Original Article

Adversarial training for prostate cancer classification using magnetic resonance imaging

Lei Hu[1], Da-Wei Zhou[2], Xiang-Yu Guo[3], Wen-Hao Xu[1], Li-Ming Wei[1], Jun-Gong Zhao[1]

[1]Department of Diagnostic and Interventional Radiology, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China; [2]State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an, China; [3]Xi'an OUR United Co., Ltd., Xi'an, China

Contributions: (I) Conception and design: L Hu, DW Zhou, JG Zhao; (II) Administrative support: JG Zhao; (III) Provision of study materials or patients: WH Xu; (IV) Collection and assembly of data: WH Xu; (V) Data analysis and interpretation: L Hu, DW Zhou; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Jun-Gong Zhao, MD, PhD. Department of Diagnostic and Interventional Radiology, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, No. 600, Yi Shan Road, Shanghai 200233, China. Email: zhaojungongradio@hotmail.com.

Background: To use adversarial training to increase the generalizability and diagnostic accuracy of deep learning models for prostate cancer diagnosis.

Methods: This multicenter study retrospectively included 396 prostate cancer patients who underwent magnetic resonance imaging (development set, 297 patients from Shanghai Jiao Tong University Affiliated Sixth People's Hospital and Eighth People's Hospital; test set, 99 patients from Renmin Hospital of Wuhan University). Two binary classification deep learning models for clinically significant prostate cancer classification [PM1, pretraining Visual Geometry Group network (VGGNet)-16-based model 1; PM2, pretraining residual network (ResNet)-50-based model 2] and two multiclass classification deep learning models for prostate cancer grading (PM3, pretraining VGGNet-16-based model 3; PM4: pretraining ResNet-50-based model 4) were built using apparent diffusion coefficient and T2-weighted images. These models were then retrained with adversarial examples starting from the initial random model parameters (AM1, adversarial training VGGNet-16 model 1; AM2, adversarial training ResNet-50 model 2; AM3, adversarial training VGGNet-16 model 3; AM4, adversarial training ResNet-50 model 4, respectively). To verify whether adversarial training can improve the diagnostic model's effectiveness, we compared the diagnostic performance of the deep learning methods before and after adversarial training. Receiver operating characteristic curve analysis was performed to evaluate significant prostate cancer classification models. Differences in areas under the curve (AUCs) were compared using Delong's tests. The quadratic weighted kappa score was used to verify the PCa grading models.

Results: AM1 and AM2 had significantly higher AUCs than PM1 and PM2 in the internal validation dataset (0.84 *vs.* 0.89 and 0.83 *vs.* 0.87) and test dataset (0.73 *vs.* 0.86 and 0.72 *vs.* 0.82). AM3 and AM4 showed higher κ values than PM3 and PM4 in the internal validation dataset {0.266 [95% confidence interval (CI): 0.152–0.379] *vs.* 0.292 (95% CI: 0.178–0.405) and 0.254 (95% CI: 0.159–0.390) *vs.* 0.279 (95% CI: 0.163–0.396)} and test set [0.196 (95% CI: 0.029–0.362) *vs.* 0.268 (95% CI: 0.109–0.427) and 0.183 (95% CI: 0.015–0.351) *vs.* 0.228 (95% CI: 0.068–0.389)].

Conclusions: Using adversarial examples to train prostate cancer classification deep learning models can improve their generalizability and classification abilities.

Keywords: Deep learning (DL); magnetic resonance imaging (MRI); prostatic neoplasms; neural networks; robotics

Submitted Nov 09, 2021. Accepted for publication Mar 16, 2022.
doi: 10.21037/qims-21-1089
View this article at: https://dx.doi.org/10.21037/qims-21-1089

Introduction

In most countries, prostate cancer (PCa) is the second most commonly diagnosed malignancy among men. Its accurate classification is critical for selecting the appropriate treatment, leading to improved outcomes and ultimately reducing overtreatment and mortality (1). Currently, the mainstream clinical method for PCa identification is systematic biopsy under transrectal ultrasound (TRUS) guidance in case of suspicious PCa due to a high prostate-specific antigen (PSA) level or an abnormal screening digital rectal examination. Since TRUS-guided biopsy for PCa detection has high false-negative results, and due to its invasiveness, it is not suitable for screening a large patient population for PCa detection. Therefore, over the past decade, multiparametric magnetic resonance imaging (mpMRI) has become increasingly important in PCa diagnosis because of its high sensitivity for detecting prostate lesions (2-4). However, the traditional assessment of prostate mpMRI is based on subjective visual assessment, leading to inter-reader variability and a suboptimal ability to assess lesion aggressiveness (5). In addition, manually interpreting mpMRI sequences requires substantial experience and labor, limiting its clinical applicability (6). Thus, how to effectively and efficiently interpret mpMRI data to achieve a satisfactory accuracy for PCa diagnosis in the clinic remains unresolved.

Deep learning (DL) has provided new potential methods to solve these issues. Numerous DL approaches have been proposed for PCa diagnosis, leading to the performance of many PCa classification tasks with remarkable accuracy (1,5-15). However, despite the good performance of many DL models in terms of PCa classification, their practical application is still controversial as these artificial neural network-based methods are vulnerable to small perturbations in the images (16-20). These subtle perturbations can be caused by changes in noise (21,22) imperceptible to the human eye but which can easily deceive DL models. Images with such intentionally added perturbation are called adversarial examples (AEs). A previous study (23) proved that AEs can easily mislead the prediction of a neural network classifier, thereby resulting in the attacked model reporting high confidence in the wrong prediction. The existence of AEs raises questions about the generalizability of DL models and a number of social security concerns; however, because these AEs are only applicable in a very specific setting (i.e., the attacker knows the DL model & can control the input image), which generally do not exist in medical images, the practical clinical application of AEs for medical image classification should not be affected (23).

Recent studies (23-26) have shown that using AEs to train machine learning models [adversarial training (AT)] can significantly improve the ability of deep neural networks to resist adversarial noise (24-26). In addition, AT improves not only the classification accuracy of the target model, but also its accuracy for the original samples (27,28). Based on these studies, we hypothesized that AT might increase the robustness and generalizability of PCa classification DL models and improve their diagnostic accuracy in different validation datasets.

In this study, binary classification DL models for clinically significant PCa (csPCa) detection and multivariate classification DL models for PCa grading were built and then trained by AEs. The diagnostic performances of the DL models before and after AT were compared and evaluated. We present the following article in accordance with the STARD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-21-1089/rc).

