ORIGINAL PAPER

A Comprehensive Study of Data Augmentation Strategies for Prostate Cancer Detection in Diffusion-Weighted MRI Using Convolutional Neural Networks

Ruqian Hao[1,2,3] · Khashayar Namdar[3] · Lin Liu[1] · Masoom A. Haider[4,5,6] · Farzad Khalvati[2,3,7]

Abstract

Data augmentation refers to a group of techniques whose goal is to battle limited amount of available data to improve model generalization and push sample distribution toward the true distribution. While different augmentation strategies and their combinations have been investigated for various computer vision tasks in the context of deep learning, a specific work in the domain of medical imaging is rare and to the best of our knowledge, there has been no dedicated work on exploring the effects of various augmentation methods on the performance of deep learning models in prostate cancer detection. In this work, we have statically applied five most frequently used augmentation techniques (random rotation, horizontal flip, vertical flip, random crop, and translation) to prostate diffusion-weighted magnetic resonance imaging training dataset of 217 patients separately and evaluated the effect of each method on the accuracy of prostate cancer detection. The augmentation algorithms were applied independently to each data channel and a shallow as well as a deep convolutional neural network (CNN) was trained on the five augmented sets separately. We used area under receiver operating characteristic (ROC) curve (AUC) to evaluate the performance of the trained CNNs on a separate test set of 95 patients, using a validation set of 102 patients for finetuning. The shallow network outperformed the deep network with the best 2D slice-based AUC of 0.85 obtained by the rotation method.

Introduction

Prostate cancer is the second most common type of cancer among men. Traditional screening methods such as Prostate Specific Antigen test and Digital Rectal Examination are usually at a high risk of low accuracy and overdiagnosis [1]. Wide application of diffusion-weighted magnetic resonance imaging (DW-MRI) for prostate cancer detection has resulted in a higher sensitivity of predictive models. However, specificity is still low which is caused by dependence of examination results to experience and preference of the radiologist. Unnecessary biopsies are a major consequence of the ongoing practice [2]. The recent advent of convolutional neural networks (CNNs) has resulted in more focus on using machine learning (ML) for classification and diagnosis of prostate cancer. Deep learning algorithms which rely on CNNs with more layers require a large amount of data. Nonetheless, it is difficult to get access to significant amounts of prostate cancer data due to the limited available

✉ Farzad Khalvati
farzad.khalvati@utoronto.ca

1 School of Optoelectronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

2 Institute of Medical Science, University of Toronto, Toronto, ON, Canada

3 Department of Diagnostic Imaging, The Hospital for Sick Children (SickKids), University of Toronto, 555 University Avenue, Toronto, ON, Canada

4 Joint Department of Medical Imaging, Sinai Health System, University Health Network, University of Toronto, Toronto, ON, Canada

5 Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada

6 Sunnybrook Research Institute, Toronto, ON, Canada

7 Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada

data and existing regulations and protocols of medical images [3].

In order to accumulate enough data and improve the performance of deep learning models for medical imaging classification tasks, multiple data augmentation strategies have been proposed. Cao et al. performed intensity normalization and basic augmentation methods (translate, scale, and flip) on a prostate multi-parametric MRI (mp-MRI) dataset and then trained a CNN to detect prostate lesions. As a result, sensitivity improved by 9.8% [4]. Campanella et al. used a prostate core biopsy dataset consisting of 24,859 slides to classify prostate cancer. A ResNet34 model was trained without and with augmentation on the fly which consisted of 90° rotations, horizontal flips, and color jitter. Nonetheless, the experiment results showed that data augmentation was not effective on the large-scale dataset. The highest balanced accuracy, which is the accuracy scaled by size of each individual class, on the model trained without augmentation was 0.95% higher than that with augmentation [5].

Using data augmentation methods in the field of medical imaging is not quite novel; however, most studies have not quantified the effects of these techniques. Ishioka et al. utilized a series of transformations which included cropping prostate area of 261×261 pixels randomly and shaping them into parallelogram, and then generated a total of 2 million prostate MR images training dataset to train CNNs for prostate cancer detection. The experiments results showed that the best validation results for the area under the receiver operating characteristic (ROC) curve (AUC) increased from below 0.4 to 0.793 as the number of augmented training images increased from 0 to 2 million, and the AUC value of this model for the test dataset was 0.636 [6]. Wang et al. cropped each training T2-weighted prostate image of 360×360 into multiple subimages of 288×288 pixels randomly. A deep convolutional neural network (DCNN) was trained for the automatic classification of prostate cancer, and the AUC result was 0.84 [7]. Liu et al. sliced 3D multiparametric MRI data at 7 different orientations and performed in-plane rotation, random shearing, and one pixel translation of the lesions for each slice. At the end, 207,144 training samples were prepared. When training their proposed XmasNet, random mirroring was also performed on the fly. The same augmentation procedure was also applied to the validation and test sets. The AUC result was 0.84 for slices with biopsy in the augmented test set [8]. Mehrtash et al. used flipping and translation to generate 5-fold cross-validation prostate mpMRI datasets with 10,000 training and 2000 validation samples for each fold, and then they trained a CNN for prostate cancer classification. AUC on the test set was 0.80 [9]. Esteva et al. rotated each skin cancer lesion image randomly between 0° and 359°. The largest upright inscribed rectangle was then cropped from the image and was flipped vertically with a probability of 0.5. As a result, the number of samples in training data size was enlarged by a factor of

720. Then, a CNN was trained to classify skin cancer and the ultimate AUC reached over 91% [10]. Bae et al. used Perlin noise as an augmentation method to train a CNN for 2D high-resolution CT in diffuse interstitial lung disease image classification. In their work, the accuracy with data augmentation using Perlin noise (89.5%) was significantly higher than that with conventional data augmentation (82.1%) [11].

In the natural image domain, multiple augmentation methods have been investigated to improve recognition and classification accuracy. Ding et al. compared three types of data augmentation operation including translation, speckle noising (pointwise multiplying the mean filtered SAR image by random samples from exponential distribution), and pose synthesis (rotate a SAR image and combine it with the original one linearly), and the experiment results showed that combining all types of augmentation operations is a practical approach for target recognition in challenging conditions of target translation, random speckle noise, and missing pose [12]. Lv et al. proposed five data augmentation methods to face images including landmark perturbation (using landmark to do translation, rotation, shearing, and scaling) and four synthesis methods (hairstyles, glasses, poses, and illuminations), and trained a CNN with a dataset by concatenating all data augmentation features to recognize face. They achieved an accuracy of 94.08%, reducing the error by 26% compared with the one without data augmentation [13]. Zhong et al. proposed the Random Erasing method by cutting out an arbitrary region of the input image during each training iteration. For the person re-identification task, the rank 1 accuracy and mean average precision (mAP) increased by 3.10% and 2.67%, respectively, when employing Random Erasing as a data augmentation method on the Market-1501 dataset [14].

In the previous studies in prostate cancer detection using deep learning methods [4–9], as a conventional method for expanding the size of dataset, a data augmentation method or a combination of different methods were randomly picked. While some augmentation strategies and their combinations have been investigated for various medical imaging tasks, there has been no dedicated work on exploring effects of different augmentation methods on performance of deep learning models in prostate cancer detection. Our work focuses on studying the effects of several conventional data augmentation methods on the performance of a shallow and deep CNN, which can fill the gap in this research field and provide guidance on what data augmentation methods should be used with CNNs of different depths in future prostate cancer detection research.

