Physics in Medicine & Biology

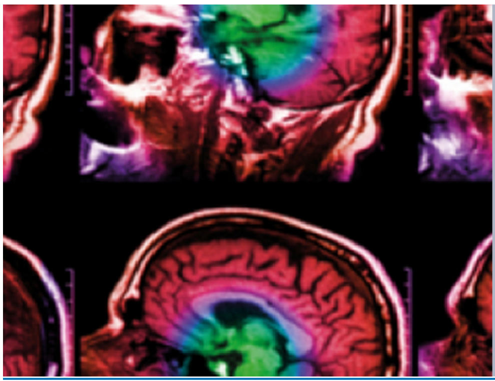
IPEM
Institute of Physics and
Engineering in Medicine

PAPER

Locoregional recurrence prediction in head and neck cancer based on multi-modality and multi-view feature expansion

To cite this article: Rongfang Wang *et al* 2022 *Phys. Med. Biol.* **67** 125004

View the article online for updates and enhancements.



IOP Publishing Phys. Med. Biol. 67 (2022) 125004 https://doi.org/10.1088/1361-6560/ac72f0

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
7 December 2021

REVISED
10 May 2022

ACCEPTED FOR PUBLICATION
24 May 2022

PUBLISHED
8 June 2022

PAPER

Locoregional recurrence prediction in head and neck cancer based on multi-modality and multi-view feature expansion

Rongfang Wang[1,2], **Jinkun Guo**[1], **Zhiguo Zhou**[3], **Kai Wang**[2], **Shuiping Gou**[1], **Rongbin Xu**[4], **David Sher**[2] **and Jing Wang**[2,*]

[1] School of Artificial Intelligence, Xidian University, Xi'an 710071, People's Republic of China
[2] Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75235, United States of America
[3] School of Computer Science and Mathematics, University of Central Missouri, Warrensburg, MO 64093, United States of America
[4] School of Information Engineering, Putian Unviersity, Putian 351100, People's Republic of China
[*] Author to whom any correspondence should be addressed.

E-mail: Jing.Wang@utsouthwestern.edu

Keywords: head and neck cancer, locoregional recurrence prediction, multi-modality, multi-view feature expansion, convolutional neural network

Abstract

Objective. Locoregional recurrence (LRR) is one of the leading causes of treatment failure in head and neck (H&N) cancer. Accurately predicting LRR after radiotherapy is essential to achieving better treatment outcomes for patients with H&N cancer through developing personalized treatment strategies. We aim to develop an end-to-end multi-modality and multi-view feature extension method (MMFE) to predict LRR in H&N cancer. *Approach.* Deep learning (DL) has been widely used for building prediction models and has achieved great success. Nevertheless, 2D-based DL models inherently fail to utilize the contextual information from adjacent slices, while complicated 3D models have a substantially larger number of parameters, which require more training samples, memory and computing resources. In the proposed MMFE scheme, through the multi-view feature expansion and projection dimension reduction operations, we are able to reduce the model complexity while preserving volumetric information. Additionally, we designed a multi-modality convolutional neural network that can be trained in an end-to-end manner and can jointly optimize the use of deep features of CT, PET and clinical data to improve the model's prediction ability. *Main results.* The dataset included 206 eligible patients, of which, 49 had LRR while 157 did not. The proposed MMFE method obtained a higher AUC value than the other four methods. The best prediction result was achieved when using all three modalities, which yielded an AUC value of 0.81. *Significance.* Comparison experiments demonstrated the superior performance of the MMFE as compared to other 2D/3D-DL-based methods. By combining CT, PET and clinical features, the MMFE could potentially identify H&N cancer patients at high risk for LRR such that personalized treatment strategy can be developed accordingly.

1. Introduction

Head and neck (H&N) cancer was the seventh most common cancer worldwide in 2018 (Bray *et al* 2018), accounting for 5%–10% of new cancer cases in developed countries (Bogowicz *et al* 2019). Radiotherapy plays an important role in H&N cancer management, and definitive chemoradiation has markedly improved treatment outcomes for patients with H&N cancer (Caudell *et al* 2017). The overall five-year survival rates of oral cancer, pharyngeal cancer and laryngeal cancer are as high as 61%, 41%, and 69%, respectively (Nahavandipour *et al* 2019). However, 15%–50% of H&N cancer patients still experience local recurrence (LRR), most of which occurs within three years of treatment (Beaumont *et al* 2019, Keek *et al* 2020, Wang *et al* 2020a). For patients at high risk of LRR, intensified treatment such as additional systemic therapy may reduce the risk of treatment failure. Therefore, identifying patients at high risk for LRR after radiotherapy could lead to better treatment outcomes for patients with H&N cancer.

With the development of machine learning, many predictive models have been proposed for patient risk stratification (Wang *et al* 2020a, 2020b, Shanthi and Rajkumar 2021). In recent years, deep learning as a powerful tool has been widely used in treatment outcome prediction and has achieved great success (Zhu *et al* 2020). Due to the constraints of GPU memory, many deep learning-based methods employ two-dimensional (2D) slices to predict treatment outcomes (Diamant *et al* 2019, Le *et al* 2020, Saha *et al* 2020, Rose *et al* 2021). However, since 2D convolutional neural networks (CNNs) take a single slice as input, they inherently fail to utilize the context from adjacent slices and therefore cannot fully use the volumetric information. Three-dimensional (3D) CNNs address this issue by using a volume patch of a scan as input. Many 3D CNN-based outcome prediction methods have been proposed that can leverage inter-slice context to improve performance. For example, Yang *et al* (2019*)* developed a deep convolutional neural network that uses 3D PET image data to classify patients who died within one year or survived more than one year after diagnosis of esophageal cancer. Zhao *et al* (2020) proposed a cross-modal deep learning system that integrates preoperative clinical knowledge and 3D CT images into the neural network to improve the accuracy of lymph node metastasis prediction. Starke *et al* (2020) explored different deep learning methods' ability to predict locoregional tumor control of H&N cancer from CT images. Their work proposed and tested different deep learning methods, among which a 3D-CNN-based method achieved the best results, as it can learn 3D environmental information around the tumor. However, 3D CNNs come with a high computational cost resulting from the increased number of parameters, which also have a high demand for memory and computing resources.

To solve the dilemma posed by 2D CNNs lacking volumetric information and 3D CNNs yielding too high a model complexity, in this paper, we proposed an end-to-end multi-modality and multi-view feature extension CNN method (MMFE) to predict LRR in H&N cancer. Multi-modality data are more comprehensive than single-modality data and can provide complementary information on different aspects of the patient (Branstetter *et al* 2005, Baltrušaitis *et al* 2018, Lv *et al* 2019, Rose *et al* 2021). For example, positron emission tomography (PET) images reveal the molecular metabolism activities within the human body, while computed tomography (CT) reflects the attenuation coefficient to x-rays (Guo *et al* 2019). In addition, clinical data such as age, primary site, and T- and N-stage can provide patient-specific features beyond what images reveal, which may further improve the performance of the prediction model (Chang *et al* 2017, Beesley *et al* 2019).

