International Journal of Neural Systems, Vol. 32, No. 1 (2022) 2150055 (14 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0129065721500556

Deep Multimodal Neural Network Based on Data-Feature Fusion for Patient-Specific Quality Assurance

Ting Hu[*,‡], Lizhang Xie[*,§], Lei Zhang[*,¶], Guangjun Li[†,‖] and Zhang Yi[*,**]
[*]*Department of Computer Science and Technology*
Sichuan University, Section 4, Southern 1st Ring Rd
Chengdu, Sichuan, P. R. China
[†]*Department of Radiation Oncology*
Cancer Center and State Key Laboratory of Biotherapy
West China Hospital, Sichuan University
Chengdu, Sichuan, P. R. China
[‡]*huting@stu.scu.edu.cn*
[§]*xielizhang1994@stu.scu.edu.cn*
[¶]*leizhang@scu.edu.cn*
[‖]*gjnick829@sina.com*
[**]*zhangyi@scu.edu.cn*

Accepted 5 September 2021
Published Online 10 December 2021

Patient-specific quality assurance (QA) for Volumetric Modulated Arc Therapy (VMAT) plans is routinely performed in the clinical. However, it is labor-intensive and time-consuming for medical physicists. QA prediction models can address these shortcomings and improve efficiency. Current approaches mainly focus on single cancer and single modality data. They are not applicable to clinical practice. To assess the accuracy of QA results for VMAT plans, this paper presents a new model that learns complementary features from the multi-modal data to predict the gamma passing rate (GPR). According to the characteristics of VMAT plans, a feature-data fusion approach is designed to fuse the features of imaging and non-imaging information in the model. In this study, 690 VMAT plans are collected encompassing more than ten diseases. The model can accurately predict the most VMAT plans at all three gamma criteria: 2%/2 mm, 3%/2 mm and 3%/3 mm. The mean absolute error between the predicted and measured GPR is 2.17%, 1.16% and 0.71%, respectively. The maximum deviation between the predicted and measured GPR is 3.46%, 4.6%, 8.56%, respectively. The proposed model is effective, and the features of the two modalities significantly influence QA results.

Keywords: Radiation therapy; VMAT plan; multimodal model; GPR prediction; quality assurance.

1. Introduction

Volumetric Modulated Arc Therapy (VMAT) is the system about intensity-modulated radiotherapy treatment delivery by dynamic arcs with continuous variation of multileaf collimator (MLC) shapes, dose rates, and gantry rotations.[1,2] Radiotherapy systems record the information (such as MLC and monitor unit (MU)) during VMAT treatment. VMAT can shorten treatment time and generate highly conformal dose distributions delivered with superior

[¶]Corresponding author.

dosimetric accuracy.[3] In recent years, it has become a critical radiotherapy technique due to these advantages. However, discrepancies exist between the delivered and planned dose distributions because VMAT plans cannot be perfectly implemented during the actual plan delivery.[4] Hence, it is necessary to verify the delivered dose distributions (i.e. patient-specific quality assurance (QA)). In the clinical, QA for VMAT plans is routinely performed before treatment. The gamma passing rate (GPR) is a metric to evaluate the goodness of a treatment plan.

In radiotherapy, patient-specific QA is highly time-consuming and labor-intensive. Patients' radiotherapy treatment plans tend to be complicated, which may result in patient-specific QA failure.[5] Hence, researchers display considerable interest in prediction models of patient-specific QA results. Machine learning has been studied in many years, and there are many effective machine learning methods are constantly emerging.[6–9] Besides, they have been applied to all aspects of radiotherapy, encompassing target volume delineation, dose prediction, treatment planning optimization, MLC positioning error, and linear accelerator performance.[10,11] In addition, machine learning can be applied in the QA to improve the quality and efficiency of treatment plans and implementation. It can make radiotherapy decision-making more simplified, individualized and accurate, improve the automation of radiotherapy plan design and QA, and promote individualized precision treatment. The strength of taking machine learning models into the QA prediction is that it can inform the radiologists about which treatment plans might fail before QA measurements. These models identify the correlation between the radiotherapy plan and GPR. They can evaluate treatment plans in detail and predict individualized QA passing rates. Utilizing GPR prediction models could help to improve the efficiency of VMAT QA, increase patient satisfaction, reduce the risks, and save time and resources.

Hand-crafted features and two-dimensional (2D) images are two common data using machine learning models to predict GPR. In previous related research, many works utilize hand-craft features as input. Hand-craft features extracted from the treatment plan by the professional physician are represented as a vector. Abstracting hand-crafted features also requires strong domain knowledge. Besides, a few works with 2D convolutional neural networks (CNN) have been explored in QA in clinical radiotherapy, which takes images gained from plans as input. All the existing studies are based on individual imaging or non-imaging modalities. However, a single modality may be short of information that is vital for predicting GPR. Without adequate information, the effect of the model will be affected. So, this paper presents an alternative perspective to solve these problems.

This paper focuses on two modalities in the VMAT plans. Hence, the GPR prediction problem can be characterized as multimodal since it encompasses imaging and non-imaging data. Multimodal machine learning fuses information from different modalities to perform a prediction. It is conducive to reduce the loss of data in a single modality. In this study, a three-dimensional multimodal ResNet (3D-MResNet) model that integrates different modalities is proposed. The contributions of this study are summarized as follows: (1) A VMAT QA dataset, including 690 cases with more than ten cancers, is constructed to build fully automated models of GPR prediction. To the best of our knowledge, our dataset is the only VMAT QA plan dataset containing the imaging and non-imaging modalities. (2) According to the characteristics of VMAT data, the feature-data fusion (FDF) approach, which deals with the relationship between MLC images and MU values, fuses each image's result and the corresponding MU value. It can learn the features of one-to-one correlation between the images with MU values, rather than just fusing the features of modalities. This work is the first study that considers the imaging and non-imaging data to predict the GPR. (3) The 3D-MResNet model that fuses the two modalities aiming to process and associate features from each modality is proposed. There is currently no such method as joining two modalities in VMAT plans to perform GPR prediction.

The remaining sections of this paper are organized as follows. Section 2 displays the related works for GPR prediction, 3D CNN, and multimodal learning in medical images analysis. The image processing method and dataset are described in Sec. 3. The proposed model is discussed in Sec. 4. Section 5 demonstrates the results and analysis of experiments, and Sec. 6 gives a conclusion to this paper.

2. Related Work

2.1. *CNNs in medical images analysis*

CNNs are the primary machine learning method currently used in visual object recognition and classification.[12–17] Besides, they are widely applied in the medical field.[18–24] However, 2D CNNs adopt a single image as input, and they are inherently unable to take advantage of the context from adjacent slices in medical images. They are not applicable for some 3D medical images such as CT, MRI.

For 3D medical data analysis, CNNs with spatio-temporal 3D convolutional kernels are more effective than CNNs with 2D. Compared with 2D CNNs, 3D CNNs have the capability of encoding representations from volumetric receptive fields. Hence, they can abstract more discriminative characteristics through more abundant 3D spatial information. Because of the strong feature extraction capability of 3D CNNs, they have become the popular medical image analysis approach.[25–27] For example, a 3D CNN was used to identify Parkinsons disease in 3D nuclear imaging data.[28] Yang *et al.*[29] presented a 3D model to classify Alzheimers disease by 3D MRI images. Inspired by these methods, a 3D ResNet module is used to extract the features of 3D image data in VMAT plans.