In this work, we have statically applied five most popular augmentation techniques (i.e., random rotation, horizontal flip, vertical flip, random crop, and translation) to the prostate DW-MRI training dataset of 217 patients separately and trained a shallow as well as a deep 2D CNN on the five

augmented datasets, respectively. Finally, we used AUC which is a commonly used metric for binary classification of unbalanced medical image datasets [15], to evaluate performance of the trained CNNs of different depths on a separate test set of 95 patients, using a validation set of 102 patients for finetuning. The classification was done on 2D DW-MRI slices, and the highest AUC on the test set was 85.04% achieved by training the shallow CNN on the augmented dataset using random rotation method. To gain a deeper insight into augmentation methods, we used a CNN heatmap generator to investigate the best way to augment images in different DW-MRI sequences, whether to treat images of all sequences as a single unit and augment as such or apply the augmentation methods to images of each sequence independently.

Our proposed method is fully automated with no user intervention. The 2D slices along with their labels (cancer or no cancer) were fed to the CNNs for training. In the test phase, the 2D slices were fed into the CNN regardless whether they contained prostate and hence, eliminating the need to manually (or automatically) segment prostate. The fully automated nature of our algorithm makes it a better candidate for clinical integration of prostate cancer management. It can process a large set of prostate DW-MRI images and produce results in minutes.

Methods

In this section, we present dataset description, different augmentation methods that we have used for the classification of prostate cancer in DW-MRI dataset, augmentation process for each DW-MRI sequences, an off-the-shelf deep CNN and our proposed shallow CNN architectures, and the GradCAM attention network [16] we used to evaluate robustness of the results.

Dataset

The dataset was obtained as part of a retrospective single institution study and the institutional review board approval was received. All patients had a positive MRI. All exams were performed on a 3T MRI system without an endorectal coil MRI. Our proprietary dataset included DW-MRI images of 414 patients, corresponding to 10,128 2D 6-channel slices. Each slice consisted of a DWI sequence (i.e., 6-channels in the domain of image processing) comprising of an apparent diffusion coefficient (ADC) map and five b-values (0, 100, 400, 1000, and 1600 smm $s\ mm^{-2}$). The DWI data was acquired using a single-slot spin-echo echo-planar imaging sequence, repetition time (TR) 5000~7000 ms, echo time (TE) 61 ms, slice thickness 3 mm, field of view (FOV) 240 mm × 240 mm, and matrix of 140 × 140. Four b values (0,

100, 400, and 1000 smm^{-2}) were acquired, which were then used for ADC map and b1600 image calculation [17].

The four b0, b100, b400, and b1000 images contain various signal intensities showing the amount of water diffusion in the tissue [18]. Thus, although using all b-value images along with ADC and computed b1600 as 6 input channels to CNNs may introduce noise, it has been shown that they may contain complementary information, which may help the classification of prostate cancer [19, 20]. To mitigate the problem of added noise, we added the dropout regularization method to the proposed CNNs, which randomly drops out nodes during training to reduce overfitting caused by noise [21]. In addition, adding noise sometimes is considered a useful data augmentation method to improve CNN performance [11, 22]. Our experiment results also confirmed that the performance of the CNNs is the highest when all 6 image sets were used. Therefore, we selected all four b-values along with computed b1600 and ADC map as input channels to our CNN.

Every DW-MRI 2D slice was labeled according to the Gleason score (GS) from targeted biopsy results. When the GS was larger than 6 (GS > 6), the case was considered as a clinically significant prostate cancer and the slice was labeled as positive. According to the International Society of Urological Pathology (ISUP), there are 5 grade groups for prostate cancer based on the modified GS groups: grade group (GG) 1 = GS ≤ 6, GG 2 = GS 3 + 4, GG 3 = GS 4 + 3, GG 4 = GS 4 + 4, GG 5 = GS 9 and 10 [23]. The European Association of Urology (EAU) guidelines define clinically significant prostate cancer as GG ≥ 2 [24], which means GS ≥ 3 + 4 or GS > 6.

The dataset was split into training set (217 patients 5300 2D slices), validation set (102 patients, 2500 2D slices), and test cohort (95 patients, 2328 2D slices). The detailed patient and slice distribution of dataset is shown in Table 1. Between training and test datasets, the ratios of positive samples and negative samples were similar.

Augmentation Methods

Augmentation methods in deep learning can be applied either statically or on the fly. Static data augmentation refers to appending the augmented data to the training dataset and using the augmented dataset for training the model. On the other hand, when we randomly augment the data in each batch while the model is being trained, it is, in fact, on the fly data augmentation applied. Although on-the-fly augmentation is more efficient in terms of computational and storage resources, the static approach provides more flexibility to data augmentation research. For example, with static augmentation, we can go back and manually check each augmented example. Therefore, we chose to implement our algorithms statically.

Table 1 Patient and slice distribution of the train dataset, validation dataset, and test dataset

Data	Training set	Validation set	Test set
Number of patients	217	102	95
Number of slices	5300	2500	2328
Patient level: number of positive cases	84	49	37
Patient level: number of negative cases	133	53	58
Patient level: ratio of positive/negative	0.632	0.925	0.638
Slice level: number of positive cases	357	232	159
Slice level: number negative cases	4943	2268	2169
Slice level: ratio of positive/negative	0.072	0.102	0.073
Gleason score < 7 (n/%)	133 (61.3)	53 (52.0)	58 (60.7)
Gleason score = 3 + 4 (n/%)	46 (21.2)	27 (26.5)	18 (19.1)
Gleason score = 4 + 3 (n/%)	27 (12.4)	14 (13.7)	14 (14.9)
Gleason score > 7 (n/%)	11 (5.1)	8 (7.8)	5 (5.3)
TZ lesions (n/%)	102 (47.0)	48 (47.1)	33 (35.1)
PZ lesions (n/%)	115 (53.0)	54 (52.9)	62 (64.9)
Mean age (Years)	64.08	64.41	65.16

To reduce bias, achieve better generalization, and compensate for a relatively small amount of training dataset, random rotation, horizontal flip, vertical flip, random crop, and translation are the most frequently used augmentation techniques in medical imaging tasks. For each image $I(x, y)$ in the training set, we utilize an augmentation strategy to obtain transformed image $I(x', y')$.

Random Rotation

Our random rotation augmentation acts clockwise by an angle selected from the range of degrees (−degrees, +degrees) randomly. The selected angle is an integer number. Pixels outside the rotated area will be filled with 0. The rotation formula is given by Eq. (1) and Eq. (2).

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta & 0 \\ -sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

$$x' = x * cos\theta - y * sin\theta, \; y' = x * cos\theta + y * sin\theta \tag{2}$$

Horizontal Flip

The horizontal flip augmentation flips the input image along its vertical (left to right) axis randomly with a given probability. The horizontal flip formula is given by Eq. (3) and Eq. (4).

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3}$$

$$x' = -x, \; y' = y \tag{4}$$

Vertical Flip

The vertical flip augmentation flips the input image along its horizontal (top to bottom) axis randomly with a given probability. The vertical flip formula is given by Eq. (5) and Eq. (6).

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{5}$$

$$x' = x, \; y' = -y \tag{6}$$

Random Crop

The sub-image is a square, and the side length of the square (i.e., crop size) is given in advance which is smaller than the minimum side length of the original image. The sub-image is included in the original image, and its center position is randomly selected. Using this method, our random crop augmentation crops the original image into a given size sub-image randomly. Then the cropped images are scaled up to the original image size by an upsampling technique called nearest-neighbor interpolation [25].

