Knowledge-Based Systems 271 (2023) 110549

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Multilayer perceptron neural network with regression and ranking loss for patient-specific quality assurance

Wenjie Liu [a], Lei Zhang [a,*], Lizhang Xie [a], Ting Hu [a], Guangjun Li [b], Sen Bai [b], Zhang Yi [a]

[a] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, China
[b] Department of Radiation Oncology, Cancer Center and State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China

ARTICLE INFO

Article history:
Received 30 September 2022
Received in revised form 28 March 2023
Accepted 4 April 2023
Available online 21 April 2023

Keywords:
Patient-specific quality assurance
Volumetric modulated arc therapy
Multilayer perceptron
Gamma passing rate
Convolutional neural network

ABSTRACT

Patient-specific quality assurance (PSQA) of volumetric modulated arc therapy (VMAT) treatment plans is crucial to enable the plans to be validated for clinical acceptance. However, performing PSQA for clinical delivery is labor-intensive and time-consuming. The existing prediction models do not take into account the dynamic delivery process of VMAT plans. To solve the above problems and improve accuracy of PSQA, this paper presents a multilayer perceptron (MLP) neural network model with regression and ranking loss to predict the gamma passing rate (GPR). The proposed model combines a convolutional neural network with multiple MLP blocks for extracting inter-image correlation features of plan files during dynamic delivery. To focus on the similarity and specificity of multiple VMAT plans, a regression and ranking loss function with dynamic weights is proposed to optimize the training process. In addition, a clinical workflow is proposed to combine the designed model with measurement-based PSQA to screen potential risky plans better. A total of 690 VMAT plans from multiple treatment sites are collected to validate the performance. For 2%/2 mm, 3%/2 mm and 3%/3 mm, the best result of mean absolute error and max error between measured and predicted GPR are 2.17%, 1.25%, 0.74%, and 7.89%, 4.29%, 3.05%, respectively. Experimental results demonstrate that the proposed method has a state-of-the-art performance and can improve the VMAT PSQA process and reduce PSQA workloads.

1. Introduction

Volumetric modulated arc therapy (VMAT) [1] is a commonly used modern radiotherapy technique. Compared to conventional intensity-modulated radiotherapy, VMAT improves dosimetry and reduces treatment time [2]. However, VMAT plans consist of highly modulated apertures with increased dosimetric uncertainty and pose a great challenge to the dosimetric accuracy of complex RapidArc plans [3,4]. So a safe and comprehensive quality assurance (QA) must be performed before a patient can undergo surgery.

Patient-specific quality assurance (PSQA) is a technique using gamma evaluation for the quantitative evaluation of dose distributions to solve the above problem, and it is of vital importance in the pretreatment process [5]. In radiation therapy, doctors carefully design and determine the dose-intensity graphic that best fits the tumor shape and then use computer-controlled linear accelerators to send precise doses of radiation to the malignant tumor or specific areas within the tumor. Normal human tissues and organs have a certain tolerance to the radiation dose. A lower dose may result in a poor tumor treatment effect, while a higher dose may reduce the survival rate of patients [1]. Therefore, finding the most appropriate radiation dose intensity is crucial for patient safety and treatment. PSQA enables the designed radiotherapy plan to be validated for clinical acceptance. Although PSQA has some problems, such as being insensitive to errors [6–8], it is still widely used in a number of institutional guidelines [9].

Gamma analysis [5] are commonly used to assess the integrity of deliveries in undertaking VMAT PSQA. But the problem occurs that it is time-consuming and labor intensive and adds a lot of burden to the clinical treatment [10]. So a number of algorithms and models have been proposed to assist the calculation of gamma passing rate (GPR) in ensuring that the dose distribution meets clinical standards [11–16]. These methods can be divided into two categories: traditional machine learning methods that use expert-designed features for training, and deep neural network methods that do not require domain knowledge. While some of these methods perform well, some issues remain and deserve to be addressed.

* Corresponding author.
E-mail address: leizhang@scu.edu.cn (L. Zhang).

https://doi.org/10.1016/j.knosys.2023.110549
0950-7051/

Firstly, most traditional machine learning models are trained with complex features designed by domain experts. These features contain limited information and may overlook some VMAT plan file information. Secondly, previous deep neural network models are trained using only regression loss function, ignoring the correlation and specificity between different plans. Thirdly, none of these models consider the dynamic delivery process of the VMAT plan, i.e., the correlation between the multileaf collimator (MLC) aperture images of the control points. Finally, it is unclear how these models can be combined with the clinical workflow. An explicit integration process needs to be proposed to assist PSQA better.

To address the aforementioned problems, a multilayer perceptron neural network model is proposed for PSQA of VMAT treatment plans. The inputs to the model are MLC aperture images as well as the monitor unit (MU) weights. The model consists of three modules: a feature extraction module, a correlation extraction module, and a feature fusion module. The feature extraction module consists of a convolutional neural network (CNN) for extracting image features. The correlation extraction module consists of multiple multilayer perceptron (MLP) blocks for extracting correlation features between aperture images. The feature fusion module integrates the image features with the MU features. To focus on the similarity and specificity of different VMAT treatment plans, a regression and ranking loss function is proposed to optimize the training process. Meanwhile, we experimentally present a clinical workflow to combine the designed model with measurement-based PSQA. To validate the proposed method and the clinical workflow, 690 VMAT plans, including 125580 aperture images, were collected from the West China Hospital of Sichuan University and used for the experiments. The major contributions of this paper are summarized as follows:

(1) A MLP network model is proposed for PSQA of VMAT treatment plans. This model takes into account the dynamic nature of VMAT plans during delivery and is the first to employ MLP blocks to extract correlation features between MLC aperture images.

(2) A loss function combining regression and ranking is proposed for training. It can minimize the average error while focusing on the similarity and specificity between multiple VMAT plans.

(3) A clinical workflow is designed to combine the proposed prediction model with the measurement-based PSQA, which can better assist physicians in identifying risky plans.

2. Related works

This section presents the development of MLP networks in radiotherapy, followed by an overview of previous studies on PSQA of intensity-modulated radiation therapy (IMRT) or VMAT plans using traditional machine learning methods and deep neural network methods.

2.1. MLP networks in radiotherapy

Conventional MLP networks mainly consist of multiple fully connected layers, and such networks can be used to extract features from the input data. Sun et al. proposed a three-layer MLP network for respiratory signals prediction in gated treatment of moving target in radiation therapy [17]. A three-layer MLP network was also used to predict intrafraction lung tumor motion [18]. Zhu et al. combined the MLP network with long–short term memory structure to improve the prediction of grade 4 radiotherapy-induced lymphopenia [19]. However, the ability of these networks to extract features is unsatisfactory, most notably because of the small number of layers and the lack of a structured connection to integrate multiple layers. In addition, MLP networks are often used as feature classifiers rather than feature extractors [20–22], which discards the ability to extract correlation features. So a deeper MLP network needs to be proposed for exploring the correlation features of radiotherapy data.

Recently, significant breakthroughs have been made in the research of MLP network models [23,24]. The structured blocks make even simple fully connected structures have powerful feature extraction capabilities. Most importantly, by exploring the correlation between multiple patches, the MLP model has excellent ability to extract correlation features. Correlations naturally exist between the MLC aperture maps of VMAT treatment plan files, and it is worth exploring how to use the MLP model to extract such features.

2.2. Traditional machine learning methods for PSQA

Before treating, the dose distribution is usually measured in a phantom. Then the measured dose and dose distribution(s) are compared with those predicted by the planning system to assess the agreement of the two distributions [5]. The usual metrics used include point-by-point percent dose difference, distance-to-agreement (DTA), and the gamma index, which combines both DTA and dose evaluation [5,25,26]. But performing PSQA for the clinical delivery of IMRT/VMAT plans is time-consuming and not altogether instructive due to the myriad sources that may produce a failing result. Thus Valdes et al. developed a mathematical framework using Poisson regression with Lasso regularization to predict IMRT QA passing rates [11]. This method used 78 metrics describing the difference between calculated and measured values to train the model and identified the correlation between IMRT plan complexity metrics and GPR. Later, they verified this method using the data obtained by different measurement approaches at various institutions, proving that this algorithm can be accurately applied to the IMRT quality assurance [12].

