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