2.2. *Multimodal deep learning*

Ordinarily, there are various representations of entities, as a specific object can be expressed by a picture, paragraph, or symbol.[30] Hence, the research problem is regarded as a multimodal problem when it encompasses multiple modalities. There are wide applications in multimodal learning such as image registration, image reconstruction, and medical images.[31–34] An unsupervised model that can extract the cross-modality correlations was proposed for cross-modality element-level feature learning.[35] Besides, a scalable multimodal CNN was presented for brain tumor segmentation.[36] Le *et al.*[37] applied multimodal CNN to the automated diagnosis of prostate cancer. Xu *et al.*[38] used a multimodal deep learning method to diagnose diseases. Fusion as an import task in multimodal has been used for many years. It is commonly executed at two levels for these models: feature level (early fusion) and decision level (late fusion).[39,40] The feature-level models first combine the features abstracted from input data and then analyze the fusion features to make decisions. For the decision-level fusion, the local decisions of different modalities are provided first, and then the model makes a final decision result by the fused local decisions.

Multimodal data can provide complementary information to improve the effectiveness of models. Similar approaches have been employed in this paper to integrate two modalities in VMAT plans. The FDF method, which contains the early and late fusion strategies, is presented. Naturally, it belongs to the hybrid fusion that exploits the advantages of both strategies. The imaging and non-imaging data can be obtained from VMAT plans. The contents of different modalities are not necessarily the same, and some essential metrics may be lacking in the single modality. There is a difference in prediction result as the different form of modality determines the information interpretation. For making good use of VMAT plans, the FDF method fuses two modalities of information into a representation that merges them. A multimodal model is applied to learn imaging and non-imaging features from plans, and the fusion module amalgamates the nonlinear correlations across all sources of modality information.

2.3. *Prediction of QA results based on machine learning*

At present, machine learning models have been introduced into radiation oncology as popular prediction technologies.[41] They solve several long-standing issues and raise working efficiency in QA workflows.[42–44] In general, radiologists need to assess and evaluate every patient's VMAT plan in detail before radiotherapy. Systems defect or dosimetric errors may lead to patient injury. Therefore, making use of the results of prediction models can prevent adverse outcomes during the treatment.

Models based on machine learning have been widely applied to predict individualized Intensity Modulated Radiation Therapy (IMRT) and VMAT QA passing rates. Machine learning methods are always used for IMRT/VMAT QA prediction through the complexity metrics abstracting from the treatment plans.[45–47] These approaches use handcrafted features extracted from plans. However, they rely heavily on professional knowledge or experience and may miss some vital information.

There is increasing interest in applying CNNs for GPR prediction in recent years, and models use images extracted in plans as input instead of hand-crafted features. VGG-16 used the fluence map image of IMRT plans as input and mapped it to the QA passing rate.[48] Tomori *et al.*[49] proposed a CNN model to predict GPR for patient-specific QA results in prostate treatment. Besides, they developed a deep learning-based GPR prediction model for VMAT.[50] Nyflota *et al.*[51] adopted CNNs to predict the mistakes in radiation therapy plans delivery by patient-specific gamma images.

All the previous approaches exploit single modality data, and the data abstracted from plans leave out some necessary information. For example, hand-crafted features, they lack MLC aperture shapes that are crucial in radiation treatment because the shape of the tumor is closely related to MLC aperture. For image data, dose distributions and MU information in plans cannot be embodied directly in images. Based on these findings, an innovative GPR prediction model named 3D-MResNet that fully exploits the inherent correlations across imaging and non-imaging modalities is presented in this paper. Besides, information of MLC sequences has been taken into account in the model. So far as we know, this study is the first to investigate the use of 3D multimodal information in VMAT GPR prediction.

3. Dataset and Data Processing

This section introduces the data used in this paper. It first gives the data sources and details of VMAT plans. Then, the way of obtaining imaging and non-imaging data is presented. Eventually, a new VMAT plan dataset is generated after pretreatment of the raw VMAT plans data.

3.1. *Dataset*

The data in this study are from the West China School of Medicine and West China Hospital. In this study, 690 VMAT plans were collected between June 2018 and August 2019 at a single institution. The plans consist of 37 clinical sites. Table 1 presents the kinds of cancers and the number of patients in our dataset. It shows that VMAT plans mainly include the Rectum, Nasopharyngeal Carcinoma (NPC), Cervix, and Prostate four diseases.

Table 1. The disease distribution in the dataset. “Others” indicates the total of small quantity diseases.

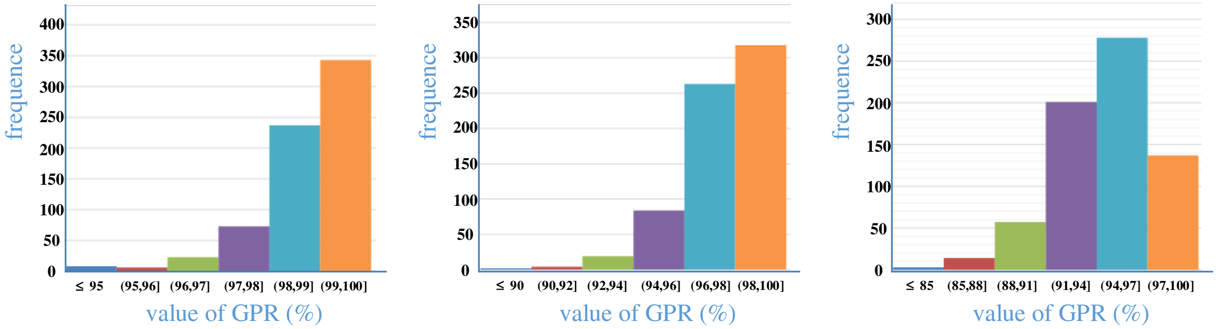
Disease	Number
Rectum	185
NPC	141
Cervix	67
Prostate	60
Uterus	28
Stomach	27
Brain	22
Larynx	19
Pharynx	10
Pancreas	9
Colon	9
Tongue	8
Others	95

To fully utilize VMAT information, two modalities (3D images and MU values) are attained from VMAT plan data. The proposed dataset consists of the imaging and non-imaging modalities from VMAT plans. In the dataset, the labels of VMAT plans are GPR values that are calculated using three common gamma criteria 3%/3 mm, 3%/2 mm, and 2%/2 mm. Each plan has three GPR values, respectively. Figure 1 shows the number of patients in different ranges of GPR values. The majority of values measured for GPR of VMAT plans are distributed in the range of 85% to 100% at 2%/2 mm, 90% to 100% at 3%/2 mm, and 3%/3 mm gamma criteria, as presented in Fig. 1.

In this study, VMAT plans include two beams, and each beam contains 91 control points (CPs). Therefore, there are 182 CPs in a plan. Each CP has corresponding parameters (such as MU weights and MLC shapes) that are obtainable from DICOM RT Plan files. MLC shape of each control point is stored in a 2D array of $400 \times 400\,\text{mm}^2$ with a pixel size of a $1 \times 1\,\text{mm}^2$.[52] Hence, they are considered as the imaging modality with a resolution of 400×400 pixels. Considering a series of MLC apertures existing in a plan, a 3D MLC aperture shape is constructed by a sequence of 2D MLC apertures.

3.2. *Data processing*

For the MLC shape images, we first extract MLC apertures' precise location and compute a bounding



(a) 3%/3 mm (b) 3%/2 mm (c) 2%/2 mm

Fig. 1. Distribution of measured GPR of VMAT plans. (a)–(c) show the distribution at the three criteria, respectively.

