Medical Image Analysis 79 (2022) 102423

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

MLDRL: Multi-loss disentangled representation learning for predicting esophageal cancer response to neoadjuvant chemoradiotherapy using longitudinal CT images

Hailin Yue[a], Jin Liu[a], Junjian Li[a], Hulin Kuang[a], Jinyi Lang[b,c], Jianhong Cheng[a], Lin Peng[d], Yongtao Han[d], Harrison Bai[e], Yuping Wang[f], Qifeng Wang[b,c,**], Jianxin Wang[a,*]

[a] Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China
[b] Department of Radiation Oncology, Sichuan Cancer Hospital and Institution, Sichuan Cancer Center, School of Medicine, Radiation Oncology Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China, Chengdu, China
[c] Department of Medical Oncology, Sichuan Cancer Hospital and Institution, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China
[d] Department of Thoracic Surgery, Sichuan Cancer Hospital and Institution, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China
[e] Department of Diagnostic Imaging, Rhode Island Hospital and Alpert Medical School of Brown University, Providence, RI 02912 USA
[f] Department of Biomedical Engineering, Tulane University, New Orleans, LA 70118 USA

ARTICLE INFO

Article history:
Received 23 April 2021
Revised 8 March 2022
Accepted 12 March 2022
Available online 2 April 2022

Keywords:
Esophageal cancer
Longitudinal CT images
Disentangled representation learning
Pathological complete response (pCR)

ABSTRACT

Accurate prediction of pathological complete response (pCR) after neoadjuvant chemoradiotherapy (nCRT) is essential for clinical precision treatment. However, the existing methods of predicting pCR in esophageal cancer are based on the single stage data, which limits the performance of these methods. Effective fusion of the longitudinal data has the potential to improve the performance of pCR prediction, thanks to the combination of complementary information. In this study, we propose a new multi-loss disentangled representation learning (MLDRL) to realize the effective fusion of complementary information in the longitudinal data. Specifically, we first disentangle the latent variables of features in each stage into inherent and variational components. Then, we define a multi-loss function to ensure the effectiveness and structure of disentanglement, which consists of a cross-cycle reconstruction loss, an inherent-variational loss and a supervised classification loss. Finally, an adaptive gradient normalization algorithm is applied to balance the training of multiple loss terms by dynamically tuning the gradient magnitudes. Due to the cooperation of the multi-loss function and the adaptive gradient normalization algorithm, MLDRL effectively restrains the potential interference and achieves effective information fusion. The proposed method is evaluated on multi-center datasets, and the experimental results show that our method not only outperforms several state-of-art methods in pCR prediction, but also achieves better performance in the prognostic analysis of multi-center unlabeled datasets.

1. Introduction

Esophageal cancer is one of the most common malignant diseases, and its five-year overall survival rate is only about 15%. In 2018, 572,000 people worldwide were diagnosed with esophageal cancer, and approximately 509,000 died of it (Bray et al., 2018; Chen et al., 2019). At present, neoadjuvant chemoradiotherapy (nCRT) followed by esophagectomy has become the standard treatment method for locally advanced retractable esophageal cancer (van Hagen et al., 2012; Shapiro et al., 2015). After nCRT, the probability of achieving pathological complete response (pCR) is 25 to 42%. Many studies have shown that patients with pCR appear to have a superior overall survival without surgery (Meguid et al., 2009; Tong et al., 2010; van Hagen et al., 2013), and the "wait and see" management approach can avoid surgery and preserve organs. However, pCR can only be confirmed after surgery. Therefore, it is significant to create a validated and noninvasive method

* Corresponding author at: Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China
** Corresponding author at: Department of Radiation Oncology, Sichuan Cancer Hospital and Institution, Sichuan Cancer Center, School of Medicine, Radiation Oncology Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China, Chengdu, China.
E-mail addresses: wangqifeng@uestc.edu.cn (Q. Wang), jxwang@mail.csu.edu.cn (J. Wang).

https://doi.org/10.1016/j.media.2022.102423
1361-8415/

to accurately predict pCR of patients with esophageal cancer after nCRT.

With the rapid development of imaging techniques, various imaging modalities have been used to predict pCR, including magnetic resonance imaging (MRI), positron emission tomography (PET) and computed tomography (CT). However, according to the economic feasibility and clinical performance, CT is the most common material for diagnosing esophageal cancer. Researches on esophageal cancer are mostly based on CT images rather than other types of medical images. Nowadays, many CT-based studies have made progress in the field of esophageal cancer. Yang et al. (2019) established radiomics signatures to predict pCR by using single CT images. Hou et al. (2019) established a single CT-based feasible model to predict the prognosis of patients who had undergone nCRT. Wu et al. (2018), Tan et al. (2019) and Larue et al. (2018) also conducted researches on esophageal cancer based on CT images. Although these studies have made breakthroughs in the diagnosis or prognosis of esophageal cancer, they ignored the complementary information in longitudinal CT images. Currently, more and more studies have used the multi-perspective data to make medical predictions. Shao et al. (2020) used the histopathological data, the gene expression data and survival information to discover the relationships between diagnosis and prognosis tasks. Zhou et al. (2020) first projected neuroimaging features from multiple modalities into a common latent space, and then exploited the cross-modality correlations as complementary information to diagnose Alzheimer's disease. Pi et al. (2020) used attention-augmented deep neural networks to fuse features from anterior and posterior whole-body bone scans for classification. Jin et al. (2021) presented a multi-task approach to simultaneously realize tumor segmentation and pCR prediction using longitudinal CT images. Despite these progress, how to use the existing longitudinal CT images to discover more useful information in pCR prediction remains a challenge.

There are three types of methods can be used to fuse longitudinal CT images, including early fusion, late fusion and whole fusion. The early fusion method concatenates features from each modality as a sequence before dimensionality reduction and then selects the most conducive features for classification. For the late fusion method, the features of each modality are selected separately and then combined to make the classification. These two ideas are widely used in the study of multi-perspective data fusion. For example, Zhu et al. (2016), Lei et al. (2017) used the idea of early fusion and late fusion respectively to achieve the diagnosis of Alzheimer's disease. Besides early fusion and late fusion, there are whole fusion methods, such as HyperDense-Net (Dolz et al., 2018). In HyperDense-Net, the dense connections were performed not only between layers in the same modality, but also between the layers in different modalities, which could yield a much more powerful feature representation. Although these methods can effectively fuse the longitudinal data, few of these methods further explores the correlations among longitudinal CT images.

As we know, pCR serves as an essential indicator of patients' prognosis. Patients with pCR generally have a superior overall survival without surgery (van Hagen et al., 2013; Meguid et al., 2009; Zanoni et al., 2013). Therefore, it is necessary to consider the performance of models for prognostic analysis on the patients without surgery. In addition, some studies have performed prognostic analysis in patients who have undergone surgery. Yokoyama et al. (2020) investigated the biomarkers that related to the prognosis of patients with pancreatic neoplasms after surgery. Chiang et al. (2012) explored the clinicopathological features that could predict the resectability of pancreatic adenocarcinoma, and conducted prognostic analysis of pancreatic adenocarcinoma after resection to identify favorable survival factors. Qian et al. (2019) used platelet-associated indicators to perform prognostic analysis of resectable colorectal cancer. All these studies illustrate the importance of prognostic analysis. Therefore, in this study, we further evaluate the reliability of our method by exploring the relationship between the predicted results of our method and the prognosis on samples that have not undergone surgery.

In general, there are at least the following issues associated with current studies. First, most studies use single stage data to predict pCR, without considering the complementary information among the longitudinal data. Second, existing methods simply concatenate features from each modality, which can not discover the complementary information between different modalities. Third, pCR is closely related to prognosis, and there are few existing studies further exploring the reliability of their methods in the prognostic analysis on samples that have not undergone surgery.

Recently, disentangled representation learning is playing an important role in deep learning. A disentangled representation can represent information as numerous independent factors, in which each factor captures some meaningful aspects of the data. In recent years, disentangled representation learning has shown great promise in the field of computer vision. Lee et al. (2018) proposed a disentangled representation based method to embed images into two spaces: domain-shared and domain-specific. The experimental results showed that this way could generate diverse and realistic images. Hu et al. (2020) used disentangled-multimodal adversarial autoencoder to find the common and specific features between multi-modal MRI, which used a common-specific distance ratio loss function to constrain the structure of the latent space. Guo et al. (2020) also used deep disentangled representation learning to predict lymphoma outcome. In our study, we inherit these advantages to realize the disentanglement of the longitudinal data.

Inspired by the above considerations, we propose a new Multi-Loss Disentangled Representation Learning (MLDRL) to accurately predict pCR by making full use of complementary information of radiomics features in longitudinal CT images. Intuitively, our method may find complementary information that is conducive to predicting pCR. To verify the effectiveness of our method, we not only evaluate MLDRL in pCR prediction, but also perform prognostic analysis on multi-center unlabeled datasets. The experimental results show that our method is not only superior to related methods in pCR prediction, but also achieves better performance in the prognostic analysis of multi-center unlabeled datasets.

Compared with the existing methods, our main contributions can be briefly summarized as follows:

- We propose a new multi-loss disentangled representation learning (MLDRL) for pCR prediction using longitudinal CT images.
- A multi-loss function is designed to make the latent learned representation structural, which consists of a cross-cycle reconstruction loss, an inherent-variational loss and a supervised classification loss.
- We adopt an adaptive gradient normalization algorithm to improve the imbalance magnitude and the inconsistent convergence of each loss term.
- We investigate the reliability of our method by performing prognostic analysis on multi-center unlabeled datasets, further demonstrating the clinical practical value of our method.

