DGUQA: Domain Generalization Uncertainty Informed Patient-Specific Quality Assurance

Xiaoyang Zeng[1[0009-0001-9928-8782]], Awais Ahmed[1[0000-0003-4410-2028]]

Rui Xi[1*[0000-0002-6400-9106]] and Mengshu Hou[1,2[0000-0002-5283-7318]]

[1] University of Electronic Science and Technology of China, Chengdu 611731, China
[2] Chengdu Technological University, China 611730, China
{202011081605, 202014080105}@std.uestc.edu.cn, ruix.ryan@gmail.com* and mshou@uestc.edu.cn

Abstract. Deep Learning Automated Patient-Specific Quality Assurance (PSQA) endeavors to diminish the reliance on clinical resources. The accurate estimation of the dose difference metric, particularly the Gamma passing rate, is paramount in ensuring the safety and efficacy of radiation therapy plans. Although current research has yielded an overall performance on par with that of experts, it fails to address the model's local performance discrepancies across diverse lesions, thereby highlighting a generalization challenge that undermines its credibility in real clinical settings. This paper introduces **D**omain **G**eneralization **U**ncertainty Informed Patient-Specific **Q**uality **A**ssurance, abbreviated as DGUQA, based on the theory of domain generalization in deep learning. DGUQA employs an adversarial loss-based regularization to address the issue of generalization. Further, since the model is biased with the most common lesion organs, relying solely on a domain-generalized model would decrease overall performance. Therefore, in conjunction with safety requirements, we also model predictive uncertainty. The domain generalization model is used only when the uncertainty exceeds a certain threshold; otherwise, a standard model is employed. Experiments demonstrate that DGUQA shows superiority in both generalization performance and overall effectiveness. DGUQA notably enhances the deep learning trustworthiness in the PSQA and has meaningful implications for the clinical significance of medical deep learning.

Keywords: PSQA, Quality Assurance, Domain Generalization, Deep Learning, Uncertainty.

1 Introduction

Radiotherapy is a pivotal modality in the comprehensive treatment of cancer, encompassing the intricate process of plan design and the essential aspect of plan verification. Within radiotherapy, Intensity-Modulated Radiation Therapy (IMRT) emerges as a prevalent plan type, characterized by a dose value distribution matrix representing the spatial dispersion of radiation. Owing to constraints imposed by machine capabilities and the inherent complexity of radiotherapy blueprints, the strict execution of designed radiotherapy plans faces practical challenges. Consequently, the imperative of quality assurance during the formal implementation of radiotherapy assumes paramount significance in safeguarding patient well-being and safety [3]. Patient-Specific Quality Assurance (PSQA) constitutes a critical process to validate radiotherapy plans before administration, thereby mitigating potential risks. Among the array of methods employed in PSQA, Gamma analysis stands out as a widely utilized measurement-based quality assurance technique [9]. This method scrutinizes the correlation between the actual administered dose and the intended dose, serving as a yardstick to ascertain plan acceptability. The Gamma Passing Rate (GPR) metric, commonly employed in clinical practice, is a quantitative indicator of plan adherence. However, the conventional gamma analysis methodology, requiring physical execution, engenders a laborious and resource-intensive endeavor [1, 12]. In response to these challenges, artificial intelligence has been integrated into PSQA frameworks to enable the prediction of GPR through computational models, thereby streamlining the assessment process [24].

Existing Artificial intelligence (AI)-based PSQA methods can be divided into two categories. The first category revolves around machine learning, where intricate metrics are meticulously crafted by hand and subsequently integrated as sequential features into models such as SVM [15], XGboost [16], the Random Forest (RF) model, the Elastic Net [5], and the Poisson Lasso (PL) model, and the Gradient Boosting Decision tree (GBDT) model [22]. On the other hand, the second category delves into deep learning approaches, treating the dose matrices of formulated plans akin to images and feeding them into neural networks to derive the ultimate GPR. Examples include Resnet-based Unet++ [4]. Furthermore, beyond conventional Convolutional Neural Network (CNN) architectures, the study by [19] introduces a transformer-based methodology called TranQA for predicting GPR. In addition to feedforward neural network strategies, the research utilizes CycleGAN [23] to model the PSQA task. AI models have demonstrated better performance levels in PSQA than expert practitioners.

