Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

Deep learning-based radiomic features for improving neoadjuvant chemoradiation response prediction in locally advanced rectal cancer

Jie Fu[1] , Xinran Zhong[2] , Ning Li[3] , Ritchell Van Dams[1] , John Lewis[4] , Kyunghyun Sung[2] ,
Ann C Raldow[1], Jing Jin[3] and X Sharon Qi[1]

[1] Department of Radiation Oncology, University of California, Los Angeles, CA 90095, United States of America
[2] Department of Radiological Sciences, University of California, Los Angeles, CA 90095, United States of America
[3] Department of Radiation Oncology, Cancer Hospital Chinese Academy of Medical Sciences, Beijing, People's Republic of China
[4] Cedars Sinai Medical Center, Los Angeles, CA 90048, United States of America

E-mail: jiefu@mednet.ucla.edu and xqi@mednet.ucla.edu

Abstract

Radiomic features achieve promising results in cancer diagnosis, treatment response prediction, and survival prediction. Our goal is to compare the handcrafted (explicitly designed) and deep learning (DL)-based radiomic features extracted from pre-treatment diffusion-weighted magnetic resonance images (DWIs) for predicting neoadjuvant chemoradiation treatment (nCRT) response in patients with locally advanced rectal cancer (LARC). 43 Patients receiving nCRT were included. All patients underwent DWIs before nCRT and total mesorectal excision surgery 6–12 weeks after completion of nCRT. Gross tumor volume (GTV) contours were drawn by an experienced radiation oncologist on DWIs. The patient-cohort was split into the responder group ($n = 22$) and the non-responder group ($n = 21$) based on the post-nCRT response assessed by postoperative pathology, MRI or colonoscopy. Handcrafted and DL-based features were extracted from the apparent diffusion coefficient (ADC) map of the DWI using conventional computer-aided diagnosis methods and a pre-trained convolution neural network, respectively. Least absolute shrinkage and selection operator (LASSO)-logistic regression models were constructed using extracted features for predicting treatment response. The model performance was evaluated with repeated 20 times stratified 4-fold cross-validation using receiver operating characteristic (ROC) curves and compared using the corrected paired t-test. The model built with handcrafted features achieved the mean area under the ROC curve (AUC) of 0.64, while the one built with DL-based features yielded the mean AUC of 0.73. The corrected paired t-test on AUC showed P-value < 0.05. DL-based features extracted from pre-treatment DWIs achieved significantly better classification performance compared with handcrafted features for predicting nCRT response in patients with LARC.

1. Introduction

Colorectal cancer is the third most common cancer diagnosed and the second most common cause of cancer deaths in the US (Siegel *et al* 2019). Rectal cancer accounts for about 30% of all colorectal cancer diagnoses (Siegel *et al* 2019). Treatment for rectal cancer is based largely on the stage at diagnosis. Locally advanced rectal cancer (LARC) is commonly treated with neoadjuvant chemoradiation therapy (nCRT) followed by total mesorectal excision (TME) and adjuvant chemotherapy (Kapiteijn *et al* 2001, van de Velde *et al* 2014). Tumor response to nCRT is associated with recurrence and survival and can serve as a prognostic factor (Quah *et al* 2008, Trakarnsanga *et al* 2014). 15%–27% of patients who undergo such treatment achieve pathologic complete response (pCR) (Maas *et al* 2010). TME is a highly invasive procedure with the potential risk of morbidity and functional complications. Achieving early prediction of tumor response using

pre-treatment noninvasive approaches may allow for design of individualized chemo-radiation treatment and potential avoidance of TME following nCRT for patients.

Magnetic resonance imaging (MRI) is widely used in rectal cancer diagnosis and staging as it provides excellent soft tissue contrast for tissue characterization. Specifically, increasing evidence has shown that diffusion-weighted images (DWIs), providing tissue cellularity information, aids the assessment of rectal cancer response to neoadjuvant treatment (Schurink *et al* 2019). DWI is recommended to be routinely acquired in clinical guidelines (Beets-Tan *et al* 2018). The interpretation of DWI has gradually shifted from qualitative evaluation to quantitative assessment. For example, the apparent diffusion coefficient (ADC) map was one major quantitative map calculated from DWI. However, several studies showed that the mean pretreatment tumor ADC value was not a reliable indicator for predicting treatment response (Kim *et al* 2011, Amodeo *et al* 2018).