The rest of this paper is organized as follows. Section 2 describes the materials and methods. Section 3 introduces the details of the experiments and results. The experimental discussion is reported in Section 4. Finally, we conclude this study in Section 5.

Table 1
Demographic information of all samples in this study.

	Dataset A		Multi-center unlabeled datasets		
	pCR	Non-pCR	Dataset B	Dataset C	Dataset D
Gender [%]					
Male	43[34.68]	64[51.61]	62[76.54]	31[65.96]	19[82.60]
Female	6[4.84]	11[8.87]	19[23.46]	16[34.04]	4[17.40]
Age(mean±std)	60.12±7.56	59.01±7.51	60.82±7.03	59.93±7.03	65.65±7.03
T Stage [%]					
T1	3[2.42]	0[0]	0[0]	0[0]	0[0]
T2	15[12.10]	3[2.42]	10[12.34]	9[19.15]	3[13.04]
T3	25[20.16]	36[29.03]	50[61.73]	28[59.57]	8[34.78]
T4	6[4.84]	36[29.03]	21[25.93]	10[21.28]	12[52.18]
N Stage [%]					
N0	15[12.10]	5[4.03]	0[0]	1[2.13]	2[8.70]
N1	21[16.94]	29[23.39]	35[43.21]	23[48.94]	21[91.30]
N2	5[4.03]	16[12.90]	38[46.91]	18[38.30]	0[0]
N3	8[6.45]	25[20.16]	8[9.88]	5[10.63]	0[0]
Total [%]	49[39.51]	75[60.49]	81[100]	47[100]	23[100]

T stage is clinical Tumor stage, N stage is clinical Node stage.

2. Materials and methods

2.1. Dataset

Our study was approved by the ethics committee of Sichuan Cancer Hospital (Approved Number SCCHEC-02-2020-015). All procedures involving human participants were under ethical standards. At the same time, informed consent was acquired from each participant.

In our study, there were four datasets from different centers. Dataset A, B and C were collected from Sichuan Cancer Center in different periods, Dataset D was collected from National Cancer Center (Beijing). The demographic information of all the samples has been summarized in Table 1. The inclusion criteria of these datasets were as follows:

- Dataset A met the following criteria: (I) All patients were diagnosed with esophageal cancer by biopsy; (II) All patients were received the complete nCRT treatment; (III) All patients underwent esophagectomy; (IV) after surgery, pCR was identified by the histopathological examination of surgically resected specimens.
- Datasets B and C met the following criteria: (I) All samples were diagnosed with esophageal cancer by biopsy; (II) All samples were received complete nCRT treatment; (III) All samples did not underwent esophagectomy after nCRT; (IV) All samples' prognostic information (survival time and survival statue) were available.
- The inclusion criteria of Dataset D were consistent with the Datasets B and C.

All patients enrolled in our study must undergo longitudinal CT examinations (before nCRT and after nCRT), and the region of interest (ROI) of two-stage were performed by two experienced radiologists.

2.2. Overview of our method

The framework of MLDRL is shown in Fig. 1, which consists of two parts. In training and testing part, we first extract multi-view radiomics features from longitudinal CT images, and analyze the importance of each feature. Second, the proposed MLDRL is used to find the shared and complementary information of the longitudinal radiomics features. Third, after the complementary information is integrated, we use it to predict the response of each sample to nCRT. In external prognostic analysis part, we further evaluate the clinical practical value of MLDRL by performing prognostic analysis on multi-center unlabeled datasets.

2.3. Multi-view feature extraction and feature selection

The importance of multi-view strategy has become more and more apparent in recent years (Kang et al., 2020; Pi et al., 2020; Carneiro et al., 2017; Jouirou et al., 2019; Carneiro et al., 2017; Khan et al., 2019), especially in the field of medical image analysis. In our study, we inherit this essence and extract multi-view features of longitudinal CT images to explore the relationship between the complementary information of longitudinal CT images and pCR.

In order to standardize the process of extracting radiomics features, we develop a medical imaging platform that can extract intratumoral and peritumoral features. The platform is free available at www.miacsu.group. The pre-processing steps of our proposed radiomics feature extraction are shown as follows. Firstly, to reduce the noisy background, we only preserve the part of the CT images that contains the full ROI. Secondly, since longitudinal CT images are captured by different CT scanners at different times, we perform intensity normalization for all CT images, in our study, the intensity values are normalized into [0, 1]. Finally, in order to train the network better, we perform data augmentation operations, which include data cropping and data rotation. After this, we first use multiple filters to process the original images, and then extract various types of features from filtered images, including Firstorder, gray level cooccurrence matrix (GLCM), gray level dependence matrix (GLDM), gray level run length matrix(GLRLM), gray level size zone matrix (GLSZM), neighboring gray tone difference matrix (NGTDM), and shape. Among of them, the distribution of voxel intensities within the ROI is described by the Firstorder, including entropy, kurtosis, skewness, etc. GLCM, GLDM, GLDM, GLSZM and NGTDM describe the texture information of ROI. Shape describes the surface features of ROI. The number of features of each type is shown in Table 2. For each patient, we can extract 2153-dimensional features from CT images of each stage. The details of each feature are summarized in the supplementary.

The dimension of the extracted radiomics features for each stage is very high, and there might be redundancy among them. Therefore, before training our network, a feature selection method based on a random forest proposed in Kawakubo and Yoshida (2012) is adopted due to its superior performance. And then, the top-K features of each stage are selected as the input to train our model. The K is determined as 81 after experiments

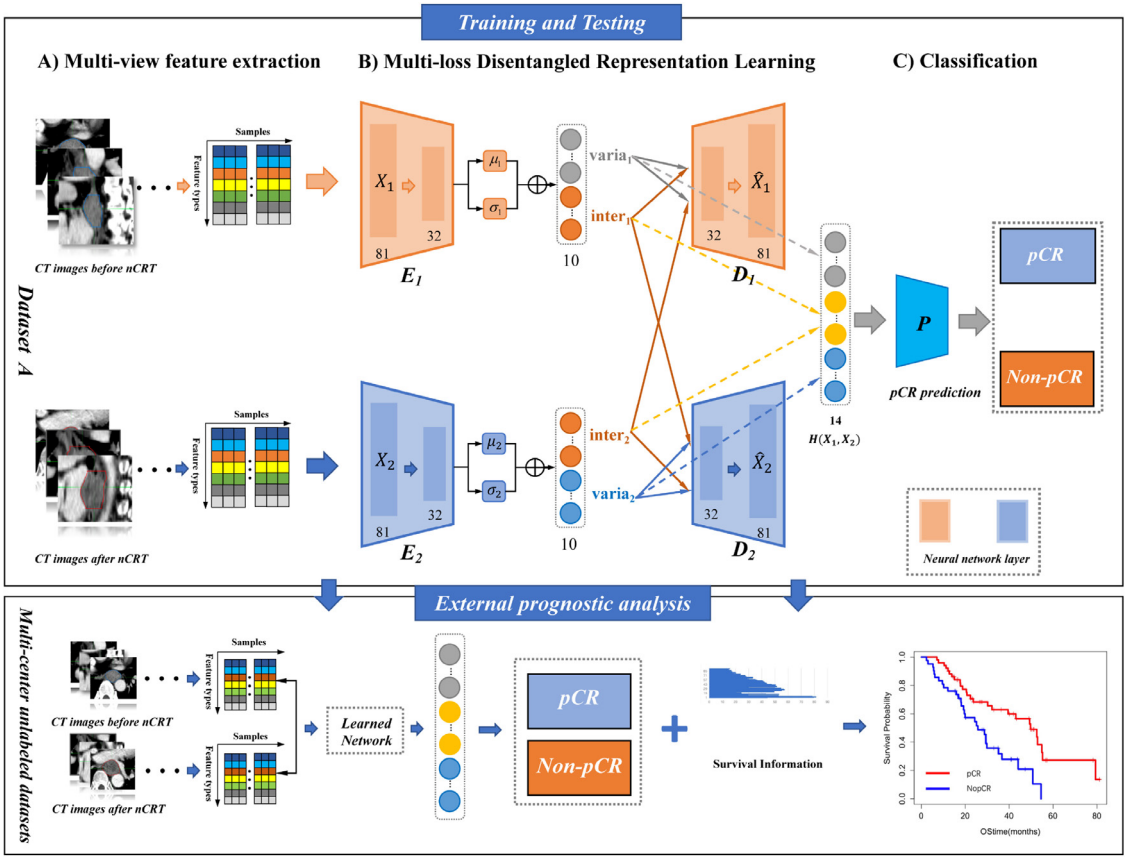

Fig. 1. An illustration of the proposed method. (A) We first extract the multi-view features of longitudinal CT images. (B) Multi-loss disentangled representation learning is used to train the structural latent representation. (C) We use the integrated latent representation $H(X_1, X_2)$ to predict pCR. Besides, in order to further investigate the reliability of our proposed method MLDRL, we perform prognostic analysis on multi-center unlabeled datasets, as shown in external prognostic analysis. E_i and D_i are the encoder and decoder of each stage, X_i and $\hat{X}_i$ are the input features and reconstructed features of each stage, the numbers in the training and testing part denote the number of neurons. P is the classification module.

Table 2
The number of features of different types.