Methods

Study design

This retrospective, multicenter study was approved by the local ethics committee of our institution. Informed consent was obtained from all patients. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Patients

We collected data from patients who underwent 3-T multiparametric prostate MRI and had a subsequent targeted MRI-TRUS fusion biopsy confirming PCa between November 2018 and December 2020. The inclusion criteria included the following: (I) complete

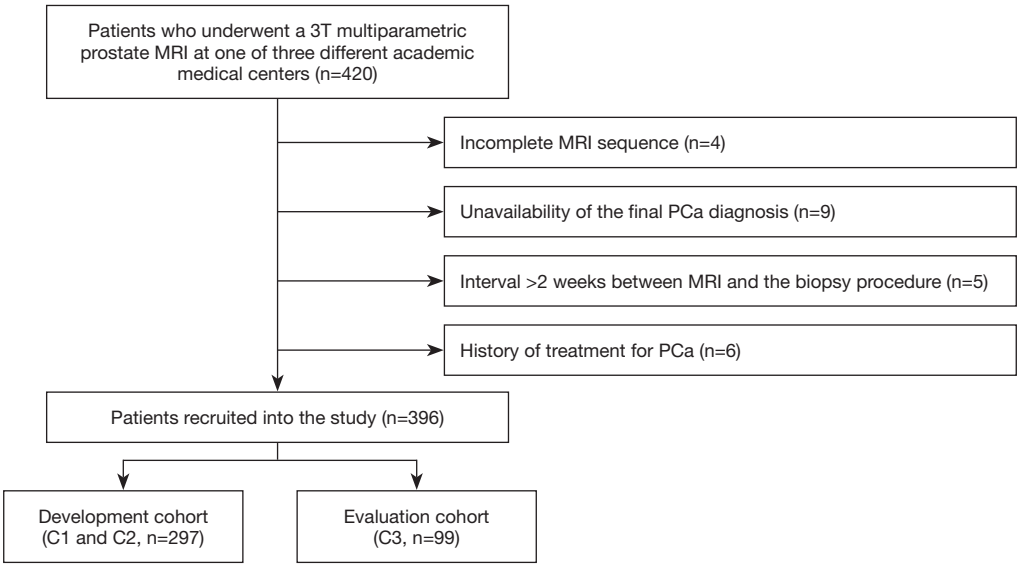

Figure 1 Diagram for patient inclusion into the study. MRI, magnetic resonance imaging; PCa, prostate cancer.

clinical information and pathologic examination results; (II) prostate lesions with definite boundaries on all magnetic resonance (MR) images. The exclusion criteria were as follows: (I) history of treatment for PCa, including surgery, hormone therapy, radiation therapy, or cryotherapy; (II) >2 weeks between MRI and the biopsy procedure; (III) unavailability of the final PCa diagnosis; and (IV) incomplete MRI sequence.

Initially, a total of 420 consecutive participants satisfying the inclusion criteria were enrolled. Of these, 24 were excluded according to the exclusion criteria. The detailed reasons for exclusion are listed in *Figure 1*. Finally, a total of 396 patients at three different academic medical centers were recruited for our study. The development set consisted of 297 patients enrolled in two medical centers (Shanghai Jiao Tong University Affiliated Sixth People's Hospital and Eighth People's Hospital) between November 2018 and November 2020. The development set was furtherly randomly divided into a training set and an internal validation set at a ratio of 4:1. The remaining 99 patients were collected from another medical center (Renmin Hospital of Wuhan University) between July 2018 and December 2020 and were used as a test set for the test of the PCa classification DL models before and after AT.

The study included four steps: MRI examinations and image preprocessing, model pretraining, AT, and model evaluation (*Figure 2*).

MRI examinations

All images were acquired using one of three 3.0-T imagers (MAGNETOM Verio, Skyra, or Prisma; Siemens Healthcare, Erlangen, Germany) and a pelvic phased-array coil. Each mpMRI scan included axial T2-weighted imaging (T2WI) (repetition time/echo time, 6,000/101 ms; section thickness, 3.5 mm; matrix, 320×256; in-plane resolution, 0.625×0.625 mm^2; number of averages, 1; field of view, 200×200 mm^2; bandwidth, 200 Hz/px; acquisition time, 2:08 min) and diffusion-weighted imaging (DWI) (repetition time/echo time, 5,600/79 ms; section thickness, 3 mm; matrix, 178×178; in-plane resolution, 2.10×1.60 mm^2; number of averages, 1/3/6; field of view, 380×281 mm^2; bandwidth, 2,160 Hz/px; b-values, 50/1,000/1,500 s/mm^2; acquisition time, 3:05 min). Apparent diffusion coefficient (ADC) maps were inline calculated by scanner software using linear fitting based on a mono-exponential model.

Datasets

Because T2WI and ADC sequences can provide different and complementary information and their fusion can improve the accuracy of PCa diagnosis (1,9,14), we used axial T2WI and ADC sequences selected for PCa classification.

A radiologist with >20 years of experience in prostate MRI reviewed the T2WI sequences and ADC maps with