Nevertheless, the current impediment to the substantial integration of AI within PSQA clinical practice does not primarily stem from overall performance issues. Instead, the crux of the challenge lies in the reliance of AI methodologies on Independent and Identically Distributed Hypotheses, commonly referred to as I.I.D., which is susceptible to significant performance degradation in specific scenarios, particularly within domains characterized by sparse data. This susceptibility poses a tangible threat to patient safety. Illustrated in Fig. 1 is an initial experiment showcasing that the model's predictive accuracy experiences a notable decline in Mean Absolute Error (MAE) concerning plans pertaining to the Pelvis, Head, and Neck regions, which exhibit lower data representation. In contrast, the model's predictions exhibit closer alignment with the patterns observed in the Chest and Abdomen regions, culminating in enhanced performance levels.

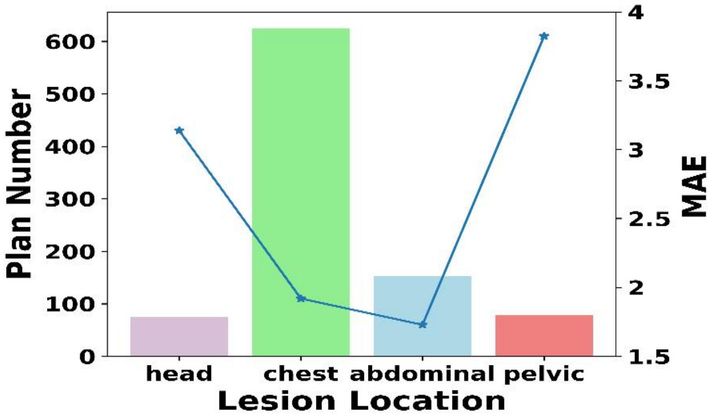

Fig. 1. MAE loss and Number in different lesion location

Experimental results demonstrate that the morphology and dimensions of the target volume, the organs' susceptibility to radiation, and the presence of surrounding organs at risk (OARs) exhibit variability contingent upon the specific organ. Consequently, variations in the probabilistic dose distribution of treatment plans for each lesion organ may arise, with treatment strategies for organs with limited representation potentially faltering under the Independent and Identically Distributed (I.I.D.) assumption. Notably, the study by [2] pioneered the concept of domain adaptation (DG) within deep learning, primarily addressing the non-I.I.D. predicament.

Specifically, DG aims to learn a model using data from a single or multiple related but distinct source domains in a way that allows the model to generalize well to any shifted target domain [21,13]. Domain generalization (DG) primarily encompasses two widely utilized categories. The first category involves data augmentation techniques, such as interpolation within the same domain across different categories or domains within the same category [18]. This approach aims to minimize the model's fit to domain biases and to learn domain-invariant features as much as possible. The second category leverages regularization terms for domain alignment, where the types of regularization can include measures such as the Maximum Mean Discrepancy (MMD) distance [8] or adversarial loss [14]. This category method's typical implementation typically applies auxiliary loss. This latter method is the approach adopted in this paper. Besides, ensemble learning [17] and meta-learning [7] are also employed for DG.

This research introduces DGUQA, a system for solving domain shifts and improving patient-specific medical picture analysis across various lesion organs. Particularly, the key contributions are the following:

- DGUQA leverages adversarial domain generalization, significantly enhancing the generalizability and reliability of deep learning-based patient-specific quality assurance.

- While maintaining generalizability, we have also further improved the overall performance by uncertainty-based filtering.
- Experiments prove our superiority in both overall and generalized performance in PSQA.

2 Dataset Collection

In this study, we employed a dataset in collaboration with the Department of "Radiation Oncology" at "Peeking Union Medical College Hospital"[1] that encompasses 154 FF-IMRT treatment plans (containing 932 beam fields) for various treatment sites that were collected retrospectively from December 2020 to July 2021. The dataset includes (19h & n short head and neck) plans, while (82c, chest) plans, (31a, abdominal) plans, and (22p, pelvic) plans resulting in a total of 154 treatment plans. All plans used the sliding window technique and were generated based on Eclipse TPS version 15.6 and delivered by Halcyon 2.0 linac equipped with SX2 dual-layer MLC (Varian Medical System, Palo Alto, CA). The dose distribution was calculated using the anisotropic analytic algorithm (AAA, version. 15.6.06, Varian Medical Systems, Palo Alto, CA) with a dose calculation grid of 2.5 mm, and the plan optimization algorithm was photon optimization (PO, version. 15.6.06, Varian Medical Systems, Palo Alto, CA) algorithm.