	Feature type	Number
Intensity features	First order	414
Texture features	GLCM	552
	GLDM	322
	GLRLM	368
	GLSZM	368
	NGTDM	115
Surface features	Shape	14

where we change the dimension of the input selected features while keeping other parts of the proposed method unchanged.

2.4. Multi-loss disentangled representation learning

In order to effectively disentangle the inherent and variational components of the latent variables, we propose a multi-loss disentangled representation learning. In this section, we introduce the architecture of our proposed MLDRL, the structure for latent representation and the adaptive gradient normalization algorithm for optimization.

2.4.1. Architecture of our proposed MLDRL

Our proposed MLDRL is mainly composed of two variational autoencoders (VAE). Each VAE consists of an encoder and a decoder. The encoder maps the input features into a Gaussian probability distribution, which is used to generate the latent space Z. The decoder uses the latent space Z to reconstruct the input features. In our study, for radiomics features of each stage, we first use a multi-layer perceptron neural network as the encoder to generate the latent variables. Then, a multi-layer perceptron neural network is used as the decoder to reconstruct the input features. Finally, we disentangle the latent variables Z_i of each stage into inherent and variational components. In our study, the number of the neurons in each layer is based on additional experiments. After experiments where we change one parameter while keeping others unchanged, the numbers of neurons in other two layers of the encoder are set as 32 and 10, the numbers of the inherent and variational components are set as 6 and 4. Moreover, our code is publicly available at https://github.com/yuehailin/MLDRL.git.

2.4.2. Structure for latent representation

In our study, we hope to make full use of complementary information from longitudinal features. The specific details are as follows: we disentangle the latent variables Z_i of each stage into $Inher(Z_i)$ and $Varia(Z_i)$. $Inher(Z_i)$ represents the shared information of the longitudinal features, while $Varia(Z_i)$ represents the complementary information of the longitudinal features. Intuitively, we hope that the disentangled latent variables of longitudinal features meet the following requirements: (1) $Inher(Z_1)$ and $Inher(Z_2)$ are as similar as possible; $Varia(Z_1)$ and $Varia(Z_2)$ are as different as possible; (2) The disentangled latent variables Z_i can reconstruct the i-th input well; (3) The latent representation $H(X_1, X_2)$ can distinguish pCR and Non-pCR well. In order to satisfy the require-

ments of (1) and (2), we define a cross-cycle reconstruction loss and an inherent-variational loss. In order to satisfy requirement (3), we define a supervised classification loss. The details of these three losses are as follows.

Cross-cycle reconstruction loss. VAE is one component of our proposed method, which has been widely applied in disentangled representation learning (Chen et al., 2018a). The loss function of VAE consists of a reconstruction loss and a Kullback-Leibler (KL) divergence, which can be defined as follows:

$$Loss_{vae}(\phi,\theta) = -\mathbb{E}_{Z\sim q_\phi(Z|X)}[log p_\theta(X|Z)] + KL(q_\phi(Z|X) \parallel p_\theta(Z)), \quad (1)$$

where $p_\theta(Z)$ is the prior distribution imposed on the latent space Z. $p_\theta(Z) = \mathcal{N}(Z|0, I)$, I is the identity matrix, $q_\theta(Z|X)$ is the encoding distribution and $p_\phi(X|Z)$ is the decoding distribution. $KL(\bullet|\bullet)$ is used to measure the KL divergence between two distributions.

In order to make the latent variables structural, we design a cross-cycle reconstruction loss. For each stage, $Z_i = [Inher(Z_i), Varia(Z_i)]$ is the latent variables for X_i, we hope the i-th latent variables Z_i can reconstruct i-th input X_i well. In addition, we hope the shared information $Inher(Z_i)$ can provide support to reconstruct the features of the other stage. Specifically, $Inher(Z_1)$ joint $Varia(Z_2)$ can reconstruct X_2, and $Inher(Z_2)$ joint $Varia(Z_1)$ can reconstruct X_1. To achieve the above objective, we introduce the cross-cycle reconstruction loss to ensure the effectiveness of the disentanglement, which is defined as follows:

$$Loss_{recon} = \sum_{i=1}^{2}\sum_{j=1}^{2}\left\|X_i - D_i(Inher(E_j(X_j)), Varia(E_i(X_i)))\right\|^2 + KL(q_\theta(Z_i|X_i) \parallel p(Z_i)), \quad (2)$$

where E_i is the ith encoder, D_i is the ith decoder. Through this loss, the latent variables is disentangled initially.

Inherent-variational loss. In order to make the latent variables can be disentangled more thoroughly, we define an inherent-variational loss, which is defined as follows:

$$Loss_{inher-varia} = \frac{Loss_{inher}}{Loss_{varia}}, \quad (3)$$

$$Loss_{inher} = \mathbb{E}_{X_1,X_2}\|Inher(E_1(X_1)), Inher(E_2(X_2))\|_2, \quad (4)$$

$$Loss_{varia} = \mathbb{E}_{X_1,X_2}\|Varia(E_1(X_1)), Varia(E_2(X_2))\|_2, \quad (5)$$

where $Loss_{inher}$ and $Loss_{varia}$ are designed to measure the difference between the two inherent components (i.e., $Inher(E_1(X_1))$ and $Inher(E_2(X_2))$, and the difference between the two variational components (i.e., $Varia(E_1(X_1))$ and $Varia(E_2(X_2))$) from the longitudinal data, respectively. When $Loss_{inher}$ becomes smaller and $Loss_{varia}$ becomes larger, $Loss_{inher-varia}$ becomes smaller. This implies that the optimization of $Loss_{inher-varia}$ can make our proposed method disentangle the latent variables of longitudinal features well. The function form of the proposed inherent-variational loss in Eq. (3) is determined via extensive experiments using other forms of function such as subtraction.

Supervised classification loss. To ensure latent representation can learn the effective fusion of the disentangled features from the longitudinal data, we design a multi-layer perceptron neural network (MLP) with the fusion of disentangled features as input for supervised classification and propose a supervised classification loss using the supervised classification results. The definition of the supervised classification loss is given as follows:

$$Loss_{class} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} Y_n^{(m)} log\hat{Y}_n^{(m)}, \quad (6)$$

where $\hat{Y}_n^{(m)}$ is the probability of nth sample for class m. M is the number of classes. N is the number of samples. In this study, the value of M is 2.

Joint loss. Based on the above analyses, the joint loss can be defined as:

$$Loss_{joint} = \lambda Loss_{recon} + \beta Loss_{inher-varia} + \gamma Loss_{class}. \quad (7)$$

where λ, β and γ are the weights of the cross-cycle reconstruction loss, the inherent-variational loss and the supervised classification loss, respectively.

2.4.3. Adaptive gradient normalization algorithm for optimization

The multi-loss function in our study consists of three independent loss terms, but the loss function with multiple loss terms is difficult to train. We hope that the magnitude of each loss term in the loss function can be well balanced, so that each of them can play a respective role. In multi-task learning, many studies have been devoted to solve the imbalance problem of different tasks (Liu et al., 2019; Guo et al., 2018; Kendall et al., 2018). Inspired by Chen et al. (2018b), we adopt an adaptive gradient normalization algorithm to optimize λ, β and γ in MLDRL. The joint loss in MLDRL can be summarized as follows:

$$Loss_{joint} = \sum w_k(t) Loss_k(t), \quad (8)$$

where $w_k(t)$ is an adaptive weight which is used to balance the magnitude and convergence of different loss terms. In order to better describe the adaptive gradient normalization algorithm, we first define some variables.

$$G_W^{(k)} = \|\nabla_W w_k(t) L_k(t)\|_2, \quad (9)$$

$$\overline{G}_W(t) = E_{task}\left[G_W^{(k)}(t)\right], \quad (10)$$

$$\tilde{L}_k = L_k(t)/L_k(0), \quad (11)$$

$$r_k(t) = \tilde{L}_k(t)/E_{task}\left[\tilde{L}_k(t)\right]. \quad (12)$$

where $G_W^{(k)}$ is the L_2 norm of the gradient of the weighted kth loss $w_k(t)L_k(t)$ with respect to the shared weights W. Since our network consists of two parts that process images at different stages, we need to calculate the gradients separately, and then take the average. $\overline{G}_W(t)$ is the average gradient norm in all losses at step t, $L_k(0)$ is the kth loss value at step 0, $L_k(t)$ is the kth loss value at step t, $\tilde{L}_k$ is used to measure the inverse training rate of the kth loss at step t, $r_k(t)$ is the relative inverse training rate for kth loss at step t. We hope that the gradient norm for each loss can closely approximate to the average gradient norm. The desired gradient norm for each loss term is defined as follows:

$$G_W^{(k)}(t) \mapsto \overline{G}_W(t) \times [r_k(t)]^\alpha, \quad (13)$$

where α is an extra hyperparameter. When multiple loss terms are different in their complexity, we set α with a higher value. When multiple loss terms are symmetric, a lower α is more suitable. In our study, $\alpha = 0.16$. For all losses, we use the Least Absolute Deviations (LAD) to calculate the loss between the actual and the expected gradient norms at each step for all losses. The L_{grad} can be summed as follows:

$$L_{grad}(t; w_k(t)) = \sum_k \left|G_W^{(k)}(t) - \overline{G}_W(t) \times [r_k(t)]^\alpha\right|. \quad (14)$$

We use $\nabla_{w_k} L_{grad}$ to update w_k at ecah step. The adaptive strategy is summarized in Algorithm 1.

Algorithm 1 The optimization procedure for our proposed MLDRL.