Inspired by this work, one Poisson Lasso regression model was developed by Li et al. to assess the accuracy of machine learning to predict quality assurance results for VMAT plans [27]. This model used 54 metrics for training, and a random forest classification model was developed to classify QA results as "pass" or "fail". The mean prediction error is 1.81%, 2.39%, and 4.18% at 2%/2 mm, 3%/2 mm, and 3%/3 mm, respectively.

Lam et al. used three tree-based machine learning algorithms (AdaBoost, Random Forest, and XGBoost) to train the models and predict GPR values [28]. They demonstrated that the proposed methods allowed physicists to better identify the failures of IMRT QA measurements and to develop proactive QA approaches. Granville et al. trained a linear support vector to classify the results of VMAT plans instead of predicting the GPR values [14]. They divided the RT plans into three classes based on median dose variance: $>1\%$, $<1\%$, and within $\pm 1\%$, achieving a macro-averaged area under the ROC curve of 0.88.

These machine learning methods use a large number of complexity metrics or features to construct models. But these features designed by domain experts contain limited information and may overlook some plan file information. Thus, a new model automatically extracting more useful information is warranted.

2.3. Deep neural network methods for PSQA

Neural networks have been studied for many years [29–35], and recently they have achieved important breakthroughs using CNN models in various medical fields, including breast cancer diagnosis [36–39], thyroid diagnosis [40], radiotherapy error classification [41], and medical image segmentation [42–44]. Sahiner et al. reviewed the development of radiotherapy and concluded

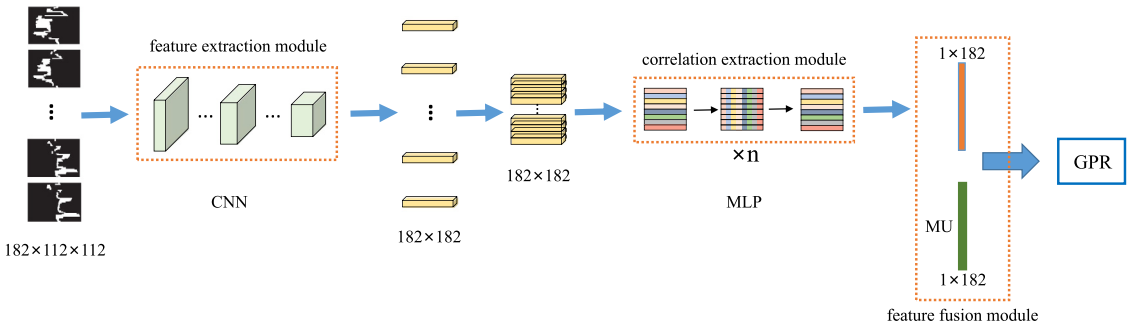

Fig. 1. Diagram of the overall network architecture of the proposed method. First, the MLC aperture images will be fed individually to the feature extraction module to extract image features. After that, all features are stacked and sent to the correlation extraction module to extract relevant features. Finally, the extracted features are concatenated with MU features and sent to the feature fusion module to obtain the predicted GPR.

that neural network methods are commonly used in radiotherapy [45]. But using the deep learning model to evaluate VMAT plans is still challenging. Wolfs and Bedford et al. used portal images as input to a neural network to identify treatment errors, and the results demonstrated the great potential of neural networks for PSQA [46,47]. Interian et al. compared the performance of the proposed deep neural network model against a technique designed by domain experts in the prediction of GPR for IMRT quality assurance [13]. They used fluence maps calculated for each plan as input to the CNN model, and the predicting results were similar to a system carefully designed by physicist experts. Similarly, Mahdavi et al. fed fluence maps into an artificial neural network to validate the dose of the IMRT plans [48].

Tomori et al. trained a CNN model using IMRT plans, which got the Spearman rank correlation coefficients of 0.62 and the MAE value 1.93 in the 2%(global)/2 mm of the test dataset [15]. Their network contained three convolution operations, each specifically including Con + ReLU + Maxpooling. They also used MU and volume as additional input information.

Nyflot et al. provided an alternative perspective for quality assessment of radiotherapy plans [20]. They investigated a deep learning approach to classify the presence or absence of introduced radiotherapy treatment delivery errors from patient-specific QA. The results suggested that the performance of the deep learning approach is better than the performance of the handcrafted method with texture features.

Ono et al. used three machine learning models, regression tree analysis (RTA), multiple regression analysis (MRA), and neural networks (NNs), to predict the dosimetric accuracy using 28 metrics [16]. The results showed that NNs performed slightly better than RTA and MRA.

Although these neural network methods effectively assess the quality of IMRT/VMAT plans, none of them take into account the dynamic delivery process of VMAT plans, i.e., the existence of a correlation between MLC aperture images. It is necessary to introduce such correlations in the model to improve performance. Moreover, these models cannot focus on the similarity between multiple plans if trained using only the regression loss function, which leads to significant prediction errors for individual outlier plans.

3. Method

In this section, the proposed overall network structure is introduced first. Then the structural MLP block is presented. In addition, section 3.3 illustrates the details of the proposed regression and ranking loss function. Finally, how the proposed model can be combined with clinical workflow to monitor risky VMAT treatment plans is shown in Section 3.4.

3.1. Overall network architecture

The purpose of this work is to predict the GPRs of VMAT treatment plans to be as close as possible to the measured GPR labels. Let the training set with N VMAT plans be noted as $D = \{((x_i, m_i), y_i); i = 1, 2, \ldots, N\}, x_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,I}\}$. I is the number of control points of one VMAT plan (I = 182 in this study). $x_{i,j} \in \mathbb{R}^{w \times h \times c}$ is the jth MLC aperture image of ith RT plan, where w, h and c represent the width, height and channel (RGB values) of the image, respectively. $m_i \in \mathbb{R}^I$ is the MU values of ith plan and $y_i \in [0, 1]$ is the measured GPR label of ith plan. Our architecture is designed to learn a robust model $f(y_i|(x_i, m_i), \phi)$ that can predict the GPR value which is equal to or very close to its corresponding label. Here, ϕ denotes the parameters in the proposed model.

The overall network structure of the proposed model is shown in Fig. 1. It consists of the following three parts: the feature extraction module, the correlation extraction module, and the features fusion module. The feature extraction module consists of a pre-trained CNN network, which is designed to extract features in one single aperture image. These features include intuitive features such as the shape and size of the aperture as well as some high-dimensional non-intuitive features. Since there are sequential relationships between multiple aperture images of one VMAT plan, the correlation extraction module is proposed to extract relevant features between them. It consists of several structural MLP blocks containing only linear layers and activation functions. Because MU values measure machine output from a clinical accelerator for radiation therapy, the features fusion module fuses the extracted features and the MU features to obtain the final predictions.

Formulaically, given a training sample (x_i, m_i), we first feed x_i into the feature extraction module. For brevity, the subscript i has been removed in the latter part. The image features $V_c \in \mathbb{R}^{I \times I}$ will be extracted by $V_c = f_c(x|\varphi_c)$, where f_c stands for a pre-trained CNN with parameters φ_c. The dimension is $I \times I$ in order to enable a linear layer of fixed dimension to be applied to both V_c and V_c^T. Then V_c will be fed into the correlation extraction module and the relevance features $v_r \in \mathbb{R}^I$ are obtained by $v_r = f_r(V_c|\varphi_r)$, where f_r is a multilayer perceptron network with parameters φ_r. The structure of this network is described in the next subsection. The final features $v_f \in \mathbb{R}^{2I}$ are obtained by concatenating v_r and m. Finally, the predicted GPRs are calculated in the feature fusion module by Eq. (1):

$$GPR = \sigma(v_f w^T + b), \tag{1}$$

where w and b are the weights and bias of a fully connected layer, and σ is a sigmoid activate function that keeps the predictions in the [0, 1].