In MMFE, we first proposed two strategies to generate clinical and image representations from clinical data and CT/PET, respectively: (1) an attribute-based hybrid coding strategy and (2) a multi-view-based feature expansion and dimensionality reduction strategy. In the first strategy, we designed different ways to code according to different attributes, so that more reasonable clinical representations can be generated. The strategy of generating image representations is motivated by a multi-view CNN method (Su *et al* 2015). By constructing a descriptor based on the aggregation of multi-view information according to the 3D shape, we obtained better classification results than with models using 2D or 3D data. The multi-view CNN method has been explored for medical imaging analyses, as medical imaging is naturally represented as a 3D volume, and the sample size is often smaller than natural images. Most existing studies use three orthogonal views (transaxial, coronal, and sagittal) as multi-view data (Wang *et al* 2017, Wei *et al* 2019, Xia *et al* 2020, Zhou *et al* 2020) and merge information at the decision-making level, or they use a splicing method to fuse information extracted from different views at the feature level. Hu *et al* (2020) proposed a model that contains multiple 2D and 3D CNN branches, where three orthogonal views are used as the inputs for three 2D networks, and the original 3D PET images are used as the inputs of 3D networks, which results in a highly complex model. In the multi-view method proposed by El-Regaily *et al* (2020), the single slices with the largest tumor area in the transaxial, coronal, and sagittal directions are used as the three views, which may lose spatial information from a 3D tumor.

Unlike the aforementioned methods, our strategy includes vertical rotation for feature expansion and horizontal projection to reduce dimensionality. In the vertical rotation step, we rotate the 3D CT/PET volumetric data along the vertical axis, then collect a projection image at every rotated degree. This process is similar to the process of generating projection data during CT, where the x-ray source and detector rotate around the patient (Townsend 2004, Seeram and Tomography-E-book 2009). Therefore, image representations generated at these different vertical angle views could effectively expand the features. In the horizontal projection step, all the slices of one view are projected by an average operator along the horizontal axis. Compared with directly using a certain single slice, this approach can reduce dimensionality and significantly decrease the number of model parameters while preserving the spatial structure information. Finally, we designed a multi-modality deep neural network that can be trained in an end-to-end manner and can jointly optimize the use of deep features of CT, PET and clinical data to improve the model's prediction ability. Comparison experiments demonstrated that the proposed MMFE obtains more accurate prediction results than 3D CNNs under the same structure and outperforms 2D CNN-based methods that focus on a single slice of an image.

Table 1. Illustration of the coding of six clinical variables.

Age		Primary site		T-stage		N-stage		HPV status		Therapy	
Value	Coding	Value	Coding	Value	Coding	Value	Coding	Value	Coding	Value	Coding
Age	Age/100	Larynx	1 0 0 0	T1	1/4	N0	1/4	Negative	1 0 0	Radiation	1 0
		Nasopharynx	0 1 0 0	T2	2/4	N1	2/4	Positive	0 1 0	Chemo radiation	0 1
		Oropharynx	0 0 1 0	T3	3/4	N2	3/4	Unknown	0 0 1		
		Hypopharynx	0 0 0 1	T4	4/4	N3	4/4				

2. Materials and pre-processing

2.1. Materials

The dataset used in this study contains the medical images and clinical variables of patients with H&N cancer who received radiotherapy from the university of texas southwestern medical center (UTSW) (Dallas, TX, USA) between september 2005 and november 2015. The median follow-up duration was 37 months. Tumor contours used for the analysis were based on clinically approved clinical target volumes. Clinical information, such as age, gender, T-stage, N-stage and disease site status, were collected from patient charts to build the model. Patients with a follow-up period of less than one year were excluded. The remaining dataset included 206 eligible patients, of which, 49 had LRR during the follow-up period and 157 did not. We use all 206 samples for five-fold cross-validation in our method and the comparison experiments.

2.2. Clinical data pre-processing

A total of six clinical variables were used in this study: *Age*, *primary site* (larynx, nasopharynx, oropharynx, hypopharynx), *T-stage* (T1-T4), *N-stage* (N0-N3), human papillomavirus status (*HPV status*) (negative, positive, unknown) and *Therapy* (radiation, chemoradiation).

Firstly, we handled the missing values in the clinical data for a few cases. Approaches to address missing values are usually categorized as deletion-based methods and imputation-based methods (Soley-Bori 2013). Deletion methods simply exclude cases with missing values, which may remove a large amount of potentially usable information. Imputation methods usually replace each missing value with another value determined from a reasonable guess. In this work, we adopted a K-nearest neighbor (KNN)-based imputation method (Faisal and Tutz 2017) to fill in each missing attribute value. The detailed steps are as follows: (1) all samples X are divided into two subsets: $\mathbf{X} = \mathbf{X}^{\mathrm{m}} + \mathbf{X}^{\mathrm{nm}}$, where $\mathbf{X}^{\mathrm{m}}$ denotes the sample set with missing value, and $\mathbf{X}^{\mathrm{nm}}$ denotes the sample set without missing value; (2) for each $\boldsymbol{x}_i \in \mathbf{X}^{\mathrm{m}}$, we use the KNN method to find the K ($K = 3$) nearest neighbors from the set $\mathbf{X}^{\mathrm{nm}}$, and only consider the items without missing value to measure the distance; (3) we replace the missing value of $\boldsymbol{x}_i \in \mathbf{X}^{\mathrm{m}}$ with the median value of corresponding item from the three nearest neighbors.

We adopted an attribute-based hybrid coding strategy to generate the clinical representations. Specifically, the clinical variables were first divided into three categories: numerical, nominal, and ordinal attributes. Then, we used three different coding methods to code the corresponding clinical variables:

(1) Numerical variable. In this work, there is only one numerical feature: age. The value of this feature is divided by 100 directly and normalized as one-bit code.

(2) Nominal variables. The nominal features in this work are primary site, HPV status and therapy. For nominal features, the variable comprises a finite set of discrete values with no ordinal relationship and no overlap between values, so we used One-Hot Encoding (Zheng and Casari 2018) to represent each item. According to the number of values, these three nominal features were converted to four-, three-, and two-bit code, respectively.

(3) Ordinal variables. The ordinal variables in this work are T-Stage and N-Stage. For ordinal features, the variable can be ordered and ranked, and the value has a partial order relationship. For example, the higher the stage, the worse the prognosis of the tumor will generally be. We first converted the original order to a numerical order using integer encoding; then, each unique value was divided by the maximum value of the corresponding feature and normalized as one-bit code.