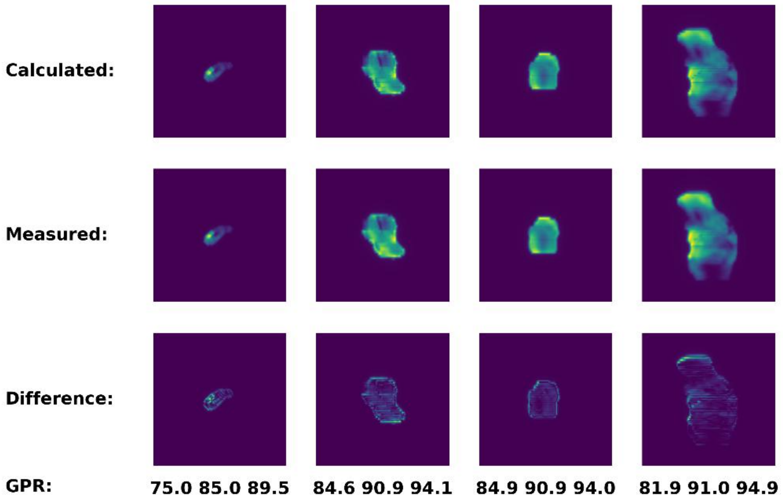

Fig. 2. Some Samples Presentation, the designed plan is the first row), real measurement is the second row, dose difference is the third row, and GPR based on dose difference calculated result is in the fourth row with the order of $[1\%/1mm, 2\%2mm, and\ 2\%3mm]$

[1] https://www.pumch.cn/en/introduction.html

Following the recommendations in the TG-218 report [11], PSQA measurements were conducted before treatment delivery using actual angles for each beam, employing Portal Dosimetry. Dose calibration was performed every day before data collection. All gamma analyses were performed with 1%/1mm, 2%/2mm, and 2%/3mm criteria at a 10% threshold of the maximum dose (only points with doses greater than 10% of the global maximum dose per beam were analyzed). The gamma analyses were performed in absolute dose mode, and global normalization was applied to the results. The Treatment Planning System (TPS) calculated fluence maps, exported them in DICOM format, and utilized them as input for the Deep Learning network. This robust dataset and rigorous PSQA measurements form the foundation for the proposed Virtual Dose Verification.

The Raw fluence maps of 932 beam fields exhibited varied spatial resolutions and sizes, necessitating several pre-processing steps. Firstly, 2D fluence maps were resampled to the consistent spatial resolution of $1mm \times 1mm$ and then cropped to 224×224 pixels to remove the redundant background. Finally, before being fed into the network, the pixel values of the input images were re-scaled to $[0,1]$ by Min-Max normalization, ensuring uniformity in the input data. Fig. 2 visually underscores the need for meticulous data balancing due to evident imbalances in the gamma passing rate values individually. Generally, GPR values above 90 are considered passing, while those below 90 must be re-conducted. However, the current study does not mitigate the existing imbalance as the primary objective is to study multi-granularity prior networks to propose a robust and efficient PSQA framework.

Table 1. Mathematical Symbol Table

Symbol	Description
X	Input dose plan matrix with the shape of [224,224]
Y	The GPR target with the dimension of 3, since there exist 3 common GPR criteria, namely $[1\%/1mm, 2\%2mm, and 2\%3mm]$
D	Selected domain index, namely lesion
Ep	The expectation with the integration of P
$N(y \mid \mu, \sigma)$	The Normal distribution, parameterized by its mean μ and variance σ , Our implementation apply trivariate independent Gaussian distribution
k	The inverse of σ
Log P	The log-likelihood of distribution P
γ	The coefficient of adversarial loss
β	Whether we use uncertain loss, in the ungeneralized model, it would be set to be 0, otherwise 1
f	The Backbone Network Function, the output is the feature

To further influence the PSQA values, this study presented a Fig. 2 that summarizes PSQA values. The figure is used to present GPR values from the designed plan (the first row), real measurement (the second row), dose difference (the third row), and GPR (the fourth row).