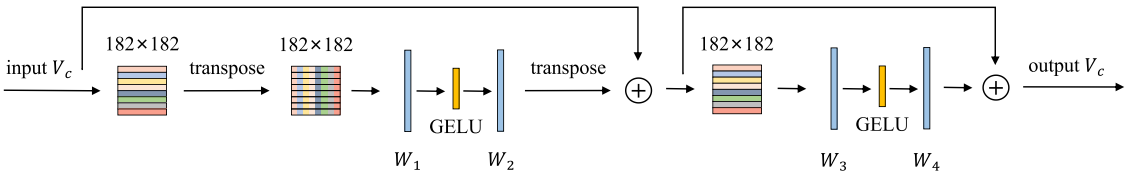

Fig. 2. Diagram of the architecture of an MLP block. W_1, W_2, W_3, and W_4 denote the weights of the four fully connected layers, respectively. Given the input V, it will first be fed into two fully connected layers to be computed in the column direction, after which it will be summed with the original matrix by the residual structure. The same computation is then performed in the row direction of the matrix, and finally, the output is obtained.

In contrast to previous studies, our inputs are only MLC aperture images as well as MU weights, and they can both be derived directly from the planning system. This allows the data preprocessing process without requiring domain knowledge. Most importantly, the proposed model combines the advantages of CNN and MLP, taking into account both the specificity of a single image and the correlation between images. In this way, the model can capture deeper features, improving accuracy.

3.2. Multi-layer perceptron network

This section describes the details of the multi-layer perceptron network. In previous studies, features would be fed directly into the logistic regression layer to obtain predicted values, but this approach ignores the correlation between ordered images. Typically, a VMAT treatment plan contains consecutive control points, and there may be some similarity between their corresponding aperture images. The connection between adjacent control points will be greatly lost if only the features of each aperture image are extracted separately. Hence, the purpose of the proposed method is to extract such correlation features using a lightweight and straightforward network and incorporate these features into the model. One possible approach is to use recurrent neural networks (RNNs), such as LSTM [49] or GRU [50]. However, these networks are too complex and have limited ability to extract long-range features. Inspired by the work of [23,24], the multi-layer perceptrons are used to extract the correlation features between the MLC aperture images of VMAT treatment plans. To the best of our knowledge, we are the first to use the MLP network to explore correlations between MLC aperture images and use them to predict GPR values, which provides a new perspective for future studies.

Specifically, the MLP network consists of multiple repetitive linear structured blocks. The structure of an MLP block is shown in Fig. 2. Each block contains four fully-connected layers, two GERU nonlinearity functions, and two skip connections. After feeding the aperture images into the feature extraction module, $V_c \in \mathbb{R}^{I\times I}$ is obtained. Then it is updated by an MLP block according to

$$\begin{aligned} T_c &= V_c + (\sigma(Norm(V_c)^T W_1)W_2)^T, \\ V_c &= T_c + (\sigma(Norm(T_c)W_3)W_4), \end{aligned} \tag{2}$$

where $Norm$ is the layer norm function, σ is the GELU nonlinearity functions [51], and W_1, W_2, W_3, W_4 are the weights of four fully-connected layers, respectively. After computation by multiple identical MLP blocks, the features are averaged to obtain the final features for regression.

Most MLP models first cut images into multiple patches, and later extract patch-wise correlation features. Unlike them, the proposed model does not slice the images, but lets the model learn the image-wise correlation features directly. The MLC aperture images of VMAT plans are inherently highly correlated with each other, and this MLP structure maximizes the ability to capture these features and use them for prediction. Moreover, single convolution can only capture local domain information of ordered images, and capturing long-distance dependencies requires repeated local computations, which is inefficient. The distance here refers to the interval step in the sequence of two different aperture pictures of one plan. In contrast, the MLP model can extract long-distance dependency features among ordered pictures by global computation of linear layers, which improves performance.

3.3. Regression and ranking loss function

The task in this paper can be considered as a regression task. The model can be trained using classical regression loss functions, such as MSEloss. In clinical PSQA, physicians are particularly concerned about plans with low prediction values because these plans are most likely to be risky. However, as described in Section 4.1, the distribution of labels is concentrated and using only the regression loss function to train the model results in more concentrated predictive values. Moreover, there are similarities in the design of some plans, and the model will lose the global view using only the regression loss function. Hence, a regression and ranking loss (RRLoss) function is proposed to overcome these problems. Based on regression, the proposed loss function also treats the sample space as a whole and pulls back a small number of samples with low GPR values to the space it belongs to by ranking. The following three subsections describe the proposed process of RRLoss.

3.3.1. Regression loss

Let define y as the measured labels and y' as the predicted labels. According to [15], the Huber loss function [52] can enable stable training. It is defined as:

$$L_m = \begin{cases} \frac{1}{2}(y-y')^2, & if\ |y-y'| \le \delta, \\ \delta\cdot(|y-y'| - \frac{1}{2}\delta), & otherwise, \end{cases} \tag{3}$$

where δ is the empirical parameter ($\delta = 1$ in this study). By minimizing the distance between y and y', this loss function can be well used for the regression task to obtain small mean errors.

3.3.2. Ranking loss

Since the GPR labels are concentrated, using only the regression loss function to train the model will result in concentrated predicted GPRs, which leads to edge samples with large errors. Inspired by ListNet [53], measuring the agreement between the predicted ranking list and the ground truth labels enables fine-tuning of individual samples with large deviations, thus improving model performance. Treating multiple samples as a whole by ranking also allows the model to learn the similarities between plans. However, it is impossible to rank all the predicted GPRs in one batch. So a loss function that only ranks for batch size samples is proposed to solve the problem.

Assume that the batch size is n. The permutation is written as $\pi = \langle \pi(1), \pi(2), \ldots, \pi(n)\rangle$, where $\pi(i)$ refers to the VMAT treatment plan at the ith position in the permutation. As aforementioned, the predicted GPR of the VMAT plan pointed by $\pi(i)$ is $y'_{\pi(i)}$. Any permutation is possible, so the set of all possible

permutations is denoted as Ω_n. One permutation probability in Ω_n is defined as Eq. (4).

$$P_{y'}(\pi) = \prod_{j=1}^{n} \frac{\Phi(y'_{\pi(j)})}{\sum_{k=j}^{n} \Phi(y'_{\pi(k)})}, \tag{4}$$

where $\Phi(.)$ is an increasing and strictly positive function. The top-1 probability $P_{y'}(i)$ is defined as:

$$P_{y'}(i) = \frac{\Phi(y'_i)}{\sum_{k=1}^{n} \Phi(y'_k)}, \tag{5}$$

where y'_i is the ith predicted GPR in the batch. For convenience, we use the softmax function to calculate Eq. (5), i.e., $P_{y'}(i) = softmax(y'_i)$. Cross Entropy can be used to calculate the distance. So given two lists of GPRs y and y', the ranking loss is computed by:

$$L_r = -\sum_{i=1}^{n} P_y(i) \log(P_{y'}(i)). \tag{6}$$

This ranking loss function introduces relative positions between the predicted values, thus exploring the correlation between the samples. Batch-wise ranking also allows the model to treat a batch of samples as a whole and learn similarities between samples.

3.3.3. RRLoss

The mean absolute error is an important metric to assess the performance of the model, so the regression loss function is particularly important. The idea of loss combination is that the weights of ranking loss gradually decrease and the weights of regression loss gradually increase during training. Referring to [54], the dynamic RRLoss function is designed as follows:

$$\begin{aligned} L &= \tau_e \times L_m + (1 - \tau_e) \times L_r, \\ \tau_e &= \frac{1}{1 + \exp(\gamma(E/2 - e))}, \end{aligned} \tag{7}$$

where e is the value of the current epoch, E is the total number of epochs, and γ is a hyper-parameter with a value of 0.01 in this paper.