All the codes were spliced together to generate a final clinical code of total twelve bits, which served as the input samples for the clinical modality of the prediction model. An illustration of the coding for the six clinical variables is shown in table 1.

2.3. Image pre-processing

For the CT and PET image pre-processing, we used a multi-view-based feature expansion and dimensionality reduction strategy to generate the image representations. This strategy involves multiple steps: (1) resolution standardization and normalization; (2) vertical rotation for feature expansion; (3) horizontal projection for dimensionality reduction; and (4) tumor region sample construction on PET and CT. The details of each step are described in the following.

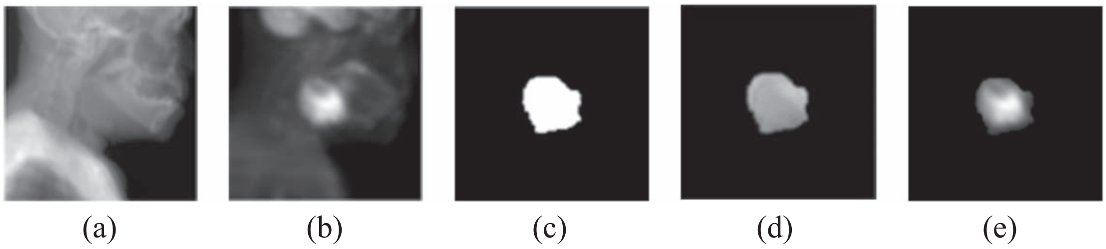


Figure 1. One patient example at the twelfth view (90 degrees). (a) Projection of averaged CT; (b) projection of averaged PET; (c) projection of averaged contour; (d) tumor region on CT; (e) tumor region on PET.

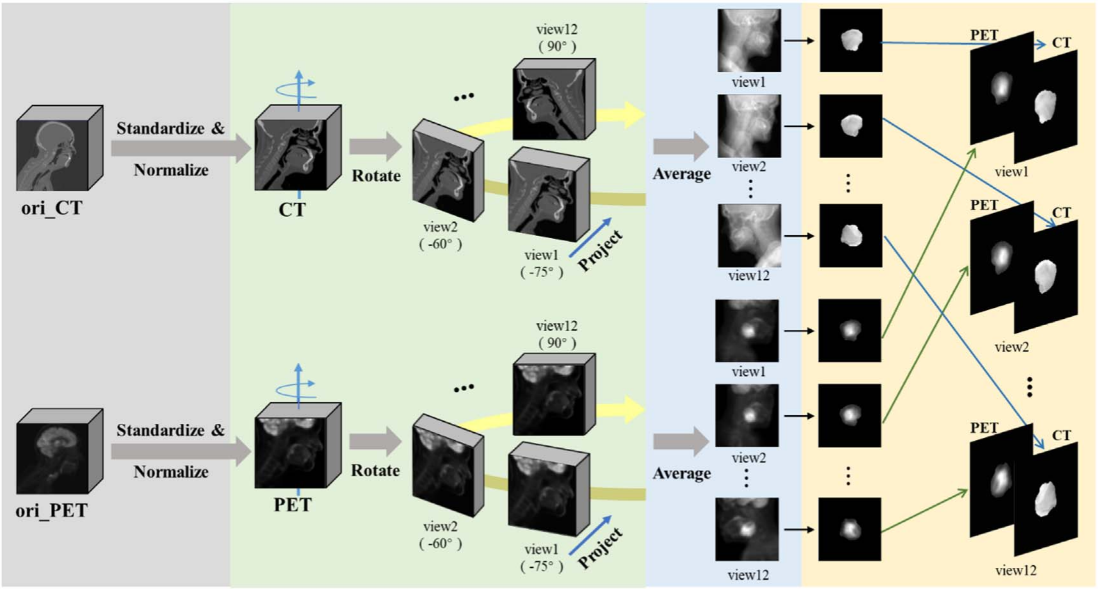


Figure 2. Flowchart of image pre-processing.

2.3.1. Resolution standardization and normalization

Because subsequent operations need images of different modalities to be aligned and superimposed, we resampled CT and PET images to the same resolution, $1 \times 1 \times 1$ mm^3, according to typical processes for pre-processing CT-PET data (Zhao *et al* 2018, Kumar *et al* 2019, Zhong *et al* 2019). Then, we converted pixel values in CT and PET images to CT numbers and standard uptake values (SUVs), respectively. The calculation equations for CT numbers and SUVs can be found in Paquet *et al*'s work (Paquet *et al* 2004). Because inputs with large integer values can disrupt or slow down the learning process, we normalized the CT numbers and SUVs to the interval [0, 1].

2.3.2. Vertical rotation for feature expansion

After standardizing the resolution and normalizing the image, we separately rotated the 3D CT/PET volume data, which consist of image and contour, along the Z axis (vertical axis) from −75 to 90 degrees. We then collected one view of volumetric images every 15 degrees to obtain a total of twelve views of PET and CT images. The reason for collecting views every 15 degrees is based on the experimental results in section 4.3.

2.3.3. Horizontal projection for dimension reduction

To reduce the model complexity, we projected all the slices at each view by averaging the corresponding pixel values of all slices along the horizontal axis This reduces all the 3D CT/PET images and contours to 2D images. We used a threshold of 0.5 to identify the tumor area on the averaged tumor masks, i.e. only pixels with values larger than 0.5 can be marked as a tumor area on an averaged 2D contour image.

2.3.4. Constructing tumor region samples

By following the tumor region identified on the average contour image, the MMFE extracted the tumor parts of the CT and PET images. All the images were uniformly sized to 200×200 through zero padding, and the tumor