Radiomics is an emerging field of studies where a large number of medical image features are extracted in order to achieve better clinical diagnosis or decision support (Lambin *et al* 2017). The conventional radiomics analysis typically involves extraction and analyzing quantitative imaging features from the previously defined region of interests (ROI) on one or multiple image modalities with the ultimate goal to obtain predictive or prognostic models. Previous studies showed that handcrafted, or explicitly designed, features extracted from the ADC ROI have predictive power for early nCRT treatment response in LARC patients (Nie *et al* 2016, Horvat *et al* 2018). However, handcrafted features are lower-order image features and limited to current expert knowledge. Another type of radiomic feature is deep learning (DL)-based extracted from the pre-trained convolutional neural networks (CNN) via transfer learning (LeCun *et al* 2015, Afshar *et al* 2019). Several studies have demonstrated that the DL-based features showed promising performance in breast cancer diagnosis, ovarian cancer recurrence prediction, and glioblastoma multiforme survival prediction (Lao *et al* 2017, Antropova *et al* 2017, Wang *et al* 2019). To our knowledge, no published study has investigated the DL-based features for managing LARC patients.

In this work, we first aimed to construct radiomics classifiers based on the handcrafted and DL-based radiomic features extracted from pre-treatment DWIs. Then, we compared the performance of the two classifiers to predict post-nCRT response in patients with LARC.

2. Materials and methods

2.1. Dataset

We identified 43 consecutive patients with locally advanced rectal cancer (LARC) treated from December 2015 to December 2016 at a single institution. All patients received concurrent capecitabine with a total prescription dose of 50 Gy in 25 fractions, followed by the TME surgery after 6–12 weeks of the nCRT completion. The resection specimens were evaluated by an expert pathologist. Patients were separated into good responders (GR) and non-GR groups based on the postoperative pathology report, MRI or colonoscopy. The GR group consisted of patients with either complete response (evaluated by pathology or MRI and colonoscopy) or partial response (assessed by pathology), and the non-GR group consisted of patients with stable disease (assessed by pathology) and progressive disease (confirmed by CT/MR).

All patients underwent pre-treatment DWIs before the nCRT. The DWI images were acquired using single-shot echo planar imaging (ssEPI) sequence on two 3-Tesla MR scanners (Discovery MR750 and Signa HDxt, GE Healthcare). MR imaging parameters are summarized in table 1. For each patient, the ADC map was computed using the equation $ADC = -\frac{1}{800} \ln\left(\frac{S}{S_0}\right)$, where S_0 and S correspond to MR voxel intensities at b-values of 0 s mm^{-2} and 800 s mm^{-2}. Gross tumor volume (GTV) of the primary tumor was manually delineated on the DWI image with the b-value of 800 s mm^{-2} by a board-certified oncologist with 5 year experience.
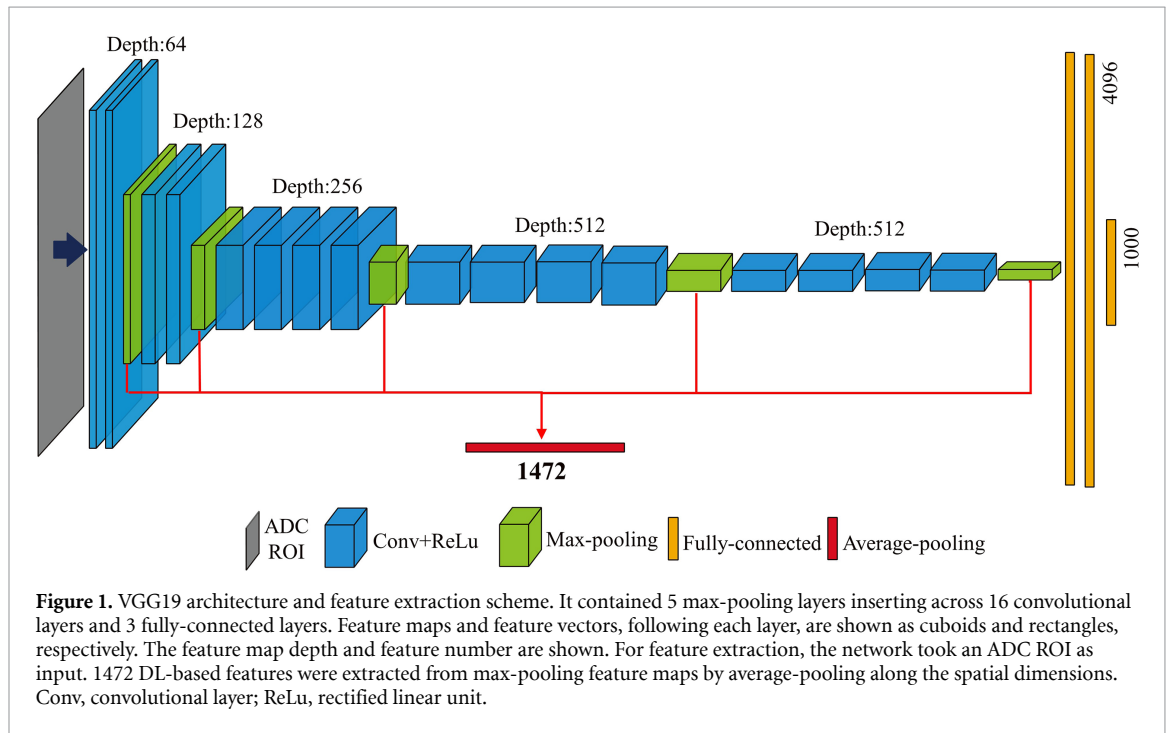
2.2. Feature extraction