Algorithm 1 Training process of the proposed model.

Input:
The input dataset: $D = \{((x_i, m_i), y_i); i = 1, 2, \cdots, N\}$ Ending epoch = E

Output:
The predicted GPRs y_i'

```
Initializing the CNN with pre-trained parameters
for training epoch e = 1 : E do
    for i-th VMAT plan in dataset D do
        Calculating image features V_c ← f_c(x_i|φ_c)
        Calculating correlation features v_r ← f_r(V_c|φ_r) by Equation (2)
        Computing final features v_f by concatenating v_r and m_i
        Computing GPRs y_i' by Equation (1)
        Computing L by Equation (7)
        Updating gradients with BP algorithm
    end for
end for
```

The overall training process of the proposed network is shown in Algorithm 1. By dynamically adjusting the weight of L_m and L_r during training, the model takes into account the specificity of a single sample and the correlation between batch-wise samples.

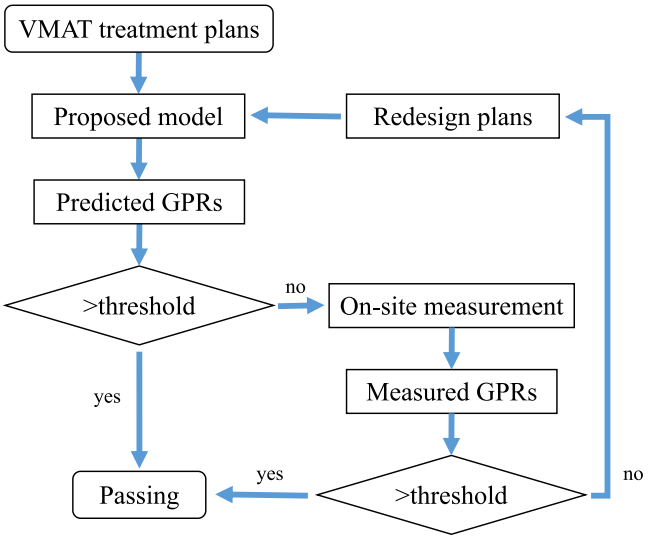

Fig. 3. Diagram of the clinical workflow using the proposed model. Passing means that the plan will not require further device-based physical measurements.

3.4. Clinical workflow

This section describes how the proposed model can be applied to clinical patient-specific QA. Measurement-based PSQA is labor-intensive and time-consuming, while neural network models provide instant prediction with high accuracy. The main idea of the designed workflow is to combine the prediction model with the measurement-based QA. After the Kolmogorov–Smirnov test, the p-values of the collected dataset are 0.0287, 6e-5, and 3e-12 for 2%/2 mm, 3%/2 mm, and 3%/3 mm, respectively. It indicates that the dataset distribution is a part of a normal distribution ($p < 0.05$). According to the AAPM TG-218 report [55], the tolerance and action limits of 0.90% and 0.95% for 2%/2 mm and 3%/2 mm criteria in VMAT QA are recommended based on a part of a normal distribution, respectively. So 0.9 and 0.95 are chosen as thresholds to judge whether a VMAT plan is safe or not.

The clinical workflow using the proposed model is shown in Fig. 3. Firstly, the VMAT treatment plans are fed into the proposed model to get the predicted GPRs. Then they are compared with the threshold (0.9/0.95), and plans with GPRs higher than the threshold will be considered as safety plans, while plans with GPRs lower than the threshold will be considered as risky plans. On-site physical measurements are required for the risky plans, and plans with GPRs higher than the threshold will be reconsidered as safety plans. Plans that are still risky should be fixed or re-planed by the dosimetrists or physicists. The proposed model acts as a prediagnosis, saving time and increasing efficiency. In addition, by adjusting the thresholds or changing the gamma criteria, different institutions can customize the tolerance of VMAT treatment plans.

4. Experiment

4.1. Data set

The data used in this experiment included a total of 125580 MLC aperture images of 690 VMAT plans collected from the Department of Radiation Oncology of the West China Hospital of Sichuan University from June 2018 to June 2020. The plans consist of 37 clinical sites (Rectum (185), Nasopharyngeal Carcinoma (141), Cervix (67), Prostate (60), Uterus (28), Stomach

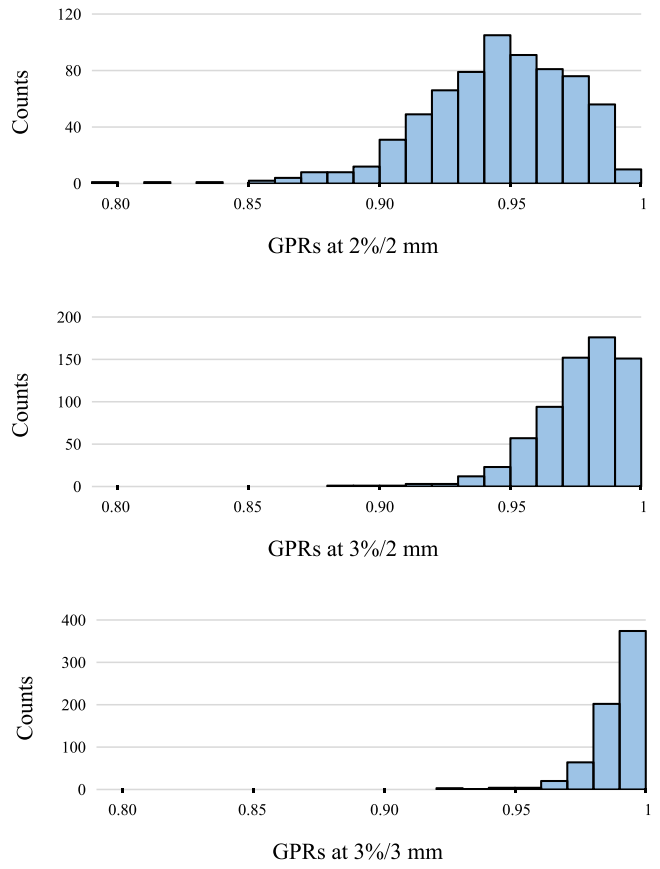

Fig. 4. Distribution of GPR labels for 690 VMAT plans at 3%/3 mm, 3%/2 mm, and 2%/2 mm gamma criteria.

(27), Brain (22), Larynx (19), Pharynx (10), Pancreas (9), Colon (9), Tongue (8), other diseases (95)). Three types of GPR labels are calculated by comparing the measured and calculated dose plans in 3%/3 mm, 3%/2 mm, and 2%/2 mm criteria, respectively. The range and distribution of the measured GPR values are shown in Fig. 4.

There are two beams in each VMAT plan file, and each beam contains 91 control points (CPs). Therefore, a VMAT plan consists of 182 CPs. Each CP has corresponding MLC positions and MU weights. The shape of an MLC aperture depends on the coordinates of the MLC leaf pairs extracted from the VMAT treatment plan files. The MLC aperture image size of the CP is 400×400. To extract features better, the useless background of each MLC aperture image is removed, and a 112×112 image is obtained. Because MU values are a factor for assessing beam and overall plan complexity and deliverability [11,56], they are fed into the model as features.

4.2. Training details

The proposed model is implemented using the Pytorch framework, and all experiments are conducted on a workstation with a Linux OS and an NVIDIA GeForce RTX 3090 GPU. The input data dimension is $182 \times 3 \times 112 \times 112$, i.e., number of control points, image channels, image length, and image width. All the MU values are normalized to scalars in $[-1, 1]$ using Z-score normalization to ensure that different types of features have the same scale. The learning rate and batch size are set as 0.01 and 4. The SGD optimizer is adopted with the parameters of weight decay, momentum, and dampening set as 0.0001, 0.9, and 0.9, respectively. The loss function used in the experiment is shown in Eq. (7), where the hyper-parameter τ is 0.01.