3 Methodology

First, we define some mathematical symbols in the Table 1. Then, we present our initial general GPR regression loss, namely MSELoss.

$$L_{gpr} = E_D E_{(Y,X)|D} (Y - Regress\_head(f(X)))^2$$

As we have demonstrated in the introduction, the performance of deep learning models significantly decreases when faced with scarce lesions (domains) in the dataset, proving the limitation of the current work's generalization ability. To address this issue, we hope the network can obtain domain-independent features, more specifically, to enable the model to acquire effective knowledge across all lesions (domains). We have adopted an approach that utilizes adversarial loss-based regularization to achieve domain generalization.

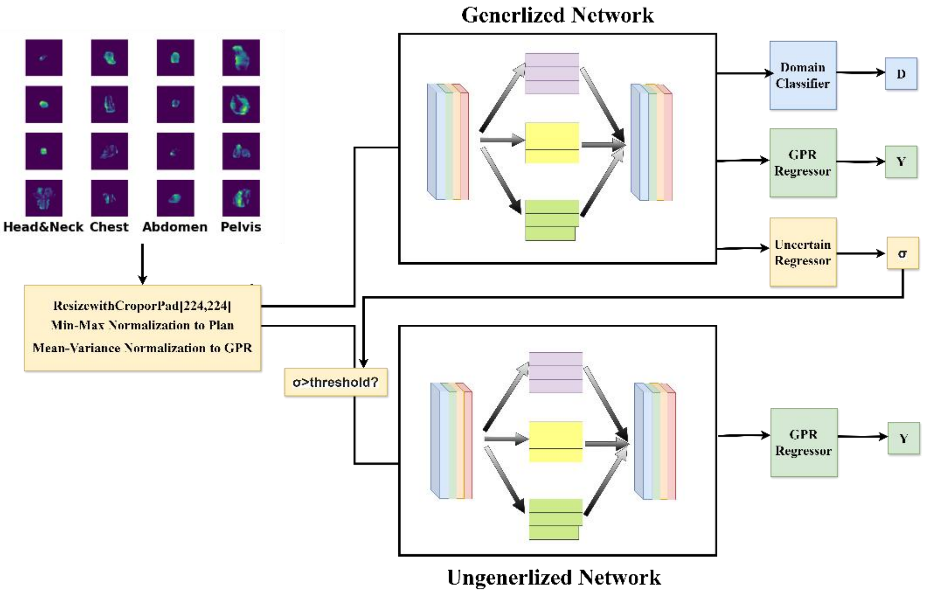

Fig. 3. The DGUQA Framework Overview, with the Backbone of MobileVit

$$L_{adv} = E_D E_{X|D} (D \cdot \log Dis(f(X))$$

Then the total loss would be as follows:

$$L_{total} = L_{gpr} + \gamma \cdot maxL_{adv}$$

Where Dis is the domain classifier head.

Although domain-generalization models exhibit superior generalization performance on patterns unfamiliar to the model, our experiments indicate that this may diminish the effectiveness of familiar patterns. Thus, we introduce uncertainty modeling to assess how much the model grasps the data patterns. Only when the model identifies encountering unfamiliar patterns do we employ domain-generalization models. We utilize models processed through domain generalization for patterns deemed familiar by the model.

The network output would be modeled as a normal distribution parameter [6], with the original GPR regress serving as the mean and the additional output head as the variance, namely the uncertain head.

Uncertainty quantification integration with the initial GPR regression loss would be:

$$\begin{aligned} L_{gpr}^{*} &= -ElogN(y \mid f(x), \sigma) \\ &= E_D E_{(Y,X)|D} \frac{(Y - f(X))^2}{\sigma^2} + log\,\sigma \\ &\overset{k=\frac{1}{\sigma^2}}{=} E_D E_{(Y,X)|D} (Y - f(X))^2 * k - 1/2logk \end{aligned}$$

To ensure the mathematical and implementation simplicity, we implement k (the inverse of σ) rather than the original σ . In an implementation, the threshold is set to be 0.9 quantiles of the calibration dataset uncertainty score σ .

Since k has to be greater than 0, we used a modified Softplus as the activation of k, named as $Softplus^{*}$:

$$Softplus^{*}(x) = log(e + \beta * e^{x})$$

We modified the original softplus to maintain the mathematical consistency to loss without uncertainty for the ungeneralized model loss. We just need to set the β to 0 in ungeneralized model training, and the uncertain output would remain at 1.

$$k = Softplus^{*}(unc\_head(f(x)))$$

Therefore, the final total loss of the generalized model is:

$$L_{total}^{*} = L_{gpr}^{*} + \gamma \cdot maxL_{adv}$$

And the final loss of the ungeneralized model is L_{gpr} .