2.2.1. Handcrafted features

105 handcrafted features were extracted from the ADC map within the GTV contour for each patient using PyRadiomics package (version 2.1.2) (van Griethuysen *et al* 2017). Extracted features consisted of 14 shape-based features, 18 first-order statistic features, and 73 textural (second-order statistic) features. The methods used for extracting textural features were gray level co-occurrence matrix, gray-level size zone matrix, gray level run length matrix, gray level dependence matrix, and neighborhood gray-tone difference matrix. Shape-based features describe the shape characteristics of the GTV contour. First-order statistic features describe the distribution of voxel intensities within the GTV contour. Textural features describe the patterns or second-order spatial distributions of the voxel intensities.

Table 1. MR imaging parameters.

Scanner model	Patient number	TR/TE (ms)	Matrix	Field of view (mm^2)	Transvers spatial resolution (mm^2)	Slice thickness (mm)	*b* value (s/mm^2)
Discovery MR750	36	2600/74	256×256	380^2 or 400^2	1.48^2 or 1.56^2	5	0, 500, 800, 1000
Signa HDxt	7	4500–6000/64–67	256×256	320^2–400^2	1.25^2–1.56^2	5 or 6	0, 800

Figure 1. VGG19 architecture and feature extraction scheme. It contained 5 max-pooling layers inserting across 16 convolutional layers and 3 fully-connected layers. Feature maps and feature vectors, following each layer, are shown as cuboids and rectangles, respectively. The feature map depth and feature number are shown. For feature extraction, the network took an ADC ROI as input. 1472 DL-based features were extracted from max-pooling feature maps by average-pooling along the spatial dimensions. Conv, convolutional layer; ReLu, rectified linear unit.

2.2.2. DL-based features

The publicly-available pre-trained CNN, VGG19 (Simonyan and Zisserman 2014), was used to extract DL-based features. The network was trained using approximately 1.2 million images from the ImageNet database (Russakovsky *et al* 2015) for classifying nature images into 1000 objects. As the natural objects used for training VGG19 varied in their physical size, the extracted DL-based features using pre-trained VGG19 may be less sensitive to image spatial resolution (pixel size) compared to other factors like image gradients. Figure 1 shows the network architecture. It contained 16 convolutional layers followed by 3 fully-connected layers. 5 max-pooling layers were inserted across convolutional and fully-connected layers to reduce model parameter number for controlling overfitting and help achieve partial invariance to small translations. For each patient, a 2D square region of interest (ROI) was selected from the transverse slice that contains the largest tumor area. The ROI center was set as the center of the smallest bounding box covering the 2D tumor. The ROI size was set as the maximum dimension of the smallest bounding box. The ADC ROI was extrapolated to 224 by 224 for matching the original VGG19 design. The intensities of the ROI were converted to the range [0, 255]. Resampled ROI was copied into a 3-channel image and then inputted into the pre-trained model for feature extraction. We adopted the feature extraction method proposed by Antropova *et al* (2017). As shown in figure 1, five DL-based feature vectors were extracted by average-pooling the feature maps after max-pooling layers. Each feature vector was normalized with its Euclidean norm and then concatenated to one feature vector, which was normalized again to acquire the feature vector consisting of 1472 features. After extracting features for all patients, a cutoff on feature variance was used to pre-select 105 DL-based features out of 1472 features with the highest variance to train the prediction models.

