text{rCNN}_{\text{PCa\_ADC}}$ using the majority voting operator. Most critically, we observe that false negative rate (FNR) of patients with GS $\geq 3 + 4$ drops to 9% and false positive rate (FPR) of patients with Normal/Benign gland drops to 8%, both due to the creation of the "Equivocal" class. It should be noted that introducing the "Equivocal" class also causes a drop in true positive rate (TPR) and true negative rate (TNR), but as is the case with PCa imaging reporting systems, this is tolerated to reduce costly misclassifications. Figure 4d shows the impact of the max operator. The idea of the max operator is to upgrade the class of a patient if the rCNNs trained on the CSPCa vs. nCSPCa task present evidence that the patient may have CSPCa. We observe a drop in FNR of patients with GS $\geq 3+4$ from 9% to 3% without a proportionally large drop in TNR.

4.3.2 Comparison to radiologist

A radiologist of greater than 10 years experience in reading prostate MR scored each patient using a 1-5 scale for the presence of GS $\geq 3+3$ PCa. By combining the groups 1 and 2 and the groups 4 and 5 shown in Table 1, we can directly compare the radiologists performance with the performance of our system.



(a) (b)

Figure 5: Figure (a) shows the classification performance of an experienced radiologist. Figure (b) shows the performance of our system set to match the FNR on GS $\geq 3+4$ patients obtained by the radiologist (2%). Red font indicates incorrect group allocation.

Figure 5 shows the performance of the radiologist and our system. By selecting high sensitivity cut-off scores for rCNNs, we match the radiologists FNR on patients with GS $\geq 3+4$ (2%). We observe a superior TPR (85%) and TNR (31%), though we also observe a higher incidence of misclassified indolent cancer patients (8%). Table 3 converts Figure 5 into sensitivities and specificities by considering two thresholds, (i) where the "Equivocal" class and "Likely PCa" class indicates that the patient has cancer and (ii) where the "Likely PCa" group alone indicates that the patient has cancer.

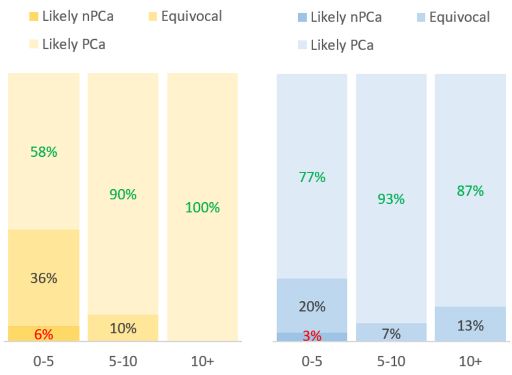
Table 3: Triage output sensitivity and specificity at two thresholds for the experienced radiologist and our system.

	Threshold	SN (GS $\geq 3+3$)	SN (GS $\geq 3+4$)	SP
Radiologist	Equivocal	0.99	0.98	0.12
	Likely PCa	0.76	0.83	0.77
System	Equivocal	0.97	0.99	0.31
	Likely PCa	0.78	0.85	0.77

The most important goal of PCa diagnosis is to correctly identify clinically significant cancers. In figure 6, we analyse why a small proportion of patients with clinically significant cancer have been classified as "Equivocal" or "Likely nPCa" by both the radiologist and our system. Figure 6 shows the majority of misclassifications are on small tumors with biopsy max core length from 0-5mm.