All VMAT treatment plans are randomly divided into a 7:3 ratio for training and testing. The maximum training epochs is 300, and parameters of the network with the best results will be saved to evaluate the performance. The proposed method is evaluated in terms of mean absolute error (MAE), standard deviation (SD), and max error (ME). The accuracy of the threshold-based clinical workflow is also considered.

4.3. Comparison among different backbones

In this study, the features of the aperture images will be extracted by a feature extraction module, which consists of a pre-trained network. So some common pre-trained models are compared, such as AlexNet [57], Resnet101 [58], DenseNet121 [59], VGG16 [60]. For a fairer comparison, all parameters are set to be the same, including learning rate, optimizer, batch size, and the maximum number of epochs. Also, the number of layers of MLP modules are eight.

Table 1 shows the performance of different networks under different criteria. For both MAE and ME metrics, the lower the value, the better the performance of the model. DenseNet121 shows superb feature extraction ability and achieves the smallest MAE and ME values under all three criteria. Therefore, it is selected as the backbone network for the CNN module in the subsequent experiments. It is noteworthy that the proposed method can be well combined with various networks, which also shows its generality.

4.4. Comparison among different MLP layers

The MLP network is used in this paper to extract the correlation features between multiple aperture images of a single VMAT treatment plan. Fig. 2 shows how an MLP block is computed, which means the number of MLP blocks may affect the performance of the model. So ablation experiments for a different number of MLP blocks are performed to illustrate its effect on the performance. The results are illustrated in Table 2. The number of parameters and the throughput of the model are also listed for comparison.

Despite the different number of blocks, all these models achieved satisfactory performance. As the number of blocks increases, the performance of the model does not improve significantly. On the contrary, the performance of the model decreases when the number of blocks reaches 32. It may be because an excessive focus on the correlation between images affects the performance of the feature extraction module, and balancing the feature extraction module and the correlation extraction model is the key to improving the performance. Increasing the number of blocks also leads to a rise in the number of parameters as well as a decrease in throughput. Considering performance, the number of parameters, and throughput, a model with 8 MLP blocks has the most potential. It achieves the lowest MAE and ME for the 3%/2 mm criterion while obtaining a competitive performance for the other two criteria. Most importantly, the smaller number of parameters and higher throughput of the 8-layer MLP network makes it more suitable for deployment. Therefore it is used to validate the performance of the proposed clinical workflow.

4.5. Comparison between 2D and 3D networks

In this paper, an MLP network is proposed to extract the correlation features between aperture images. In fact, all aperture images of one VMAT plan can also be regarded as time-series

Table 1
Comparison of different backbones under different criteria.

Method	2%/2 mm (%)			3%/2 mm (%)			3%/3 mm (%)		
	MAE	SD	ME	MAE	SD	ME	MAE	SD	ME
AlexNet+MLP	2.24	1.96	9.86	1.37	1.16	4.41	0.75	0.87	4.79
Resnet101+MLP	2.27	**1.87**	8.04	1.35	1.20	5.20	0.77	0.82	4.61
DenseNet121+MLP	**2.17**	1.88	**7.89**	**1.25**	**1.08**	**4.29**	**0.74**	**0.76**	**3.05**
VGG16+MLP	2.21	1.91	8.03	1.27	1.11	4.94	0.75	0.83	4.51

MAE: mean absolute error; SD: standard deviation; ME: max error.
The bold value is the optimal value for a metric.

Table 2
Comparison of different number of MLP Blocks under different criteria.

Number of layers	2%/2 mm (%)			3%/2 mm (%)			3%/3 mm (%)			Parameters (M)	Throughput (plans/s)
	MAE	SD	ME	MAE	SD	ME	MAE	SD	ME		
4	2.20	2.00	8.91	1.28	1.17	4.79	0.73	0.82	4.35	34.0	56
8	**2.17**	1.88	7.89	**1.25**	**1.08**	**4.29**	0.74	0.76	**3.05**	38.7	45
12	2.27	1.97	7.51	1.26	1.19	4.66	0.75	0.79	3.28	43.5	39
24	2.38	**1.87**	**7.46**	1.35	1.17	4.98	**0.71**	0.82	3.33	57.8	38
32	2.23	1.90	8.04	1.31	1.14	5.18	0.75	**0.75**	3.58	67.3	36

MAE: mean absolute error; SD: standard deviation; ME: max error.
The bold value is the optimal value for a metric.

Table 3
Performance comparison between 2D and 3D models.

Method	Convolution	2%/2 mm (%)			3%/2 mm (%)			3%/3 mm (%)		
		MAE	SD	ME	MAE	SD	ME	MAE	SD	ME
AlexNet [57]	3D	2.99	2.30	11.88	2.32	1.49	7.29	0.95	0.97	4.43
Resnet101 [58]	3D	2.31	2.01	8.22	1.36	1.22	5.11	0.76	0.86	4.42
DenseNet121 [59]	3D	2.21	1.88	8.47	1.33	1.20	4.89	0.75	0.82	4.06
VGG16 [60]	3D	3.40	3.12	13.18	2.27	2.40	12.57	1.57	2.19	14.42
ResNet101+LSTM	2D	2.81	2.43	12.9	1.71	1.41	6.38	1.04	1.11	5.46
ResNet101+GRU	2D	2.61	2.26	8.80	1.59	1.45	6.43	0.85	0.94	4.62
DenseNet121+LSTM	2D	3.27	3.00	17.28	2.01	2.10	10.13	1.31	1.69	8.07
DenseNet121+GRU	2D	3.21	3.11	16.96	1.96	2.21	12.95	1.14	1.33	6.33
Our model	2D	**2.17**	**1.88**	**7.89**	**1.25**	**1.08**	**4.29**	**0.74**	**0.76**	**3.05**

Convolution: Convolution kernel dimension; MAE: mean absolute error; SD: standard deviation; ME: max error.
The bold value is the optimal value for a metric.

input, and RNN-based models can be applied to such data. Therefore, we replace the MLP network with LSTM [49] or GRU [50] to compare the performance comprehensively, while the feature extraction module remains unchanged. On the other hand, stacking images together in a sequence of CPs can form 3D aperture images. Directly using 3D networks to extract features from 3D aperture images is also a feasible approach. Therefore, with all settings being the same, we modified the convolution kernel of the commonly used network into a 3D convolution kernel for comparison with the proposed method. The extracted features are likewise concatenated with the MU values, and the predicted GPRs are calculated by Eq. (1). The results are shown in Table 3.

It can be seen that the proposed method achieves MAE of 2.17%, 1.25%, and 0.74% in 2%/2 mm, 3%/2 mm, and 3%/3 mm, respectively, showing state-of-the-art performance. For 2D convolutional models, the performance of RNN-based models does not meet expectations, and some models even obtain high ME. It may be because the aperture images are not strictly time-series data. Still, a correlation exists, and the MLP network captures this correlation sufficiently to achieve the desired performance compared with RNN-based models. For the models with 3D convolution, DenseNet achieves the best performance, followed by ResNet. The more dense the connections between layers of the network, the easier it is to extract the correlation features of the aperture images. In addition, our model has a substantial reduction in ME compared to other models, which demonstrates its high robustness.

4.6. Ablation study with different loss

To verify whether the proposed RRLoss can combine the features of regression and ranking, we train the model using regression loss or ranking loss alone. The results are shown in Table 4. It illustrates that it is challenging to train the model using only ranking loss because patient-specific QA is not a ranking task per se. Focusing only on the relative positions between different plans can make it difficult for the model to converge. In contrast, regression loss demonstrates superior performance on this task. The performance of the model is further improved when combining regression loss with ranking loss. It is possible to train the model using only the regression loss function, but then each plan is isolated. The model would only focus on reducing the overall error and thus ignore the specificity of individual plans. Ranking, on the other hand, allows the model to expand its field of view during training by first observing one batch number of plans. In this case, the model pays extra attention to the correlation between these plans. By combining regression loss with batch-wise ranking loss, the model will train steadily while taking into account the similarity and specificity between plans.