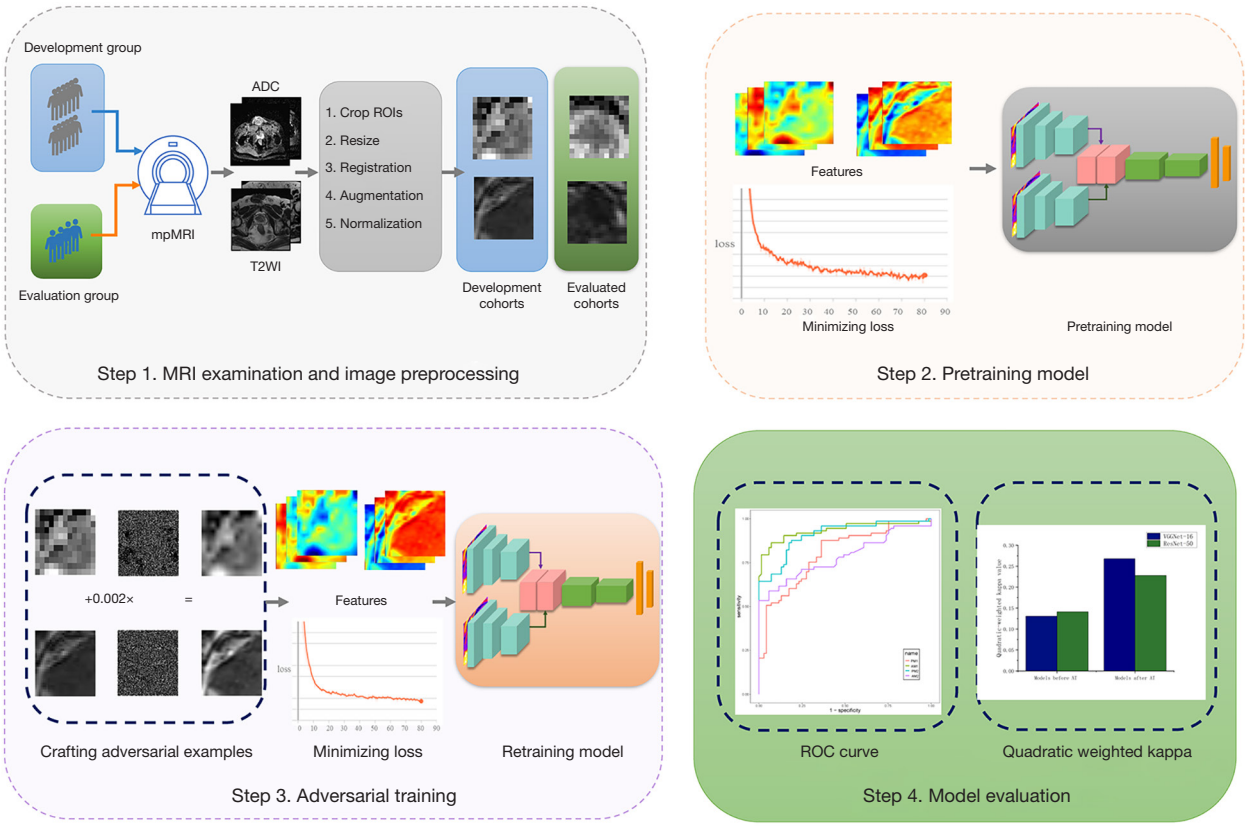

Figure 2 Overall study flow diagram. MRI, magnetic resonance imaging; ROI, region of interest; mpMRI, multiparametric MRI; ADC, apparent diffusion coefficient; T2WI, T2-weight imaging; ROC, receiver operating characteristic.

the reference of ultrasound (US) fusion-guided biopsies results and other clinical information. When MRI-suspected PCa and the target lesion for MRI/transrectal US fusion-guided biopsies were in the same sector, the image slice containing the largest lesion extent was selected and the lesion coordinates on the selected map slice were recorded.

The Gleason score on base of biopsy result is the single most powerful predictor of PCa prognosis. A pathologist with >20 years of experience blind to clinical information analyzed biopsies and defined the Gleason grade grouping (GGG) of the selected lesions (29). A total of 571 lesions with definite coordinates on the images were identified. Two cohorts were generated for the two tasks: Cohort 1 was used for csPCa classification, and Cohort 2 was used for PCa GGG grading. Specifically, for Cohort 1, the development set contained 429 lesions, including 99 lesions with a GGG of 1 and 330 lesions with a GGG >1. The test set had 69 lesions with a GGG of 1 and 73 lesions with a GGG >1. For Cohort 2, the development set contained 99, 76, 95, 47, and 112 lesions with GGGs of 1, 2, 3, 4, and 5, respectively. The test set contained 69, 20, 22, 21, and 10 lesions with GGGs of 1, 2, 3, 4, and 5, respectively.

Image preprocessing

Details on image preprocessing procedures are shown in Appendix 1. In brief, given the lesion coordinates, rectangular region of interest (ROI) patches around the lesions were cropped from T2WI sequences and ADC maps and resized to an image resolution of 224×224. Next, the ADC ROI patches were aligned to the T2WI patches. To avoid the imbalance of biased classification results toward the class with the most cases (9), we balanced the number of training images in the development cohorts for both binary and multivariate classification tasks by random translation and rotation to enhance the generalizability of the classification DL models (30,31). In addition, we flipped each ROI patch horizontally and vertically to augment the development sets. After data augmentation, there were 1,980 ROI patches for each image sequence in the development set