5. DISCUSSION

We have proposed a CAD system that applies decision fusion to the output of multiple trained CNNs to classify patients with suspected PCa into three classes "Likely PCa", "Equivocal" and "Likely nPCa". Critically, we obtain satisfactory binary classification performance from our trained rCNNs. A significant reason for this was the inclusion of the publicly available ProstateX dataset for training. Of most importance was increasing the quantity of benign patients, which the trial dataset lacked. Moreover, we harmonised T2WI in both datasets using histogram standardisation as described in subsection 3.1. We further increased the size of the datasets for



(a) (b)

Figure 6: Breakdown of (a) radiologist classification and (b) our system classification by max tumour core length (mm) for patients with clinically significant cancer. Red font indicates incorrect group allocation.

training by using flip and random deformation augmentations. Our performance may also be explained by the use of volumetric input processed by volumetric convolutions, giving 3D context. 3D context can be important in lesion characterisation. Furthermore, we cropped image volumes forcing features to be learned from the region of interest and not from background. Finally, we conducted an extensive hyperparameter search; in particular, we tuned the number of network parameters, batch size and learning rate to limit overfitting and underfitting and ensure stable weight updates.

As described in subsection 3.3, the binary outputs of trained rCNNs are transformed into a three class classification through majority voting and max operators. It should be noted that a three class classification could be achieved by training networks on the first task (PCa vs. nPCa) only and using the majority voting operator only. However, including the second task (CSPCa vs. nCSPCa) increases our sensitivity to clinically significant cancers through the max operator. This relies on the observation that some clinically significant cancers are misclassified by at least one of the networks trained on the first task, but are correctly classified by at least one of the networks trained on the second task, which effectively upgrades these patients through the max operator.

The CAD system described in this work recognises and addresses some of the common problems associated with developing CAD systems for PCa diagnosis. A considerable advantage of our system is the lack of need for voxel-level lesion annotations. These can be difficult to obtain for multimodal 3D datasets due to the time and effort required of radiologists. Removing the need for manually annotated datasets allows the use of much larger datasets, limited only by the availability of associated ground truth biopsy or prostatectomy.

A current limitation of our system is the inability to localise tumors. While localisation is desirable, it is not strictly necessary. Many systems in the literature output tumor locations to help less experienced radiologists identify tumours or to provide locations for targeted biopsy.[33,34] Our system as currently constructed aims to provide a second opinion in screening applications. However, attention mechanisms/saliency maps/relevance propagation can be explored as future work to introduce explainability.

6. CONCLUSION

We have proposed a CAD system that uses decision fusion of multiple CNNs to classify patients with suspected PCa into three classes. A particularly novel aspect is the introduction of an "Equivocal" class that can reduce misclassifications by allowing uncertainty. We obtained performance arguably superior to an experienced radiologist. Therefore, following further evaluation on other datasets, it is feasible that our system be placed alongside a radiologist as a second reader in screening applications.

ACKNOWLEDGMENTS

EPSRC-funded Centre for Doctoral Training in Intelligent, Integrated Imaging In Healthcare (EP/L016478/1)

REFERENCES

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A., "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians* **68**, 394–424 (2018).

[2] Cancer Research UK, "Prostate Cancer Statistics." https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer. (Accessed: 18 November 2018).

[3] Wang, S., Burtt, K., Turkbey, B., Choyke, P., and Summers, R., "Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research," *BioMed Research International* **2014** (2014).

[4] American College of Radiology, [*PI-RADS version 2*] (2015).

[5] Brizmohun Appayya, M., Adshead, J., Ahmed, H. U., Allen, C., Bainbridge, A., Barrett, T., Giganti, F., Graham, J., Haslam, P., Johnston, E. W., Kastner, C., Kirkham, A. P., Lipton, A., McNeill, A., Moniz, L., Moore, C. M., Nabi, G., Padhani, A. R., Parker, C., Patel, A., Pursey, J., Richenberg, J., Staffurth, J., van der Meulen, J., Walls, D., and Punwani, S., "National implementation of multi-parametric magnetic resonance imaging for prostate cancer detection – recommendations from a UK consensus meeting," *BJU International* **122**, 13–25 (2018).

[6] Woo, S., Suh, C. H., Kim, S. Y., Cho, J. Y., Kim, S. H., and Moon, M. H., "Head-to-head comparison between biparametric and multiparametric MRI for the diagnosis of prostate cancer: A systematic review and meta-analysis," *American Journal of Roentgenology* **211**, 226–241 (2018).