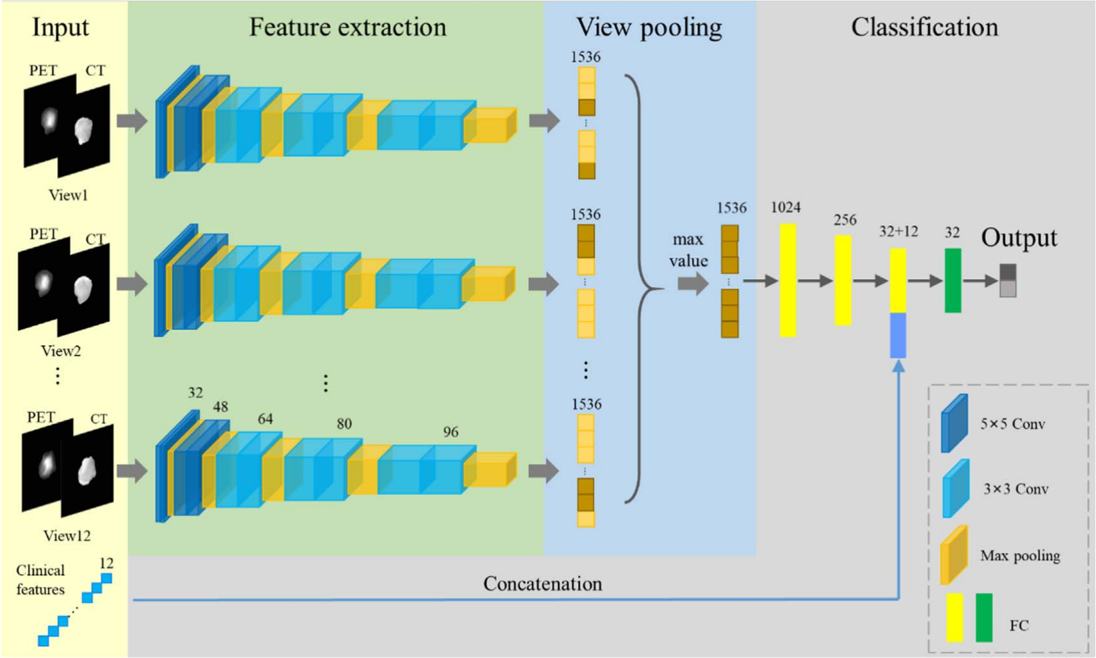


Figure 3. Illustration of the proposed MMFE framework.

part was placed at the center of the image. One patient example at the twelfth view (90 degrees) is shown in figure 1. Finally, the corresponding CT and PET tumor part images obtained were superimposed into a matrix of $200 \times 200 \times 2$ as the input samples for the prediction model.

Figure 2 shows the image pre-processing procedure for the above four steps and illustrates the flowchart of the proposed multi-view-based feature expansion and dimensionality reduction strategy.

2.4. Augmentation and balancing operation

To increase the model's generalization and reduce the training bias, we performed image augmentation for 2D CT/PET tumor images generated at each view. The augmentation consists of random flipping (horizontal and/or vertical) and rotation alone the clockwise or counterclockwise direction around its central point in an angle ranging from -30 to 30 degrees. Since augmented samples generated by a rotation in the range of $[-30^\circ, 30^\circ]$ have a higher affinity than a random rotation and are large enough to generate new invariant samples, we adopted this common setting according to several previous studies (Taylor and Geoff 2018, Lei *et al* 2019, Gontijo-Lopes *et al* 2020). Each sample was augmented 30 times. If the number of augmented samples is too small, the positive samples (with LRR) will be insufficient during the training as there are fewer positive samples in our dataset; if the number of augmented samples is too large, it may bring redundant information for the training. Our exploratory experiments showed that 30 times was a suitable setting.

The dataset used in our study has 49 positive samples and 157 negative samples. If a patient had LRR during the follow-up period (more than one year), which was defined as the positive sample, otherwise as the negative sample. Among patients with LRR, 18, 6 and 25 patients had negative, positive, and unknown HPV status, respectively. Among patients without LRR, 24, 38 and 95 patients had negative, positive, and unknown HPV status, respectively. To reduce the adverse effects of sample imbalance, we designed the strategy of selecting samples by category to ensure that the positive and negative samples are balanced in each batch. Let N be the batch size, $N = N_{\text{po}}+N_{\text{no}}+N_{\text{pa}}+N_{\text{na}}$, where $N_{\text{po}} = N_{\text{no}}$, $N_{\text{pa}} = N_{\text{na}}$, and N_{po}, N_{no}, N_{pa}, and N_{na} denote the selected number of positive original samples, negative original samples, positive augmented samples, and negative augmented samples, respectively. In this way, among the N samples for training in each batch, the positive and negative samples remain balanced. Meanwhile, both the original and the augmented samples are included. In an epoch, through random selection of multiple batches, positive samples are oversampled (Buda *et al* 2018, Byrd and Lipton 2019) to make up the minority class samples.

Table 2. A coding example of clinical feature for one patient.

Variable value						Clinical feature Code
Age	Primary site	T-stage	N-stage	HPV status	Therapy	
42	Larynx	T2	N2	Positive	Chemo radiation	[0.42, 1, 0, 0, 0, 0.5, 0.75, 0, 1, 0, 0, 1]

3. Methodology

3.1. Framework

The framework of our proposed MMFE method is shown in figure 3. The network consists of four layers: an input layer, a multi-view feature extraction layer, a view pooling layer, and a classification layer. The input layer contains multi-modal samples obtained after pre-processing, including twelve views of CT and PET dual-channel images with a size of 200 × 200 × 2, and a 12-dimensional clinical feature vector generated from six clinical variables. An example of proposed attribute-based hybrid coding strategy is shown in table 2. Note that primary site, HPV status and therapy are considered as nominal variable; and 4, 3, 2-dimension vectors are coded for these clinical features, respectively, using the One-Hot Encoding strategy.

The feature extraction layer consists of convolutional layers and max-pooling layers to extract features from multiple views, where all views use the same network architecture, and model parameters are shared among views. The architecture is a stack structure similar to VGG-Net (Simonyan and Zisserman 2014). It adopts multiple convolutional layers with smaller convolution kernels instead of one convolutional layer with larger convolution kernels. This multi-layer convolutional stacking method increases nonlinearity, which can better describe the characteristics of the input image. Selecting a small stride can prevent a larger stride from causing loss of detailed information. In each view, 1536 features are extracted.

The output features of each view in the feature extraction layer are aggregated in the view pooling layer. The view pooling layer uses an element-wise maximum operation on all views, which reduces the number of parameters while retaining the extracted texture information as much as possible. We expanded all view elements into a one-dimensional array and took the maximum value of the corresponding position of all views in each position, which resulted in a 1536-dimensional vector.

The input for the classification layer is the output of the view pooling layer and the 12-dimensional clinical features. The classification layer consists of four fully connected layers. The clinical parameter is concatenated with the third fully connected layer FC3 in the classification layer. The FC3 layer contains 32 features, which are slightly different from the clinical data feature dimensions; this can increase the clinical features' influence on the classification results. After the output layer passes through the softmax function, the probability values of the two categories are obtained to predict whether the patient is at high or low risk of LRR. The multi-view convolutional neural network is a directed acyclic graph structure, which was trained through stochastic gradient descent backpropagation in an end-to-end fashion.