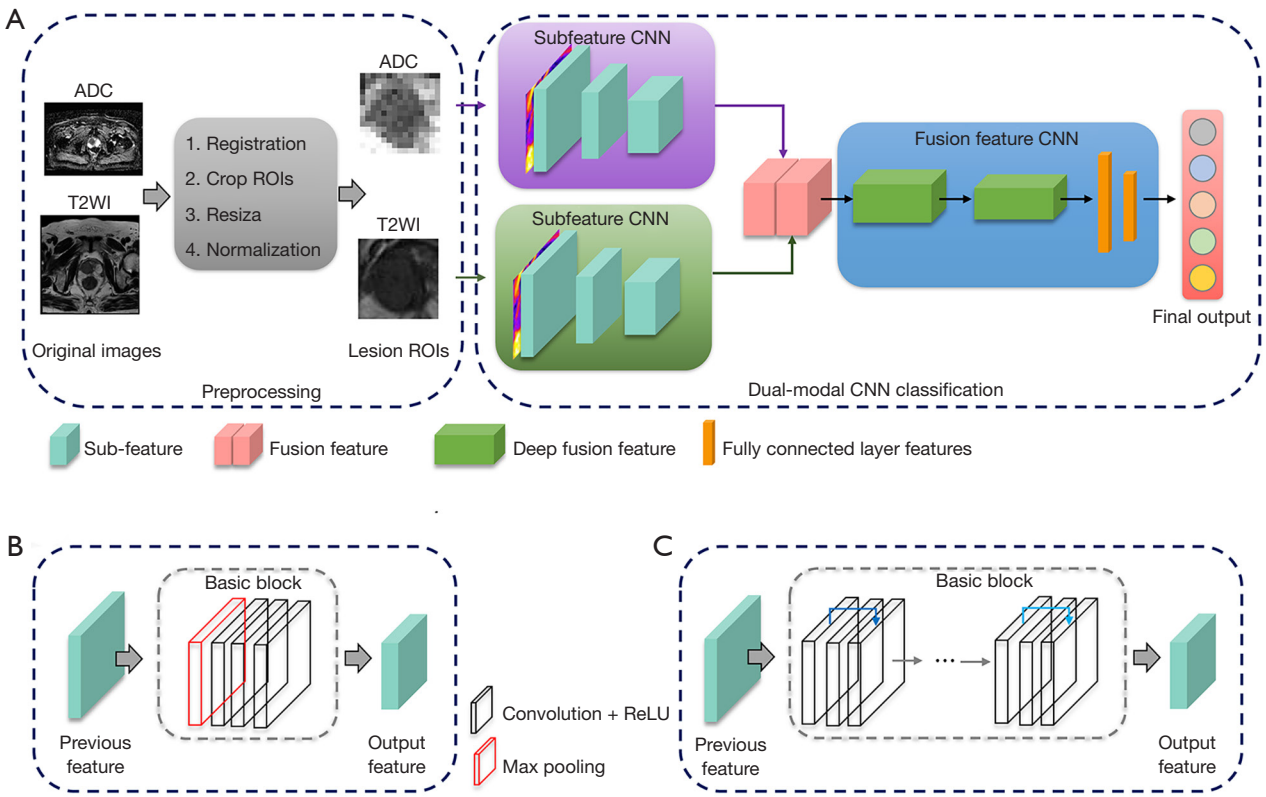

Figure 3 Illustration of our framework. (A) Workflow of the classification model for prostate cancer diagnosis based on dual-modal CNNs. We used VGGNet-based (B) and ResNet-based (C) basic blocks to extract deep features. ADC, apparent diffusion coefficient; CNNs, convolutional neural networks; ROI, region of interest; ReLU, rectified linear unit; T2WI, T2-weighted imaging; VGGNet, Visual Geometry Group network.

for the csPCa detection task and 1,680 for the PCa grading task. Finally, the intensities of both ADC and T2WI patches were normalized to handle the problem of inhomogeneous intensities for each modality among patients.

Network architecture

The workflow of the classification model for PCa diagnosis is shown in *Figure 3A*. Visual Geometry Group network (VGGNet)-16 (*Figure 3B*) and residual network (ResNet)-50 (*Figure 3C*) were the two baseline network architectures used to train the models in this study. In contrast to the standard VGGNet-16 or ResNet-50, we designed two parallel subnetworks for extracting sub-features for ADC and T2WI sequences, respectively, by using multiple basic blocks of VGGNet-16 or ResNet-50. The two subnetworks were then connected to fuse sub-features from ADC and T2WI sequences. Multiple deep basic blocks and fully connected layers were used to further extract deep fusion features. Finally, the predicted probabilities of the input patch pair for the classification tasks were obtained using a *softmax* function.

The experiments were conducted on four NVIDIIA RTX 2080 GPUs, and all procedures were implemented using PyTorch.

Model pretraining

Based on the two network architectures described above, we pretrained two binary classification DL models (PM1: pretraining VGGNet-16-based model 1 and PM2: pretraining ResNet-50-based model 2) for the detection of csPCa lesions (GGG =1 *vs.* GGG >1). Using the same training mechanism, we also trained two multivariate DL classification models (PM3: pretraining VGGNet-16-based model 3 and PM4: pretraining ResNet-50-based model 4) to identify the GGG of PCa lesions (GGG =1–5). The training process is described in Appendix 2.

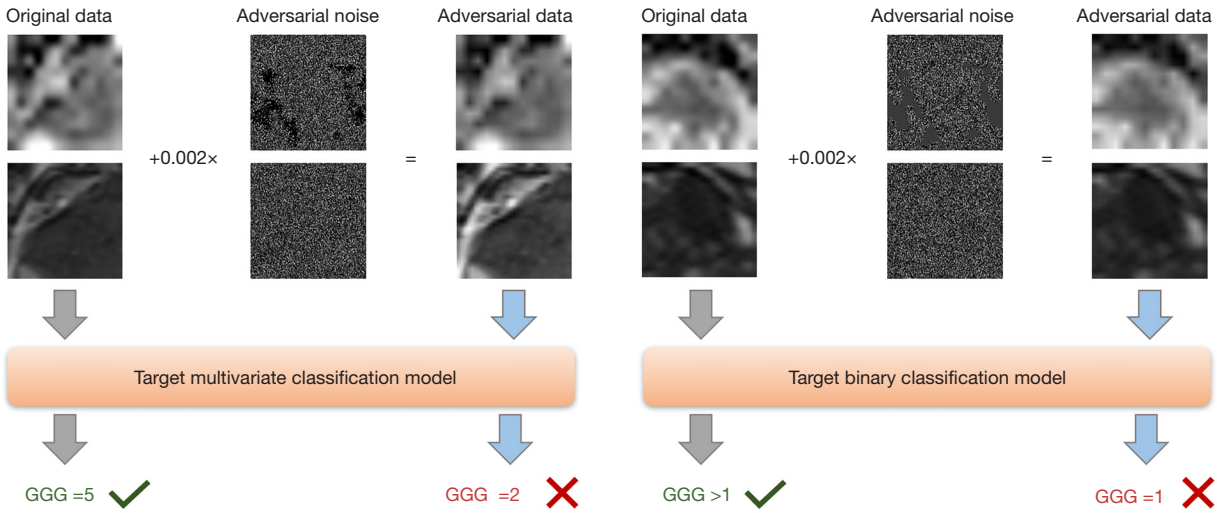

Figure 4 AEs generated for the proposed models. AEs were generated by adding subtle noise to correctly predicted images, which caused both the binary and multivariate classification models to incorrectly predict the classification. It is difficult to distinguish AEs from ground-truth natural examples with the naked eye, but the produced AEs did cause the model to produce misleading predictions. AEs, adversarial examples; GGG, Gleason grade group.

Adversarial training

After completing the model pretraining, we used AEs crafted by the decoupling direction and norm (DDN) method to implement AT. The DDN method, which won the Neural Information Processing Systems Adversarial Vision Challenge (2018) on non-targeted black-box attacks, can generate gradient-based AEs that induce misclassifications with small L2 norm distances by decoupling the direction and adding adversarial perturbations to the images (32).

Examples of the AEs used for AT are shown in *Figure 4*. Using the DDN-based AEs as new training sets, we retrained the binary classification DL models (AM1: VGGNet-16-based AT model 1 and AM2: ResNet-50-based AT model 2) and multivariate DL classification models (AM3: VGGNet-16-based AT model 3 and AM4: ResNet-50-based AT model 4) in the same manner as described in the model pretraining section.

Model evaluation

To verify whether AT can improve the effectiveness of the diagnostic model, we compared the diagnostic effectiveness of the DL methods before and after AT in the internal validation set and the test set. In addition, we evaluated the differences in performance of the internal validation set and the test set for each model, and used this as an index to evaluate the generalizability of the model.

Statistical analyses

A one-sample Kolmogorov-Smirnov test was used to check the assumption of normal distribution. An independent *t*-test was used for normally distributed data. A Mann-Whitney U test was used to assess non-normally distributed continuous variables. To evaluate the binary classification DL models for csPCa detection, the area under the receiver operating characteristic curve (AUC) was calculated. Using the cutoff value at the top left corner of the ROC curve, the accuracies, specificities, and sensitivities were identified. A comparison of sensitivity and specificity was performed using McNemar test. Delong's tests were conducted to compare differences in AUCs between models. The quadratic weighted kappa score κ (14) was used to verify the multivariate classification of DL models for PCa grading. This metric regards the GGG as the ordinal multiclass variable; an incorrectly estimated GGG, which is further from the ground truth, is more strongly penalized (7). The κ coefficients were assessed as follows: 0.01–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–0.99, almost perfect agreement.