[7] Tofts, P. S., "Modeling Tracer Kinetics in Dynamic Gd-DTPA MR Imaging," *Journal of Magnetic Resonance Imaging* **7**, 91–101 (1997).

[8] Shah, V., Turkbey, B., Mani, H., Pang, Y., Pohida, T., Merino, M., Pinto, P., Choyke, P., and Bernardo, M., "Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging," *Medical Physics* **39**, 4093–4103 (2012).

[9] Niaf, E., Flamary, R., Rouviere, O., Lartizien, C., and Canu, S., "Kernel-based learning from both qualitative and quantitative labels: application to prostate cancer diagnosis based on multiparametric MR imaging," *IEEE Transactions of Image Processing* **23**, 979–991 (2014).

[10] Antonelli, M., Johnston, E. W., Dikaios, N., Cheung, K. K., Sidhu, H. S., Appayya, M. B., Giganti, F., Simmons, L. A. M., Freeman, A., Allen, C., Ahmed, H. U., Atkinson, D., Ourselin, S., and Punwani, S., "Machine learning classifiers can predict Gleason pattern 4 prostate cancer with greater accuracy than experienced radiologists," *European Radiology* **29**, 4754–4764 (2019).

[11] Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., and Huisman, H., "Computer-Aided Detection of Prostate Cancer in MRI," *IEEE Transactions on Medical Imaging* **33**, 1083–1092 (2014).

[12] Giannini, V., Mazzetti, S., Vignati, A., Russo, F., Bollito, E., Porpiglia, F., Stasi, M., and Regge, D., "A fully automatic computer aided diagnosis system for peripheral zone prostate cancer detection using multi-parametric magnetic resonance imaging," *Computerized Medical Imaging and Graphics* **46**, 219–226 (2015).

[13] Lay, N., Tsehay, Y., Greer, M. D., Turkbey, B., Kwak, J. T., Choyke, P. L., Pinto, P., Wood, B. J., and Summers, R. M., "Detection of prostate cancer in multiparametric MRI using random forest with instance weighting," *Journal of Medical Imaging* **4** (2017).

[14] Feldman, A., Dai, Z., Carver, E., Liu, C., Lee, J., Pantelic, M., Elshaikh, M., and Wen, N., "Utilizing a Deep Learning-Based Object Detection and Instance Segmentation Algorithm for the Delineation of Prostate and Prostate Cancer Segmentation," *International Journal of Radiation Oncology Biology Physics* **105**, 197–198 (2019).

[15] Ishioka, J., Matsuoka, Y., Uehara, S., Yasuda, Y., Kijima, T., Yoshida, S., Yokoyama, M., Saito, K., Kihara, K., Numao, N., Kimura, T., Kudo, K., Kumazawa, I., and Fujii, Y., "Computer-aided diagnosis of prostate cancer on magnetic resonance imaging using a convolutional neural network algorithm," *BJU International* **122**, 411–417 (2018).

[16] Sumathipala, Y., Lay, N., Turkbey, B., Smith, C., Choyke, P. L., and Summers, R. M., "Prostate cancer detection from multi- institution multiparametric MRIs using deep convolutional neural networks," *Journal of Medical Imaging* **5** (2018).

[17] Aldoj, N., Lukas, S., Dewey, M., and Penzkofer, T., "Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network," *European Radiology.* (2019).

[18] Song, Y., Zhang, Y. D., Yan, X., Liu, H., Zhou, M., Hu, B., and Yang, G., "Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI," *Journal of Magnetic Resonance Imaging* **48**, 1570–1577 (2018).

[19] Yuan, Y., Qin, W., Buyyounouski, M., Ibragimov, B., Hancock, S., Han, B., and Xing, L., "Prostate cancer classification with multiparametric MRI transfer learning model," *Medical Physics* **46**, 756–765 (2019).