2.3. Classification and evaluation

The least absolute shrinkage and selection operator (LASSO) penalized logistic regression (Tibshirani 1996, Wu *et al* 2009) was used for classification using radiomic features (Python version 2.7.13). The LASSO regularization was selected to handle the high feature dimension. The handcrafted classifier and DL-based

Table 2. Patient clinical characteristics; GR, good responder, nGR, non-good responder, SD, standard deviation.

Characteristic	GR ($n = 22$)	nGR ($n = 21$)	Total ($n = 43$)
Gender (male/female)	14/8	14/7	28/15
Age (mean, SD, in years)	53.7 (9.1)	54.9 (10.9)	54.3 (10.3)
Pre-nCRT TNM staging			
T stage (2/3/4)	1/18/3	1/16/4	2/34/7
N stage (0/1/2)	5/11/6	0/9/12	5/20/18
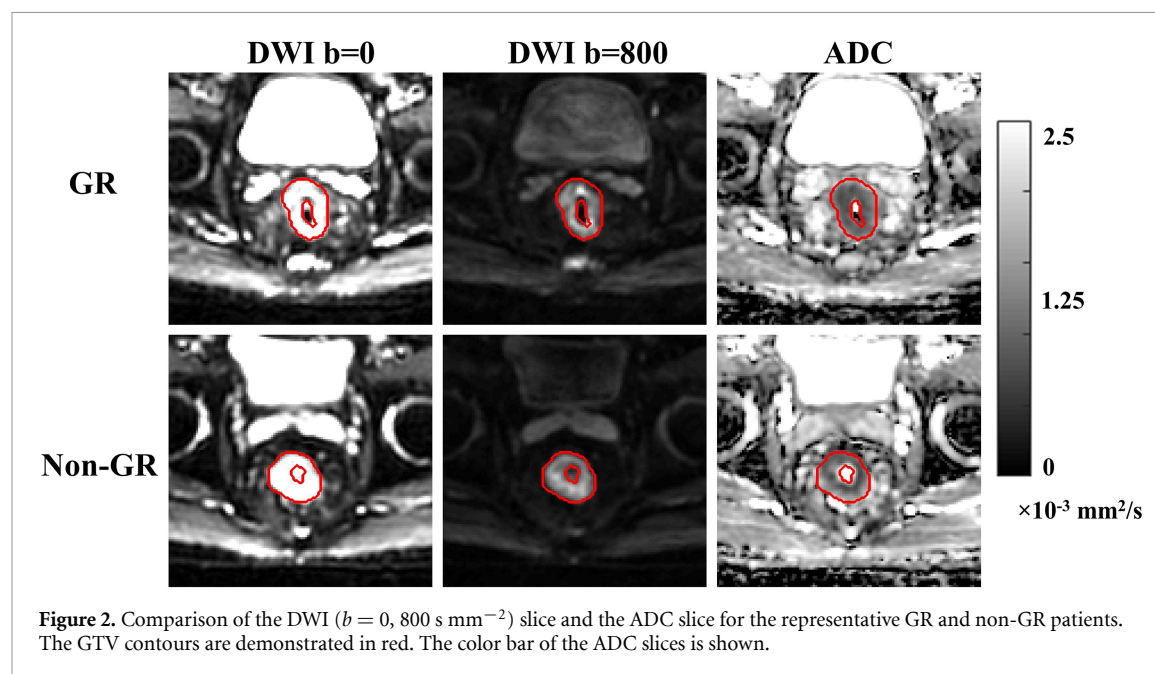
Figure 2. Comparison of the DWI ($b = 0, 800$ s mm^{-2}) slice and the ADC slice for the representative GR and non-GR patients. The GTV contours are demonstrated in red. The color bar of the ADC slices is shown.

classifier were trained using handcrafted features and DL-based features, respectively. Regularization parameter was optimized by grid searching with repeated 20 times stratified 4-fold cross-validation. For each cross-validation, stratified random sampling was used to split the patient cohort was into 4 folds, where 3 folds were used as the training set to train the classifier and the remaining one as the testing set for evaluation.