Initialize $w_1(0) = 1.4$, $w_2(0) = 3$, $w_3(0) = 0.3$
Initialize network weights W
for $t = 0$ to *epoch* **do**
  Input features of X to compute
  $L_k(t)\ \forall k$ and $loss_{joint} = \sum w_k(t) loss_k(t)$
  Compute $G_W^{(k)}(t)$ and $r_k(t)$
  Compute $\overline{G}_W(t)$ by averaging $G_W^{(k)}(t)$
  Compute $L_{grad}(t; w_k(t)) = \sum_k \left| G_W^{(k)}(t) - \overline{G}_W(t) \times [r_k(t)]^{\alpha} \right|$
  Compute $\nabla_{w_k} L_{grad}$ and $\nabla_W L_t$
  Update $w_k(t) \mapsto w_k(t+1)$ using $\nabla_{w_k} L_{grad}$
  Update $W(t) \mapsto W(t+1)$ using $\nabla_W L_t$
end for

2.5. Classification

After the latent variables of each stage have been disentangled into $Inher(E_i(X_i))$ and $Varia(E_i(X_i))$, we integrate these structural components into a learned latent representation $H(X_1, X_2)$, which is defined as follows:

$$H(X_1, X_2) = [inherent_{1,2}, variational_1, variational_2], \tag{15}$$

where $inherent_{1,2}$ is the mean of $inherent_1$ and $inherent_2$.

After obtaining the learned latent representation $H(X_1; X_2)$, we input it into a two-layer MLP for achieving the final classification results, making the proposed MLDRL be an end-to-end classification network. The dimension of this input layer is decided by the numbers of the inherent and varaiational components and the Eq. (15). The number of neurons in the output layer is 2. In our study, the threshold is set to 0.5. If the probability is less than 0.5, the sample is classified as pCR, otherwise Non-pCR. To further validate the reliability of the model, we use the learned network to predict pCR on multi-center unlabeled datasets for prognostic analysis.

3. Experiments and results

3.1. Experimental settings

In our study, Dataset A is used to train and test our proposed MLDRL. Multi-center unlabeled datasets are used to validate the practical value of MLDRL on samples without esophagectomy. The network in our study is implemented with Python and optimized using Adam with a learning rate of 0.0001. Since the adaptive gradient normalization algorithm used in this study requires initial weights of each loss as the basis for step-by-step weight updating, we set the initial weights of λ, β and γ in Eq. (7) to 1.4, 3 and 0.3 respectively via additional experiments where we use different initial weights of each loss while keeping other parts unchanged. Then, we also need to evaluate the influence of the numbers of neurons in the inherent component and variational component on the performance of pCR prediction. Finally, additional experimental results show that the numbers of neurons in inherent component and variational component that are most suitable for this study are 6 and 4, respectively. In addition, in order to address the imbalanced data, we apply a synthetic minority oversampling technique (SMOTE) for data augmentation to achieve class balance.

We conduct extensive experiments on both Dataset A and multi-center unlabeled datasets to evaluate the performance of our proposed method in predicting pCR. For Dataset A, we perform 5-fold cross validation and calculate the average of classification accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the curve (AUC) to measure the performance of MLDRL and other compared methods. The definitions of ACC, SEN and SPE are as follows:

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{16}$$

$$SEN = \frac{TP}{TP + FN} \tag{17}$$

$$SPE = \frac{TN}{TN + FP} \tag{18}$$

where TP, FN, FP and TN are true positive, false negative, false positive, true negative, respectively. For multi-center unlabeled datasets, we use the established model to predict the pCR, and analyze the difference in survival time between the two groups (pCR and Non-pCR). We compare the performance of different methods in prognostic analysis by calculating the Brier score (BS) and drawing the Kaplan-Meier curve. In addition, we also calculate 95% confidence interval (95% CI) for all metrics.

3.2. Comparison of MLDRL with other methods for pCR prediction

- Comparison with the state-of-the-art methods using the single stage data. Most radiomics methods only conduct experiments on the single stage data. To demonstrate the superiority of our method, we compare our proposed method MLDRL with following radiomics methods only using the single stage data (before nCRT). Cho et al. (2018) used minimum redundancy maximum relevance and random forest (RF) (Liaw et al., 2002) for classification. Cheng et al. (2020) used the least absolute shrinkage and selection operator to select features, and input the selected features into RF, logistic regression (LR) (Peng et al., 2002), XGBoost (XGB) (Chen and Guestrin, 2016) and support vector machine (SVM) (Cortes and Vapnik, 1995) to make classification.
- Comparison with other fusion methods. We compare our proposed MLDRL with the following 4 fusion methods: (1) Early fusion method (EFM). Zhu et al. (2016) used this idea to fuse PET and MRI data for Alzheimer's disease diagnosis; (2) Late fusion method (LFM). Zhang et al. (2012) used this idea of late fusion and multi-task feature selection to realize the diagnosis of cancer; (3) Disentangled-Multimodal adversarial autoencoder (D-AAE). Hu et al. (2020) used disentangled learning to realize fusion for the regression task; (4) HyperDense-Net (HD-Net). Dolz et al. (2018) used HyperDense-Net to realize the whole fusion between different modalities.

Table 3 shows the performances of MLDRL and the state-of-the-art methods using the single stage data (before nCRT) on Dataset A. Our method has an improvement in AUC compared with other single stage methods. Specifically, our method can obtain an AUC of 86.55%, which is higher than the best AUC (75.12%) of the state-of-the-art methods using the single stage data. Compared with the state-of-the-art single stage methods, our proposed method also has an improvement in ACC. Therefore, we can conclude that MLDRL can achieve better performance than the state-of-the-art methods using the single stage data.

In this section, we introduce the results of our proposed method and other fusion methods using the longitudinal data. As shown in Table 4, the AUC of EFM, LFM and D-AAE are 78.43%, 80.37% and 80.29% respectively. Although HD-Net is poorer than our proposed MLDRL, it achieves a better AUC than EFM, LFM and D-AAE. In addition, MLDRL outperforms all the comparison methods. Specifically, the AUC (86.55%) and ACC (80.96%) of our proposed MLDRL are superior to other fusion methods. For SPE and SEN, MLDRL also performs well. These results clearly demonstrate that MLDRL not only achieves better performance than the state-of-the-art methods using single stage data, but also achieves better performance than other fusion methods.

Table 3
The performance of our proposed method and the state-of-the-art methods using single stage data (before nCRT) for the prediction of pCR.

Method	AUC [95% CI]	ACC [95% CI]	SPE [95% CI]	SEN [95% CI]
Cho et al. (2018) + RF	62.48 [54.14-70.82]	60.77 [55.05–66.49]	48.75 [44.34–53.17]	69.18 [64.76–73.61]
Cheng et al. (2020) + LR	75.12 [71.32–78.91]	71.43 [63.41–79.44]	65.80 [61.94–69.66]	77.90 [76.63–79.17]
Cheng et al. (2020) + RF	70.57 [62.58–78.56]	66.35 [62.22–70.48]	59.40 [52.06–66.74]	72.16 [64.48–79.84]
Cheng et al. (2020) + XGB	70.24 [65.66–74.82]	65.03 [58.19–71.88]	58.56 [53.97–63.15]	71.21 [66.35–76.06]
(Cheng et al., 2020) + SVM	73.05 [67.21–78.90]	68.68 [64.50–72.87]	58.68 [53.49–63.87]	75.52 [72.41–78.64]
MLDRL (Ours)	**86.55 [83.54–89.56]**	**80.96 [76.74–85.19]**	**68.38 [57.20–79.56]**	**87.45 [84.58–90.32]**

Table 4
The performance of our proposed method and other fusion methods for pCR prediction.

Method	AUC [95% CI]	ACC [95% CI]	SPE [95% CI]	SEN [95% CI]
EFM (Zhu et al., 2016)	78.43 [74.70–82.16]	74.90 [70.76–79.03]	68.75 [67.74–69.75]	81.11 [77.47–84.76]
LFM (Zhang et al., 2012)	80.37 [73.86–86.88]	74.87 [71.43–78.30]	65.02 [61.12–68.92]	78.86 [75.64–82.07]
D-AAE (Hu et al., 2020)	80.29 [75.44–85.14]	73.80 [71.16–76.45]	**68.92 [63.31–74.53]**	78.35 [76.76–79.93]
HD-Net (Dolz et al., 2018)	82.83 [79.28–86.38]	72.50 [70.80–74.20]	66.22 [59.05–73.38]	76.28 [72.67–79.90]
MLDRL (Ours)	**86.55 [83.54–89.56]**	**80.96 [76.74–85.19]**	68.38 [57.20–79.56]	**87.45 [84.58–90.32]**

Table 5
The performances of latent representation obtained by our proposed method and other four methods using LR, RF, XGB and SVM as classifiers.