Translation

The translation augmentation shifts the input image in horizontal and vertical directions with a given maximum absolute fraction. For the example of translate fraction = (a, b) on an image of width W and height H, the horizontal shift (dx) and vertical shift (dy) are randomly selected from the uniform distributions over ranges [−W*a, W*a] and [−H*b,

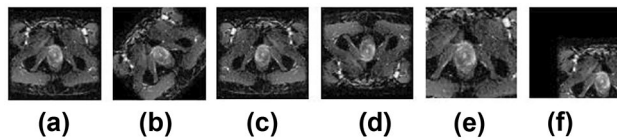**(a)** **(b)** **(c)** **(d)** **(e)** **(f)**

Fig. 1 Effects of the 5 augmentation methods applied to a prostate ADC image. **a** Original image. **b** Rotation (degree = 45). **c** Horizontal flip. **d** Vertical flip. **e** Random crop (crop size = 100). **f** Translation (translate fraction = (0.3, 0.3))

H*b], respectively. The selected dx and dy are both integer numbers. The translation formula is given by Eq. (7) and Eq. (8).

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & dx \\ 0 & 1 & dy \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{7}$$

$$x' = x + dx, \ y' = y + dy \tag{8}$$

For a more intuitive display effect, we have applied these five augmentation methods on an apparent diffusion coefficient (ADC) image of prostate cancer. Augmented images are shown in Fig. 1.

Independent Channel Augmentation Process

In the field of natural image augmentation, augmentation methods are usually applied to all three red/green/blue (RGB) channels at the same time. While the RGB channels represent different color components of a natural image, the concept of channel in medical imaging is different. In this work, each input image consists of 6 channels (sequences): ADC, b0, b1000, b100, b400, and b1600 as shown in Fig. 2. These channels indicate the amount of water diffusion in prostate gland and have different characteristics and features which can be used to distinguish between normal tissue and tumor region [18].

To implement our augmentation algorithms, two different options were considered. We could either apply our augmentation algorithm to each channel independently or stack channels together in order to have all channels manipulated similarly. Intuitively, the independent approach imposes more variance to the data which translates to a richer data. It should be noted that the same reasoning does not apply to natural images with RGB
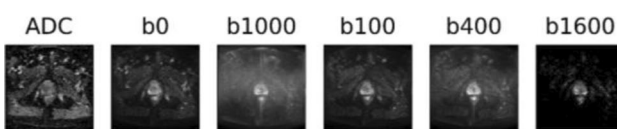

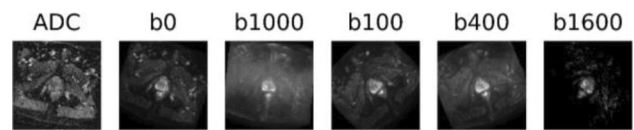**Fig. 2** Original image of 6 channels

Fig. 3 Applying random rotation in each channel independently

channels as all three channels measure texture in a similar way. We evaluated the two scenarios for our best augmentation setting. As it is depicted in Figs. 2, 3, and 4, in one trial, all the 6 channel images were rotated as a unit identity at a rotation degree from the range of [−50, 50], and in other trials, they were rotated independently. Our experiment showed that when each channel is augmented separately, the test AUC result improved by 1.75%.

Gradient-Weighted Class Activation Mapping

In addition to AUC results for different augmentation methods, we used a CNN heatmap generator to investigate whether treating all images of six channels as a single unit in applying augmentation performs better than when each channel is augmented independently. The hypothesis for using CNN visualization was that if the original sample is misclassified and augmentation helped to classify it correctly, this will be reflected in the heatmap. To do so, we input the original and augmented images to the trained CNN and obtain class predictions, and then computed Grad-CAM visualizations for each of the predicted classes. Proposed by Selvaraju et al., Grad-CAM determines which regions of input are more important for predictions from CNN models. Grad-CAM uses the class-specific gradient information flowing into the final convolutional layer of a CNN to generate a coarse heatmap which indicates the important regions in the image [16].

CNN Architecture

To investigate the effect of network depth on optimal augmentation settings, we used two different CNN architectures for our experiments: a shallow CNN [26] and a deep CNN whose design is inspired by VGGNet [27]. The shallow CNN consists of 4 convolutional layers, 3 max-pooling layers, 3 dropout layers, and 3 fully connected layers. The deep CNN is composed of 10 convolutional layers, 4 max-pooling
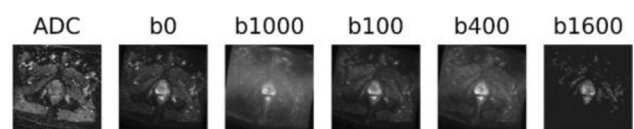

Fig. 4 Applying random rotation in all channels at the same time
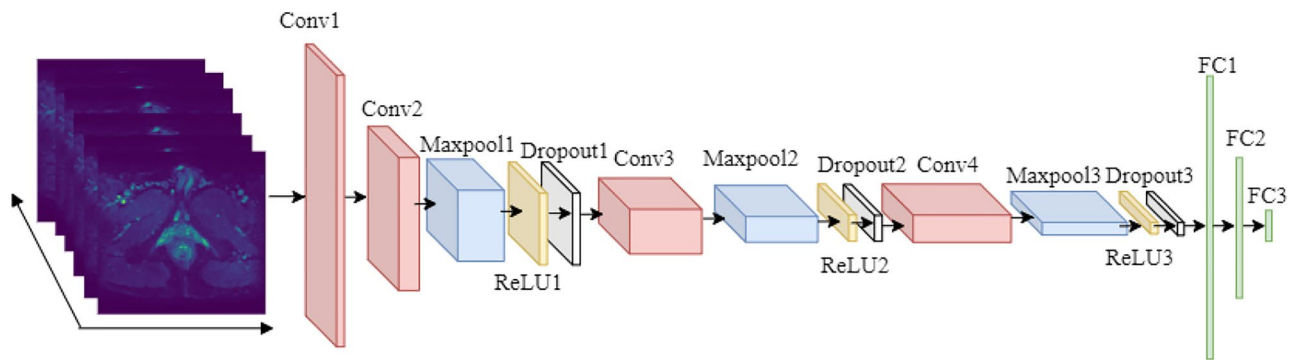
Fig. 5 Architecture of the shallow CNN

layers, and 4 fully connected layers. Architectures of the shallow and deep CNN are shown in Figs. 5 and 6, respectively. Configurations and output size of each layer in the shallow and deep CNN are listed in Table 2 and Table 3, respectively.

The CNNs proposed in this paper are both trained with Cross Entropy loss function and optimized by stochastic gradient descent (SGD) with momentum [28]. Weights of layers were initialized by the Xavier method [29], and the biases were initialized to random values picked from a uniform distribution over 0 to 1.

Settings of Training

In this work, we varied the augmentation hyperparameters across specified ranges. Each set of hyperparameters resulted in a new classification problem. Augmentation was only performed on the training set. To create the ultimate training data, the augmented slices were statically appended to the original training data. In extreme cases such as augmenting by zero degrees of rotation, this method results in simple oversampling.

Our workflow is shown in Fig. 7, which consists of the following main steps: First, we varied the augmentation hyperparameters across specified ranges for each of the 5 augmentations: random rotation, horizontal flip, vertical flip, random crop, and translation. In random rotation, we applied a rotation angle ranging from 0° to 180°, with interval of 5°. In horizontal and vertical flip augmentations, we worked

within the probability range 0 to 1, with 0.05 steps. Images in the training set were assigned random numbers as probability, which were used to decide whether to flip the image. When applying random crop, we took the crop size from 70 to 140 pixels, with intervals of 5. In translation, the translate fraction was selected from 0.0 to 0.5, with 0.01 steps. It must be noted that augmentation was only performed on the training set.

Second, we used an oversampling strategy by adding the augmented sets to the original training dataset. By performing this step, we produced five different augmented training datasets each with 10,600 slices.

Third, after performing normalization across the entire dataset, we trained the shallow CNN and deep CNN independently on the five augmented training datasets. When training the shallow CNN, the learning rate was set to 0.001, batch size was set to 1, L2 regularization penalty was set to 0.001, and momentum in SGD was 0.8 [26]. The hypermeters in the deep CNN were almost the same except for the learning rate; we decreased the learning rate to avoid divergence. Finally, we calculated AUC as a measure to evaluate performance of the trained CNNs.