Table 2. Parameter Settings

Parameter	Description	Value
Resized Dose Plan Array Sample Shape	To organize these dose plan array samples into a batch, we preprocess the ResizeWithPadOrCrop in the Monai package. And here is the resized shape.	[224,224]

Backbone Architecture	The pretrained backbone network is used as the feature extractor.	MobileVIT(Imagenet)
		ResNet18(Imagenet)

Regression head Architecture	Architecture of GPR prediction output, the output dimension is 3 since there we select three GPR criteria $[1\%/1mm, 2\%2mm, and 2\%3mm]$	• AdaptiveAvgPool 2d[1,1] • Flatten() • Dropout(p=0.1) • Linear[512,3]
Uncertain Head Architecture	Architecture of Uncertainty prediction head output, the output dimension	Same as Regress Head but with the additional *Softplus* activation
Domain Classifier Head Architecture	Architecture of domain classifier head, also denoted as Dis	the same as regression head, but the output dimension is lesion number, namely 4 in our implementation
Dropout Rate	The dropout rate implemented in network	0.1
Optimizer	Type of optimizer used	Adam
Adv_Optimizer	Type of optimizer for adversarial Domain Classifier used	SGD
Epochs	Number of training epochs	20
Batch Size	Size of the training batches	32
Learning Rate	Learning rate for the optimizer	1e-3
Dataset Split Ratio	Experiment Method dataset split ratio, namely Training Dataset Size: Validation Dataset Size: Calibration Dataset size: Test Dataset Size, Calibration Dataset is used to determine the uncertain score threshold. For the method without UQ, the calibration dataset is not used	7:1:1:1

To achieve the highest possible performance, we have adopted the advanced MobileViT [10] as the backbone. We also implemented the ResNet backbone in the experiment to make a fair comparison. Unlike the common Vit, MobileViT integrates

CNN and VIT to ensure the complexity and capture the implicit image pixel relationship. The attention function in MobileVit is:

$$Attention = Softmax(Query \cdot Key)$$

$$\text{Query} = \text{Seq}(\text{Conv}(W_q, X))$$

$$\text{Key} = \text{Seq}(\text{Conv}(W_k, X))$$

$$\text{Value} = \text{Seq}(\text{Conv}(W_v, X))$$

Seq means the operation to fold the convolution feature map into a sequence with the length of patches N_{sp} .

And the final output of MobileVit is:

$$X_{patch} = \text{Img}(Attention \cdot Value)$$

To ensure Reproducibility, Table 2 presents the specific architecture and hyperparameter setting.

4 Experiment and Discussion

4.1 Baseline Methods and Metric for PSQA

This section introduces the baseline PSQA method compared with the proposed DGUQA method. Random Forest (RF) Poisson Lasso (PL) and Gradient Boosted Decision Trees (GBDT) models were selected as traditional machine learning baseline methods. These techniques use manually designed complexity metrics as features to predict the gamma passing rate (GPR). The implementation details of these machine-learning-based methods are thoroughly described in [22]. For deep learning-based baseline methods, the study in [4] utilized a ResNet-based UNet++ architecture to predict GPR. Besides the conventional CNN architecture, the research introduced in [20] proposed a transformer-based method, TranQA, for predicting GPR. As part of the feedforward neural network methods [19] employed CycleGAN [23] to model the dose difference prediction task and then obtain GPR, hereafter referred to as CycleGAN. Besides the previous methods, the DGQA denotes the method of applying domain generalization without the uncertainty threshold filter.

To illustrate generalization, we predefine a new metric called Generalized Mean Absolute Error (abbreviated as GMAE), defined as follows:

$$GMAE = \max_{D} |Y - f(X)|$$

Which is the maximum MAE of the domains (lesion organs). The original MAE is:

$$MAE = \underset{D}{\mathrm{E}}|Y - f(X)|$$

4.2 Exp-A: Comparison with Baseline Methods

In EXP-A, we analyze DGUQA PSQA performance using traditional and benchmark works.

Table 3 lists the Mean Absolute Error (MAE) comparative results for different criteria. The``Criteria" column has three criteria, such as ``1%/1 mm", ``2%/2 mm", and ``2%/3 mm". These parameters indicate varying tolerance levels for differences in dose, measured in terms of % percentage. 1%/1 mm is a strict standard among them. The performance of PSQA models in predicting GPR is evaluated across different lesions (H & N, C, A, P) and criteria (1%/1 mm, 2%/2 mm, 2%/3 mm), as illustrated in Table 3.