Classifier	Metric	EFM* [95% CI]	LFM* [95% CI]	D-AAE* [95% CI]	HD-Net* [95% CI]	MLDRL* [95% CI]
LR	AUC	74.88 [71.78–77.98]	80.35 [75.17–85.53]	78.67 [74.04–83.31]	78.98 [76.65–81.31]	**83.79 [80.74–86.84]**
	ACC	67.51 [61.53–73.50]	72.53 [67.22–77.85]	74.56 [72.11–77.01]	70.00 [67.12–72.87]	**74.96 [71.59–78.34]**
	SPE	64.33 [56.93–71.73]	69.16 [62.41–75.90]	64.10 [61.47–66.74]	68.00 [60.26–75.73]	**72.78 [68.99-76.56]**
	SEN	72.41 [62.27–82.55]	74.17 [69.61–78.72]	**81.12 [76.25–85.99]**	70.25 [65.28–75.22]	81.02 [77.67–84.37]
RF	AUC	72.23 [66.16–78.29]	74.75 [71.98–77.53]	74.12 [68.18–80.07]	76.09 [73.09–79.10]	**82.75 [77.73–87.77]**
	ACC	66.51 [57.44–75.58]	70.78 [66.39–75.17]	71.25 [68.19–74.30]	68.60 [66.89-70.30]	**74.56 [69.51–79.62]**
	SPE	53.04 [50.40–55.68]	57.89 [54.09–61.68]	56.99 [51.30–62.69]	56.11 [50.85–61.36]	**64.11 [61.04–67.18]**
	SEN	75.83 [60.59–91.08]	78.25 [73.87–82.62]	79.33 [77.14–81.52]	76.96 [75.20–78.71]	**84.02 [79.09–88.95]**
XGB	AUC	71.97 [68.06–75.89]	72.77 [67.60–77.94]	74.78 [71.13–78.43]	76.94 [71.31–82.58]	**77.42 [68.30–86.54]**
	ACC	66.51 [57.44–75.58]	67.23 [64.17–70.29]	67.36 [58.26–76.47]	71.36 [66.31–76.42]	**71.51 [68.64–74.39]**
	SPE	61.78 [58.46–65.09]	57.44 [44.47–70.41]	59.66 [47.24–72.08]	**62.44 [52.50–72.38]**	62.18 [52.95-71.40]
	SEN	69.79 [56.23–83.35]	72.24 [65.62–78.87]	72.50 [65.67–79.33]	77.62 [74.27–80.97]	**78.82 [74.55–83.10]**
SVM	AUC	75.61 [71.52–79.69]	80.50 [76.81–84.19]	74.27 [60.75–87.80]	77.87 [76.52–79.23]	**84.15 [79.76–88.54]**
	ACC	68.57 [66.30–70.84]	69.92 [65.81–74.03]	68.95 [55.82–82.07]	70.95 [66.85–75.05]	**75.27 [72.37–78.18]**
	SPE	60.77 [51.11–70.43]	68.97 [63.33–74.61]	56.44 [45.41–67.47]	68.88 [63.91–73.85]	**73.16 [68.91–77.41]**
	SEN	73.04 [67.76–78.31]	73.35 [65.98–80.71]	77.44 [63.43–91.45]	72.24 [68.93–75.56]	**78.32 [76.13–80.51]**

* The symbol represents the latent representation obtained by our proposed method and other fusion methods.

3.3. Validity of latent representation in MLDRL

In order to further evaluate the validity of latent representation in our method, we input the final latent representation of different fusion methods into different classifiers. The classifiers in our study including LR, RF, XGB and SVM. Table 5 shows the performance of latent representation obtained by our method and other methods using LR, RF, XGB and SVM. The latent representation of HD-Net is poorer than our proposed method but when using RF and XGB, it achieves better performance than that of the latent representations generated by EFM, LFM and D-AAE. The latent representation of our proposed method obtains the best AUC on all four classifiers when compared the latent representation of other fusion methods. These results demonstrate the validity of the latent representation generated by our proposed method from the longitudinal CT images.

3.4. Comparison between before fusion and after fusion

In order to better prove the effectiveness of the fusion mechanism proposed in this study, we conduct additional experiments before fusion, i.e., using the single stage data before nCRT and after nCRT respectively. Since the proposed cross-cycle reconstruction loss and the inherent-variational loss are computed based on the interactions and comparisons between the disentangled inherent and variational features from the longitudinal data. For the single stage data, due to lack of the other stage data, no inherent and variational features can be disentangled from this single stage data, leading to that no cross-cycle reconstruction and inherent-variational losses can be computed. Therefore, the single stage data is unsuited to the proposed MLDRL. In our study, to implement supervised pCR classification using each single stage data, we remove the multi-loss disentangled representation learning part from our proposed MLDRL, directly input the selected 81 features after feature selection from each single data into the multi-layer perceptron neural network used in our proposed MLDRL, and then compute a supervised classification loss using the supervised classification results via Eq. (6) to promote the training of the model. Besides, we use the DeLong test to decide whether there are significant differences between the AUC values of the fusion methods and their corresponding single stage methods: only using images before nCRT or after nCRT. As we can see from Fig. 2, our proposed MLDRL with fusion achieves significantly higher AUC (85.66%) than that of only using the images after nCRT (AUC=79.22%, $P < 0.05$), and obtains insignificantly higher AUC than that of only using the images before nCRT (AUC=81.39%, $P > 0.05$). The EFM, LFM, D-AAE and HD-Net with fusion achieve insignificantly higher AUC than those of their corresponding single-stage methods before fusion (all $P > 0.05$). These comparisons imply that there might be complementary information between the longitudinal data that is useful for pCR prediction. At the same time, these comparisons can also illustrate the effectiveness of our proposed method MLDRL in longitudinal data fusion. In addition, we also find that for all the fusion methods, the AUC of using images before nCRT is higher than that of using images after nCRT, which shows that images before nCRT is more effective than images after nCRT for the pCR prediction.

Table 6
The performance of our proposed method using the single loss and different combinations of losses.

CCR Loss	IV Loss	SC Loss	AUC [95% CI]	ACC [95% CI]	SPE [95% CI]	SEN [95% CI]
✓	–	–	68.48 [56.57–80.38]	64.87 [53.88-75.86]	55.53 [47.37–63.68]	70.69 [62.09–79.29]
–	✓	–	67.49 [57.04–77.95]	62.04 [52.48–71.60]	59.04 [49.93–68.14]	71.65 [56.58–86.73]
–	–	✓	75.82 [69.49–82.16]	72.10 [67.11–77.09]	62.72 [53.42–72.03]	75.41 [67.84–82.97]
✓	–	✓	80.04 [74.98–85.11]	73.16 [66.06–80.26]	62.61 [53.27–71.96]	79.15 [76.22–82.09]
–	✓	✓	79.54 [72.97–86.11]	74.42 [67.76–81.09]	63.31 [55.60–71.02]	78.51 [73.83–83.20]
✓	✓	–	71.11 [64.01–78.21]	67.77 [60.76–74.79]	60.82 [55.39–66.25]	70.04 [60.71–79.36]
✓	✓	✓	**86.55 [83.54–89.56]**	**80.96 [76.74–85.19]**	**68.38 [57.20–79.56]**	**87.45 [84.58–90.32]**

Note. CCR Loss, IV Loss and SC Loss represent cross-cycle reconstruction loss, inherent-variationa loss and supervised classification loss respectively.

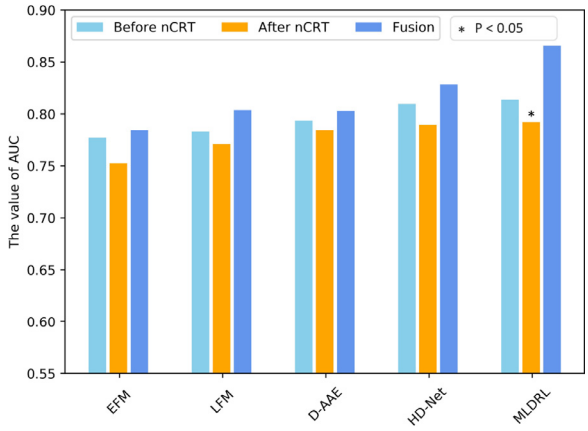

Fig. 2. The comparison between the five fusion methods and their single stage methods: only using images before nCRT or after nCRT.

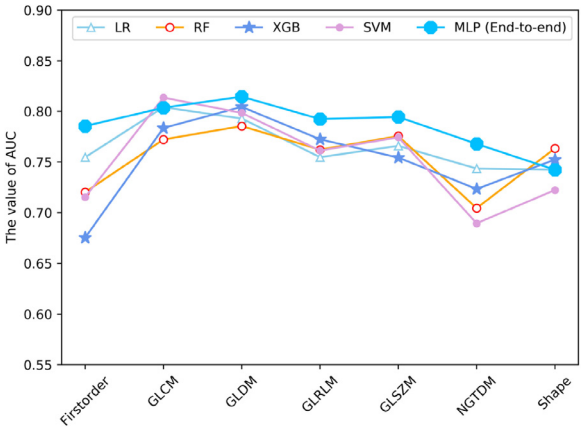

Fig. 3. The performance of our proposed method using features of each type.

3.5. Effectiveness analysis of each loss

In this section, we conduct the ablation experiments using the single loss and different combinations of the inherent-variational loss (IV Loss), the cross-cycle reconstruction loss (CCR Loss), and the supervised classification loss (SC Loss). As we can see from Table 6, compared with using only the CCR Loss, the IV Loss or the SC Loss, adding any loss that is different from itself can improve the performance of the model. For example, compared with the CCR Loss and the SC Loss alone, the AUC values are increased by 2.63% and 3.72%, respectively after adding the IV Loss. Compared with only using the CCR Loss or the IV Loss, the AUC values are increased by

11.56% and 12.05%, respectively after adding the SC Loss. These comparisons demonstrate that each of the three losses is effective and plays an important role in achieving better performance. Besides, compared with combinations of any two losses and the single loss, the combination of all the three losses achieves best performance. For example, compared with the combination of the IV Loss and the SC Loss, the AUC is increased by 7.01% after adding the CCR Loss; Compared with the combination of the CCR Loss and the SC Loss, the AUC is increased by 6.51% after adding the IV Loss; On the basis of the combination of the CCR Loss and the IV Loss, adding the SC Loss, the AUC is increased by 15.44%. Above analyses show that the combination of the three losses designed in this study is effective.