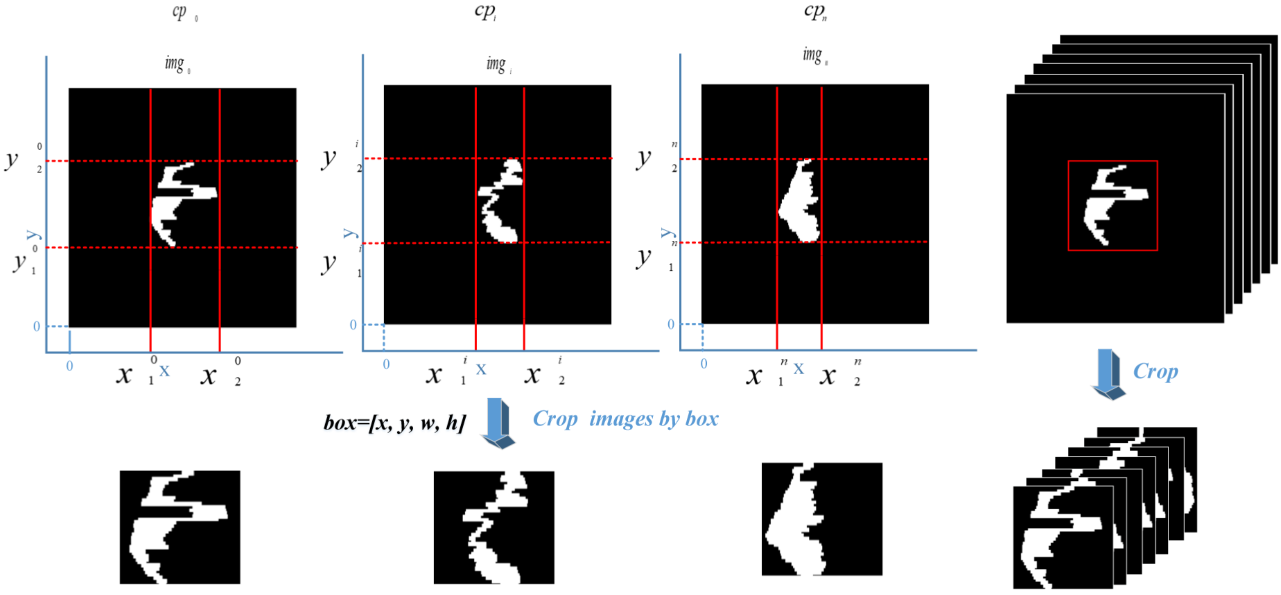


Fig. 2. The process of MLC shape images. The images in the first row are the original MLC shape images with the size of 400 × 400 pixels. We compute a bounding box that contains the MLC aperture in images. Then, the original images are cropped according to the bounding box, and the second row presents the cropped images.

box that contains all MLC apertures at every CP. Then, the bounding box is used to crop the MLC shape images because we focus on the image information inside the bounding box and ignore information outside. Figure 2 demonstrates the images processing, where the cp_i presents the ith control point in the plan, and img_i is the MLC aperture image at cp_i. To attain the aperture shapes in images, a bounding box ($\mathrm{box} = [x, y, w, h]$) is selected to crop all MLC shape images, removing the useless background information. For the box, x is the minimum value of the left edge of the aperture shape, and y is the minimum value of the bottom edge of the aperture shape. They can be presented as follows:

$$x = \min\{x_1^i\}, \quad y = \min\{y_1^i\}, \quad 0 \le i < n, \quad (1)$$

where n indicates the total number of control points, x_1^i and y_1^i denote the left and bottom edge of the aperture shape in cp_i, respectively. The width w of the box is the maximum difference between right and left edge, formulated as follows:

$$w = \max\{x_2^j\} - \min\{x_1^i\}, \quad 0 \le i, j < n, \quad (2)$$

where x_2^j denotes the right edge of aperture shape in cp_j. The height h of the box is the maximum difference between the bottom and up edge, formulated as follows:

$$h = \max\{y_2^j\} - \min\{y_1^i\}, \quad 0 \leq i, j < n, \tag{3}$$

where y_2^j denotes the up edge of the aperture shape in cp_j. The region in the box contains shape information of all MLC sequences.

MU values of CPs can be calculated by the MU weights and total MU values. MU weights ($W = [w_0, w_1, \ldots, w_n], w_0 = 0, w_n = 1$) and total MU values ($t_{\mathrm{mu}}$) are parameters obtained from DICOM RT Plan files. The w_i, which is a cumulative value, denotes the MU weight at cp_i. It is equivalent to denote the total weight between cp_0 and cp_i. The value $(w_{i+1} - w_i)$ is the weight in the interval $[cp_i, cp_{i+1}]$, and the value v_i is the total MU of the interval $[cp_i, cp_{i+1}]$, expressed as follows:

$$v_i = (w_{i+1} - w_i) \cdot t_{\mathrm{mu}}. \tag{4}$$

The value m_i denotes the MU value at control point cp_i, and half of v_{i-1} plus half of v_i equals m_i. The MU values ($M = [m_0, m_1, \ldots, m_n]$) of control points can be computed as follows:

$$m_i = \begin{cases} \dfrac{(w_{i+1} - w_{i-1})}{2} \cdot t_{\mathrm{mu}}, & 0 < i < n, \\ \dfrac{(w_{i+1} - w_i)}{2} \cdot t_{\mathrm{mu}}, & i = 0, \\ \dfrac{(w_i - w_{i-1})}{2} \cdot t_{\mathrm{mu}}, & i = n. \end{cases} \tag{5}$$

4. Methods

This section introduces the proposed method, and Fig. 3 illustrates the whole network structure. The network takes two modalities in the VMAT plan as input. The 3D-MResNet consists of two main components: 3D ResNet and feature fusion module. In the model, the 3D ResNet is employed to convert the image data into feature vectors, and the fusion module integrates them with MU information. The model learns the correlations between imaging and non-imaging modalities in a deep neural network.

4.1. 3*D ResNet*

The MLC aperture shapes can be regarded as images in a timed sequence. Therefore, 3D images attained from the VMAT plan are the temporal sequence of