The MAE values are given for each combination of criteria and Lesion Organ. The table also categorizes the methods into several main categories. The machine learning category records three traditional methods: GBDT, RF, and PL. Deep learning records three further methods, including UNET++, TransQA, and CycleGAN. Generalized deep learning methods are presented, including proposed DGUQA and DGQA and the combination with the backbone.

Based on the observation, Head & Neck radiation therapy plans are the scarcest. Consequently, the generalized mean absolute error (GMAE) is predominantly determined by the mean absolute error (MAE) of Head & Neck plans. Most plans experience significant degradation in performance in Head & Neck radiation therapy if not utilizing deep learning methods. Specifically, employing DG (Deep Generative) methods substantially enhances predictive performance in head and neck radiation therapy, resulting in only marginal differences compared to other regions. However, without uncertainty filtering, overall performance experiences a significant decline. Only when both methods are utilized can improvements in both generalization and overall performance be achieved. In most metrics, adopting DGUQA (Deep Generative Uncertainty Quantification and Assessment) with MobileVit achieves optimal results, whereas for Head & Neck GPR (Generalized Perturbation-based Radiotherapy) MAE, i.e., GMAE, the version employing ResNet in DGUQA demonstrates superior generalization performance.

For overall MAE, Comparing the results with these categories of broader methods from machine learning, GBDT performs better than the other two. However, extreme (<85) values show potential for vulnerabilities. In contrast, TransQA performed better for deep learning than the other two, and with extreme (<85) values, it performs worse than machine learning. Lastly, the proposed generalization deep learning methods show stable performance across all methods with some variations among each.

Table 3 Mean Absolute Error (MAE) Comparison in Different Leisons. Here, H & N, C, A, and P represent ``Head and Neck", ``Chest," ``Abdomen," and ``Pelvis," respectively.

			GPR MAE (%) in Different Lesions				
Methods Category	Method	Criteria	All	H&N	C	A	P
ML	GBDT	1%/1 mm	2.207	3.673	2.015	1.857	2.852
		2%/2 mm	1.652	2.749	1.467	1.516	2.222
		2%/3 mm	1.056	1.852	0.93	1.04	1.274
	RF	1%/1 mm	2.566	3.342	2.502	2.591	2.289
		2%/2 mm	1.717	2.184	1.624	1.683	2.027
		2%/3 mm	1.188	1.574	1.109	1.128	1.495
	PL	1%/1 mm	2.464	3.27	2.536	2.07	1.899
		2%/2 mm	1.652	2.434	1.528	1.425	2.203
		2%/3 mm	1.059	1.549	1.019	0.886	1.195
DL	UNET++	1%/1 mm	2.392	3.729	2.441	1.721	1.979
		2%/2 mm	1.48	2.804	1.365	1.226	1.54
		2%/3 mm	1.126	1.772	1.061	1.041	1.155
	TransQA	1%/1 mm	2.2	3.42	2.063	1.994	2.439
		2%/2 mm	1.411	2.159	1.284	1.279	1.869
		2%/3 mm	1.076	1.615	1.001	1.065	1.147
	CycleGAN	1%/1 mm	2.142	2.797	2.25	1.58	1.724
		2%/2 mm	1.423	1.862	1.448	0.976	1.633
		2%/3 mm	0.976	1.728	0.813	0.982	1.417
GDL	DGQA-ResNet	1%/1 mm	2.701	2.87	2.833	2.659	1.676
		2%/2 mm	1.639	2.315	1.582	1.235	2.11
		2%/3 mm	1.348	1.63	1.382	1.031	1.384
	DGQA-MobileVit	1%/1 mm	2.548	2.914	2.513	1.66	3.068
		2%/2 mm	1.572	2.554	1.458	1.159	2.192
		2%/3 mm	1.085	1.77	0.997	0.861	1.477
	DGUQA-ResNet	1%/1 mm	2.209	2.423	2.366	1.179	2.617
		2%/2 mm	1.505	1.851	1.416	0.875	2.878
		2%/3 mm	1.04	1.335	0.961	0.672	1.95
	DGUQA-MobileVit	1%/1 mm	1.966	2.401	1.927	1.608	1.744
		2%/2 mm	1.371	2.138	1.24	1.21	1.891
		2%/3 mm	0.939	1.236	0.922	0.815	0.967

The improved performance of DGUQA models suggests that incorporating generalized deep learning architectures and attention mechanisms can significantly enhance the predictive accuracy for PSQA tasks. The adaptability of DGUQA-MobileVit across different criteria and lesions further emphasizes the potential of mobile vision transformers in medical imaging tasks, offering a balance between computational efficiency and predictive performance.