3.6. Effectiveness analysis for features of each type

In this section, we explore which type of features contributes the most to the pCR prediction. We evaluate the performance of our proposed method using features of each type. As we can see from Fig. 3, on average, GLDM performs best, followed by GLCM. Moreover, we find that GLRLM and GLSZM also perform well. At the same time, the performances of Firstorder and NGTDM are unstable in different classification methods. After the above analyses, we can conclude that the four types of features that contribute the most to pCR prediction are GLDM, GLCM, GLRLM and GLSZM.

3.7. Comparison of MLDRL with other methods for prognostic analysis

From the above analysis, we can know that MLDRL is superior to other fusion methods on Dataset A. As we know, pCR serves an essential indicator of prognosis, the fundamental purpose of predicting pCR is to make a judgment on the prognosis after nCRT. Therefore, we believe that it is important to explore the practical value of our method on the samples without surgery. In our study, we first use the established model to stratify these unlabeled samples into two groups (pCR and Non-pCR). Secondly, we calculate the BS values of the compared methods and our method. Last, we use the predicted results, the survival time and the survival statue to draw the Kaplan-Meier curves. In this section, we compare the ability of different methods for prognostic analysis. Specifically, we choose three comparison methods with the better performance on Dataset A, including LFM, D-AAE and HD-Net. Table 7 shows the BS values of our proposed method and the three compared methods on multi-center unlabeled datasets. As we can see from Table 7, the BS values of our proposed method on Dataset B, C and D are 0.2723, 0.3124, and 0.2984, respectively, are all smaller than those of the three compared methods. Besides, the smaller 95% CI demonstrates that our proposed method is robust for prognostic analysis. In addition, as shown in Fig. 4, we also find that the p-values of our method are also superior to its competitors on three unlabeled datasets. Overall, compared with other methods, MLDRL can obtain batter performance on all unlabeled datasets for prognosis analysis. These results also demonstrate that MLDRL has the

Table 7
Comparison of different methods on Brier Score (BS) for prognostic analysis, the smaller BS indicates better prognostic results.

Dataset	LFM [95% CI]	D-AAE [95% CI]	HD-Net [95% CI]	MLDRL [95% CI]
Dataset B	0.3625 [0.3354–0.3895]	0.3192 [0.3151–0.3233]	0.3587 [0.2986–0.4188]	**0.2723 [0.2604–0.2842]**
Dataset C	0.4024 [0.3753–0.4295]	0.3925 [0.3802–0.4048]	0.3986 [0.3776–0.4196]	**0.3124 [0.3051–0.3198]**
Dataset D	0.3664 [0.3382–0.3946]	0.3724 [0.3689–0.3759]	0.3387 [0.3350–0.3424]	**0.2984 [0.2949–0.3019]**

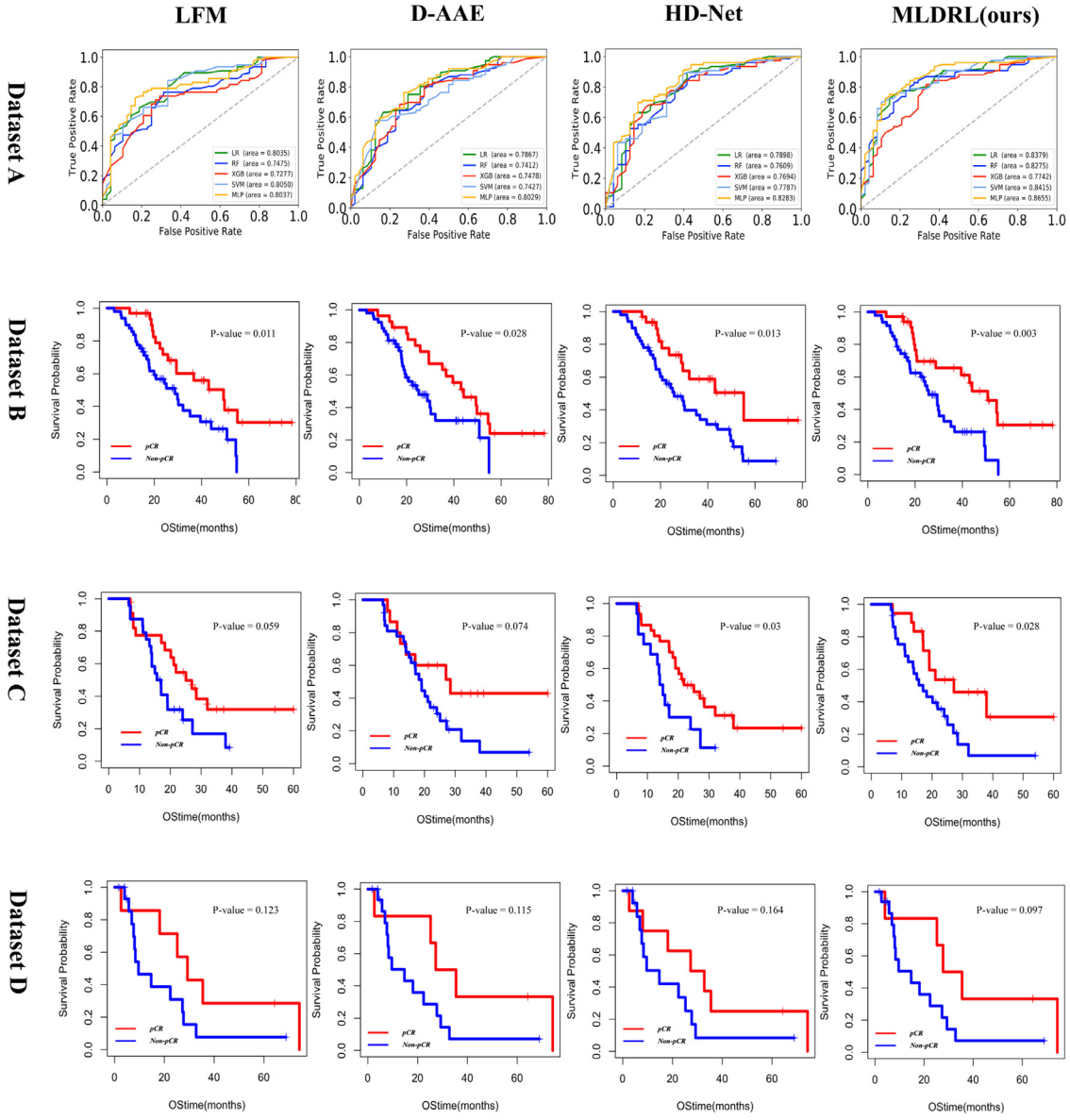

Fig. 4. The AUC and Kaplan-Meier curve of different methods on multi-center datasets. The first row shows the performance of different methods on dataset A, and the second, third and fourth rows are the survival analysis of different methods on multi-center unlabeled datasets.

potential to accurately predict pCR for esophageal patients after nCRT.

From what has been discussed above, our proposed MLDRL not only achieves the best performance for pCR prediction on Dataset A, but also achieves the best performance for prognostic analysis on the multi-center unlabeled datasets. In this section, we visualize the relationship between pCR prediction performance and prognostic analysis, as shown in Fig. 4. It is not difficult to find that the classification performance and prognostic analysis are positively correlated. This phenomenon is consistent with some previous studies (Meguid et al., 2009; Tong et al., 2010; van Hagen et al., 2013).

4. Discussion

Early identification of pCR offers an important basis for precision treatment and avoids unnecessary surgery-associated morbidity. In this study, we establish a novel method to accurately predict pCR. At the same time, we evaluate the feasibility of MLDRL on multi-center unlabeled datasets for prognosis analysis. Compared with other fusion methods, our proposed method not only achieves the best performance on Dataset A, but also achieves the best performance for prognostic analysis on the multi-center unlabeled datasets.

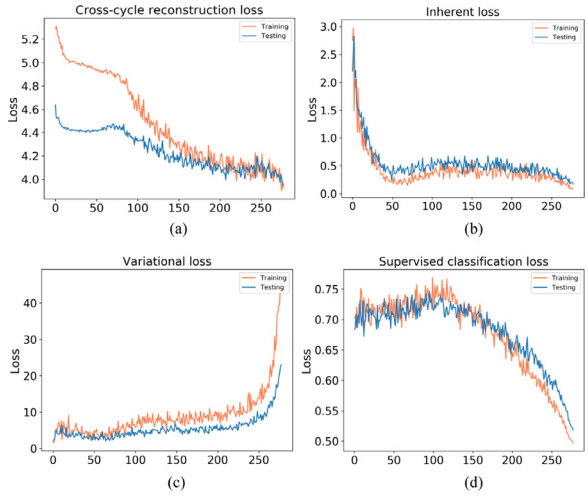

Fig. 5. Iterative process of the cross-cycle reconstruction loss (a), the inherent loss (b), the variational loss (c) and supervised classification loss (d) of our proposed method MLDRL.

In order to effectively fuse radiomics features of longitudinal CT images, we define a multi-loss function, which consists of a cross-cycle reconstruction loss, an inherent-variational loss and a supervised classification loss. Among them, the cross-cycle reconstruction loss and the inherent-variational loss are used to explore the shared and complementary information of radiomics features in longitudinal CT images, while the supervised classification loss is used to make the latent representation more separable. The experimental results show that the multi-loss function we designed can effectively explore the complementary information of features in longitudinal CT images. At the same time, we find that the three proposed losses can complement each other and promote the learning of the model.