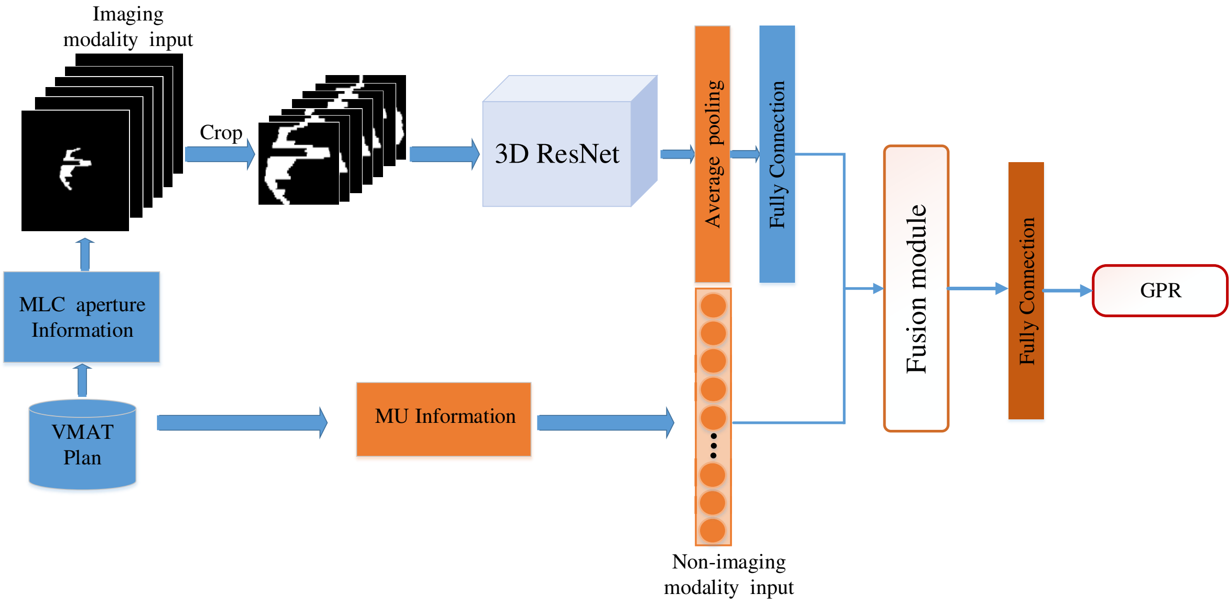


Fig. 3. The structure of 3D-MResNet. The model's inputs are the 3D MLC aperture images and the vector of MU values obtained from VMAT plans. The 3D ResNet extracts the image feature information of image sequences, and the feature fusion module concentrates on the fusion of imaging and non-imaging modalities characteristics. The dimensions of MU values are 182, and the image features are mapped to a vector with a length of 182. The fusion module concatenates these two vectors and makes a prediction. The number of neurons is 182 in the fully connected layer.

the MLC aperture. Consequently, this paper proposes a 3D model for analyzing images in plans. The 3D convolutions are applied in convolutional layers so that discriminative characteristics of the temporal and spatial dimensions are caught.

In our network, the 3D ResNet[53] captures temporal feature information of MLC aperture images well, which is more accurate in description with various aspects of image content. It takes a sequence of images as inputs, as shown in Fig. 3. The basic block in the model is shown in Fig. 4. Each residual block contains two 3D convolutional layers followed by the batch normalization and rectified linear unit layers. The kernel size for convolutional layers is $3 \times 3 \times 3$, and the stride is 1. Extracted feature maps combine information from all images through the CNN structure, and 3D average pooling is applied on the feature maps to cast image features into a feature vector. A fully connected layer with size 182 is then appended to the features extraction module for mapping the features to a one-dimensional vector with a fixed length of 182. For our image data, the input size of the image is $(182, 112, 112)$, and the feature maps of size $(512, 12, 4, 4)$ are obtained from 3D ResNet. Then, the 3D average pooling is performed on the feature maps to produce a feature vector of size $(512, 1, 1)$. Finally, the fully connected layer maps the flattened feature vector to a new vector with a length of 182.

4.2. *Feature-data fusion module*

Imaging and non-imaging modalities obtained from VMAT plans are used to predict GPR. Therefore, the QA prediction model is characterized as a multimodal model. Data fusion methods are essential means of multimodal analysis and mining. At present, feature-level and decision-level fusion are major fusion strategies. The existing fusion techniques combine the features of modalities, and they are difficult to have a one-to-one correlation between the modalities. Hence, they are not suitable for the two modalities of data in VMAT plans.

For VMAT data, the imaging data and non-imaging data have strong relevance, and one MLC image corresponds to a MU value. Besides, the 3D image data contains redundant information and high dimensions. However, there is no redundancy in MU values with lower dimensions, and all of them are vital to GPR prediction. Characteristics of two-modal data make it difficult to fuse data effectively. Enlightened by these, this paper presents an FDF approach that is regarded as a hybrid fusion method according to the VMAT data characteristics.

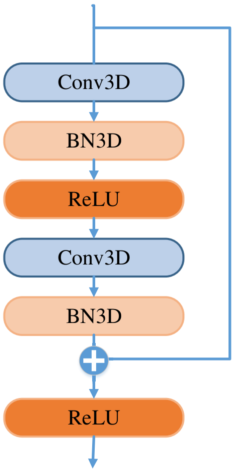
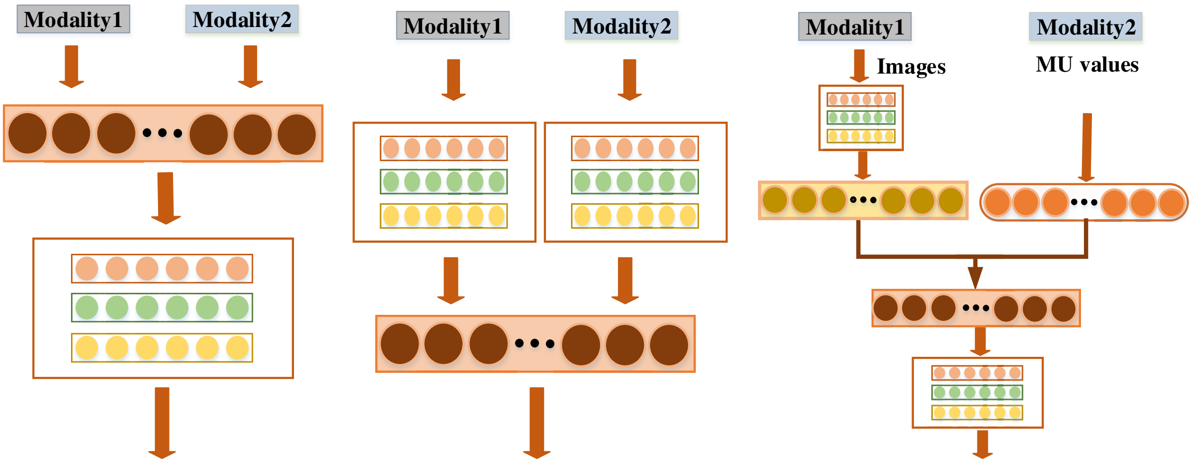


Fig. 4. The basic block of 3D ResNet in this paper. Here, + denotes the add operation.

The FDF approach differs from the feature-level and decision-level fusion. It combines both input data and decision level strategies, and the illustration is presented in Fig. 5. In the early-fusion method, the abstracted features of different modalities are integrated as a combined eigenvector first, and then a neural network model predicts the result through the combined eigenvector, as shown in Fig. 5(a). In the late fusion method, the neural network models is applied on each modality and give the local results of all modalities. The local results are combined to make a final result, as shown in Fig. 5(b). In the FDF method presented in Fig. 5(c), a neural network model extracts features from one modality data and analyzes the features to attain the local decisions of data. However, the local decisions are not the results of the modality, and they are the results of all images in 3D images, namely, each image in the modality has a result. Furthermore, the other modality data and the local decisions are fused as feature vectors. Finally, a neural network model handles the feature vectors to obtain a final decision.

There are significant contributions to GPR prediction for two modalities. It is crucial to make full use of the data during the fusion. The FDF module makes the fusion on equal terms to prevent the predominance of imaging data over non-imaging data.



(a) early-fusion

(b) late-fusion

(c) feature-data fusion (FDF)

Fig. 5. The fusion methods. (a) and (b) are the common fusion methods to merge two different modalities and (c) is the proposed method that combines the images and non-imaging features of a plan.