Statistical analyses were performed using R (version

Table 1 Patient characteristics

Variables	Development cohort (n=297)	Evaluation cohort (n=99)
Median age [IQR] (years)	62 [55–72]	61 [51–71]
Median PSA [IQR] (ng/mL)	6.7 [4.6–10.1]	6.9 [5.1–12.1]
Scanner		
Verio	77	27
Skyra	175	58
Prisma	45	14
Number of patients with MRI-detected lesions		
1 lesion	212	70
2 lesions	45	18
3 lesions	33	8
4 lesions	7	3
Number of MRI-detected lesions		
Total	429	142
Peripheral zone	300	102
Transition zone	129	40
Gleason grade group (Gleason score)		
Gleason grade group 1 (GS 3 + 3)	99	69
Gleason grade group 2 (GS 3 + 4)	76	20
Gleason grade group 3 (GS 4 + 3)	95	22
Gleason grade group 4 (GS =8)	47	21
Gleason grade group 5 (GS >8)	112	10

PSA, prostate-specific antigen; IQR, interquartile range; MRI, magnetic resonance imaging; GS, Gleason score.

4.0.1, R Project for Statistical Computing, Vienna, Austria). Statistical significance was set at P<0.05.

Results

Patient characteristics

No adverse events occurred in this retrospective study. Patients in both development and test cohorts had no disease symptoms that could influence test accuracy. Detailed clinical and tumor characteristics, including age, PSA level, and lesion location, are summarized in *Table 1*. No significant differences in age or PSA level were found between the development and test cohorts (age: P=0.541; PSA: P=0.342).

Performance of binary classification DL models

The performance of the binary classification DL models before and after AT is shown in *Figure 5* and *Table 2*. For csPCa classification, the DL models after AT had significantly higher AUCs, sensitivities, specificities, and accuracies than those before AT in both the internal validation sets and test sets (P<0.001 for all comparisons). This suggests that AT can increase the diagnostic efficiency of the DL models for csPCa.

The diagnostic efficacies of PM1 and PM2 in the test set decreased by 10.6% and 4.6% in accuracy, 8.0% and 3.8% in sensitivity, and 11.3% and 4.4% in specificity, respectively, compared with those in the internal validation set. The AUCs were both 11% lower than those in the internal validation dataset. Conversely, for AM1 and AM2, the diagnostic efficacy in the test set decreased by approximately 2.0% in accuracy for both, 2.3% and 2.8% in sensitivity, respectively, and 0.3% and 1.4% in specificity, respectively, compared with the internal validation set. The corresponding AUCs were 3% and 5% lower, respectively, than those in the internal validation set.

Performance of the multivariate classification DL models

The performance of the multivariate classification DL models before and after AT is shown in *Figure 6*. In the internal validation set, the DL models before AT reached fair consistency between the predicted and true values {PM3: κ, 0.266 [95% confidence interval (CI): 0.152–0.379]; PM4: κ, 0.254 (95% CI: 0.159–0.390)}, whereas in the test set, these DL models only reached slight consistency between the predictions and the ground truth [PM3: κ, 0.196 (95% CI: 0.029–0.362); PM4: κ, 0.183 (95% CI: 0.015–0.351)]. The DL models after AT showed higher κ values than the DL models before AT in both the internal validation sets [AM3: 0.292 (95% CI: 0.178–0.405); AM4: 0.279 (95% CI: 0.163–0.396)] and test sets [AM3: 0.268 (95% CI: 0.109–0.427); AM4: 0.228 (95% CI: 0.068–0.389)].

The differences in κ values of AM3 and AM4 between the internal and test sets were much smaller than those of PM3 and PM4 (PM3: 0.070, PM4: 0.071, AM3: 0.024, and AM4: 0.051).

Discussion

The main finding of our study is that using AEs to train PCa DL models can effectively improve their PCa

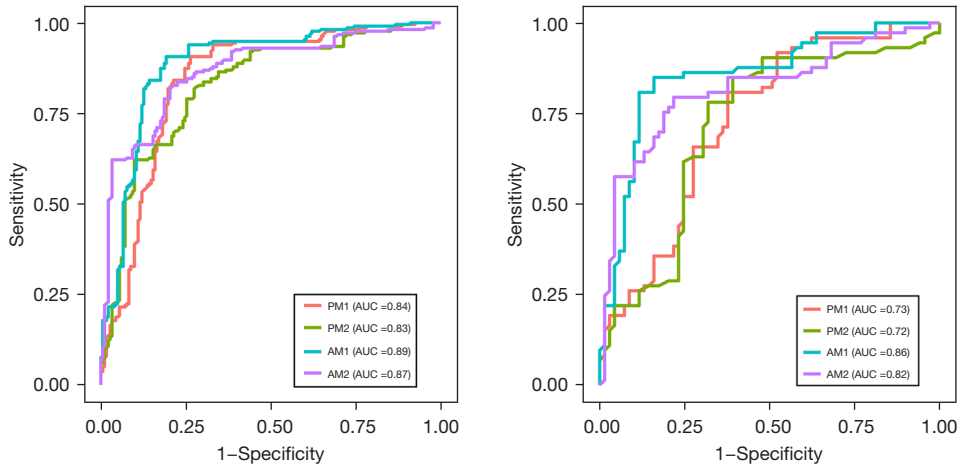

Figure 5 Performance of the binary classification deep learning models (PM1: VGGNet-16-based pretraining model; PM2: ResNet-50-based pretraining model; AM1: VGGNet-16-based AT model; AM2: ResNet-50-based AT model) in the internal verification (left) and external verification (right) datasets. AT, adversarial training; AUC, area under the curve; ResNet, residual network; VGGNet, Visual Geometry Group network.