3.2. Network settings

In our network structure, the number of views is twelve. The feature extraction layer comprises a stack of ten convolutional layers and five max-pooling layers, and the classification layer consists of five fully connected layers. The rectified linear units (ReLU) is employed as nonlinear activation functions in the convolutional layers and the fully connected layers, and the softmax function is used in the output layer. The network weights are optimized using the Adam algorithm based on the cross entropy loss function. The fixed learning rate is 1×10^{-5}, and the batch size $N = N_{po} + N_{no} + N_{pa} + N_{na} = 40$, where $N_{po} = N_{no} = 4$, $N_{pa} = N_{na} = 4N_{po} = 16$. The maximum epoch is set to 300, which ensures adequate training. In all experiments of proposed MMFE, two variants of MMFE, and the ablation experiments, the training epoch was set at 300. After 300 epochs, all experiments related to MMFE have converged. The L2 regularization term is added to the loss function to prevent overfitting and to improve the generalizability of the model.

4. Experiments

4.1. Experimental setup

4.1.1. Methods

The experimental setup of the proposed method MMFE, two variants of MMFE, and two comparison methods is as follows:

4.1.1.1. MMFE (2D-CNN)

We tested the performance of our MMFE by using different input combinations—clinical data, CT, PET, CT + PET and CT + PET + clinical data—in our experiment, For the input using clinical data only, we adopted a deep neural network (DNN) (Wang *et al* 2019). For the other input combinations, we used MMFE with the same parameters as all other MMFE-based strategies.

4.1.1.2. vMMFE-SV(2D-CNN)

To analyze the impact of the multi-view strategy on MMFE performance, we modified MMFE by using only a single view (vMMFE-SV) as input, while leaving everything else the same. The single view was obtained by rotating the patient image along the vertical axis at −75 degrees. The average image along the horizontal axis was then used as the input.

4.1.1.3. vMMFE-w/oRP(3D-CNN)

To compare the multi-view feature expansion with the original 3D data, we modified the MMFE by using 3D image data as inputs without the vertical rotation or horizontal projection steps in image pre-processing (vMMFE-w/oRP), while the rest of the network structure remained the same. Similar to the step of constructing tumor region samples, all the 3D CT and PET tumor part images were uniformly sized to $135 \times 135 \times 135$ through zero padding, then superimposed into a matrix of $135 \times 135 \times 135 \times 2$ as the input samples.

4.1.1.4. 2D-CNN (Diamant et al 2019)

Diamant *et al*'s original method used only CT data and chose one central slice of the tumor as the input sample, so the original network contained only a single channel. Its network consists of three consecutive 2D convolutional blocks (each of which contains a convolutional layer, a max-pooling layer and a parametric rectified linear unit) and two fully connected layers. This network structure is simpler than vMMFE-SV, which is also a 2D-CNN.

4.1.1.5. 3D-CNN (Starke et al 2020)

Starke *et al*'s original method is based on a 3D-CNN, where only CT volume data are used as single-channel inputs. Its network consists of five consecutive 3D convolutional blocks (each of which contains a convolutional layer, a max-pooling layer and a ReLU nonlinear activation unit) and two fully connected layers. The number of convolution kernels increases sequentially from 16 to 32, 64, 128, and 256. This network structure is more complex than that of vMMFE-w/oRP, which is also a 3D-CNN.

In our comparison study, for Diamant *et al*'s and Starke *et al*'s methods, we first used CT and PET separately as the input data, the same as with the original methods. Then, we separately modified them to a dual-channel network for comparison with inputs using both CT and PET. To be consistent with MMFE, the input samples contained segmented tumor parts only in all experiments.

4.1.2. Metrics

We evaluated the performance of the model through five-fold cross-validation. All positive and negative samples were randomly divided into five folds of approximately equal size, respectively. A fold of positive samples and a fold of negative samples were selected in turn as the testing set, while the remaining four folds of positive and negative samples were used as the training set. The same 5-fold partition was in all the experiments presented in this work. The area under the receiver operating characteristic curve (AUC) (Ling *et al* 2003) was used as the selection criterion for the optimal model. During each training process, a model with the highest AUC value on the testing set was saved as the optimal model. After five times cross-validation, we got the prediction probability of five test sets from the corresponding optimal model $\{P_i, i{=}1,\ldots,5\}$. The evaluation criteria were calculated on the probability values of all 206 samples $\boldsymbol{P} = [P_1; P_2; P_3; P_4; P_5]$, not by taking the average values of criteria from all folds.

For all comparison and ablation experiments in this work, the criteria were evaluated in the same way as described above. For all methods, the optimal model was determined based on the highest AUC value on the testing set. For comparison methods of Diamant *et al* (2019) and Starke *et al* (2020), the setting of maximum epoch in the original work was 100 and 200, respectively. For a fair comparison, the maximum epoch of both comparison methods was set at 300, which was consistent with MMFE. The other hyper-parameter of these comparison methods were set according to the original work.

Sensitivity, specificity, accuracy and AUC were used as the criteria to evaluate the performance of the different methods. These indicators were calculated as follows:

$$\mathrm{SEN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}), \tag{1}$$

$$\mathrm{SPE} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}), \tag{2}$$

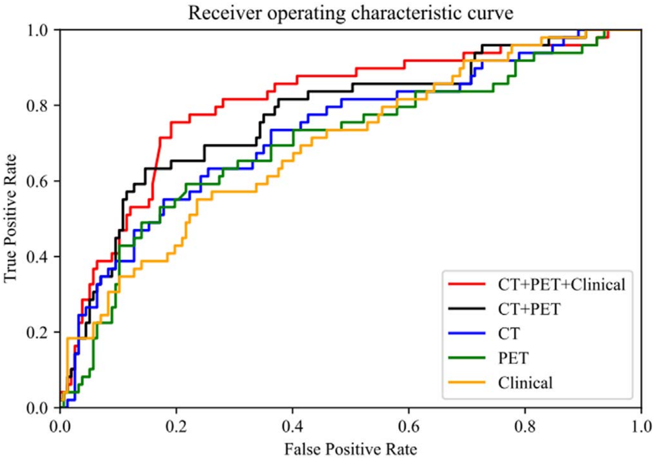


Figure 4. Receiver operating characteristic curves (ROCs) of MMFE with different input data.

Table 3. Performance of MMFE and comparison methods on different input data.