Results

The Original AUC and Baseline AUC

For the purpose of comparing the effect of different augmentation methods from the same starting point with minimal
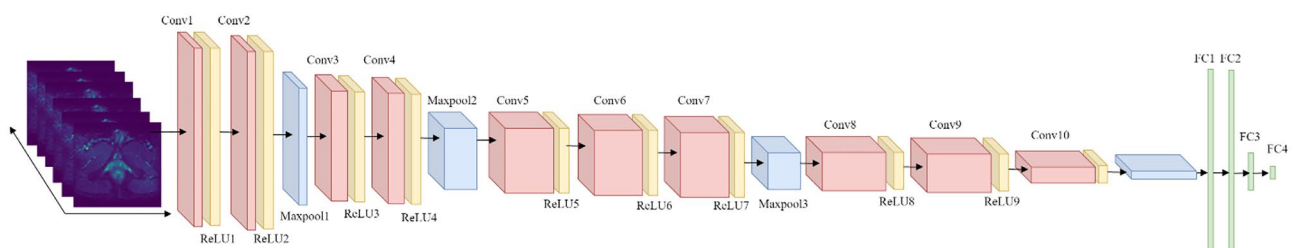

Fig. 6 Architecture of the deep CNN

Table 2 Configuration of the shallow CNN

Layer	Kernel size	Stride	Padding	Probability	Output size
Conv1	3 × 3	1	0		128 × 142 × 142
Conv2	3 × 3	2	0		256 × 64 × 64
Maxpool1	2 × 2	2			256 × 32 × 32
Dropout1				0.1	
Conv3	3 × 3	2	0		512 × 15 × 15
Maxpool2	2 × 2	2			512 × 7 × 7
Dropout2				0.1	
Conv4	3 × 3	2	0		1024 × 3 × 3
Maxpool3	2 × 2	2			1024 × 1 × 1
Dropout3				0.1	
FC1					256 × 1
FC2					64 × 1
FC3					2 × 1

computational cost, all experiments were performed only once with a fixed random seed. The original AUC and baseline AUC are presented in Table 4. The original AUC was calculated without oversampling technique, which means CNNs were trained on the original training dataset. The baseline AUC was computed when there were no augmentation methods applied to the training dataset, except for duplicating the training set. As it is seen from Table 4, the AUC improvement was negligible or nonexistent when oversampling technique was applied to train the shallow and deep CNNs, respectively. All the experiments were conducted on an Nvidia DGX station platform, using Python 3.7.3 and PyTorch 1.2.0.

Table 3 Configuration of the deep CNN

Layer	Kernel size	Stride	Padding	Output size
Conv1	3 × 3	1	1	64 × 144 × 144
Conv2	3 × 3	1	1	64 × 144 × 144
Maxpool1	2 × 2	2		64 × 72 × 72
Conv3	3 × 3	1	1	128 × 72 × 72
Conv4	3 × 3	1	1	128 × 72 × 72
Maxpool2	2 × 2	2		128 × 36 × 36
Conv5	3 × 3	1	1	256 × 36 × 36
Conv6	3 × 3	1	1	256 × 36 × 36
Conv7	3 × 3	1	1	256 × 36 × 36
Maxpool3	2 × 2	2		256 × 18 × 18
Conv8	3 × 3	1	1	512 × 18 × 18
Conv9	3 × 3	1	1	512 × 18 × 18
Conv10	3 × 3	2	1	512 × 9 × 9
Maxpool4	2 × 2	2		512 × 4 × 4
FC1				1024 × 1
FC2				1024 × 1
FC3				256 × 1
FC4				2 × 1

AUC Results of Using Different Augmentation Methods

Figure 8 shows the visualizations of validation and test AUC results on shallow CNN and deep CNN when varying the augmentation hyperparameters across specified ranges for five augmentation methods (random rotation, horizontal flip, vertical flip, random crop, and translation). There are five rows and two columns in Fig. 8, a total of ten subfigures. In each subfigure, the solid lines indicate the AUC values obtained by using the data augmentation method corresponding to the parameters of the horizontal axis, and the dotted lines represent the baseline AUC which was computed when there were no augmentation methods applied to the training dataset (training data was only duplicated). The two blue and red colors in each subfigure represent AUC results for the validation set and test set, respectively.

By comparing each row in Fig. 8, it is seen how different depths of CNN architecture affected the AUC results when applying the same data augmentation method to the training dataset; overall, it can be observed that the shallow CNN performs better than the deep CNN. We can also compare the subfigures in each column, which presents the AUC results of training the same CNN with different data augmentation methods. The experiment results show that random rotation and translation were the most efficient augmentation methods when using the shallow CNN and the deep CNN, respectively.

Table 5 lists the highest validation and test AUC results of both shallow CNN and deep CNNs for the five augmentation methods. As it can be seen, the highest validation AUC of 88.93% and test AUC of 85.04% were obtained when the random rotation technique (with [−100, 100] and [−50, 50] degree) was applied to train shallow CNN for prostate cancer classification.

Fig. 7 Workflow chart

Figure 9 visualizes the comparison of the test AUC results of shallow CNN and deep CNN. For each of the augmentation methods, it shows the highest AUC, the lowest AUC, and the average AUC in the test dataset among the range of hyperparameters for the augmentation methods (e.g., rotation degree).

To assess the significance of differences among the proposed five augmentation techniques and 2 types of CNNs, the Tukey-Kramer approach and the Benjamini-Hochberg procedure were used for multiple comparisons [30, 31]. When comparing different augmentation methods, ten pairwise comparisons were not independent. For example, if the test AUC under random rotation technique is higher than the test AUC under horizontal flip technique, there is a good chance that the random rotation method performs better than vertical flip method on the test dataset as well [32]. Thus, the Tukey-Kramer approach was used for this scenario. On the other hand, the Benjamini-Hochberg procedure assumes that the individual tests are independent of each other [32], which is suitable for the scenario of comparing the two types of CNNs.

Table 6 lists the results of multiple comparisons for five augmentation methods. As it is seen, for the shallow CNN, the random rotation method, which yielded the higher AUC, is significantly different compared to the other four

methods on both validation dataset and test dataset. For the deep CNN, the translation method has a significant difference compared to other methods on the validation dataset, and on the test dataset, the translation method was significantly different from other methods, except for the vertical flip method.

Table 7 lists the result for multiple comparisons for two types of CNNs. As it is seen, except for the translation technique on the validation dataset, there is a statistically significant difference between the shallow CNN and the deep CNN on both validation and test datasets when using different augmentation methods in training procedure.

AUC Results on Different Devices

For the purpose of comparing the effect of the randomization, the same augmentation method with different seedings on different devices was tested. For this purpose, Nvidia TITAN X (Pascal) GPU was used to run the code. It would be computationally prohibitive to try multiple seeds for each experiment. Hence, we used the augmented dataset using the random rotation method to train the shallow CNN and compared the test AUC results with the results obtained from the previous run using the DGX platform. The comparison of test AUC results on different devices is visualized in Fig. 10.