Table 2 Performance of binary classification DL models

Datasets	Method	Positives	Negatives	TP	TN	FP	FN	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC
Internal validation set (n=396)	PM1	238	158	190	134	48	24	88.8	73.6	81.8	0.84
	AM1	229	167	194	147	35	20	90.7	80.8	86.1	0.89
	PM2	225	171	175	132	50	39	81.9	72.5	77.8	0.83
	AM2	213	183	176	145	37	38	82.2	79.7	80.8	0.87
Test set (n=142)	PM1	85	57	59	43	26	14	80.8	62.3	71.2	0.73
	AM1	78	64	65	56	13	8	88.4	80.5	84.5	0.86
	PM2	79	63	57	47	22	16	78.1	68.1	73.2	0.72
	AM2	73	69	58	54	15	15	79.5	78.3	78.9	0.82

DL, deep learning; TP, true positive; TN, true negative; FP, false positive; FN, false negative; AUC, area under the receiver operating characteristic curve; AM1, VGGNet-16-based AT model 1; AM2, ResNet-50-based AT model 2; PM1, pretraining VGGNet-16-based model 1; PM2, pretraining ResNet-50-based model 2; ResNet, residual network; VGGNet, Visual Geometry Group network.

classification ability and generalizability.

In our study, AT-based DL models for PCa classification showed better performance than those without AT in both the internal and test sets. For csPCa classification, AM1 and AM2 had comparable AUCs of 0.86 and 0.82 on external evaluation, whereas for the task of PCa GGG, AM3 and AM4 had fair agreement, with κ values of 0.268 and 0.228, which are within the range of previously reported results (–0.245 to 0.277) in the PROSTATEx-2 2017 challenge specially designed for PCa GGG (6,7).

Although various DL methods for PCa classification have been proposed (1,5,6,11,14,33,34), most studies only conducted internal validation of their proposed methods (1,5,14,34,35). Therefore, the generalizability of these models is unclear. Several studies performed external verification (6,33,36), but the training and test sets consisted of patients from the same medical center or images acquired from a single manufacturer, which does not consider potential differences in scanners or between medical centers. Thus, the generalizability of the proposed models may still be overestimated. In contrast to previous studies, to better evaluate the generalizability of our proposed models, our

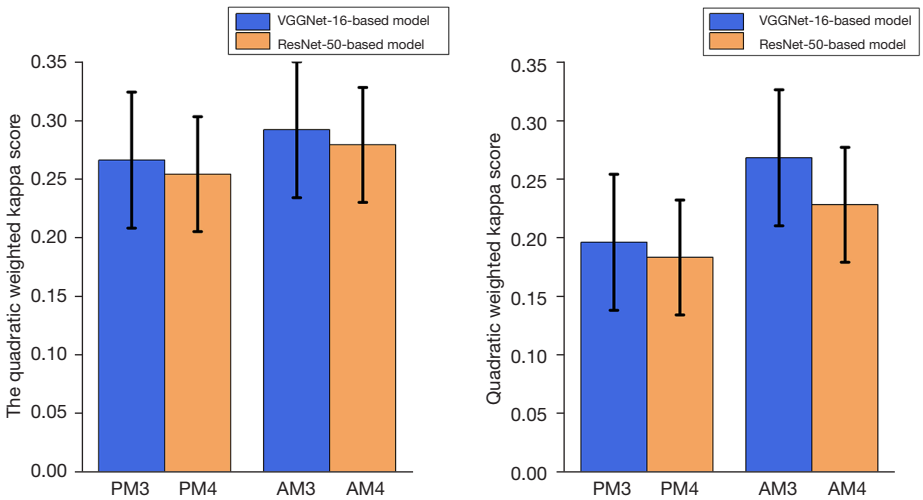

Figure 6 Performance of the multivariate classification deep learning models (PM3: VGGNet-16-based pretraining model; PM4: ResNet-50-based pretraining model; AM3: VGGNet-16-based AT model; AM4: ResNet-50-based AT model) in the internal validation datasets (left) and test sets (right). The kappa calculation results range from −1 to 1, but κ ranges usually between 0 and 1 and is further divided into five groups to express different consistency levels (0–0.20: slight consistency, 0.21–0.40: fair consistency, 0.41–0.60: moderate consistency, 0.61–0.80: substantial consistency, and 0.81–0.99: almost perfect consistency). Error bars show the standard error of κ values. AT, adversarial training; ResNet, residual network; VGGNet, Visual Geometry Group network.

study had both internal and external verifications that were performed to evaluate the performance of the DL models. Moreover, the development set data and test set data came from different medical centers; thus, both the diagnostic efficacy and generalizability of the model can be better evaluated.

We found that the DL models before AT, which performed well in the internal verification, had reduced diagnostic efficiencies in the external verification. This indicates that data augmentation and normalization strategies to improve the generalizability of DL models (35) are insufficient. Compared with those of the pretrained DL models, the performance differences between the internal validation and test sets of the DL models after AT were much smaller. This implies that AT can potentially improve the generalizability of PCa classification DL models.

Since good generalizability is the premise by which these DL models can be applied in the clinic (23,37), improving the generalizability and robustness of DL models is a core issue not yet resolved (19). The quality of prostate MR images is easily affected by various factors, such as metal artifacts, magnetic field inhomogeneity, involuntary patient movement, and differences between software and hardware (38). All of these factors may cause noise (18) and are, therefore, potential disturbances to classification models that result in reduction in the PCa classification accuracy. Using adversarial attack methods can identify the noises that maximize the classification error loss and add them to the original examples to generate AEs (17-19). The target models retrained with these AEs may be able to more precisely classify test examples with different noises. In the present study, we chose a DDN attack to craft AEs. This method effectively and quickly crafts AEs to retrain the target model for improving the adversarial generalizability of DL models. Moreover, using this method, the L2 norm distances between the original images and their corresponding AEs are relatively small, largely avoiding affecting the retrained model's predictions on examples without noise (32). This could be why AT can improve the generalizability of the PCa classification DL models.

As a tentative study, our study has several limitations. First, because of technical and equipment limitations, targeted biopsy was used as a reference in the development cohort rather than prostatectomy. Whole-mount serial sections may improve the accuracy of the agreement between MR images and histopathology, and minimize biases for the assessment of PCa detection performance of DL models. Second, various studies indicated that DL models, including DWI and DCE-MRI, can provide

additional different and complementary information to improve the accuracy of PCa diagnosis. However, the optimal b-value of DWI and the best phase of DCE-MRI are unclear; therefore, for the choice of data, as in many previous studies, we selected ADC and T2WI sequences for model training and evaluation. Third, in the selection of the network framework and construction of the model, this study only examined the role of AT in 2D DL models for PCa classification based on VGGNet-16 and ResNet-50. The influence of AT on 3D models based on other types of networks for more tasks, such as PCa detection and segmentation, still requires further experimental demonstration. Finally, in the selection of AT methods, we used the representative L2-norm adversarial attack to craft adversarial noise without introducing additional constraints to simulate the unique noise of MR images, such as artifacts and deformation. In the future, we will consider adding style transfer supervision to craft MRI-specific adversarial noise, which may be used to further improve the adversarial robustness of the classification model.