The performances of the handcrafted and DL-based classifiers were evaluated using the average area under the receiver operating curve (AUC) of 20 cross-validation repetitions. The corrected paired *t*-test (Bouckaert and Frank 2004) was conducted to compare the AUC results for two classifiers. *P*-value < 0.05 was considered to indicate a significant difference.
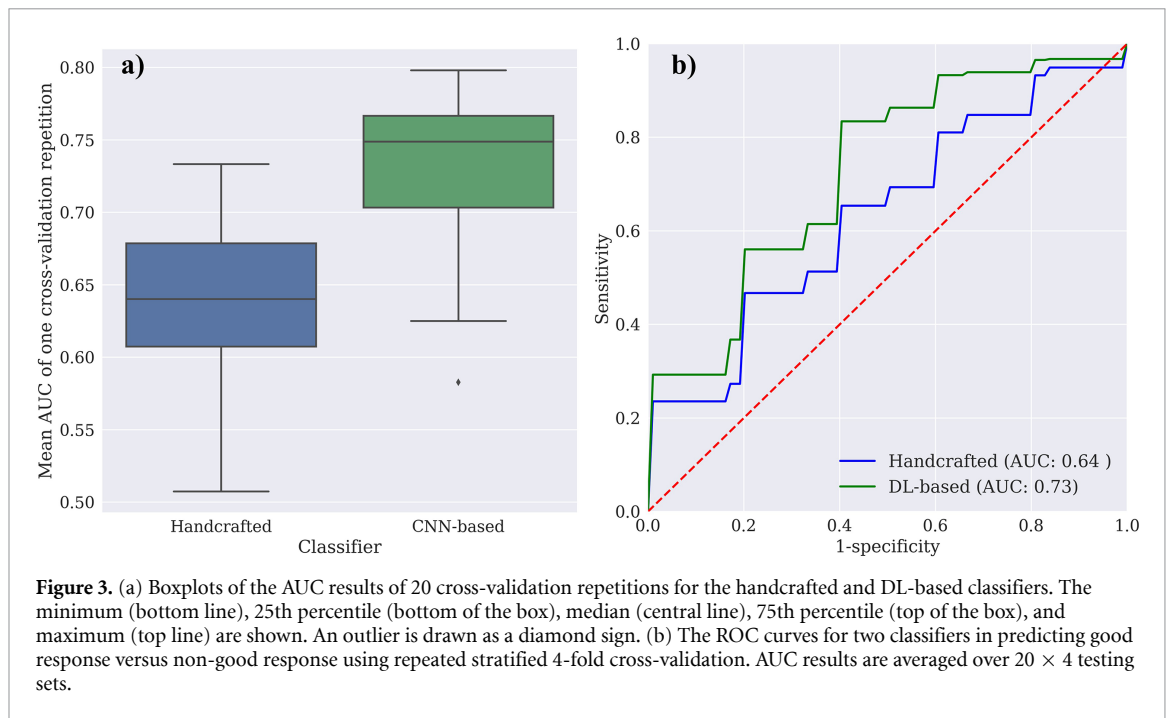
3. Results

3.1. Patient characteristics

Table 2 summarizes the clinical characteristics of our patient cohort. 22 (51.2%) patients achieved GR after nCRT. Among the 22 GR patients, there were 14 (63.6%) men and 8 (36.4%) women. Among 21 non-GR patients, there were 14 (66.7%) men and 7 (33.3%) women.

Figure 2 shows the transverse slices of DWIs and ADC maps for the representative GR and non-GR patients. Both patients are male with rectal cancer at the same clinical stage of T3N1. No significant visual differences were observed.