4.7. Analysis of predictions

The prediction results for the three criteria on the test set are shown in Fig. 5. During training, the model minimizes the overall

Table 4
Performance comparison between different loss functions.

Method	2%/2 mm (%)			3%/2 mm (%)			3%/3 mm (%)		
	MAE	SD	ME	MAE	SD	ME	MAE	SD	ME
Ranking loss	21.84	3.57	28.58	56.98	2.10	60.75	21.16	1.61	24.94
Regression loss	2.22	1.91	8.63	1.30	1.17	4.77	0.74	0.80	4.04
RRLoss	**2.17**	**1.88**	**7.89**	**1.25**	**1.08**	**4.29**	**0.74**	**0.76**	**3.05**

MAE: mean absolute error; SD: standard deviation; ME: max error.
The bold value is the optimal value for a metric.

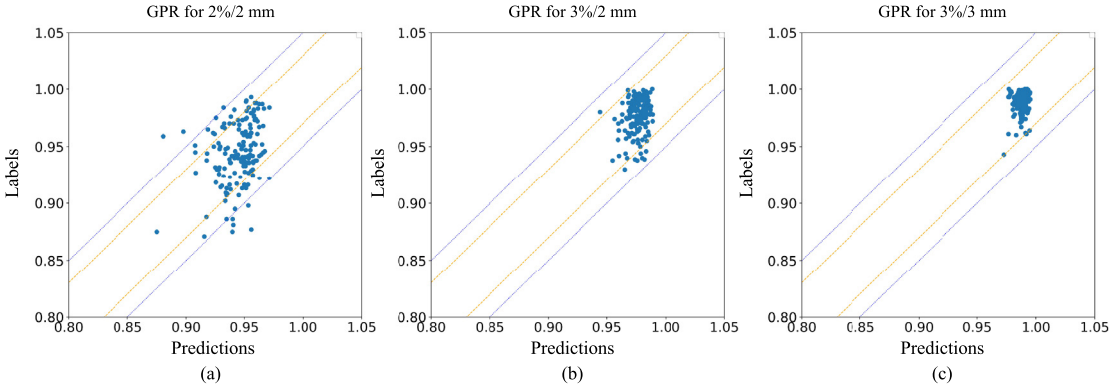

Fig. 5. Scatter plot of predicted GPR values for three gamma criteria on the test set. The yellow or blue lines represent ±3% or ±5% difference between the predicted and measured values, respectively.

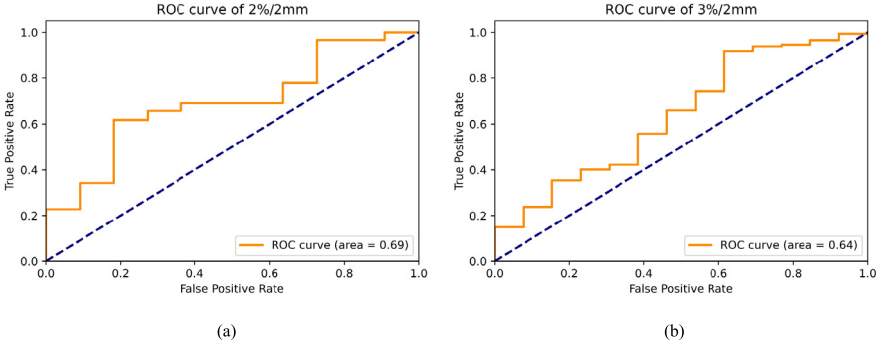

Fig. 6. ROC curve of 2%/2 mm and 3%/2 mm criteria on the test set. The figures are plotted with two parameters, a fixed threshold (0.90 or 0.95) for determining whether one plan is a positive or negative sample and a varying threshold (0% to 100%) for calculating the true/false positive rates.

error, which weakens the effect of individual outlier samples, thus leading to overfitting. Therefore, under the 3%/3 mm criterion, the prediction values are mostly at a specific level (98%–100%), and some plans with low measured values have higher prediction errors. In the future, we will collect more plans with low measured values to alleviate this problem. In addition, the ROC curves for the 2%/2 mm and 3%/2 mm criteria on the test set are shown in Fig. 6. With thresholds of 0.90 and 0.95, these two types of criteria obtained area under curves of 0.69 and 0.64, respectively. It is worth mentioning that the thresholds are adjustable so that different thresholds can meet the needs of various institutions for the tolerance and action limits in QA.

5. Discussion

As shown in Table 3, the proposed model achieves state-of-the-art results in all evaluation metrics. To better demonstrate the performance of the model, we visualize the prediction results on the training and test sets and show them in Fig. 7. It can be seen that the absolute error between the predicted and measured GPRs is less than 4% for a large number of plans, and there are only a few plans with larger errors. For different gamma criteria, the less accurate the measurement criterion, the smaller the prediction errors. In general, an error less than 5% is within the acceptable level between predicted and measured GPR. In Fig. 7(b), the absolute errors of all predictions under 3%/2 and 3%/3 mm criteria are less than 5%. The errors of predictions under the 3%/3 mm criterion are more significant, which may be due to data imbalance, and we will collect more low-GPR data in the future to improve the performance.

In section III-D, a method is proposed for combining our model with clinical work, and the process is shown in Fig. 3. As suggested by the AAPM TG-218 report [55], 0.90 and 0.95 are set as the threshold values for the 2%/2 mm and 3%/2 mm criteria, respectively. In order to verify whether this combined approach can screen out risky VMAT plans, the prediction results of the training and test sets are counted and presented in Table 5. The accuracy is calculated through the process in Fig. 3, i.e., $Acc = num(right)/num(total)$, where $num(right)$ denotes the number of plans that are correctly passing and $num(total)$ denotes the total number of plans. With both 2%/2 mm and 3%/2 mm criteria, plans are successfully detected as safe or dangerous with an accuracy of more than 90%, which meets physician expectations. It indicate that the proposed method can be well integrated with clinical

Table 5
Prediction performance of the proposed model on training and test sets under different criteria.

Metrics	2%/2 mm (n)		3%/2 mm (n)		3%/3 mm (n)	
	Training	Test	Training	Test	Training	Test
Abs Err < 3%	445	124	513	151	523	158
3% <= Abs Err < 5%	70	28	13	9	5	2
Abs Err > 5%	15	8	4	0	2	0
Acc	95.8%	92.5%	94.1%	91.2%	–	–

Abs Err: absolute error; Acc: accuracy in clinical workflow.

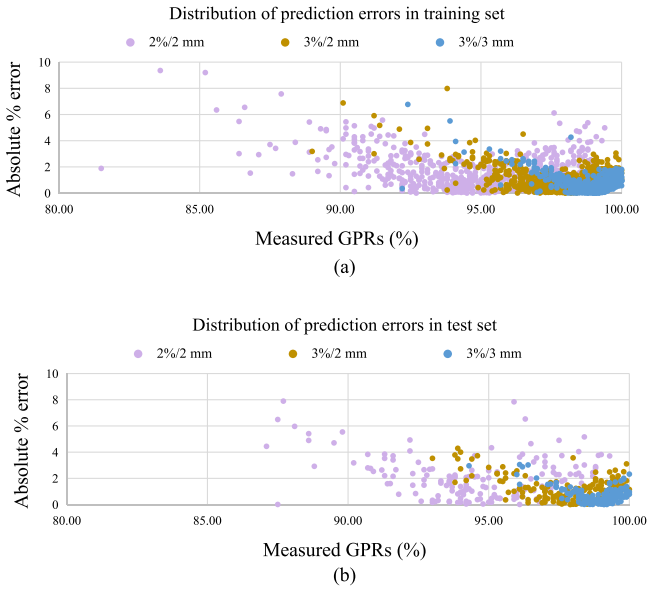

Fig. 7. Distribution of prediction errors in training set and test set. The *x*-axis represents the measured GPR labels, and the *y*-axis represents the absolute errors between the predicted and measured values.

workflow to screen out risky plans, which can assist physicians in better controlling the quality of VMAT plans.