Hence, we should reduce the dimensions of image data and retain meaningful features. Image features and MU values have the same length when integrating features. For imaging data in VMAT, local decision vectors of 182 dimensions are generated from images by 3D ResNet before fusion. It is equivalent to make a decision on each image. Because each MU value is essential, the MU feature vector of 182 dimensions is eventually retained in the model. In the FDF method, a concatenation layer with 364 units is applied to fuse image features and MU feature vectors. There is a one-to-one relation between the MLC shape image and MU value for each control point in VMAT plans. Therefore, one MU value corresponds to one image features' decision, and they are associated with each other. The image features and MU values are projected to a shared vector, which contains the complementarity between multiple modal data and builds the relationship between multimodal features, and then the GPR prediction is performed based on the vector directly.

4.3. 3D-MResNet

In the 3D-MResNet model, there are three main modules, including the data process module, feature extractor module, and fusion module. The MLC shape images and MU values are generated from raw VMAT plans by the data process method as Eqs. (1)–(5). The feature extractor uses 3D ResNet that abstracts the MLC shape image features. It removes the redundant features effectively and retains the significant features of 3D images. The fusion module employs the FDF approach that integrates the 3D image features with original MU values data. In this study, the image features and MU vectors extracted from VMAT plans are required in the fusion module, which is applied to fuse multiple features. It needs to process and relate information from these two modalities. The MU values are expressed as a feature vector with length 182. Here, the feature vector of images and MU values is concatenated as new eigenvectors with more content.

For gaining much richer information of plan, the 3D-MResNet is introduced in the prediction technique of GPR. It has access to obtain complementary information about different modalities. Thus, it can avoid the problem that some effective information is missing in individual modalities. The advantage is that multimodal data offers sufficient information. The model can learn adequate features at one time and combine features into a joint representation to make a more accurate prediction.

For the VMAT plan dataset ($D = \{(X^{\text{img}}, X^{\text{mu}}), Y\}$), each plan contains the 3D images and MU values two modalities, and it has a label (the value of GPR). For the ith plan $D_i = (X_i^{\text{img}}, X_i^{\text{mu}})$ in the dataset, the GPR prediction problem can be formally denoted as

$$G : D_i \rightarrow P_i, \tag{6}$$

where P_i is the prediction value of ith plan. For the 3D images X_i^{img} in plan D_i, the feature maps M_i of images X_i^{img} are extracted by a 3D ResNet. This process can be expressed as follows:

$$M_i = 3D\_\text{ResNet}(X_i^{\text{img}}). \tag{7}$$

Then, decisions on each image are executed on the feature maps M_i, and one-dimensional vectors F_i are obtained, formally denoted as

$$G_{\text{map}} : M_i \rightarrow F_i. \tag{8}$$

The fusion layer concatenates the image features F_i and MU vector X_i^{mu} to compose a new feature vector with the length of 182×2, and then a neural network predicts a GPR value of the plan D_i. This process is expressed as follows:

$$P_i = G_{fdf}(F_i, X_i^{\text{mu}}), \tag{9}$$

where G_{fdf} denotes the FDF method. The predicted value of GPR is the output of the 3D-MResNet model. Mean square error (MSE) measures the deviation between the predicted value and the true value. The formula of the loss function is expressed as follows:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}(P_i - Y_i)^2, \tag{10}$$

where Y_i is the label, and N is the total number of plans. This process of model is demonstrated in Algorithm 1.

Algorithm 1. 3D multimodal ResNet model

```
Input: The VMAT plan dataset $D$
Output: The prediction $P$ of GPR value
for $i$th VMAT plan in dataset $D$ do
    extract MLC shape images and MU information
    calculate box = $[x, y, w, h]$ by Eqs. (1)–(3)
    crop images: $X_i^{\text{img}} = \text{crop}(\text{box}, X_i^{\text{img}})$
    calculate MU values $X_i^{\text{mu}}$ by Eq. (5)
    the input of model: $D_i = \{X_i^{\text{img}}, X_i^{\text{mu}}\}$
    use 3D ResNet to abstract the feature maps of $X_i^{\text{img}}$: $M_i = 3D\_\text{ResNet}(X_i^{\text{img}})$
    map image features $M_i$ to a vector $F_i$
    fuse image features $F_i$ and $X_i^{\text{mu}}$ and predict GPR : $P_i = G_{fdf}(F_i, X_i^{\text{mu}})$
    update gradients with back propagation algorithm
end for
```

5. Experiments and Results

In this section, we first introduce the details and implementation of the networks. Then, the experiment results and analysis are given. Besides, the performances of the model are discussed.

5.1. *Details of implementation*

The following experiments have been conducted using the multimodal data abstracted from the VMAT plans, described in more detail in Sec. 3. The dataset consists of 690 VMAT plans, and the statistics of the data used in the experiments are summarized in Table 1. The dataset has been randomly divided into the training set and testing set, with 530 and 160 VMAT plans separately.

In this study, the 3D-MResNet maps features of 3D images and MU values of plans to predict GPR. The model is implemented with the Pytorch library. We train the proposed model with weights initialized with kaiming_normal.[54] The network uses MLC aperture images with a specific size ($182 \times 112 \times 112$) and MU values with size (1×182). In the experiments, the learning rate is initialized as 0.01, and the number of epochs is 200, and the mini-batch size is 1. The Adam optimizer[55] is used to optimize the weight of the model. All training is performed using the GPU of NVIDIA Tesla P100.

5.2. *Results and analysis*

The experiments are executed to verify the validity of the proposed method. For the prediction model, the outputs are the predicted GPR of VMAT plans. The predicted GPRs are evaluated under three criteria (3%/3 mm, 3%/2 mm, 2%/2 mm). The prediction error between the predicted and measured GPR is assessed by mean absolute error (MAE). Standard deviation (SD) can better reflect the discrete degree of difference between predicted and target GPR values for each case. Max error (ME) means the maximum deviation between the predicted and measured GPR. MAE, SD, ME are professional assessment indicators, and they are used as model performance indicators in this study.

To assess the predicted GPR, classification accuracy (ACC) of predicted GPR is given. In this study, we classify the predicted GPR values into two categories: correct and incorrect results. If the difference between the target and predicted GPR is less than

3%, the predicted GPR is considered equivalent to the target, so the predicted GPR is defined as a correct classification result. Otherwise, the classification result is incorrect. The sensitivity and stability of Linacs may be different, and consequently the GPR values of VMAT plans are influenced due to the differences. Studies indicate that the mean deviation of the GPR for the Linacs is <1%, and the maximum deviation is 2.6%.[45] Hence, we adopt 3% as the threshold to measure the prediction accuracy of GPR. What is more, the 3% threshold is clinically relevant.

Different 3D CNN networks are applied to determine the most appropriate model to predict GPR in this study. We compare the results on 3D AlexNet,[56] 3D VGG,[57] 3D Inception,[58] 3D ResNet,[53] respectively. Table 2 demonstrates the experimental results (MAE ± SD). The 3D ResNet model is the most suitable network which achieves the best outcomes at the three gamma criteria. Therefore, it is utilized to extract the features from images.