In our study, the multi-loss function is composed of multiple loss terms with different effects. To better train our network, an adaptive gradient normalization algorithm is applied to balance the training of multiple loss terms by dynamically tuning the gradient magnitudes. Specifically, the adaptive gradient normalization algorithm can achieve the following goals: (1) make the cross-cycle reconstruction loss be reduced; (2) make the inherent features of longitudinal images become more and more similar; (3) make the variational features of longitudinal images become more and more different; (4) make the latent representation more separable. As shown in Fig. 5, the variations of these losses are consistent as we expect. Specifically, as the training process goes on, the cross-cycle reconstruction loss, the inherent loss, and the supervised classification loss decrease, and the variational loss increases, which also shows the effectiveness of the adaptive gradient normalization algorithm. In addition, by applying the algorithm, we only need to adjust one parameter and three initial weights. Compared with manually adjusting multiple weights from a wide range, this method has better efficiency.

In our study, we use radiomics features as input to train our network. In order to illustrate the reasons why we use the radiomics features to train the model, we conduct the following experiments. We use 3D-AlexNet (Krizhevsky et al., 2012) and 3D-ResNet (He et al., 2016) to automatically learn features from the longitudinal data, and directly input these features into our proposed MLDRL (denoted as 3D-AlexNet* and 3D-ResNet*, respectively) via replacing the fully connected layer in 3D-AlexNet and 3D-ResNet by the input layer in our proposed MLDRL while keeping other layers unchanged, resulting in end-to-end networks. Besides, in order to further explore the effectiveness of our proposed method and the necessity of radiomics features, we compare our proposed MLDRL with the following three multi-modality or multi-input end-to-end deep learning methods which can directly use the original longitudinal CT images to make the final prediction. Dey et al. (2018) proposed a two-pathway 3D DenseNet (denoted as 3D-DenseNet) and a 3D augmented multi-output DenseNet (denoted as 3D-MODenseNet) to classify the benign and malignant lung nodules in CT images. Xu et al. (2020) used multi-scale cost-sensitive ResNet-based neural networks (denoted as MSCS) to evaluate lung nodule malignancy. It should be noted that the MSCS in our experiment is slightly different from the original MSCS. We use longitudinal data of the same size as input instead of inputting multi-scale data. The rest of the network structures are the same as the original MSCS. For more fair comparison, we load the pre-trained weights provided by Med3D Chen et al. (2019) to train the ResNet-based compared methods: 3D-ResNet* and MSCS, respectively (named 3D-ResNet*† and MSCS†). Since the fully connected layer of the original 3D-ResNet is changed, we only load the pre-trained weights outside the fully connected layer and randomly initialize the weights for other layers to ensure the effective loading of the pre-trained weights. Similarly, for MSCS, we only load the pre-trained weights for the layers that are the same in the original 3D ResNet and MSCS to successfully implement pre-trained. All above comparisons are shown in Table 8. As we can see from the Table 8, the performance of our proposed method using radiomics features is better than those of methods using deep learning features extracted by 3D-AlexNet and 3D-ResNet. The results demonstrate the radiomics features extracted in our study are better than the deep learning features learned by 3D-AlexNet and 3D-ResNet. We also find that our proposed MLDRL outperforms all the three multi-modality or multi-input end-to-end deep learning methods in terms of AUC, ACC, SPE and SEN. The AUC of 3D-ResNet*† and MSCS† are 78.79% and 77.35%, respectively, which are higher than those of 3D-ResNet* and MSCS. These results validate that the methods with pre-trained are indeed better than the models trained from scratch. Besides, our proposed MLDRL still outperforms 3D-ResNet*†, MSCS† and other multi-input deep learning methods such as 3D-MODenseNet in terms of all metrics, which show the superiority of our proposed MLDRL. All above comparisons with end-to-end deep learning methods on the same longitudinal CT images show that the effectiveness of our proposed method and the necessity of the usage of radiomics features to accurate predict the response of esophageal cancer after nCRT. Therefore, in our study, we use radiomics features as input to train our network.

Although our proposed method has improved results on both Dataset A and multi-center unlabeled datasets, there are still some limitations. First, due to the lack of prognostic information of patients with surgery, we do not perform prognostic analysis on Dataset A. In the future, we will collect the prognostic information of patients with surgery and conduct prognostic analysis on them for further validating the effectiveness of the proposed method. Second, although our method has been verified on multi-center unlabeled datasets, the number of samples of multi-center is limited, and we need to collect more data to prove the reliability of our method. In addition, hand-crafted radiomics features rather than deep learning features are used as input of the proposed method. Since the deep learning features have achieved good performance for many image classification tasks, we will collect more data and develop a deep learning network which can automatically learn the latent features from the original longitudinal images for predicting pCR. Finally, our research only base on pCR to make the prognostic analysis, does not directly predict survival time. In future, our goal is to directly predict survival time following this research.

Table 8
The performance of our proposed method and other compared methods using original CT images.

Method	AUC [95% CI]	ACC [95% CI]	SPE [95% CI]	SEN [95% CI]
3D-AlexNet* (Krizhevsky et al., 2012)	76.80 [68.99–84.61]	74.03 [66.76–79.30]	67.37 [62.01–72.73]	78.73 [74.37–83.10]
3D-ResNet* (He et al., 2016)	77.60 [74.91–80.29]	73.46 [68.87–78.06]	67.93 [60.32–75.55]	76.73 [69.73–83.73]
3D-DenseNet (Dey et al., 2018)	76.15 [68.60–83.69]	72.83 [68.76–76.89]	64.41 [57.59–71.23]	78.08 [73.43–82.74]
3D-MODenseNet (Dey et al., 2018)	76.63 [72.51–80.75]	73.09 [69.29–76.89]	65.59 [60.14–71.05]	78.56 [72.50–84.63]
MSCS (Xu et al., 2020)	77.16 [70.87–83.46]	73.75 [70.91–76.59]	66.63 [58.41–74.85]	81.44 [74.71–88.17]
3D-ResNet*† (He et al., 2016)	78.79 [73.71–83.87]	75.34 [67.29–83.38]	67.51 [58.61–76.41]	77.13 [68.57–85.68]
MSCS† (Xu et al., 2020)	77.35 [72.34–82.36]	73.97 [69.20–78.74]	67.56 [58.19–76.93]	80.78 [70.60–90.96]
MLDRL (Ours)	**86.55 [83.54–89.56]**	**80.96 [76.74–85.19]**	**68.38 [57.20–79.56]**	**87.45 [84.58–90.32]**

* The symbol represents the method using 3D deep learning methods to automatically learn features, and then input these features into our proposed MLDRL, † represents the method trained using the pre-trained model.

5. Conclusion

In our study, we propose a multi-loss disentangled representation learning to accurately predict the pCR of esophageal cancer patients using longitudinal CT images. Specifically, we first define a cross-cycle reconstruction loss and an inherent-variational loss to find the shared and complementary features in longitudinal CT images. Then, we propose a supervised classification loss to make latent representations separable. Finally, we use an adaptive gradient normalization algorithm to dynamically tune the weights of different loss terms. Besides, our proposed method is also validated on multi-center unlabeled datasets. The experimental results show that MLDRL is not only superior in pCR prediction, but also achieves better performance in the prognostic analysis of multi-center unlabeled datasets. Overall, our proposed method can accurately predict pCR of esophageal cancer after nCRT.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Hailin Yue: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Jin Liu:** Conceptualization, Methodology, Writing – review & editing, Formal analysis. **Junjian Li:** Methodology, Writing – review & editing, Formal analysis. **Hulin Kuang:** Conceptualization, Methodology, Writing – review & editing, Formal analysis. **Jinyi Lang:** Resources, Data curation, Investigation. **Jianhong Cheng:** Methodology, Writing – review & editing, Formal analysis. **Lin Peng:** Resources, Data curation, Investigation. **Yongtao Han:** Resources, Data curation, Investigation. **Harrison Bai:** Conceptualization, Formal analysis, Writing – review & editing. **Yuping Wang:** Conceptualization, Formal analysis, Writing – review & editing. **Qifeng Wang:** Resources, Data curation, Investigation, Supervision. **Jianxin Wang:** Conceptualization, Methodology, Funding acquisition, Supervision, Project administration.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China under Grant (Nos. 62102454, 62172444, and 61877059), the Natural Science Foundation of Hunan Province under Grant No. 2019JJ50775, the 111 Project (No.B18059), the Hunan Provincial Science and Technology Innovation Leading Plan (No.2020GK2019), the Fundamental Research Funds for the Central Universities of Central South University (2021zzts0741) and Department of Science and Technology of Sichuan Province (2019YFS0378).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2022.102423.

References

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 68 (6), 394–424.

Carneiro, G., Nascimento, J., Bradley, A.P., 2017. Automated analysis of unregistered multi-view mammograms with deep learning. IEEE Trans. Med. Imaging 36 (11), 2355–2365.

Chen, R.T.Q., Li, X., Grosse, R.B., Duvenaud, D.K., 2018. Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 2610–2620.

Chen, S., Ma, K., Zheng, Y., Med3d: transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A., 2018. Gradnorm: gradient normalization for adaptive loss balancing in deep multitask networks. In: Proceedings of the International Conference on Machine Learning, pp. 794–803.