3.2. Classification performance

Figure 3(a) compares the boxplots of the mean AUC results of 20 cross-validation repetitions for two classifiers. Large deviations were observed due to the small sample size. The AUC of a single repetition varies from 0.51 to 0.73 for the handcrafted classifier, and from 0.58 to 0.80 for the DL-based classifier. The average ROC curves of the two classifiers are shown in figure 3(b). The handcrafted classifier achieved the mean AUC of 0.64 (standard error [SE], 0.08) using repeated 20 times 4-fold cross-validation, while the DL-based classifier achieved 0.73 (SE, 0.05). The *p*-value of the corrected paired *t*-test was 0.049, suggesting a significant difference in the AUC results for the handcrafted classifier and DL-based classifier.

Figure 3. (a) Boxplots of the AUC results of 20 cross-validation repetitions for the handcrafted and DL-based classifiers. The minimum (bottom line), 25th percentile (bottom of the box), median (central line), 75th percentile (top of the box), and maximum (top line) are shown. An outlier is drawn as a diamond sign. (b) The ROC curves for two classifiers in predicting good response versus non-good response using repeated stratified 4-fold cross-validation. AUC results are averaged over 20 × 4 testing sets.

4. Discussion

In this study, we compared the performance of the classifiers built with the handcrafted and DL-based features, extracted from pre-treatment DWI, for predicting the post-nCRT treatment response for a cohort of LARC patients. To our knowledge, this is the first study investigating DL-based features for this application. Compared to the handcrafted features, the DL-based features consisted of more abstract high-level information extracted from DWI images. Our study indicated that the DL-based classifier achieved a significantly better predictive performance than the handcrafted classifier in nCRT response prediction for rectal cancer. Studies showed that the DL-based features achieved better performance in breast cancer diagnosis and glioblastoma survival prediction than the handcrafted features (Lao *et al* 2017, Antropova *et al* 2017). The DL-based features are expected to achieve better performance and more generalizable results in diagnosis, recurrence and survival prediction for other sites as well.

We conducted repeated 4-fold cross-validation for evaluating the model performance as it stabilizes the accuracy estimation (Bouckaert and Frank 2004, Kim 2009). The handcrafted classifier achieved the mean AUC of 0.64 for predicting GR vs non-GR, while the DL-based classifier achieved an improved mean AUC of 0.73. Additionally, a fused classifier was constructed by averaging prediction scores of two classifiers. The fused classifier achieved the mean AUC of 0.71, which is better than that for the handcrafted classifier. Nie *et al* (2016), using a single run of 4-fold cross-validation, reported the mean AUC of 0.73 for GR and non-GR prediction using DWI handcrafted features on a similar size cohort of 48 patients. The standard error of the mean AUC was not reported. To investigate the cross-validation variation caused by the different data partitions for a small dataset, we conducted 20 independent cross-validation trials using our dataset. It should be noted that 20 independent cross-validation trials are different from the repeated 20 cross-validation since each cross-validation trial has its own optimal hyperparameters, while all 20 cross-validation repetitions need to have the same hyperparameters. The mean AUC of each cross-validation trial ranged from 0.56 to 0.79 for the classifier built with the handcrafted features, and from 0.63 to 0.82 for the one built with the DL-based features. Given a relatively small patient size, a single run of cross-validation may have large bias. Also, different classification models, evaluation protocols, patient number, and response label ratio may result in different prediction accuracy.

We investigated the radiomic features extracted from a single imaging modality of DWI in this study. Several studies showed that including the handcrafted features from T2-weighted MR images and dynamic contrast-enhanced images improved predictive power (Nie *et al* 2016, Horvat *et al* 2018). The DL-based feature extraction scheme can be applied to other MR imaging modalities and may further help improve the prediction accuracy. Comparing the handcrafted and DL-based features extracted from multiparametric MR images for treatment response prediction would be an interesting study to work on in the future.

The GTV contours used for feature extraction were manually delineated by a single radiation oncology. The effect of inter-observer delineation variabilities on the extracted features were not investigated in this work. Several studies suggested that the inter-observer delineation variability resulted in many unstable handcrafted radiomic features (Pavic *et al* 2018, Fiset *et al* 2019) and hence possibly less robust prediction models. Such delineation uncertainty may also lead to unstable DL-based features. The robustness of prediction models generated from the handcrafted and DL-based features can be investigated and compared using the intraclass correlation coefficient (ICC). A higher ICC indicates a better reproducibility. A cutoff on ICC could be used to select stable handcrafted and DL-based features that may result in more robust models. Alternatively, automatic tumor segmentation methods may be utilized to establish robust prediction models by reducing delineation variability.