Conclusions

Using adversarial samples to retrain machine learning models for PCa classification on MR images can effectively improve the generalizability of these models and improve their classification abilities.

Acknowledgments

Funding: This study received funding from the National Natural Science Foundation of China (Nos. 81901845 and 81671791), Science Foundation of Shanghai Jiao Tong University Affiliated Sixth People's Hospital (No. 201818), and Shanghai Key Discipline of Medical Imaging (No. 2017ZZ02005).

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-21-1089/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-21-1089/coif). XYG is employed by Xi'an OUR United Co., Ltd. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This retrospective, multicenter study was approved by the local ethics committee of our institution. Informed consent was obtained from all patients. The name of registry and registration number: Application of artificial intelligence in MRI of prostate (ChiCTR2100041834).

References

1. Yuan Y, Qin W, Buyyounouski M, Ibragimov B, Hancock S, Han B, Xing L. Prostate cancer classification with multiparametric MRI transfer learning model. Med Phys 2019;46:756-65.
2. Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, Tempany CM, Choyke PL, Cornud F, Margolis DJ, Thoeny HC, Verma S, Barentsz J, Weinreb JC. Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. Eur Urol 2019;76:340-51.
3. Park H, Kim SH, Kim JY. Dynamic contrast-enhanced magnetic resonance imaging for risk stratification in patients with prostate cancer. Quant Imaging Med Surg 2022;12:742-51.
4. Mayer R, Simone CB 2nd, Turkbey B, Choyke P. Correlation of prostate tumor eccentricity and Gleason scoring from prostatectomy and multi-parametric-magnetic resonance imaging. Quant Imaging Med Surg 2021;11:4235-44.
5. Cao R, Mohammadian Bajgiran A, Afshari Mirak S, Shakeri S, Zhong X, Enzmann D, Raman S, Sung K. Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet. IEEE Trans Med Imaging 2019;38:2496-506.
6. Abraham B, Nair MS. Computer-aided classification of

prostate cancer grade groups from MRI images using texture features and stacked sparse autoencoder. Comput Med Imaging Graph 2018;69:60-8.
7. Armato SG 3rd, Huisman H, Drukker K, Hadjiiski L, Kirby JS, Petrick N, Redmond G, Giger ML, Cha K, Mamonov A, Kalpathy-Cramer J, Farahani K. PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. J Med Imaging (Bellingham) 2018;5:044501.
8. Chen Q, Xu X, Hu S, Li X, Zou Q, Li Y. A Transfer learning approach for classification of clinical significant prostate cancers from mpMRI scans. Proceedings Volume 10134, Medical Imaging 2017: Computer-Aided Diagnosis. SPIE, 2017:abstr 101344F.
9. Le MH, Chen J, Wang L, Wang Z, Liu W, Cheng KT, Yang X. Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. Phys Med Biol 2017;62:6497-514.
10. Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H. Computer-aided detection of prostate cancer in MRI. IEEE Trans Med Imaging 2014;33:1083-92.
11. Liu S, Zheng H, Feng Y, Li W. Prostate cancer diagnosis using deep learning with 3D multiparametric MRI. Armato SG, Petrick NA. editors. Proceedings Volume 10134, Medical Imaging 2017: Computer-Aided Diagnosis. SPIE, 2017:abstr 1013428.
12. Mehrtash A, Sedghi A, Ghafoorian M, Taghipour M, Tempany CM, Wells WM 3rd, Kapur T, Mousavi P, Abolmaesumi P, Fedorov A. Classification of Clinical Significance of MRI Prostate Findings Using 3D Convolutional Neural Networks. Proc SPIE Int Soc Opt Eng 2017;10134:101342A.
13. Tsehay YK, Lay NS, Roth HR, Wang X, Kwak JT, Turkbey BI, Pinto PA, Wood BJ, Summers RM. Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images. Proceedings Volume 10134, Medical Imaging 2017: Computer-Aided Diagnosis. SPIE, 2017:abstr 1013405.
14. Vente C, Vos P, Hosseinzadeh M, Pluim J, Veta M. Deep Learning Regression for Prostate Cancer Detection and Grading in Bi-Parametric MRI. IEEE Trans Biomed Eng 2021;68:374-83.
15. Vos PC, Hambrock T, Barenstz JO, Huisman HJ. Computer-assisted analysis of peripheral zone prostate lesions using T2-weighted and dynamic contrast enhanced T1-weighted MRI. Phys Med Biol 2010;55:1719-34.
16. Jin L, Tan F, Jiang S. Generative Adversarial Network Technologies and Applications in Computer Vision. Comput Intell Neurosci 2020;2020:1459107.
17. Zhang Y, Tian X, Li Y, Wang X, Tao D. Principal Component Adversarial Example. IEEE Trans Image Process 2020;29:4804-15.
18. Zhang J, Li C. Adversarial Examples: Opportunities and Challenges. IEEE Trans Neural Netw Learn Syst 2020;31:2578-93.
19. Yuan X, He P, Zhu Q, Li X. Adversarial Examples: Attacks and Defenses for Deep Learning. IEEE Trans Neural Netw Learn Syst 2019;30:2805-24.
20. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363:1287-9.
21. Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access 2018;6:14410-30.
22. Mustafa A, Khan SH, Hayat M, Shen J, Shao L. Image Super-Resolution as a Defense Against Adversarial Attacks. IEEE Trans Image Process 2019;29:1711-24.
23. Hirano H, Minagi A, Takemoto K. Universal adversarial attacks on deep neural networks for medical image classification. BMC Med Imaging 2021;21:9.
24. Bai X, Yang M, Liu Z. On the robustness of skeleton detection against adversarial attacks. Neural Netw 2020;132:416-27.
25. Arnab A, Miksik O, Torr PHS. On the Robustness of Semantic Segmentation Models to Adversarial Attacks. IEEE Trans Pattern Anal Mach Intell 2020;42:3040-53.
26. Allyn J, Allou N, Vidal C, Renou A, Ferdynus C. Adversarial attack on deep learning-based dermatoscopic image recognition systems: Risk of misdiagnosis due to undetectable image perturbations. Medicine (Baltimore) 2020;99:e23568.
27. Liao F, Liang M, Dong Y, Pang T, Hu X, Zhu J. Defense against adversarial attacks using high-level representation guided denoiser. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018:1778-87.
28. Wang X, Chen H, Xiang H, Lin H, Lin X, Heng PA. Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. Med Image Anal 2021;70:102010.
29. Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, Vickers AJ, Parwani AV, Reuter VE, Fine SW, Eastham JA, Wiklund P, Han M, Reddy CA, Ciezki JP, Nyberg T, Klein EA. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the