Table 4 The original AUC and baseline AUC

AUC (%)	Original AUC		Baseline AUC	
	Validation dataset	Test dataset	Validation dataset	Test dataset
The shallow CNN	82.59	78.61	82.63	80.23
The deep CNN	80.23	74.81	80.23	74.81

Fig. 8 Comparison of AUC results of using five augmentation methods to train shallow and deep CNNs. **a** Shallow CNN: random rotation. **b** Deep CNN: random rotation. **c** Shallow CNN: horizontal flip. **d** Deep CNN: horizontal flip. **e** Shallow CNN: vertical flip. **f** Deep CNN: vertical flip. **g** Shallow CNN: random crop. **h** Deep CNN: random crop. **i** Shallow CNN: translation. **j** Deep CNN: translation
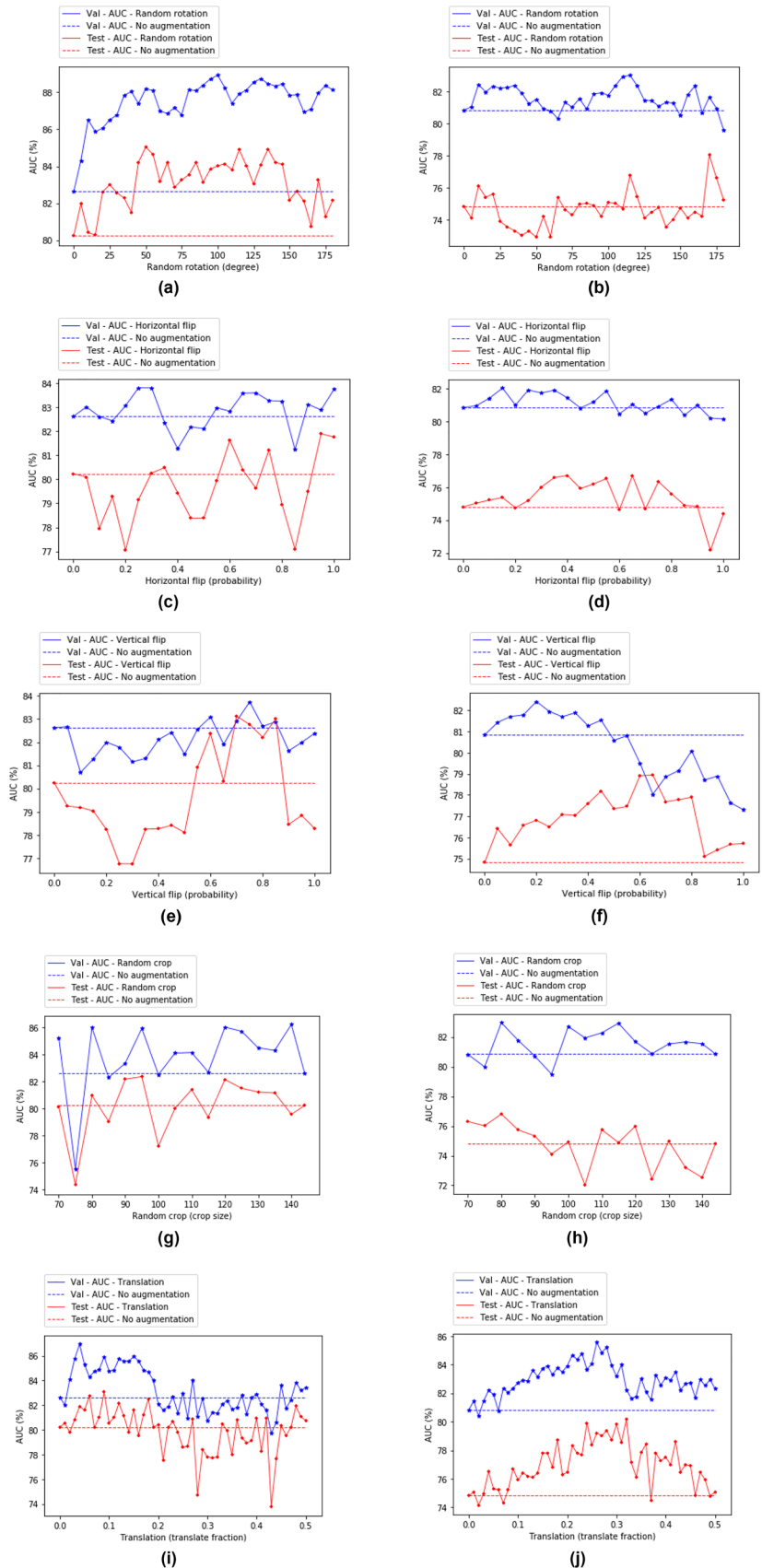
Table 5 Comparison of the highest AUC results on shallow and deep CNNs

The highest AUC (%)	Augmentation methods									
	Random rotation		Horizontal flip		Vertical flip		Random crop		Translation	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
The shallow CNN	**88.93**	**85.04**	83.81	81.90	83.74	83.12	86.23	82.37	86.95	83.06
The deep CNN	83.01	78.04	82.03	76.72	82.43	78.94	82.95	76.80	85.56	80.16

The bold numbers indicate the highest validation and test AUC results for the CNNs among the five augmentation methods

Grad-CAM Visualizations of Samples and Corresponding Augmented Samples

In order to deepen our understanding of the data augmentation process, we designed a test using a CNN heatmap generator. From the training set, examples and their corresponding augmented instances were fed to the trained shallow CNN and their Grad-CAM visualizations were calculated. Although it was not practical to manually check every single case, through examining 50 random cases, we realized that when the model is able to correctly classify an example, the heatmap correctly corresponds to the location of the prostate in the image. Otherwise, the heatmap points to an area outside prostate.

When the original sample is misclassified while the augmented sample is classified correctly, it can be observed from the visualization map that the augmentation led the network to detect correct information in the image. We used the GradCAM setting to provide more evidence to our earlier claim about independent channel augmentation in medical imaging. Shown in Fig. 11, we randomly picked a misclassified positive case whose channel-independent augmented version was correctly classified and calculated their GradCAMs. Because it gives a better representation, only ADC and b1600 channels are plotted in the figure. In the next step, we rotated the same image; however, it was not channel independent. As reflected in Fig. 12, fixed rotation (treating all channels as a single unit) is not capable of correcting CNN attention and hence the augmented sample is also misclassified. The zoomed in detail on the lesion of interest on ADC and b1600 channels is presented in Fig. 13.

To better visualize the CNN classification performance, five correctly classified positive and five correctly classified negative examples in ADC and b1600 images with corresponding GradCAM results are presented in Figs. 14 and 15, respectively.

Discussion

Based on the presented results in Table 4 and Fig. 8, random rotation is the most efficient augmentation method for prostate cancer detection when the CNN architecture was shallow. After training the shallow CNN on the augmented
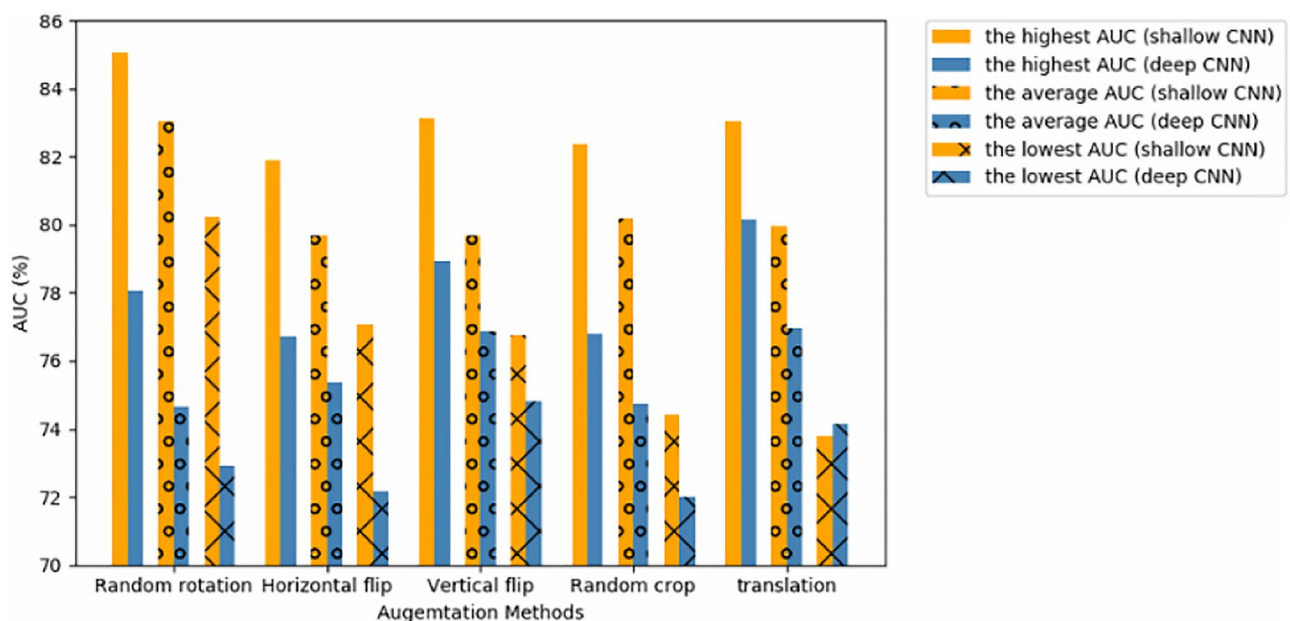