The ROI used for extracting DL-based features was set based on tumor size, so the resampled ROIs would have different spatial resolutions across patients even if the spatial normalization was applied before the feature extraction. We believe it is unnecessary to conduct spatial normalization before extracting DL-based features in this study. To investigate the effect of the spatial normalization on handcrafted features, we resampled ADC maps to $1.56 \times 1.56 \times 5.00$ mm^3 and re-extracted handcrafted features from the resampled ADC maps. The mean AUC achieved by the classifier trained using the updated handcrafted features was 0.65. Corrected paired *t*-test showed that no significant difference in AUC was observed between the classifiers trained with original or updated handcrafted features ($p = 0.49$).

Our study has several limitations. First, the study sample is relatively small, which may lead to unstable estimation and suboptimal model performance. A repeated cross-validation method was utilized to reduce the bias, and LASSO regularization was implemented to reduce overfitting. In this work, we investigated DL-based features extracted via transfer learning. Another popular DL approach for response prediction is to train CNNs from scratch or using finetuning. However, overfitting may become a major issue in this method especially when patient size is small. Our results, in concordance with other studies (Anon 2016, Huynh *et al* 2016, Lao *et al* 2017), showed that DL-based features extracted via transfer learning achieved promising results in various prediction tasks in the medical field. Second, our dataset only contained 9 patients with pathological complete response (pCR). The pCR is defined as the absence of viable tumor cells in the primary and lymph nodes. The small number of pCR patients and unbalanced labels resulted in a large standard deviation on the AUCs using either handcrafted features or DL-based features for predicting pCR vs non-pCR. We chose to construct and evaluate the predictive model with the classification labels of GR and non-GR in this preliminary study. A larger dataset is desirable to provide a more reliable estimation for the AUC of pCR and non-pCR prediction. We expect to see better performance from the DL-based features than the handcrafted features in predicting pCR on a larger dataset. Lastly, the current study focused on the pre-treatment prediction of tumor response based on a single time point, due to the unavailability of during- and post-nCRT images for some patients. Given the primary focus of this work is mainly on comparing handcrafted features to DL-based features, we illustrated the earlier prediction for post-nCRT response, based on pre-treatment images, such early prediction will provide advantages for chemo-radiation treatment design and schedule. It may be beneficial to assess the response by combining images at other time points, such as during- and post-nCRT images for higher prediction accuracy. We expect to see better performance from the DL-based features than the handcrafted features using other MR images acquired at different time points.

5. Conclusion

Our preliminary study showed that the DL-based radiomic features extracted via transfer learning from pretreatment DWIs achieved significantly better classification performance for predicting post-nCRT treatment response in LARC patients, in comparison to the handcrafted radiomic features. Future work involves validation with a larger dataset and investigating the predictive power of the DL-based features extracted from multiparametric MR images (pre-, during-, and post-nCRT).

ORCID iDs

Jie Fu ⬤ https://orcid.org/0000-0002-5489-2732
Xinran Zhong ⬤ https://orcid.org/0000-0002-2530-5228
Ning Li ⬤ https://orcid.org/0000-0002-7968-7364
Ritchell Van Dams ⬤ https://orcid.org/0000-0002-1124-2676
John Lewis ⬤ https://orcid.org/0000-0002-5111-1865
Kyunghyun Sung ⬤ https://orcid.org/0000-0003-4175-5322

References

Afshar P, Mohammadi A, Plataniotis K N, Oikonomou A and Benali H 2019 From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities *IEEE Signal Process. Mag.* **36** 132–60

Amodeo S, Rosman A S, Desiato V, Hindman N M, Newman E, Berman R, Pachter H L and Melis M 2018 MRI-based apparent diffusion coefficient for predicting pathologic response of rectal cancer after neoadjuvant therapy: systematic review and meta-analysis *Am. J. Roentgenol.* **211** W205–16

Anon 2016 Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma *Tomography* **2** 388–95

Antropova N, Huynh B Q and Giger M L 2017 A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets *Med. Phys.* **44** 5162–71