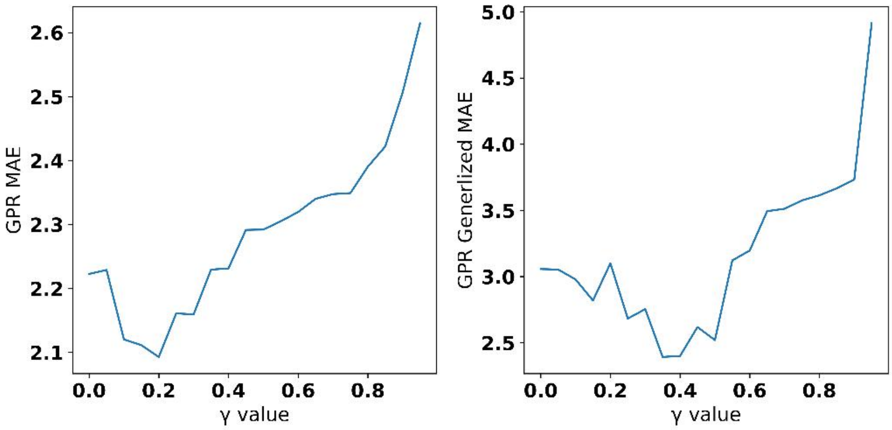

Fig. 4. The MAE and generalized MAE vary with γ

4.3 Exp-B: Ablation Study

Here, we visualize the relationship between the regularization coefficient γ, the Mean Absolute Error (MAE), and the Generalized MAE. The experiment is implemented in the DGUQA-MobileVit setting. It can be observed that as the regularization coefficient increases, the overall performance first improves and then declines. The minimum value is attained at $\gamma = 0.2$ for MAE. Similarly, the Generalized MAE increases and then decreases, but it reaches its minimum value at $\gamma = 0.45$. The optimal points for the generalized MAE and MAE do not coincide; the optimal point for the generalized MAE tends to be relatively larger. At this juncture, it can be posited that mitigating bias knowledge can mitigate the disparities between the training and testing datasets before reaching the optimal overall MAE. This can be considered as a method to alleviate overfitting. However, before reaching the optimal generalized MAE, bias mitigation continues to enhance generalization, albeit negatively impacting overall performance. At this point, bias mitigation can be seen as correcting the extent of domain shift. However, excessive bias mitigation is detrimental to both overall and generalization performance. This is primarily because the auxiliary adversarial loss excessively depletes the model's capacity.

5 Conclusion

In this paper, we summarize current methods of Deep Learning Automated Patient-Specific Quality Assurance (PSQA) and analyze the fact that, in terms of overall performance, PSQA has achieved results comparable to those of human experts. However, the challenge lies in the local performance across different lessons, which is the challenge of generalization. Based on the theory of domain generalization in deep learning, this paper proposes DGUQA. DGUQA utilizes regularization based on adversarial loss to address the issue of generalization, which significantly decreases performance in domains with scarce data. On the other hand, to prevent the model from applying biased knowledge on common data, leading to a decrease in overall performance, we also employ uncertainty-assisted modeling. The domain generalization model is only used when uncertainty exceeds a certain threshold; otherwise, a non-generalized model is utilized. Experiments demonstrate DGUQA's superiority in both global and local performance. DGUQA significantly enhances the clinical trustworthiness of deep learning in the PSQA domain, marking a meaningful contribution to the clinical importance of medical deep learning. However, despite the improvement in trustworthiness, some limitations still exist in DGUQA. The performance of this work is unsteady with γ (the coefficient of regulation loss) since there are some other steadier methods in domain generalization, like Lisa [18].