The detailed results of the proposed model are shown in Table 3, and it illustrates MAE ± SD, ACC and ME on three criteria. The MAE between target and predicted GPR values in the test set is 0.71% at 3%/3 mm, 1.16% at 3%/2 mm, and 2.17% at 2%/2 mm, respectively. The evaluation criterion is that the lower MAE values the better performance, because the predicted values are closer to true values when MAE is smaller. For 3%/3 mm, 157 (98.13%) plans have absolute prediction error lower than 3%. 150 (93.75%) plans have absolute prediction error lower than 3% at 3%/2 mm. 126 (78.75%) plans have absolute prediction error lower than 3% at 2%/2 mm. Moreover, the max absolute prediction error is 3.46%, 4.60% and 8.56% at the three criteria separately. All metrics perform best at 3%/3 mm. There is a similar effect on 3%/2 mm. For the strictest criterion 2%/2 mm, the results are slightly inferior to the other two. However, the 3D-MResNet model has a MAE value of 2.17% with a standard error of 1.72%, which satisfies the doctor's expectation.

Table 2. Comparison of the results obtained using different 3D neural networks.

3D models MAE ± SD (%)	3%/3 mm	3%/2 mm	2%/2 mm
AlexNet	0.71 ± 0.66	1.18 ± 1.04	2.37 ± 1.87
VGG	0.75 ± 0.73	1.30 ± 1.28	2.34 ± 1.89
Inception	0.78 ± 0.71	1.34 ± 1.00	2.13 ± 1.58
ResNet	0.71 ± 0.68	1.16 ± 0.93	2.17 ± 1.72

Table 3. The results of 3D-MResNet under different gamma criteria.

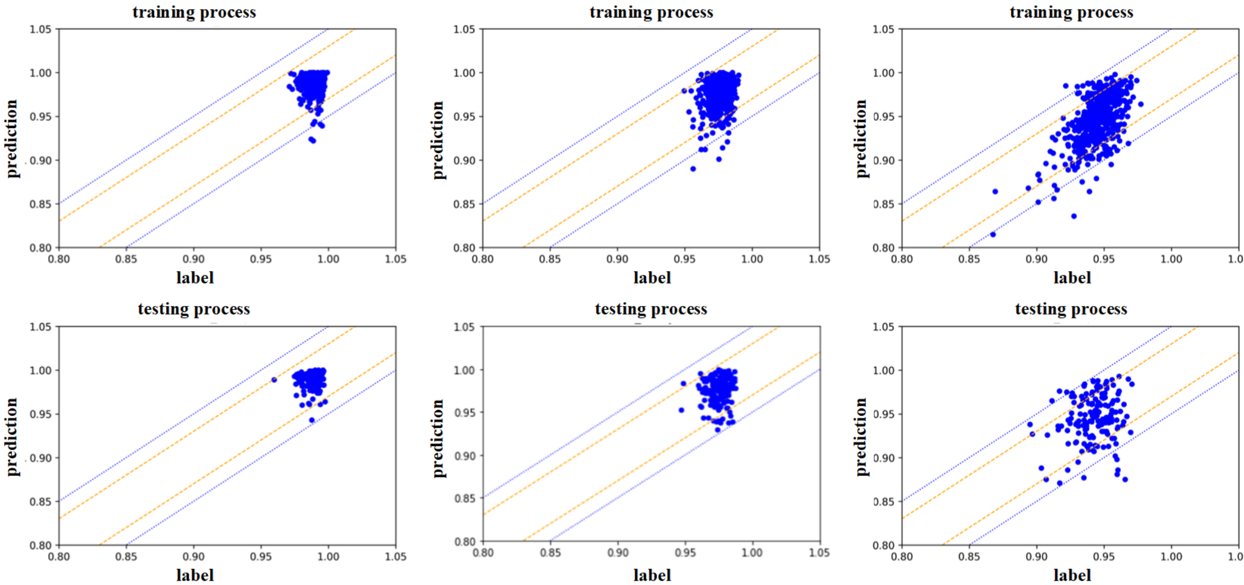
Gamma criteria	MAE ± SD (%)	ACC (%)	ME (%)
3%/3 mm	0.71 ± 0.68	98.13	3.46
3%/2 mm	1.16 ± 0.93	93.75	4.60
2%/2 mm	2.17 ± 1.72	78.75	8.56

This study also carried experiments on a single modality, and Table 4 presents the experiment results. As observed through the qualitative indicator MAE, the 3D-MResNet model combining two modalities can obtain experimental results that are clearly better than those obtained using a single-modality as input. In addition, results also show that the model with imaging data achieves better effects than non-imaging data at 3%/2 mm and 2%/2 mm. Furthermore, the best results on ME are gained when images are the inputs of networks at 3%/2 mm and 2%/2 mm. In all, results on the imaging data are better than non-imaging data.

To illustrate our results in a more intuitive way, we visualize the predicted values in the training and testing set. According to the scatter plots as shown in Fig. 6, the 3D-MResNet model has preferable prediction performance on VMAT plans, and it performs slightly bad in only a few cases. In general, the difference in predicted and measured GPR less than 5% is within the acceptable level. As presented in the results, the great majority of VMAT plan's predicted errors are within 5% at the three gamma criteria in the test set, and the maximum absolute error between target and predicted GPR is within 5% at 3%/2 mm and 3%/3 mm. Moreover, almost all plans can be accurately predicted at 3%/2 mm and 3%/3 mm within 3% error, and only a few cases' absolute errors are more than 3%. In short, the proposed model achieves the desired performance, and it is considered to be feasible for predicting GPR. For a stricter gamma index 2%/2 mm, although the prediction results of plans are inferior to 3%/2 mm and 3%/3 mm, the results meet the clinical requirement.

Table 4. Comparison of the results obtained using the proposed model versus a single-modality model.

Models gamma criteria		3%/3 mm	3%/2 mm	2%/2 mm
3D-MResNet	MAE ± SD (%)	0.71 ± 0.68	1.16 ± 0.93	2.17 ± 1.72
	ME (%)	3.46	4.60	8.56
Imaging data	MAE ± SD (%)	0.95 ± 0.61	1.33 ± 0.94	2.22 ± 1.97
	ME (%)	3.73	4.32	8.21
Non-imaging data	MAE ± SD (%)	0.89 ± 0.83	1.48 ± 1.15	2.30 ± 1.69
	ME (%)	4.92	5.93	8.79



3%/3 mm

3%/2 mm

2%/2 mm

Fig. 6. Scatter plot of predicted GPR values for three gamma criteria in training and testing set.

5.3. *Ablation experiments*

To further validate the significance of the 3D-MResNet model, experiments on different networks have been carried out. We designed contrast experiments to verify the importance of imaging and non-imaging modalities in VMAT plans. Comparison experiments on 3D-MResNet and 2D-MResNet were done to confirm the 3D feature extractor's effectiveness in the VMAT plan's prediction. Meanwhile, we compared the effects of total MU values of beams and MU values of every control point in the prediction model.

The results are shown in Table 5. Experiments demonstrate that the 3D structure performs better than 2D on the VMAT plan dataset. 3D-MResNet can better deal with the GPR prediction of VMAT plans. In fact, 3D feature extractors can better capture the temporal and spatial features of MLC sequence images, but 2D feature extractors cannot extract the image sequence information. The experimental results have proved that the 3D module has a significant impact on the prediction model. The total MU values of the two beams are used to replace the MU of every control point in the model to validate that the MU information is useful for predicting QA results. The research shows that the prediction model's performance depends on the MU values of every control point, which are necessary features for GPR prediction. Similarly, the predictive results of 3D-MResNet are superior to the model based on

Table 5. Comparison of the results obtained with the models based on 3D-MResNet, 2D-MResNet and 3D-MResNet with total MU values, respectively.