Fig. 9 Comparison of test AUC results on shallow and deep CNNs

Table 6 Multiple comparisons for five augmentation methods

Augmentation methods comparison		The shallow CNN		The deep CNN	
		Val AUC	Test AUC	Val AUC	Test AUC
		Adjusted p value	Adjusted p value	Adjusted p value	Adjusted p value
Random rotation	Horizontal flip	<0.001*	<0.001*	0.526	0.294
Random rotation	Vertical flip	<0.001*	<0.001*	<0.001*	<0.001*
Random rotation	Random crop	<0.001*	<0.001*	0.999	0.999
Random rotation	Translation	<0.001*	<0.001*	<0.001*	<0.001*
Horizontal flip	Vertical flip	0.588	1.000	0.099	0.003*
Horizontal flip	Random crop	0.314	0.877	0.817	0.620
Horizontal flip	Translation	0.916	0.967	<0.001*	<0.001*
Vertical flip	Random crop	0.011*	0.878	0.008*	<0.001*
Vertical flip	Translation	0.078	0.966	<0.001*	1.000
Random Crop	Translation	0.594	0.986	<0.001*	<0.001*

Applying the Tukey-Kramer multiple comparison procedure. Adjusted p value < 0.05 is statistically significant

*Statistically significant

set using random rotation, we achieved the highest AUC of 85.04% on the test set when the degree range was set to (−50, 50), which was 5.2% higher than the baseline AUC (80.84%). As mentioned before, baseline refers to oversampling whose AUC is higher than the original dataset (i.e., not oversampling and no augmentation).

In terms of CNN performance on both test set and validation set, translation was the most efficient augmentation method when the deep CNN was applied to the prostate cancer classification task. With the translation fraction set to (0.32, 0.32), the AUC on the test set was the highest (80.16%)—7.15% higher than the baseline AUC (74.81%). Meanwhile, vertical flip was also a useful augmentation strategy in deep CNN scenarios, with the highest AUC of 78.94% on the test set when the probability was set to 0.65, which was 5.25% higher than the baseline AUC (74.81%).

As shown in Fig. 9, except for the worst translation case, the shallow CNN performs better than the deep CNN. In other words, the effect of data augmentation was more prominent for the shallower architecture. The higher performance of the shallower network may be explained by

the noisy nature of MRI images. Due to the environment, equipment, and the performance of different doctors, MRI images usually contain a significant amount of noise [33]. We hypothesize that our proposed shallow CNN architecture is rich enough in terms of number of network weights, to not get stuck in the region of underfitting and light enough to avoid overfitting due to noise. However, the deep CNN architecture is prone to overfitting due to inheriting noise in MRI images.

For the shallow CNN, random rotation worked best because prostate is symmetrical. As a result, horizontal flip should not be able to add any value to the dataset other than a semi-oversampling effect. The nature of convolution operation in each operation minimizes the effect of vertical flip on the other hand. The padding used in translation negatively affects the convolution operation [34]. Therefore, translation is not a good option for prostate, especially in cases where the tumor is small. Rotation is an effective augmentation method which creates asymmetric images. In the case of prostate cancer detection, asymmetry pushes the network to learn decisive patterns of the given image. For the deeper

**Table 7** Multiple comparisons for two types of CNNs

Augmentation methods	Two types of CNNs comparison		AUC on validation dataset Adjusted p value	AUC on test dataset Adjusted p value
Random rotation	The shallow CNN	The deep CNN	<0.001*	<0.001*
Horizontal flip	The shallow CNN	The deep CNN	<0.001*	<0.001*
Vertical flip	The shallow CNN	The deep CNN	<0.001*	<0.001*
Random crop	The shallow CNN	The deep CNN	0.004*	<0.001*
Translation	The shallow CNN	The deep CNN	0.382	<0.001*

Applying the Benjamini-Hochberg multiple comparison procedure. Adjusted p value < 0.05 is statistically significant

*Statistically significant
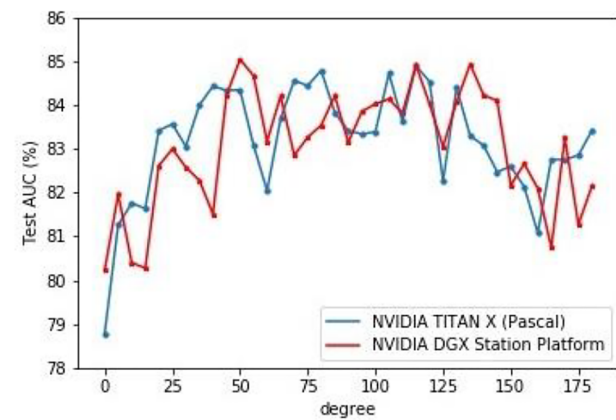
(a)
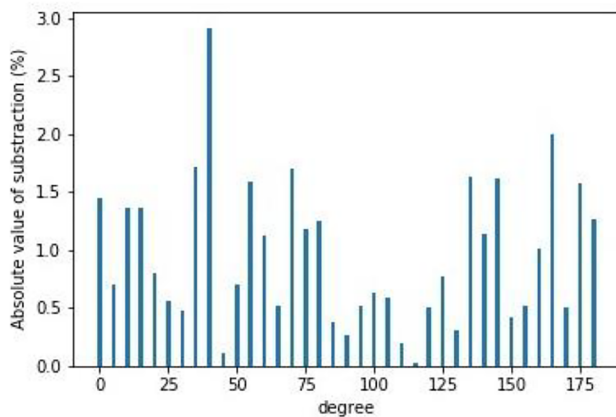
(b)

Fig. 10 Comparison of test AUC results on different devices. **a** Trend graph of the test AUC results. **b** Histogram of absolute values of AUC subtraction

CNN, while random rotation is still promising, translation in some special configurations surpasses other methods. The reason may be because due to arbitrary noise reduction which helps the network to void overfitting.