Cheng, J., Liu, J., Yue, H., Bai, H., Pan, Y., Wang, J., 2020. Prediction of glioma grade using intratumoral and peritumoral radiomic features from multiparametric MRI images. IEEE ACM Trans. Comput. Biol. Bioinf..

Chiang, K.C., Yeh, C.N., Ueng, S.H., Hsu, J.T., Yeh, T.S., Jan, Y.Y., Hwang, T.L., Chen, M.F., 2012. Clinicodemographic aspect of resectable pancreatic cancer and prognostic factors for resectable cancer. World J. Surg. Oncol. 10 (1), 1–9.

Cho, H.H., Lee, S.H., Kim, J., Park, H., 2018. Classification of the glioma grading using radiomics analysis. PeerJ 6, e5982.

Cortes, C., Vapnik, V., 1995. Support vector machine. Mach. Learn. 20 (3), 273–297.

Dey, R., Lu, Z., Hong, Y., 2018. Diagnostic classification of lung nodules using 3d neural networks. In: Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 774–778. 2018

Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ayed, I.B., 2018. Hyperdense-net: a hyper-densely connected CNN for multi-modal image segmentation. IEEE Trans. Med. Imaging 38 (5), 1116–1126.

Guo, M., Haque, A., Huang, D.A., Yeung, S., Fei-Fei, L., 2018. Dynamic task prioritization for multitask learning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 270–287.

Guo, Y., Decazes, P., Becker, S., Li, H., Ruan, S., 2020. Deep disentangled representation learning of pet images for lymphoma outcome prediction. In: Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1–4.

van Hagen, P., Hulshof, M., Van Lanschot, J., Steyerberg, E.W., Henegouwen, M.I.V.B., Wijnhoven, B., Richel, D.J., Nieuwenhuijzen, G., Hospers, G., Bonenkamp, J.J., et al., 2012. Preoperative chemoradiotherapy for esophageal or junctional cancer. N. Engl. J. Med. 366 (22), 2074–2084.

van Hagen, P., Wijnhoven, B., Nafteux, P., Moons, J., Haustermans, K., De Hertogh, G., Van Lanschot, J., Lerut, T., 2013. Recurrence pattern in patients with a pathologically complete response after neoadjuvant chemoradiotherapy and surgery for oesophageal cancer. Br. J. Surg. 100 (2), 267–273.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hou, T.C., Huang, W.C., Tai, H.C., Chen, Y.J., 2019. Integrated radiomic model for predicting the prognosis of esophageal squamous cell carcinoma patients undergoing neoadjuvant chemoradiation. Ther. Radiol. Oncol. 3 (0).

Hu, D., Zhang, H., Wu, Z., Wang, F., Wang, L., Smith, J.K., Lin, W., Li, G., Shen, D., 2020. Disentangled-multimodal adversarial autoencoder: application to infant age prediction with incomplete multimodal neuroimages. IEEE Trans. Med. Imaging 4137–4149.

Jin, C., Yu, H., Ke, J., Ding, P., Yi, Y., Jiang, X., Duan, X., Tang, J., Chang, D.T., Wu, X., et al., 2021. Predicting treatment response from longitudinal images using multi-task deep learning. Nat. Commun. 12 (1), 1–11.

Jouirou, A., Baâzaoui, A., Barhoumi, W., 2019. Multi-view information fusion in mammograms: a comprehensive overview. Inf. Fus. 52, 308–321.
Kang, H., Xia, L., Yan, F., Wan, Z., Shi, F., Yuan, H., Jiang, H., Wu, D., Sui, H., Zhang, C., et al., 2020. Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning. IEEE Trans. Med. Imaging 2606–2614.
Kawakubo, H., Yoshida, H., 2012. Rapid feature selection based on random forests for high-dimensional data. Expert Syst. Appl 40, 6241–6252.
Kendall, A., Gal, Y., Cipolla, R., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7482–7491.
Khan, H.N., Shahid, A.R., Raza, B., Dar, A.H., Alquhayz, H., 2019. Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. IEEE Access 7, 165724–165733.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.
Larue, R.T., Klaassen, R., Jochems, A., Leijenaar, R.T.H., Hulshof, M.C., van Berge Henegouwen, M.I., Schreurs, W.M.J., Sosef, M.N., van Elmpt, W., van Laarhoven, H.W.M., et al., 2018. Pre-treatment CT radiomics to predict 3-year overall survival following chemoradiotherapy of esophageal cancer. Acta Oncol. 57 (11), 1475–1481.
Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H., 2018. Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 35–51.
Lei, B., Yang, P., Wang, T., Chen, S., Ni, D., 2017. Relational-regularized discriminative sparse learning for alzheimer's disease diagnosis. IEEE Trans Cybern 47 (4), 1102–1113.
Liaw, A., Wiener, M., et al., 2002. Classification and regression by random forest. R news 2 (3), 18–22.
Liu, S., Johns, E., Davison, A.J., 2019. End-to-end multi-task learning with attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1871–1880.
Meguid, R.A., Hooker, C.M., Taylor, J.T., Kleinberg, L.R., Cattaneo II, S.M., Sussman, M.S., Yang, S.C., Heitmiller, R.F., Forastiere, A.A., Brock, M.V., 2009. Recurrence after neoadjuvant chemoradiation and surgery for esophageal cancer: does the pattern of recurrence differ for patients with complete response and those with partial or no response? J. Thorac. Cardiovasc. Surg. 138 (6), 1309–1317.
Peng, C.-Y.J., Lee, K.L., Ingersoll, G.M., 2002. An introduction to logistic regression analysis and reporting. J. Educ. Res. 96 (1), 3–14.
Pi, Y., Zhao, Z., Xiang, Y., Li, Y., Cai, H., Yi, Z., 2020. Automated diagnosis of bone metastasis based on multi-view bone scans using attention-augmented deep neural networks. Med. Image Anal. 65, 101784.
Qian, W., Ge, X.X., Wu, J., Gong, F.R., Wu, M.Y., Xu, M.D., Lian, L., Wang, W.J., Li, W., Tao, M., 2019. Prognostic evaluation of resectable colorectal cancer using platelet-associated indicators. Oncol. Lett. 18 (1), 571–580.
Shao, W., Wang, T., Sun, L., Dong, T., Han, Z., Huang, Z., Zhang, J., Zhang, D., Huang, K., 2020. Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers. Med. Image Anal. 65, 101795.
Shapiro, J., Van Lanschot, J.J.B., Hulshof, M.C., van Hagen, P., van Berge Henegouwen, M.I., Wijnhoven, B.P.L., van Laarhoven, H.W.M., Nieuwenhuijzen, G.A.P., Hospers, G.A.P., Bonenkamp, J.J., et al., 2015. Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (CROSS): long-term results of a randomised controlled trial. Lancet Oncol. 16 (9), 1090–1098.
Tan, X., Ma, Z., Yan, L., Ye, W., Liu, Z., Liang, C., 2019. Radiomics nomogram outperforms size criteria in discriminating lymph node metastasis in resectable esophageal squamous cell carcinoma. Eur. Radiol. 29 (1), 392–400.
Tong, D.K.H., Law, S., Kwong, D.L.W., Chan, K.W., Lam, A.K.Y., Wong, K.H., 2010. Histological regression of squamous esophageal carcinoma assessed by percentage of residual viable cells after neoadjuvant chemoradiation is an important prognostic factor. Ann. Surg. Oncol. 17 (8), 2184–2192.
Wu, L., Wang, C., Tan, X., Cheng, Z., Zhao, K., Yan, L., Liang, Y., Liu, Z., Liang, C., 2018. Radiomics approach for preoperative identification of stages i- II and III- IV of esophageal cancer. Chin. J. Cancer Res. 30 (4), 396.
Xu, X., Wang, C., Guo, J., Gan, Y., Wang, J., Bai, H., Zhang, L., Li, W., Yi, Z., 2020. Mscs-deepln: evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks. Med. Image Anal. 65, 101772.
Yang, Z., He, B., Zhuang, X., Gao, X., Wang, D., Li, M., Lin, Z., Luo, R., 2019. Ct-based radiomic signatures for prediction of pathologic complete response in esophageal squamous cell carcinoma after neoadjuvant chemoradiotherapy. J. Radiat. Res. 60 (4), 538–545.
Yokoyama, S., Hamada, T., Higashi, M., Matsuo, K., Maemura, K., Kurahara, H., Horinouchi, M., Hiraki, T., Sugimoto, T., Akahane, T., et al., 2020. Predicted prognosis of patients with pancreatic cancer by machine learning. Clin. Cancer Res. 26 (10), 2411–2421.
Zanoni, A., Verlato, G., Giacopuzzi, S., Weindelmayer, J., Casella, F., Pasini, F., Zhao, E., de Manzoni, G., 2013. Neoadjuvant concurrent chemoradiotherapy for locally advanced esophageal cancer in a single high-volume center. Ann. Surg. Oncol. 20 (6), 1993–1999.
Zhang, D., Shen, D., Initiative, A.D.N., et al., 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. Neuroimage 59 (2), 895–907.
Zhou, T., Thung, K.-H., Liu, M., Shi, F., Zhang, C., Shen, D., 2020. Multi-modal latent space inducing ensemble SVM classifier for early dementia diagnosis with neuroimaging data. Med. Image Anal. 60, 101630.
Zhu, X., Suk, H.I., Lee, S.W., Shen, D., 2016. Canonical feature selection for joint regression and multi-class identification in alzheimer's disease diagnosis. Brain Imaging Behav. 10 (3), 818–828.