Models gamma criteria		3%/3 mm	3%/2 mm	2%/2 mm
3D-MResNet	MAE ± SD (%)	0.71 ± 0.68	1.16 ± 0.93	2.17 ± 1.72
(MU values of CP)	ME (%)	3.46	4.60	8.56
2D-MResNet	MAE ± SD (%)	0.91 ± 0.85	1.73 ± 1.38	2.35 ± 1.90
	ME (%)	4.68	6.18	9.99
3D-MResNet	MAE ± SD (%)	1.09 ± 0.59	1.82 ± 1.07	2.18 ± 1.71
(total MU values)	ME (%)	4.07	7.61	9.03

total MU values at the three gamma criteria. Consequently, it can prove that information of the control point is essential.

The results show that it is effective to predict VMAT GPR by the 3D-MResNet, without any expert knowledge to design features. Comparative experiments show that 3D-MResNet has a much better performance than 2D-MResNet. The model can better meet the doctor's requirements and provides a new method to process VMAT plans and predict GPR.

6. Conclusions

According to the VMAT data characteristics, an automatic prediction model was proposed for the QA results of VMAT plans. It is quite different from the current QA prediction models because it utilizes multimodal data to predict results. Besides, a new fusion approach was presented to fuse the decision results of all images and the corresponding MU values. It deals with multimodal information of VMAT plans and obtains complementary characteristics in different modalities. Experiments demonstrate that the proposed model is an effective model to predict QA results. Meanwhile, the ablation experiments confirm the effectiveness of each module in the model. In addition, this study proves that the features of two modalities have a significant influence on QA results. In summary, the proposed model can accurately predict patient-specific QA results for most VMAT plans at 3%/3 mm and 3%/2 mm, and 2%/2 mm gamma criteria. This model helps radiologists to verify the delivered dose distributions and improve the quality and efficiency of treatment plans. In future works, we will attempt to study some powerful machine learning algorithms for GPR prediction, such as enhanced probabilistic neural networks, dynamic ensemble learning algorithms, neural dynamic classification algorithm, and finite element machine, because the advantages of these methods can improve the GPR prediction model.

Acknowledgments

This work was supported by the National Science Fund for Distinguished Young Scholars under Grant No. 62025601 and the National Natural Science Foundation of China under Grant No. 61772353.

References

1. E. Schreibmann, A. Dhabaan, E. Elder and T. Fox, Patient-specific quality assurance method for VMAT treatment delivery, *Med. Phys.* **36** (2009) 4530–4535.
2. F. Clemente and C. Perez-Vara, Comparison of two different setups for VMAT patient-specific QA, *Med. Phys.* **40**(6) (2013) 255.
3. K. Otto, Volumetric modulated arc therapy: IMRT in a single gantry arc, *Med. Phys.* **35**(1) (2008) 310–317.
4. J. M. Park, J. I. Kim, S. Y. Park, D. H. Oh and S. T. Kim, Reliability of the gamma index analysis as a verification method of volumetric modulated arc therapy plans, *Radia. Oncol.* **13**(1) (2018) 175.
5. V. Kearney, J. W. Chan, G. Valdes, T. D. Solberg and S. S. Yom, The application of artificial intelligence in the IMRT planning process for head and neck cancer, *Oral Oncol.* **87** (2018) 111–116.
6. M. Ahmadlou and H. Adeli, Enhanced probabilistic neural network with local decision circles: A robust classifier, *Integr. Comput.-Aided Eng.* **17**(3) (2010) 197–210.
7. M. H. Rafiei and H. Adeli, A new neural dynamic classification algorithm, *IEEE Trans. Neural Networks Learn. Syst.* **28**(12) (2017) 3074–3083.
8. K. M. R. Alam, N. Siddique and H. Adeli, A dynamic ensemble learning algorithm for neural networks, *Neural Comput. Appl.* **32**(12) (2020) 8675–8690.

9. D. R. Pereira, M. A. Piteri, A. N. Souza, J. P. Papa and H. Adeli, FEMa: A finite element machine for fast learning, *Neural Comput. Appl.* **32**(10) (2020) 6393–6404.
10. G. Valdes, C. B. Simone II, J. Chen, A. Lin, S. S. Yom, A. J. Pattison, C. M. Carpenter and T. D. Solberg, Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making, *Radiother. Oncol.* **125**(3) (2017) 392–397.
11. J. Li, L. Wang, X. Zhang, L. Liu, J. Li, M. F. Chan, J. Sui and R. Yang, Machine learning for patient-specific quality assurance of VMAT: Prediction and classification accuracy, *Int. J. Radia. Oncol. Biol. Phys.* **105**(4) (2019) 893–902.
12. J. Zhang, D. Li, L. Wang and L. Zhang, One-shot neural architecture search by dynamically pruning supernet in Hierarchical order, *Int. J. Neural Syst.* **31**(7) (2021) 2150029.
13. G. Zhang, H. Rong, P. Paul, Y. He, F. Neri and M. J. Pérez-Jiménez, A complete arithmetic calculator constructed from spiking neural P systems and its application to information fusion, *Int. J. Neural Syst.* **31**(1) (2021) 2050055.
14. J. Shen, X. Xiong, Z. Xue and Y. Bian, A convolutional neural network-based pedestrian counting model for various crowded scenes, *Comput. Aided Civil Infrastruc. Eng.* **34**(10) (2019) 897–914.
15. J. Wang, R. Ju, Y. Chen, L. Zhang, J. Hu, Y. Wu, W. Dong, J. Zhong and Z. Yi, Automated retinopathy of prematurity screening using deep neural networks, *EBio Med.* **35** (2018) 361–368.
16. L. Wang, L. Zhang, X. Qi and Z. Yi, Deep attention-based imbalanced image classification, *IEEE Trans. Neural Networks Learn. Syst.* **PP**(99) (2021) 1–11.
17. Y. Xue, P. Jiang, F. Neri and J. Liang, A multi-objective evolutionary approach based on graph-in-graph for neural architecture search of convolutional neural networks, *Int. J. Neural Syst.* **31**(9) (2021) 2150035.
18. U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli and D. P. Subha, Automated EEG-based screening of depression using deep convolutional neural network, *Comput. Methods Progr. Biomed.* **161** (2018) 103–113.
19. Y. Feng, L. Zhang and J. Mo, Deep manifold preserving autoencoder for classifying breast cancer histopathological images, *IEEE/ACM Trans. Comput. Biol. Bioinf.* **17**(1) (2020) 91–101.
20. U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan and H. Adeli, Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals, *Comput. Biol. Med.* **100** (2018) 270–278.
21. Y. Li, Z. Yu, Y. Chen, C. Yang, Y. Li, X. Allen Li and B. Li, Automatic seizure detection using fully convolutional nested LSTM, *Int. J. Neural Syst.* **30**(4) (2020) 2050019.
22. A multi-instance networks with multiple views for classification of mammograms, *Neurocomputing* **443** (2021) 320–328.
23. F. Hu, H. Wang, Q. Wang, N. Feng, J. Chen and T. Zhang, Acrophobia quantified by EEG based on CNN incorporating granger causality, *Int. J. Neural Syst.* **31**(3) (2020) 2050069.
24. G. Liu, W. Zhou and M. Geng, Automatic seizure detection based on S-Transform and deep convolutional neural network, *Int. J. Neural Syst.* **30**(4) (2020) 1950024.
25. Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin and P.-A. Heng, 3D deeply supervised network for automated segmentation of volumetric medical images, *Med. Image Anal.* **41** (2017) 40–54.
26. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox and O. Ronneberger, 3D U-Net: Learning dense volumetric segmentation from sparse annotation, in *Medical Image Computing and Computer-Assisted Intervention — MICCAI* 2016 (Springer International Publishing, Cham, 2016), pp. 424–432.
27. K. Thurnhofer-Hemsi, E. Lpez-Rubio, N. Ro-Vellv and M. A. Molina-Cabello, Multiobjective optimization of deep neural networks with combinations of Lp-norm cost functions for 3D medical image super-resolution, *Integr. Comput. Aided Eng.* **27**(1) (2020) 1–19.
28. O. Martinez Manzanera, S. Meles, K. Leenders, R. Renken, M. Pagani, D. Arnaldi, F. Nobili, J. Obeso, M. Oroz, S. Morbelli and N. Maurits, Scaled subprofile modeling and convolutional neural networks for the identification of Parkinson's disease in 3D nuclear imaging data, *Int. J. Neural Syst.* **29**(9) (2019) 1950010.
29. C. Yang, A. Rangarajan and S. Ranka, Visual explanations from deep 3D convolutional neural networks for Alzheimer's disease classification, *Amia Ann. Symp. Proc.* **2018**(12) (2018) 1571–1580.
30. T. Baltrušaitis, C. Ahuja and L.-P. Morency, Multimodal machine learning: A survey and taxonomy, *IEEE Trans. Pattern Anal. Machine Intell.* **41**(2) (2018) 423–443.
31. C. Xue, F. H. Tang, C. Lai, L. J. Grimm and J. Y. Lo, Multimodal Patient-specific registration for breast imaging using biomechanical modeling with reference to AI evaluation of breast tumor change, *Life* **11**(8) (2021) 747.
32. F. J. Vera-Olmos, E. Pardo, H. Melero and N. Malpica, DeepEye: Deep convolutional network for pupil detection in real environments, *Integr. Comput. Aided Eng.* **26** (2018) 1–11.
33. T. Yang, C. Cappelle, Y. Ruichek and M. E. Bagdouri, Multi-object tracking with discriminant correlation filter-based deep learning tracker, *Integr. Comput. Aided Eng.* **26**(3) (2019) 273–284.