It was interesting to observe that the results of augmentation methods on the validation and test sets are highly correlated. This indicates that our validation and test cohorts are
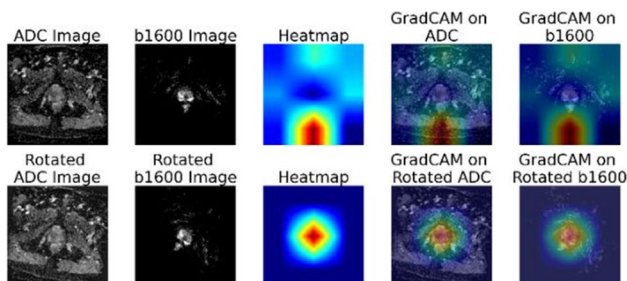
Fig. 11 GradCAM visualizations of a misclassified positive example and its channel-independent augmentation pair
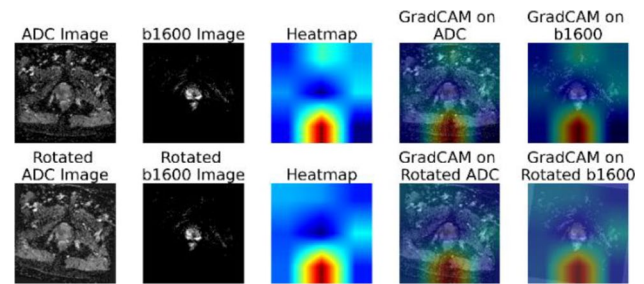

Fig. 12 GradCAM visualizations of a misclassified positive example and its channel-depended augmentation pair

both good representatives of the original population. Furthermore, it can be inferred that our augmentation technique is effective for the true distribution and is not an arbitrary fit to our sample space.

A similar DW-MRI prostate cohort was used by Yoo et al. to develop a CNN-based pipeline using a modified ResNet architecture for the same prostate cancer detection [17]. The main differences between our method and that of [17] are as follows. First, [17] used a deep CNN (ResNet) while our best performance was achieved using a shallow CNN. Second, as opposed to our method, no augmentation was used in [17]. Finally, in [17], the slices that did not contain any portion of prostate gland were manually eliminated and the remaining
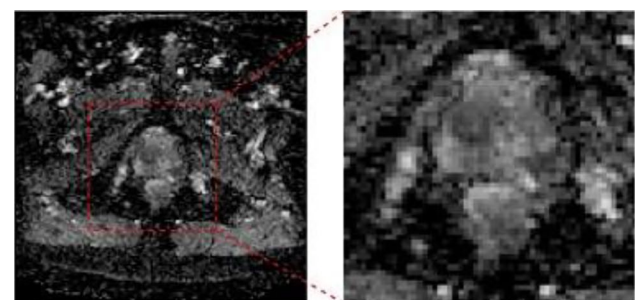
(a)
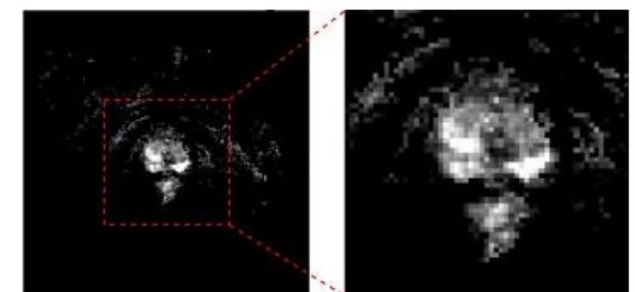
(b)

Fig. 13 A positive example and corresponding zoomed in detail on the lesion of interest. **a** ADC image and corresponding zoomed in region (prostate gland) where the lesion is located. **b** b1600 image and corresponding zoomed in region (prostate gland) where the lesion is located
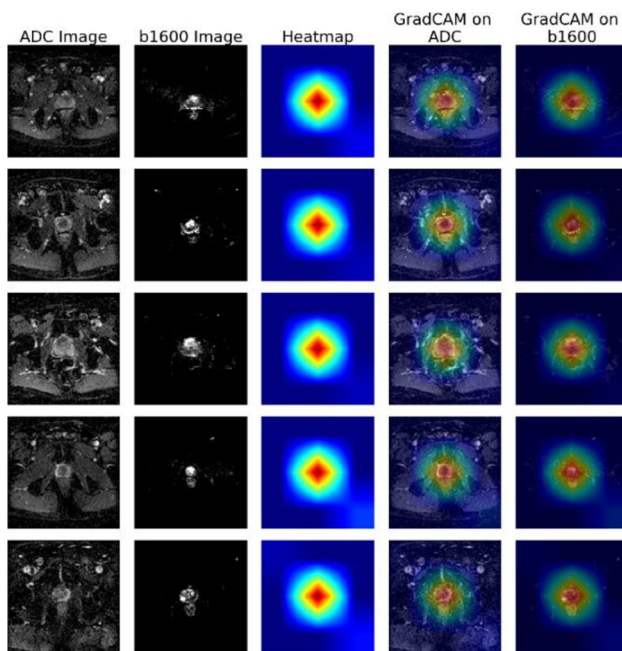
Fig. 14 Correctly classified positive examples

slices were center-cropped with a fixed size. This requires manual markings of the apex and base slices of the prostate. In our work, however, detection of prostate cancer is fully automated with no preselection of slices. The effect of our approach is reflected in our sample sizes where for example our test set contains 2328 slices as compared to their 1486 slices in [17]. Although our AUC result (85.04%) is slightly less than the average slice-level result reported in [17] (mean
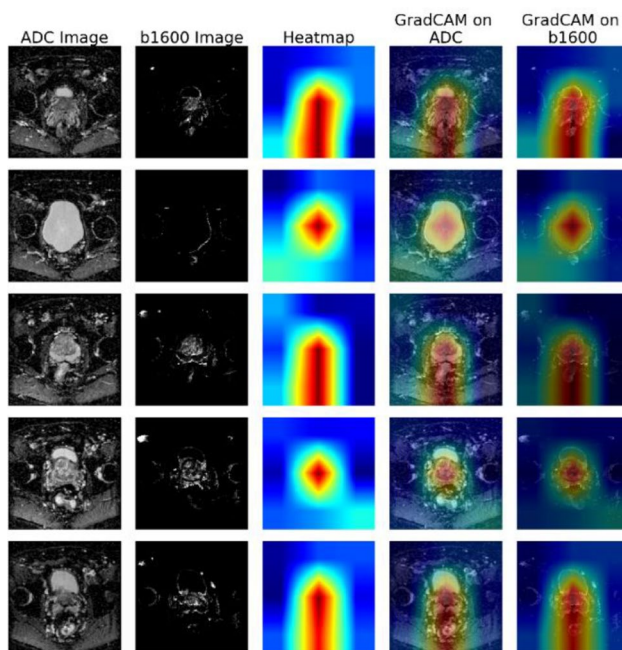

Fig. 15 Correctly classified positive examples

AUC of 86%), our method is fully automated with no need for any user intervention. The method presented in [17] also included patient-level classification, which was not a part of our work in this paper.

While it was not computationally feasible to repeat our results with different network initializations, we evaluated the results for the best configuration (rotation). Our parallel experiment on a second hardware platform well conformed with the original results. We also selected the best augmented dataset and examined GradCAMs of several random examples from the training cohort to study the effect of augmentation at the end of the training phase. It was concluded that channel-independent augmentation works better in DW-MRI image for prostate. With further investigation, this may be shown to be the case for other cancer sites. We also observed that an effective augmentation can help the CNN classify an augmented sample correctly while the original sample was misclassified.

We tried different augmentation combinations; however, we did not find them effective and the AUC decreased. For example, we combined random rotation [−50, 50] and random crop (crop size = 95) which were the most efficient candidates to the shallow CNN. We obtained AUCs of 87.86% and 82.40% on validation and test sets, respectively, which are both lower than that of using only random rotation as augmentation. Similar results were achieved for the deep CNN. With rotation, we also applied background perturbation using a mask to maintain the tumor area unchanged and only rotate background randomly to augment training dataset. The AUC results on the validation and test sets set were 0.74% and 2% lower, respectively, compared to rotation only.