[20] Wang, X., Yang, W., Weinreb, J., Han, J., Li, Q., and Kong, X., "Searching for prostate cancer by fully automated magnetic resonance imaging classification : deep learning versus non-deep learning," *Scientific Reports* **7**, 1–8 (2017).

[21] Wang, Z., Liu, C., Cheng, D., Wang, L., Yang, X., and 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**, 1127–1139 (2018).

[22] Yang, X., Liu, C., Wang, Z., Yang, J., Min, H. L., Wang, L., and 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).

[23] Simmons, L. A., Kanthabalan, A., Arya, M., Briggs, T., Barratt, D., Charman, S. C., Freeman, A., Gelister, J., Hawkes, D., Hu, Y., Jameson, C., McCartan, N., Moore, C. M., Punwani, S., Ramachandran, N., Van Der Meulen, J., Emberton, M., and Ahmed, H. U., "The PICTURE study: Diagnostic accuracy of multiparametric MRI in men requiring a repeat prostate biopsy," *British Journal of Cancer* **116**, 1159–1165 (2017).

[24] Simmons, L. A., Ahmed, H. U., Moore, C. M., Punwani, S., Freeman, A., Hu, Y., Barratt, D., Charman, S. C., Van der Meulen, J., and Emberton, M., "The PICTURE study - prostate imaging (multi-parametric MRI and Prostate HistoScanning$^{\text{TM}}$) compared to transperineal ultrasound guided biopsy for significant prostate cancer risk evaluation," *Contemporary Clinical Trials* **37**, 69–83 (2014).

[25] Nyúl, L. G., Udupa, J. K., and Zhang, X., "New variants of a method of MRI scale standardization," *IEEE Transactions on Medical Imaging* **19**, 143–150 (2000).

[26] Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M. J., and Vercauteren, T., "On the compactness, efficiency, and representation of 3d convolutional networks: Brain parcellation as a pretext task," *CoRR* **abs/1707.01992** (2017).

[27] Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., Whyntie, T., Nachev, P., Modat, M., Barratt, D. C., Ourselin, S., Cardoso, M. J., and Vercauteren, T., "NiftyNet: a deep-learning platform for medical imaging," *Computer Methods and Programs in Biomedicine* **158**, 113–122 (2018).

[28] Modat, M., Cash, D. M., Daga, P., Winston, G. P., Duncan, J. S., and Ourselin, S., "Global image registration using a symmetric block-matching approach," *Journal of Medical Imaging* **1** (2014).

[29] Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., Fox, N. C., and Ourselin, S., "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine* **98**, 278–284 (2010).

[30] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," *CoRR* **abs/1512.03385** (2015).

[31] He, K., Zhang, X., Ren, S., and Sun, J., "Identity mappings in deep residual networks," *CoRR* **abs/1603.05027** (2016).

[32] Kingma, D. P. and Ba, J., "Adam: A Method for Stochastic Optimization," in [*International Conference on Learning Representations*], (2015).

[33] Campa, R., Del Monte, M., Barchetti, G., Pecoraro, M., Salvo, V., Ceravolo, I., Indino, E. L., Ciardi, A., Catalano, C., and Panebianco, V., "Improvement of prostate cancer detection combining a computer-aided diagnostic system with TRUS-MRI targeted biopsy," *Abdominal Radiology* **44**, 264–271 (2019).

[34] Greer, M. D., Lay, N., Shih, J. H., Barrett, T., Bittencourt, L. K., Borofsky, S., Kabakus, I., Law, Y. M., Marko, J., Shebel, H., Mertan, F. V., Merino, M. J., Wood, B. J., Pinto, P. A., Summers, R. M., Choyke, P. L., and Turkbey, B., "Computer-aided diagnosis prior to conventional interpretation of prostate mpMRI: an international multi-reader study," *European Radiology* **28**, 4407–4417 (2018).