34. F. R. Lera, F. M. Rico and V. M. Olivera, Neural networks for recognizing human activities in home-like environments, *Integr. Comput. Aided Eng.* **26** (2018) 1–10.
35. Q. Guo, J. Jia, G. Shen, L. Zhang, L. Cai and Z. Yi, Learning robust uniform features for cross-media social data by using cross autoencoders, *Knowl.-Based Syst.* **102** (2016) 64–75.
36. L. Fidon, W. Li, L. C. Garcia-Peraza-Herrera, J. Ekanayake, N. Kitchen, S. Ourselin and T. Vercauteren, Scalable multimodal convolutional networks for brain tumour segmentation, in *Int. Conf. Medical Image Computing and Computer-Assisted Intervention* (Springer, 2017), pp. 285–293.
37. M. H. Le, J. Chen, L. Wang, Z. Wang, W. Liu, K.-T. T. Cheng and X. Yang, Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks, *Phys. Med. Biol.* **62**(16) (2017) 6497.
38. T. Xu, H. Zhang, X. Huang, S. Zhang and D. N. Metaxas, Multimodal deep learning for cervical dysplasia diagnosis, in *Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, (Springer, Cham, 2016), pp. 115–123.
39. P. K. Atrey, M. A. Hossain, A. E. Saddik and M. S. Kankanhalli, Multimodal fusion for multimedia analysis: A survey, *Multimedia Syst.* **16**(6) (2010) 345–379.
40. K. Liu, Y. Li, N. Xu and P. Natarajan, Learn to combine modalities in multimodal deep learning, preprint (2018), arXiv:1805.11730.
41. R. F. Thompson *et al.*, Artificial intelligence in radiation oncology: A specialty-wide disruptive transformation? *Radiotherap. Oncol.* **129**(3) (2018) 421–426.
42. L. Wang, J. Li, S. Zhang, X. Zhang, Q. Zhang, M. F. Chan, R. Yang and J. Sui, Multi-task autoencoder based classification-regression model for patient-specific VMAT QA, *Phys. Med. Biol.* **65**(23) (2020) 235023.
43. D. A. Granville, J. G. Sutherland, J. G. Belec and D. J. L. Russa, Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics, *Phys. Med. Biol.* **64**(4) (2019) 095017.
44. P. D. Wall and J. D. Fontenot, Quality assurance-based optimization (QAO): Towards improving patient-specific quality assurance in volumetric modulated arc therapy plans using machine learning, *Phys. Med.* **87** (2021) 136–143.
45. D. Lam, X. Zhang, H. Li, Y. Deshan, B. Schott, T. Zhao, W. Zhang, S. Mutic and B. Sun, Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning, *Med. Phys.* **46**(10) (2019) 4666–4675.
46. G. Valdes, R. Scheuermann, C. Hung, A. Olszanski, M. Bellerive and T. Solberg, A mathematical frame work for virtual IMRT QA using machine learning, *Med. Phys.* **43**(7) (2016) 4323–4334.
47. T. Ono, H. Hirashima, H. Iramina, N. Mukumoto, Y. Miyabe, M. Nakamura and T. Mizowaki, Prediction of dosimetric accuracy for VMAT plans using plan complexity parameters via machine learning, *Med. Phys.* **46**(9) (2019) 3823–3832.
48. Y. Interian, V. Rideout, V. P. Kearney, E. Gennatas, O. Morin, J. Cheung, T. Solberg and G. Valdes, Deep nets versus expert designed features in medical physics: An IMRT QA case study, *Med. Phys.* **45**(6) (2018) 2672–2680.
49. S. Tomori, N. Kadoya, Y. Takayama, T. Kajikawa, K. Shima, K. Narazaki and K. Jingu, A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance, *Med. Phys.* **45**(9) (2018) 4055–4065.
50. S. Tomori, N. Kadoya, T. Kajikawa, Y. Kimura, K. Narazaki, T. Ochi and K. Jingu, Systematic method for a deep learning-based prediction model for gamma evaluation in patient-specific quality assurance of volumetric modulated arc therapy, *Med. Phys.* **48**(3) (2021) 1003–1018.
51. M. J. Nyflot, P. Thammasorn, L. S. Wootton, E. C. Ford and W. A. Chaovalitwongse, Deep learning for patient-specific quality assurance: Identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks, *Med. Phys.* **46**(2) (2019) 456–464.
52. E. Shiba *et al.*, Predictive gamma passing rate by dose uncertainty potential accumulation model, *Med. Phys.* **47**(3) (2020) 1349–1356.
53. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2016), pp. 770–778.
54. K. He, X. Zhang, S. Ren and J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in *Proc. IEEE Int. Conf. Computer Vision* (IEEE Computer Society, 2015), pp. 1026–1034.
55. D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, preprint (IEEE Computer Society, 2014), arXiv: 1412.6980.
56. A. Krizhevsky, I. Sutskever and G. Hinton, ImageNet classification with deep convolutional neural networks, *Adv. Neural Inf. Process. Syst.* **25**(2) (2012) 1097–1105.
57. K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *Comput. Sci.* (2014).
58. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the inception architecture for computer vision, 2016 *IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*) (IEEE Computer Society, 2016), pp. 2818–2826.