We tried the same Perlin noise augmentation method reported in [11], but the effects were not significant and the AUC results on the validation and test sets improved by only 0.54% and 0.76%, respectively. We also tried random erasing as a data augmentation method to improve performance of CNN [14]. However, this improved the AUC only by 1.19% and 0.6% for validation and test sets, respectively.

Not all augmentation methods (e.g., vertical flip, rotation) may be meaningful from the medical point of view since MRI exams are acquired in a standardized fashion. Nevertheless, the purpose of augmentation is not for generating more routine clinical images as the augmented images have little realistic medical significance due to spatial transformation. In other words, the augmented images were not generated for disease diagnosis purposes. This is the reason why augmentation techniques were not applied to the test cohorts. The aim of an augmentation technique is to produce more augmented data to "feed" the CNNs and get better model performance during the training process [35]. While from a medical point of view less noise leads to higher image quality with potentially better diagnosis accuracy, the technique of adding noise helps CNN performance as it has been illustrated in previous work

[11]. Additionally, choosing the best type of data augmentation tightly depends on data features and model types. It needs further research to theoretically explain why a particular data augmentation method boosts a model's performance. As future research, we will investigate the interpretability of relationship between data augmentation methods and deep learning models performance.

The current augmentation strategies have been applied to prostate DWI data. While there is no mathematical proof or guarantee that our results will generalize to other medical imaging datasets, our research introduces a systematic way of data augmentation for CNNs applied to tumor classification in medical images. In future work, we will apply these strategies to other cancer sites (e.g., liver MRI) and investigate whether the results generalize to other disease sites as well.

Conclusion

Data augmentation is a useful technique for constructing enough amount of data in the limited data domain. In this paper, we proposed a fully automated method for prostate cancer detection in DW-MRI images using CNNs. We applied 5 different data augmentation methods to the training dataset of DW-MRI images of prostate. Two different deep learning models (CNNs) were trained to classify prostate cancer. Random rotation and translation were found to be the most efficient data augmentation methods for the shallow CNN and the deep CNN in prostate cancer detection, respectively.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Welch HG and Black WC: Overdiagnosis in cancer. J Natl Cancer Inst 102: 605–613, 2010.
2. Thompson JE, Van Leeuwen PJ, Moses D, Shnier R, Brenner P, Delprado W, Pulbrook M, Böhm M, Haynes AM, Hayen A and Stricker PD: The diagnostic performance of multiparametric magnetic resonance imaging to detect significant prostate cancer. J Urol 195: 1428–1435, 2016.
3. Razzak MI, Naz S and Zaib A: Deep learning for medical image processing: Overview, challenges and the future. Lect Notes Comput Vis Biomech 26: 323–350, 2018.
4. Cao R, Bajgiran AM, Mirak SA, Shakeri S, Zhong X, Enzmann D, Raman S and Sung K: Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet. IEEE Trans Med Imaging PP: 1–1, 2019.
5. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS and Fuchs TJ: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 25: 1301–1309, 2019.
6. 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 Int 122: 411–417, 2018.
7. Wang X, Yang W, Weinreb J, Han J, Li Q, Kong X, Yan Y, Ke Z, Luo B, Liu T and Wang L: Searching for prostate cancer by fully automated magnetic resonance imaging classification: Deep learning versus non-deep learning. Sci Rep 7: 1–8, 2017.
8. Liu S, Zheng H, Feng Y and Li W: Prostate cancer diagnosis using deep learning with 3D multiparametric MRI. Med Imaging 2017 Comput Diagnosis 10134: 1013428, 2017.
9. Mehrtash A, Sedghi A, Ghafoorian M, Taghipour M, Tempany CM, Wells WM, Kapur T, Mousavi P, Abolmaesumi P and Fedorov A: Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks. Med Imaging 2017 Comput Diagnosis 10134: 101342A, 2017.
10. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM and Thrun S: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542: 115–118, 2017.
11. Bae HJ, Kim CW, Kim N, Park BH, Kim N, Seo JB and Lee SM: A Perlin Noise-Based Augmentation Strategy for Deep Learning with Small Data Samples of HRCT Images. Sci Rep 8: 1–7, 2018.
12. Ding J, Chen B, Liu H and Huang M: Convolutional Neural Network with Data Augmentation for SAR Target Recognition. IEEE Geosci Remote Sens Lett 13: 364–368, 2016.
13. Lv JJ, Shao XH, Huang JS, Zhou XD and Zhou X: Data augmentation for face recognition. Neurocomputing 230: 184–196, 2017.
14. Zhong Z, Zheng L, Kang G, Li S and Yang Y: Random Erasing Data Augmentation., 2017.
15. Park SH, Goo JM and Jo CH: Receiver operating characteristic (ROC) curve: Practical review for radiologists. Korean J Radiol 5: 11–18, 2004.
16. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D and Batra D: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis 128: 336–359, 2020.
17. Yoo S, Gujrathi I, Haider MA and Khalvati F: Prostate Cancer Detection using Deep Convolutional Neural Networks. Sci Rep 9: 1–10, 2019.
18. Glaister J, Cameron A, Wong A and Haider MA: Quantitative investigative analysis of tumour separability in the prostate gland using ultra-high b-value computed diffusion imaging. Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS: 420–423, 2012.
19. Khalvati F, Wong A and Haider MA: Automated prostate cancer detection via comprehensive multi-parametric magnetic resonance imaging texture feature models. BMC Med Imaging 15: 1–14, 2015.
20. Khalvati F, Zhang J, Chung AG, Shafiee MJ, Wong A and Haider MA: MPCaD: A multi-scale radiomics-driven framework for automated prostate cancer localization and detection. BMC Med Imaging 18: 1–14, 2018.
21. Nitish S, Geoffrey H, Alex K, Ilya S and Ruslan S: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res 15: 1929–1958, 2014.
22. Audhkhasi K, Osoba O and Kosko B: Noise-enhanced convolutional neural networks. Neural Networks 78: 15–23, 2016.
23. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR and Humphrey PA: The 2014 international society of urological

pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. Am J Surg Pathol 40: 244–252, 2016.

24. Mottet N, Bellmunt J, Bolla M, … Cornford P: EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. Eur Urol 71: 618–629, 2017.

25. Parker JA, Kenyon R V. and Troxel DE: Comparison of Interpolating Methods for Image Resampling. IEEE Trans Med Imaging 2: 31–39, 1983.

26. Namdar K, Gujrathi I, Haider MA and Khalvati F: Evolution-based Fine-tuning of CNNs for Prostate Cancer Detection. Int Conf Neural Inf Syst, 2019.

27. Simonyan K and Zisserman A: Very deep convolutional networks for large-scale image recognition. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc: 1–14, 2015.

28. Ruder S: An overview of gradient descent optimization algorithms. 1–14, 2016.

29. Glorot X and Bengio Y: Understanding the difficulty of training deep feedforward neural networks. J Mach Learn Res 9: 249–256, 2010.

30. Anthon J. H: A Proof of the Conjecture That The Tukey-Kramer Multiple Comparisons Procedure Is Conservative. Ann Stat 12: 61–75, 1991.

31. Benjamini Y and Hochberg Y: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B 57: 289–300, 1995.

32. McDonald JH: Handbook of Biolological Statistics. Sparky House Publishing, Baltimore, Maryland, U.S.A., 2014.

33. Vaishali S, Rao KK and Rao GVS: A review on noise reduction methods for brain MRI images. Int Conf Signal Process Commun Eng Syst - Proc SPACES 2015, Assoc with IEEE: 363–365, 2015.

34. Islam MA, Jia S and Bruce NDB: How much Position Information Do Convolutional Neural Networks Encode? arXiv, 2020.

35. Dyk DAV and Meng XL: The art of data augmentation. J Comput Graph Stat 10: 1–50, 2001.