Gleason Score. Eur Urol 2016;69:428-35.
30. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Trans Med Imaging 2016;35:1285-98.
31. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Jianming Liang. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Trans Med Imaging 2016;35:1299-312.
32. Rony J, Hafemann LG, Oliveira LS, Ben Ayed I, Sabourin R, Granger E. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019:4317-25.
33. Cao R, Zhong X, Afshari S, Felker E, Suvannarerg V, Tubtawee T, Vangala S, Scalzo F, Raman S, Sung K. Performance of Deep Learning and Genitourinary Radiologists in Detection of Prostate Cancer Using 3-T Multiparametric Magnetic Resonance Imaging. J Magn Reson Imaging 2021;54:474-83.
34. Jensen C, Carl J, Boesen L, Langkilde NC, Østergaard LR. Assessment of prostate cancer prognostic Gleason grade group using zonal-specific features extracted from biparametric MRI using a KNN classifier. J Appl Clin Med Phys 2019;20:146-53.
35. Aldoj N, Lukas S, Dewey M, Penzkofer T. Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network. Eur Radiol 2020;30:1243-53.
36. Schelb P, Kohl S, Radtke JP, Wiesenfarth M, Kickingereder P, Bickelhaupt S, Kuder TA, Stenzinger A, Hohenfellner M, Schlemmer HP, Maier-Hein KH, Bonekamp D. Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. Radiology 2019;293:607-17.
37. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology 2018;286:800-9.
38. Hu L, Zhou DW, Fu CX, Benkert T, Jiang CY, Li RT, Wei LM, Zhao JG. Advanced zoomed diffusion-weighted imaging vs. full-field-of-view diffusion-weighted imaging in prostate cancer detection: a radiomic features study. Eur Radiol 2021;31:1760-9.

Cite this article as: Hu L, Zhou DW, Guo XY, Xu WH, Wei LM, Zhao JG. Adversarial training for prostate cancer classification using magnetic resonance imaging. Quant Imaging Med Surg 2022;12(6):3276-3287. doi: 10.21037/qims-21-1089

Appendix 1 Image preprocessing procedures

Cropping, resizing

The apparent diffusion coefficient (ADC) data has an original in-plane resolution of 2.10×1.60 mm^2 and a matrix size of 178×178, whereas T2-weighted imaging (T2WI) data has an in-plane resolution of 0.625×0.625 mm^2 and a matrix size of 320×256. The ADC data was first resampled to an in-plane resolution of 0.625×0.625 mm^2 with a matrix size of 598×456. Then, a rectangular region of interest (ROI) region with a matrix size of 40×40 around the lesions was cropped from T2WI sequences and ADC maps according to the lesion coordinates and were scaled to an image resolution of 224×224. Next, ADC ROIs were aligned to those of T2WI images using the affine transformation implemented by the Advanced Normalization Tools (ANTs) (https://github.com/ANTsX/ANTs).

Data augmentation

To avoid the imbalance issue of biased classification results toward the class with the most training samples, we balanced the number of training samples in the five classes by random translation and rotation. By this design, for multivariate classification task, all classes of the training sample had 112 ROI patches. For binary classification task, there are 330 ROI patches for the two classes of Gleason grade grouping (GGG) =1 and GGG >1. In addition, for each ROI patch, we flipped it horizontally and vertically to augment the training set. Therefore, by the above processes, we had a total of 1,680 (112×5×3) ROI patches in the training set for both modalities for multivariate classification task, and total of 1,980 (330×2×3) ROI patches in the training set for both modalities for the binary classification task.

Normalization

Normalization transforms an n-dimensional grayscale image $I:\{\mathbb{X}\subseteq\mathbb{R}^n\}\to\{Min,\cdots,Max\}$ with intensity values in the range (*Min*, *Max*), into a new image $I_N:\{\mathbb{X}\subseteq\mathbb{R}^n\}\to\{newMin,\cdots,newMax\}$ with intensity values in the range (*new Min*, *new Max*).

The normalization of a grayscale digital image is performed according to the formula:

$$I_N=(I-Min)\frac{newMax-newMin}{Max-Min}+newMin \quad [1]$$

Where *new Max* is set to 1 and *new Min* is set to 0 in this paper.

https://dx.doi.org/10.21037/qims-21-1089

Appendix 2 The training process

The DL model first used a pair of ROI patches of ADC and T2WI as inputs to obtain two sub-features. Then, at the fusion stage, an element-wise summation was performed on the corresponding sub-features of ADC and T2WI. Next, fusion features were input into the fusion feature convolutional neural network (CNN) to obtain the final output. The training process could be formulated as follows:

Given a pair of ROI patches (x_{ADC},x_{T2}) of ADC and T2WI, we first obtained the sub-features f_{ADC} and f_{T2}. Then, we fused the sub-features into fusion feature f_F via an element-wise summation. Next, f_F was further used to extract deep fusion feature to obtain the final output:

$$\hat{y} = \sigma\left(softmax\, \mathcal{C}\left(x_{ADC}, x_{T2}\right)\right) \quad [2]$$

where $\mathcal{C}(\cdot)$ denoted the DL classification model, $softmax(\cdot)$ represented the softmax function, and $\sigma(\cdot)$ the operation of selecting the item with the highest probability.

We used a cross-entropy loss to supervise the training process:

$$\mathcal{L}_c = -\left[y log\hat{y} + \left(1-y\right) log\left(1-\hat{y}\right)\right] \quad [3]$$

Where y denoted the ground-truth class label corresponding to the input (x_{ADC},x_{T2}). For training the AT model, we replaced ADC and T2WI data with their AEs and kept the other training settings unchanged.

Note that in our paper, both the AT and non-AT model are trained starting from random initial model parameters. The "retrain" in our paper means training the same model architecture starting from the initial model parameters with the same training settings. All models (i.e., VGG-16 and ResNet-50) for both AT and non-AT were trained using SGD with momentum 0.9 and weight decay 2×10^{-4}. The training epoch number was 100 for both AT and non-AT model. The initial learning rate was 0.01, divided by 10 at the 75th and 90th epoch.

We drew the accuracy curve of the test set and the training set during the training process to observe whether the model was over-fitted. In the training process, as the number of iterations increases, the accuracy of the test set and the accuracy of the training set consistently increase, and eventually tend to be stable. This shows no overfitting.

https://dx.doi.org/10.21037/qims-21-1089