Modality	Method	TYPE	SEN	SPE	ACC	AUC
Clinical	DNN	—	0.6531	0.6369	0.6408	0.6947
CT	Diamant *et al* (2019)	2D	0.5306	0.7643	0.7087	0.7019
	Starke *et al* (2020)	3D	0.5102	0.8344	0.7573	0.7162
	vMMFE-SV	2D	0.5714	0.8158	0.7573	0.7088
	vMMFE-w/oRP	3D	0.5714	0.7452	0.7039	0.7199
	MMFE	2D	0.5714	0.7707	0.7233	**0.7286**
PET	Diamant *et al* (2019)	2D	0.4694	0.7834	0.7087	0.6520
	Starke *et al* (2020)	3D	0.4082	0.8408	0.7379	0.6918
	vMMFE-SV	2D	0.5510	0.7961	0.7379	0.6576
	vMMFE-w/oRP	3D	0.5918	0.7643	0.7233	0.6779
	MMFE	2D	0.5714	0.7834	0.7330	**0.7000**
CT + PET	Diamant *et al* (2019)	2D	0.5102	0.7898	0.7233	0.7097
	Starke *et al* (2020)	3D	0.5714	0.8344	0.7718	0.7333
	vMMFE-SV	2D	0.6122	0.7898	0.7476	0.7110
	vMMFE-w/oRP	3D	0.5714	0.8280	0.7670	0.7528
	MMFE	2D	0.5918	0.8535	0.7913	**0.7717**
CT + PET + clinical	Diamant *et al* (2019)	2D	—	—	—	—
	Starke *et al* (2020)	3D	—	—	—	—
	vMMFE-SV	2D	0.6531	0.8089	0.7718	0.7273
	vMMFE-w/oRP	3D	0.6122	0.7962	0.7524	0.7816
	MMFE	2D	0.7347	0.8089	0.7913	**0.8052**

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}). \tag{3}$$

4.1.3. Environment

Our model was built using the Tensorflow framework based on Python3. All experiments were trained and tested on an NVIDIA GTX 1080 graphics processing unit (GPU).

4.2. Comprehensive comparison with other methods

We used the above evaluation criteria of SEN, SPE, ACC and AUC to quantify the performance of our proposed MMFE, MMFE variants vMMFE-SV and vMMFE-w/oRP, and two other methods from Diamant *et al* (2019) and Starke *et al* (2020). MMFE, vMMFE-SV and Diamant *et al*'s method are based on 2D-CNN; vMMFE-w/oRP and Starke *et al*'s method are based on 3D-CNN. The results of comparing the five methods under different input data—CT, PET, CT + PET, and CT + PET + clinical—are summarized in table 3. The proposed MMFE method obtained a higher AUC value than the other four methods, regardless of input. The best prediction result was achieved when using all three modalities—CT, PET and clinical features—as inputs, which yielded an AUC value of 0.81. Figure 4 shows the receiver operating characteristic curves (ROCs) corresponding to MMFE with different input data. Results in table 3 and figure 4 suggest that using multi-modal data can yield a better AUC value than using a single modality, whether CT, PET or clinical features.

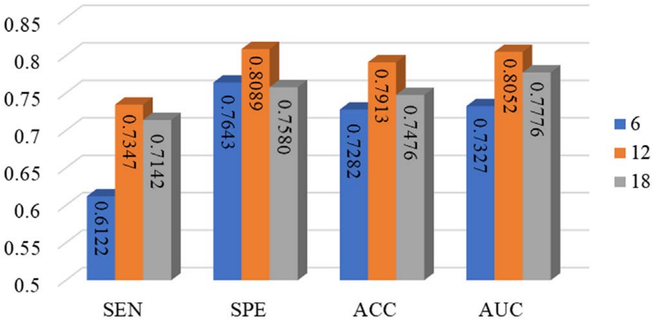


Figure 5. Comparison of MMFE performance under different numbers of views.

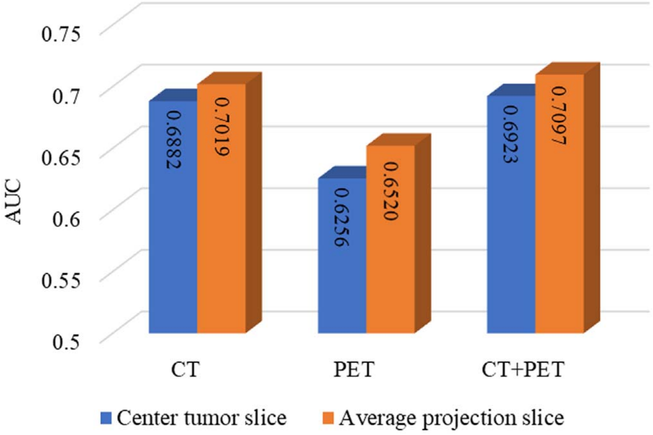


Figure 6. Performance comparison of Diamant *et al* using central tumor slice and using average projection slice as input.

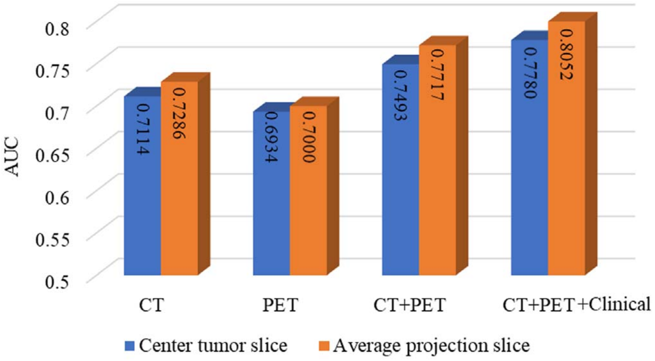


Figure 7. Performance comparison of MMFE using central tumor slice and using average projection slice as input.

4.3. Ablation studies

4.3.1. Effect of view number

To analyze the optimal number of views, we compared the model performance using views taken every 30 degrees from −60 to 90 degrees (six views in total), views taken every 15 degrees from −75 to 90 degrees (12 views in total), and views taken every 10 degrees from −80 to 90 degrees (18 views in total). All experiments used CT, PET and clinical features as input data.

Figure 5 shows that the best performance is achieved when the views are taken 15 degrees apart (total 12 views). Taking too many or too few views will cause redundancy or lack of information and affect the model's prediction ability.

Table 4. Complexity measurements of vMMFE-w/oRP and MMFE.

Method	TYPE	Memory	FLOPS	Parameters	Inference time
vMMFE-w/oRP	3D	672.37 M	32.65 M	4.50 M	2.61 s
MMFE	2D	143.04 M	28.26 M	2.29 M	1.42 s

4.3.2. *Effect of average projection*

To analyze the impact of our proposed average projection slice strategy on the performance of Diamant *et al*'s method (Diamant *et al* 2019), we replaced the original central tumor slice used as input in Diamant *et al* with our average projection slice, while the rest remained the same. As shown in figure 6, when the average projection slices were used, the CNN used by Diamant *et al* obtained higher AUC values when the input data were CT, PET or CT + PET. Furthermore, we replaced our average projection slice in MMFE with the central tumor slice from Diamant *et al* and compared the performance (figure 7). When only the central slice of tumor is used, the AUC values of MMFE with different input data—CT, PET, CT + PET and CT + PET + clinical—are all lower to varying degrees.