To the best of our knowledge, we are the first to propose a model to explore the correlation between aperture images and to apply it to VMAT QA. We are also the first to propose a combination of regression and ranking loss functions to train PSQA models. Meanwhile, we try to integrate the proposed model with clinical work to better screen for risky plans. The results also proved the validity of this combination, which provides a new idea for future QA studies.

There are some limitations to this study. First, the data are all sourced from a single institution, so the generalizability of the model needs to be verified. Second, the data distribution is unbalanced, with fewer plans having low measured GPRs. In the future, more data from different institutions should be collected to validate the generalizability and performance of the model.

6. Conclusion

In this study, a multilayer perceptron neural network model with regression and ranking loss is proposed for PSQA of VMAT treatment plans. This model is made up of a feature extraction module, a correlation extraction module, and a feature fusion module. Among them, the correlation extraction model consists of multiple MLP blocks, which can extract the correlation features between the MLC aperture images of the VMAT plans. To the best of our knowledge, we are the first to propose a model to explore the correlation between MLC aperture images and to apply it to PSQA. To focus on the similarity and specificity of different VMAT plans, a regression and ranking loss function is proposed to optimize the training process. The weights of this loss function are dynamic during training. A clinical workflow is designed to combine the proposed model with measurement-based PSQA. It can improve the VMAT PSQA process and reduce PSQA workloads. In addition, different institutions can adjust the threshold to meet the needs for tolerance and action limits in PSQA. The experimental results demonstrate the effectiveness of the proposed method. In future work, we will collect data from multiple institutions to validate the generality of the proposed model. Meanwhile, we will try to combine MLP models with some powerful models, such as vision transformers, to improve the performance of GPR prediction.

CRediT authorship contribution statement

Wenjie Liu: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Lei Zhang:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Lizhang Xie:** Methodology, Validation, Visualization. **Ting Hu:** Methodology, Validation, Visualization. **Guangjun Li:** Data annotations, Data curation, Formal analysis, Conceptualization. **Sen Bai:** Data annotations, Data curation, Formal analysis, Conceptualization. **Zhang Yi:** Supervision, Resources, Funding acquisition.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Natural Science Fund for Distinguished Young Scholar under Grant No. 62025601, the General Program of National Natural Science Foundation of China under Grants No. 61772353, the National Major Science and Technology Projects of China under Grant 2018AAA0100201, and the Sichuan University Innovation Spark Project Library under Grants No. 2018SCUH0040.

References

[1] K. Otto, Volumetric modulated arc therapy: IMRT in a single gantry arc, Med. Phys. 35 (1) (2008) 310–317.
[2] M. Popescu, Carmen C., M. Olivotto, P. Beckham, P. Ansbacher, P. Zavgorodni, F. Shaffer, M. Wai, P. Otto, Volumetric modulated arc therapy improves dosimetry and reduces treatment time compared to conventional intensity-modulated radiotherapy for locoregional radiotherapy of left-sided breast cancer and internal mammary nodes, Int. J. Radiat. Oncol. Biol. Phys. 76 (1) (2010) 287–295.