Beets-Tan R G H *et al* 2018 Magnetic resonance imaging for clinical management of rectal cancer: updated recommendations from the 2016 European Society of Gastrointestinal and Abdominal Radiology (ESGAR) consensus meeting *Eur. Radiol.* **28** 1465–75

Bouckaert R R and Frank E 2004 *Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms* (Berlin: Springer) pp 3–12

Fiset S *et al* 2019 Repeatability and reproducibility of MRI-based radiomic features in cervical cancer *Radiother. Oncol.* **135** 107–14

Horvat N, Veeraraghavan H, Khan M, Blazic I, Zheng J, Capanu M, Sala E, Garcia-Aguilar J, Gollub M J and Petkovska I 2018 MR imaging of rectal cancer: radiomics analysis to assess treatment response after neoadjuvant therapy *Radiology* **287** 833–43

Huynh B Q, Li H and Giger M L 2016 Digital mammographic tumor classification using transfer learning from deep convolutional neural networks *J. Med. Imaging* **3** 034501

Kapiteijn E *et al* 2001 Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer *N. Engl. J. Med.* **345** 638–46

Kim J-H 2009 Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap *Comput. Stat. Data Anal.* **53** 3735–45

Kim S H, Lee J Y, Lee J M, Han J K and Choi B I 2011 Apparent diffusion coefficient for evaluating tumour response to neoadjuvant chemoradiation therapy for locally advanced rectal cancer *Eur. Radiol.* **21** 987–95

Lambin P *et al* 2017 Radiomics: the bridge between medical imaging and personalized medicine *Nat. Rev. Clin. Oncol.* **14** 749–62

Lao J, Chen Y, Li Z-C, Li Q, Zhang J, Liu J and Zhai G 2017 A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme *Sci. Rep.* **7** 10353

LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44

Maas M *et al* 2010 Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data *Lancet Oncol.* **11** 835–44

Nie K, Shi L, Chen Q, Hu X, Jabbour S K, Yue N, Niu T and Sun X 2016 Rectal cancer: assessment of neoadjuvant chemoradiation outcome based on radiomics of multiparametric MRI *Clin. Cancer Res.* **22** 5256–64

Pavic M *et al* 2018 Influence of inter-observer delineation variability on radiomics stability in different tumor sites *Acta Oncol. (Madr.)* **57** 1070–4

Quah H, Chou J F, Gonen M, Shia J, Schrag D, Saltz L B, Goodman K A, Minsky B D, Wong W D and Weiser M R 2008 Pathologic stage is most prognostic of disease-free survival in locally advanced rectal cancer patients after preoperative chemoradiation *Cancer* **113** 57–64

Russakovsky O *et al* 2015 ImageNet large scale visual recognition challenge *Int. J. Comput. Vis.* **115** 211–52

Schurink N W, Lambregts D M J and Beets-Tan R G H 2019 Diffusion-weighted imaging in rectal cancer: current applications and future perspectives *Br. J. Radiol.* **92** 20180655

Siegel R L, Miller K D and Jemal A 2019 Cancer statistics, 2019 *CA Cancer J. Clin.* **69** 7–34

Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition (arXiv:1409.1556)

Tibshirani R 1996 Regression shrinkage and selection via the lasso *J. R. Stat. Soc. Ser. B* **58** 267–88

Trakarnsanga A *et al* 2014 Comparison of tumor regression grade systems for locally advanced rectal cancer after multimodality treatment *JNCI J. Natl. Cancer Inst.* **106** dju248

van de Velde C J H *et al* 2014 EURECCA colorectal: multidisciplinary management: European consensus conference colon & rectum *Eur. J. Cancer* **50** 1.e1–34

van Griethuysen J J M, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan R G H , Fillion-Robin J-C, Pieper S and Aerts H J W L 2017 Computational radiomics system to decode the radiographic phenotype *Cancer Res.* **77** e104–7

Wang S, Liu Z, Rong Y, Zhou B, Bai Y, Wei W, Wei W, Wang M, Guo Y and Tian J 2019 Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer *Radiother. Oncol.* **132** 171–7

Wu T T, Chen Y F, Hastie T, Sobel E and Lange K 2009 Genome-wide association analysis by lasso penalized logistic regression *Bioinformatics* **25** 714–21