4.4. Complexity analysis

Using CT + PET + clinical as input data, we measured the complexity of our MMFE in terms of memory, floating-point operations per second (FLOPS), parameters, and inference time. All values listed in table 4 were obtained at the inference stage on a GTX 1080 GPU card. In general, smaller numbers are better. Meanwhile, we compared the MMFE's complexity with that of the 3D-CNN-based vMMFE-w/oRP, which did not perform as well as MMFE. Table 4 shows that the proposed MMFE also has lower complexity than the vMMFE-w/oRP.

5. Discussion and conclusion

To accurately identify H&N cancer patients with high risk for LRR after definitive radiation or chemoradiation therapy, we developed an end-to-end MMFE model in this study. Two-dimensional (2D)-based DL models inherently fail to utilize the context from adjacent slices, while 3D models can fully use the volumetric information. The comparison experiments of section 4.2 show that the performance of two 3D-CNN methods (vMMFE-w/oRP and Starke *et al*'s) is better than that of two 2D-CNN methods (vMMFE-SV and Diamant *et al* 's) using different input data—CT, PET, CT + PET, and CT + PET + clinical.

The experimental results also demonstrated that our MMFE method performs better than other baseline methods based on 2D or 3D-CNN. Although the MMFE is also based on a 2D architecture, unlike conventional 2D single-view methods, the MMFE makes full use of multi-view information to compensate for the information loss that may be caused by the dimension reduction. Compared to a 2D model, the number of parameters involved in a 3D model is much larger. As such, more training data is required in 3D based models to avoid overfitting and obtain better generalization performance. The limited number of patients in this work may affect the prediction results of the 3D CNN-based methods. The MMFE strategy includes vertical rotation for feature expansion and horizontal projection to reduce the dimensionality. In the vertical rotation step, the image representations generated at these different vertical angle views could effectively expand the features. In the horizontal projection step, all the slices of one view are projected by an average operator along the horizontal axis. Although the 3D data is not directly used, MMFE utilizes compressed volumetric information. During the subsequent training, data of different views are aggregated, which eliminates redundant information of original 3D data to a certain extent and makes training easier to converge. Furthermore, our multi-view feature expansion framework incorporates other advanced feature extraction methods in the multi-view feature extraction layer to further improve the model's performance. Our experimental results demonstrated the effectiveness of our designed strategy. The advantages of multi-view representation have also been demonstrated in several applications such as for object classification (Qi *et al* 2016, Xu *et al* 2021), Alzheimer's disease diagnosis (Qiao *et al* 2021), and pose estimation (Zhang *et al* 2021). The experimental results of the aforementioned works also confirmed that view-based methods can exploit the powerful view representation, and a view-based manner with a 2D architecture can achieve better results than 3D CNN.

Furthermore, our experimental results demonstrated that MMFE can obtain better performance by combining extracted CT, PET and clinical features. Although the sensitivity and AUC values were highest in the model using combined CT, PET and clinical data as input, the specificity after adding the clinical data was not as good as the model using only CT and PET. One possible reason is that the model's ability to identify negative samples is poor when only clinical data are used. Due to the large number of negative samples, the accuracy was not that high; we observed a similar trend in the model using clinical data only. As such, the specificity in the combined model was reduced after we added clinical data to CT and PET data. However, the sensitivity was greatly improved, so the model could better focus on positive samples and find LRR patients, which led to better overall performance (AUC value) for the model.

The present work mainly focuses on methodology development and comparison studies with other state-of-the-art approaches, and it uses patient data from a single institution only. To test the model's generalizability, evaluating the model performance on an independent cohort from another institution/hospital would be highly desired. We will pursue this in a future work. Once validated in an external patient cohort, the model developed in this work could help physicians to develop optimal personalized treatment strategies for H&N cancer patients. We also plan to develop multi-modal learning methods by exploring the potential correlations between multi-modal data. This approach could potentially reduce the impact of a certain modal data anomaly on the results, which could reduce overfitting and make the prediction results more comprehensive.

Acknowledgments

This work was supported by the Engineering Research Center of Big Data Application in Private Health Medicine, Fujian Province University (No. KF2020005), the Natural Science Foundation of Shaanxi Province (No. 2019ZDLGY03-02-02), and the Research Industrialization Plan of Xi'an (No. XA2020-RGZNTJ-0075). The authors would like to thank Dr Jonathan Feinberg for scientific editing of the manuscript.

ORCID iDs

Kai Wang https://orcid.org/0000-0003-2155-1445
Jing Wang https://orcid.org/0000-0002-8491-4146

References

Baltrušaitis T, Ahuja C and Morency L P 2018 Multimodal machine learning: a survey and taxonomy *IEEE Trans. Pattern Anal. Mach. Intell.* **41** 423–43

Beaumont J *et al* 2019 Voxel-based identification of local recurrence sub-regions from pre-treatment PET/CT for locally advanced head and neck cancers *EJNMMI Res.* **9** 1–11

Beesley L J *et al* 2019 Individualized survival prediction for patients with oropharyngeal cancer in the human papillomavirus era *Cancer* **125** 68–78

Bogowicz M, Tanadini-Lang S, Guckenberger M and Riesterer O 2019 Combined CT radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer *Sci. Rep.* **9** 1–7

Branstetter B F IV *et al* 2005 Head and neck malignancy: is PET/CT more accurate than PET or CT alone *Radiology* **235** 580–6

Bray F *et al* 2018 Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries *CA: Cancer J. Clin.* **68** 394–424

Buda M, Maki A and Mazurowski M A 2018 A systematic study of the class imbalance problem in convolutional neural networks *Neural Netw.* **106** 249–59

Byrd J and Lipton Z 2019 What is the effect of importance weighting in deep learning? *Int. Conf. on Machine Learning (PMLR)*

Caudell J J *et al* 2017 The future of personalised radiotherapy for head and neck cancer *Lancet Oncol.* **18** e266–73

Chang J H *et al* 2017 Locoregionally recurrent head and neck squamous cell carcinoma: incidence, survival, prognostic factors, and treatment outcomes *Oncotarget* **8** 55600

Diamant A *et al* 2019 Deep learning in head & neck cancer outcome prediction *Sci. Rep.* **9** 1–10

El-Regaily S A *et al* 2020 Multi-view convolutional neural network for lung nodule false positive reduction *Expert Syst. Appl.* **162** 113017

Faisal S and Tutz G 2017 Nearest neighbor imputation for categorical data by weighting of attributes*Inf. Sci.* **592** 306–19

Gontijo-Lopes R *et al* 2020 Tradeoffs in data augmentation: an empirical study *Int. Conf. on Learning Representations (ICLR)*