[3] L.S. Fog, J.F.B. Rasmussen, M.C. Aznar, F. Kjaerkristoffersen, I.R. Vogelius, S.A. Engelholm, J.P. Bangsgaard, A closer look at RapidArc® radiosurgery plans using very small fields, Phys. Med. Biol. 56 (6) (2011) 1853–1863.
[4] C.L. Ong, J.P. Cuijpers, S. Senan, B.J. Slotman, W.F.A.R. Verbakel, Impact of the calculation resolution of AAA for small fields and RapidArc treatment plans, Med. Phys. 38 (8) (2011) 4471–4479.
[5] D. Low, W.B. Harms, S. Mutic, J.A. Purdy, A technique for the quantitative evaluation of dose distributions, Med. Phys. 25 (5) (1998) 656–661.
[6] B.E. Nelms, H. Zhen, W.A. Tomé, Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors, Med. Phys. 38 (2) (2011) 1037–1044.
[7] G. Yan, C. Liu, T. Simon, L. Peng, C. Fox, J.G. Li, On the sensitivity of patient-specific IMRT QA to MLC positioning errors, J. Appl. Clin. Med. Phys. 10 (1) (2009) 120–128.
[8] E.C. Ford, S. Terezakis, A. Souranis, K. Harris, H. Gay, S. Mutic, Quality control quantification (QCQ): a tool to measure the value of quality control checks in radiation oncology, Int. J. Radiat. Oncol. Biol. Phys. 84 (3) (2012) e263–e269.
[9] N. Hodapp, The ICRU Report 83: prescribing, recording and reporting photon-beam intensity-modulated radiation therapy (IMRT), Strahlentherapie Und Onkologie 188 (1) (2012) 97–99.
[10] A. Van Esch, J. Bohsung, P. Sorvari, M. Tenhunen, M. Paiusco, M. Iori, P. Engstrom, H. Nystrom, D. Huyskens, Acceptance tests and quality control (QC) procedures for the clinical implementation of intensity modulated radiotherapy (IMRT) using inverse planning and the sliding window technique: experience from five radiotherapy departments, Radiother. Oncol. 65 (1) (2002) 53–70.
[11] G. Valdes, R. Scheuermann, C.Y. Hung, A. Olszanski, M. Bellerive, T. Solberg, A mathematical framework for virtual IMRT QA using machine learning, Med. Phys. 43 (7) (2016) 4323–4334.
[12] G. Valdes, M. Chan, S. Lim, R. Scheuermann, J.O. Deasy, T. Solberg, IMRT QA using machine learning: A multi–institutional validation, J. Appl. Clin. Med. Phys. 18 (5) (2017) 279–284.
[13] Y. Interian, V. Rideout, V. Kearney, E.D. Gennatas, O. Morin, J. Cheung, T. Solberg, G. Valdes, Deep nets vs expert designed features in medical physics: An IMRT QA case study, Med. Phys. 45 (6) (2018) 2672–2680.
[14] D.A. Granville, J. Sutherland, J. Belec, D.J. La 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 (9) (2019) 095017.
[15] S. Tomori, N. Kadoya, Y. Takayama, T. Kajikawa, K. Shima, K. Narazaki, K. Jingu, A deep learning–based prediction model for gamma evaluation in patient-specific quality assurance, Med. Phys. 45 (9) (2018) 4055–4065.
[16] T. Ono, H. Hirashima, H. Iramina, N. Mukumoto, Y. Miyabe, M. Nakamura, T. Mizowaki, Prediction of dosimetric accuracy for VMAT plans using plan complexity parameters via machine learning, Med. Phys. 46 (9) (2019) 3823–3832.
[17] W. Sun, M. Jiang, L. Ren, J. Dang, T. You, F. Yin, Respiratory signal prediction based on adaptive boosting and multi-layer perceptron neural network, Phys. Med. Biol. 62 (17) (2017) 6822.
[18] T.P. Teo, S.B. Ahmed, P. Kawalec, N. Alayoubi, N. Bruce, E. Lyn, S. Pistorius, Feasibility of predicting tumor motion using online data acquired during treatment and a generalized neural network optimized with offline patient tumor trajectories, Med. Phys. 45 (2) (2018) 830–845.
[19] C. Zhu, S.H. Lin, X. Jiang, Y. Xiang, Z. Belal, G. Jun, R. Mohan, A novel deep learning model using dosimetric and clinical information for grade 4 radiotherapy-induced lymphopenia prediction, Phys. Med. Biol. 65 (3) (2020) 035014.
[20] M. Nyflot, P. Thammasorn, L.S. Wootton, E.C. Ford, 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) (2018) 456–464.
[21] S. Cui, Y. Luo, H.-H. Tseng, R.K. Ten Haken, I. El Naqa, Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage, Med. Phys. 46 (5) (2019) 2497–2511.
[22] C. Ma, R. Wang, S. Zhou, M. Wang, H. Yue, Y. Zhang, H. Wu, The structural similarity index for IMRT quality assurance: radiomics-based error classification, Med. Phys. 48 (1) (2021) 80–93.
[23] I.O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al., Mlp-mixer: An all-mlp architecture for vision, Adv. Neural Inf. Process. Syst. 34 (2021) 24261–24272.
[24] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, G. Izacard, A. Joulin, G. Synnaeve, J. Verbeek, et al., Resmlp: Feedforward networks for image classification with data-efficient training, IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[25] G.A. Ezzell, J.M. Galvin, D. Low, J.R. Palta, I.I. Rosen, M.B. Sharpe, P. Xia, Y. Xiao, L. Xing, C. Yu, Guidance document on delivery, treatment planning, and clinical implementation of IMRT : Report of the IMRT subcommittee of the AAPM radiation therapy committee, Med. Phys. 30 (8) (2003) 2089–2115.
[26] J.V. Dyk, R.B. Barnett, J.E. Cygler, P.C. Shragge, Commissioning and quality assurance of treatment planning computers, Int. J. Radiat. Oncol. Biol. Phys. 26 (2) (1993) 261–273.
[27] J. Li, L. Wang, X. Zhang, L. Liu, J. Li, M. Chan, J. Sui, R. Yang, Machine learning for patient-specific quality assurance of VMAT: Prediction and classification accuracy, Int. J. Radiat. Oncol. Biol. Phys. 105 (4) (2019) 893–902.
[28] D. Lam, X. Zhang, H. Li, Y. Deshan, B. Schott, T. Zhao, W. Zhang, S. Mutic, B. Sun, Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning, Med. Phys. 46 (10) (2019) 4666–4675.
[29] L. Zhang, Z. Yi, J. Yu, Multiperiodicity and attractivity of delayed recurrent neural networks with unsaturating piecewise linear transfer functions, IEEE Trans. Neural Netw. 19 (1) (2008) 158–167.
[30] L. Zhang, Z. Yi, S.L. Zhang, P. Heng, Activity invariant sets and exponentially stable attractors of linear threshold discrete-time recurrent neural networks, IEEE Trans. Automat. Control 54 (6) (2009) 1341–1347.
[31] L. Zhang, Z. Yi, Selectable and unselectable sets of neurons in recurrent neural networks with saturated piecewise linear transfer function, IEEE Trans. Neural Netw. 22 (7) (2011) 1021–1031.
[32] L. Zhang, Z. Yi, S. Amari, Theoretical study of oscillator neurons in recurrent neural networks, IEEE Trans. Neural Netw. 29 (11) (2018) 5242–5248.
[33] Y. Cao, Y. Cao, S. Wen, T. Huang, Z. Zeng, Passivity analysis of delayed reaction–diffusion memristor-based neural networks, Neural Netw. 109 (2019) 159–167.
[34] Y. Cao, N. Liu, C. Zhang, T. Zhang, Z.-F. Luo, Synchronization of multiple reaction–diffusion memristive neural networks with known or unknown parameters and switching topologies, Knowl.-Based Syst. 254 (2022) 109595.
[35] S. Wen, H. Wei, Y. Yang, Z. Guo, Z. Zeng, T. Huang, Y. Chen, Memristive LSTM network for sentiment analysis, IEEE Trans. Syst. Man Cybern. Syst. 51 (3) (2019) 1794–1804.
[36] X. Shu, L. Zhang, Z. Wang, Q. Lv, Z. Yi, Deep neural networks with region-based pooling structures for mammographic image classification, IEEE Trans. Med. Imaging 39 (6) (2020) 2246–2255.
[37] W. Liu, X. Shu, L. Zhang, D. Li, Q. Lv, Deep multiscale multi-instance Networks With Regional scoring for mammogram classification, IEEE Trans. Artif. Intell. 3 (3) (2021) 485–496.
[38] Y. Feng, L. Zhang, J. Mo, Deep manifold preserving autoencoder for classifying breast cancer histopathological images, IEEE/ACM Trans. Comput. Biol. Bioinform. 17 (1) (2020) 91–101.
[39] Y. Wang, Z. Wang, Y. Feng, L. Zhang, WDCCNet: Weighted double-classifier constraint neural network for mammographic image classification, IEEE Trans. Med. Imaging 41 (3) (2021) 559–570.
[40] L. Wang, L. Zhang, M. Zhu, X. Qi, Z. Yi, Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks, Med. Image Anal. 61 (2020) 101665.
[41] W. Liu, L. Zhang, G. Dai, X. Zhang, G. Li, Z. Yi, Deep neural network with structural similarity difference and orientation-based loss for position error classification in the radiotherapy of graves' ophthalmopathy patients, IEEE J. Biomed. Health Inf. 26 (6) (2021) 2606–2614.
[42] R. Gu, L. Wang, L. Zhang, DE-net: a deep edge network with boundary information for automatic skin lesion segmentation, Neurocomputing 468 (2022) 71–84.
[43] J. Mo, L. Zhang, Y. Wang, H. Huang, Iterative 3D feature enhancement network for pancreas segmentation from CT images, Neural Comput. Appl. 32 (2020) 12535–12546.
[44] Y. Yuan, L. Zhang, L. Wang, H. Huang, Multi-level attention network for retinal vessel segmentation, IEEE J. Biomed. Health Inf. 26 (1) (2021) 312–323.
[45] B. Sahiner, A. Pezeshk, L.M. Hadjiiski, X. Wang, K. Drukker, K.H. Cha, R.M. Summers, M.L. Giger, Deep learning in medical imaging and radiation therapy, Med. Phys. 46 (1) (2019) e1–e36.
[46] C.J. Wolfs, R.A. Canters, F. Verhaegen, Identification of treatment error types for lung cancer patients using convolutional neural networks and EPID dosimetry, Radiother. Oncol. 153 (2020) 243–249.
[47] J.L. Bedford, I.M. Hanson, A recurrent neural network for rapid detection of delivery errors during real-time portal dosimetry, Phys. Imag. Radiat. Oncol. 22 (2022) 36–43.
[48] S.R. Mahdavi, A. Tavakol, M. Sanei, S.H. Molana, F. Arbabi, A. Rostami, S. Barimani, Use of artificial neural network for pretreatment verification of intensity modulation radiation therapy fields, Br. J. Radiol. 92 (1102) (2019) 20190355.
[49] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, 28, 2015,
[50] K. Cho, B. Van Merrinboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014, pp. 1724–1734.

[51] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), 2016, arXiv preprint arXiv:1606.08415.
[52] P.J. Huber, Robust estimation of a location parameter, Ann. Math. Stat. 35 (1) (1964) 492–518.
[53] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, H. Li, Learning to rank: from pairwise approach to listwise approach, in: Proceedings of the 24th International Conference on Machine Learning, Vol. 227, ICML '07, ACM, 2007, pp. 129–136.
[54] M. Wu, Y. Chang, Z. Zheng, H. Zha, Smoothing DCG for learning to rank: a novel approach using smoothed hinge functions, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, ACM, 2009, pp. 1923–1926.
[55] M. Miften, A. Olch, D. Mihailidis, J. Moran, T. Pawlicki, A. Molineu, H. Li, K. Wijesooriya, J. Shi, P. Xia, et al., Tolerance limits and methodologies for IMRT measurement-based verification QA: recommendations of AAPM Task Group No. 218, Med. Phys. 45 (4) (2018) e53–e83.
[56] R. Mohan, M.R. Arnfield, S. Tong, Q. Wu, J. Siebers, The impact of fluctuations in intensity patterns on the number of monitor units and the quality and accuracy of intensity modulated radiotherapy, Med. Phys. 27 (6) (2000) 1226–1237.
[57] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, Vol. 2, 2012, pp. 1097–1105.
[58] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[59] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[60] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.