Guo Z, Li X, Huang H, Guo N and Li Q 2019 Deep learning-based image segmentation on multimodal medical imaging *IEEE Trans. Radiat. Plasma Med. Sci.* **3** 162–9

Hu H *et al* 2020 Lymphoma segmentation in PET images based on multi-view and Conv3D fusion strategy *2020 39th IEEE Int. Symp. on Biomedical Imaging (ISBI) (Piscataway, NJ)* (IEEE)

Keek S *et al* 2020 Computed tomography-derived radiomic signature of head and neck squamous cell carcinoma (peri) tumoral tissue for the prediction of locoregional recurrence and distant metastasis after concurrent chemo-radiotherapy *PLoS One* **15** e0232639

Kumar A, Fulham M, Feng D and Kim J 2019 Co-learning feature fusion maps from PET-CT images of lung cancer *IEEE Trans. Med. Imaging* **39** 204–17

Le W, Romero F P and Kadoury S 2020 A normalized fully convolutional approach to head and neck cancer outcome prediction *Medical Imaging with Deep Learning(MIDL)*

Lei C *et al* 2019 A preliminary study on data augmentation of deep learning for image classification *Proc. of the 11th Asia-Pacific Symp. on Internetware*

Ling C X, Huang J and Zhang H 2003 AUC: a better measure than accuracy in comparing learning algorithms *Conf. of the Canadian Society for Computational Studies of Intelligence (Berlin)* (Springer)

Lv W, Ashrafinia S, Ma J, Lu L and Rahmim A 2019 Multi-level multi-modality fusion radiomics: application to PET and CT imaging for prognostication of head and neck cancer *IEEE J. Biomed. Health Inf.* **24** 2268–77

Nahavandipour A *et al* 2019 Incidence and survival of laryngeal cancer in Denmark: a nation-wide study from 1980 to 2014 *Acta Oncol.* **58** 977–82

Paquet N, Albert A, Foidart J and Hustinx R 2004 Within-patient variability of (18)F-FDG: standardized uptake values in normal tissues *J. Nucl. Med.* **45** 784–88

Qi C R *et al* 2016 Volumetric and multi-view cnns for object classification on 3d data *IEEE Conf. on Computer Vision and Pattern Recognition(CVPR)*

Qiao H, Lin C and Fan Z , 2021 A fusion of multi-view 2D and 3D convolution neural network based MRI for Alzheimer's disease diagnosis *Int. Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC) (Piscataway, NJ)* (IEEE)

Rose J D, Jaspin K and Vijayakumar K 2021 *Lung Cancer Diagnosis Based on Image Fusion and Prediction Using CT and PET Image. in Signal and Image Processing Techniques for the Development of Intelligent Healthcare Systems* (Berlin: Springer)

Saha S *et al* 2020 Predicting motor outcome in preterm infants from very early brain diffusion MRI using a deep learning convolutional neural network (CNN) model *Neuroimage* **215** 116807

Seeram E 2015 *Computed Tomography-E-Book: Physical Principles, Clinical Applications, and Quality Control* (Australia: Elsevier Health Sciences) 5–6

Shanthi S and Rajkumar N 2021 Lung cancer prediction using stochastic diffusion search (SDS) based feature selection and machine learning methods *Neural Process. Lett.* **53** 1–14

Simonyan K and Zisserman A 2015 Very deep convolutional networks for large-scale image recognition *International Conference on Learning Representations (ICLR)*

Soley-Bori M 2013 Dealing with missing data: key assumptions and methods for applied analysis *Boston Univ.* **23** 1–19

Starke S *et al* 2020 2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma *Sci. Rep.* **10** 1–13

Su H, Maji S, Kalogerakis E and Learned-Miller E 2015 Multi-view convolutional neural networks for 3d shape recognition *IEEE int. Conf. on Computer Vision (ICCV) (Piscataway, NJ)* (IEEE)

Taylor L and Geoff N 2018 Improving deep learning with generic data augmentation *IEEE Symp. Series on Computational Intelligence (SSCI) (Piscataway, NJ)* (IEEE)

Townsend D W 2004 Physical principles and technology of clinical PET imaging *Annals-Acad. Med. Singap.* **33** 133–45

Wang K *et al* 2020a A multi-objective radiomics model for the prediction of locoregional recurrence in head and neck squamous cell cancer *Med. Phys.* **47** 5392–400

Wang R *et al* 2019 Multi-objective ensemble deep learning using electronic health records to predict outcomes after lung cancer radiotherapy *Phys. Med. Biol.* **64** 245005

Wang S *et al* 2017 A multi-view deep convolutional neural networks for lung nodule segmentation *2017 39th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC) (Piscataway, NJ)* (IEEE)

Wang Y *et al* 2020b CT radiomics nomogram for the preoperative prediction of lymph node metastasis in gastric cancer *Eur. Radiol.* **30** 976–86

Wei J, Xia Y and Zhang Y 2019 M3Net: a multi-model, multi-size, and multi-view deep neural network for brain magnetic resonance image segmentation *Pattern Recognit.* **91** 366–78

Xia Y *et al* 2020 Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation *Med. Image Anal.* **65** 101766

Xu J *et al* 2021 Joint multi-view 2D convolutional neural networks for 3D object classification *Int. Conf. on Int. Joint Conf. on Artificial Intelligence (IJCAI)* 3202–8

Yang C K *et al* 2019 Deep convolutional neural network-based positron emission tomography analysis predicts esophageal cancer outcome *J. Clin. Med.* **8** 844

Zhang J *et al* 2021 Direct multi-view multi-person 3D pose estimation *Neural Information Processing Systems(NeurIPS)*

Zhao X *et al* 2020 A cross-modal 3D deep learning for accurate lymph node metastasis prediction in clinical stage T1 lung adenocarcinoma *Lung Cancer* **145** 10–7

Zhao X, Li L, Lu W and Tan S 2018 Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network *Phys. Med. Biol.* **64** 015011

Zheng A and Casari A 2018 *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists* (United States of America: O'Reilly Media, Inc)

Zhong Z *et al* 2019 Simultaneous cosegmentation of tumors in PET-CT images using deep fully convolutional networks *Med. Phys.* **46** 619–33

Zhou Z *et al* 2020 M2Net: multi-modal multi-channel network for overall survival time prediction of brain tumor patients *Medical Image Computing and Computer Assisted Intervention Society (MICCAI)* (*Lecture Notes in Computer Science* 12262) (Cham: Springer) (https://doi.org/10.1007/978-3-030-59713-9_22)

Zhu W, Xie L, Han J and Guo X 2020 The application of deep learning in cancer prognosis prediction *Cancers* **12** 603–21