References

1. Abolaban, F., Zaman, S., Cashmore, J., Nisbet, A., Clark, C.: Changes in patterns of intensity-modulated radiotherapy verification and quality assur ance in the uk. Clinical Oncology 28(8), e28–e34 (2016)
2. Blanchard, G., Lee, G., Scott, C.: Generalizing from several related classi f ication tasks to a new unlabeled sample. Advances in neural information processing systems 24 (2011)
3. Hodapp, N.: The icru report 83: prescribing, recording and reporting photon-beam intensity-modulated radiation therapy (imrt). Strahlenther apie und Onkologie: Organ der Deutschen Rontgengesellschaft...[et al] 188(1), 97–99 (2012)
4. Huang, Y., Pi, Y., Ma, K., Miao, X., Fu, S., Chen, H., et al.: Virtual patient specific quality assurance of imrt using unet++: classification, gamma pass ing rates prediction, and dose difference prediction. front oncol 2021; 11: 700343 (2021). https://doi.org/https://doi.org/10.3389/fonc.2021.700343
5. Ishizaka, N., Kinoshita, T., Sakai, M., Tanabe, S., Nakano, H., Tan abe, S., Nakamura, S., Mayumi, K., Akamatsu, S., Nishikata, T., et al.: Prediction of patient-specific quality assurance for volumetric modulated arc therapy using radiomics-based machine learning with dose distribu tion. Journal of Applied Clinical Medical Physics 25(1), e14215 (2024). https://doi.org/https://doi.org/10.1002/acm2.14215
6. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learn ing for computer vision? Advances in neural information processing systems 30 (2017)
7. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.: Learning to generalize: Meta learning for domain generalization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
8. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adver sarial feature learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5400–5409 (2018)
9. Low, D.A., Harms, W.B., Mutic, S., Purdy, J.A.: A technique for the quan titative evaluation of dose distributions. Medical physics 25(5), 656–661 (1998)
10. Mehta, S., Rastegari, M.: Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021)
11. Miften, M., Olch, A., Mihailidis, D.: Tg 218: Tolerance limits and methodologies for imrt measurement-based verification qa: recom mendations of aapm task group no. 218. Med. Phys 45 (2018). https://doi.org/https://doi.org/10.1002/mp.12810

12. Pan, Y., Yang, R., Zhang, S., Li, J., Dai, J., Wang, J., Cai, J.: National sur vey of patient specific imrt quality assurance in china. Radiation oncology 14, 1–10 (2019)
13. Shah, S.A., Zeng, X., Ahmed, A., Parvez, S., Xi, R., Hou, M.: Multi-instance bias suppression for enhanced generalization in breast cancer diagnosis: Harnessing histopathological big data insights. In: 2023 IEEE International Conference on Big Data (BigData). pp. 857–864. IEEE (2023)
14. Sicilia, A., Zhao, X., Hwang, S.J.: Domain adversarial neural networks for domain generalization: When it works and how to improve. Machine Learn ing 112(7), 2685–2721 (2023)
15. Wall, P.D., Fontenot, J.D.: Quality assurance-based optimization (qao): To wards improving patient-specific quality assurance in volumetric modulated arc therapy plans using machine learning. Physica Medica 87, 136–143 (2021). https://doi.org/https://doi.org/10.1016/j.ejmp.2021.03.017
16. Wall, P.D., Hirata, E., Morin, O., Valdes, G., Witztum, A.: Prospective clinical validation of virtual patient-specific quality assurance of volumetric modulated arc therapy radiation therapy plans. International Journal of Radiation Oncology* Biology* Physics 113(5), 1091–1102 (2022)
17. Wang, S., Yu, L., Li, K., Yang, X., Fu, C.W., Heng, P.A.: Dofe: Domain oriented feature embedding for generalizable fundus image segmentation on unseen datasets. IEEE Transactions on Medical Imaging 39(12), 4237–4248 (2020)
18. Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., Finn, C.: Improving out-of-distribution robustness via selective augmentation. In: International Conference on Machine Learning. pp. 25407–25437. PMLR (2022)
19. Yoganathan, S., Ahmed, S., Paloor, S., Torfeh, T., Aouadi, S., Al-Hammadi, N., Hammoud, R.: Virtual pretreatment patient specific quality assurance of volumetric modulated arc therapy using deep learning. Medical Physics 50(12), 7891–7903 (2023). https://doi.org/https://doi.org/10.1002/mp.16567
20. Zeng, L., Zhang, M., Zhang, Y., Zou, Z., Guan, Y., Huang, B., Yu, X., Ding, S., Liu, Q., Gong, C.: Transqa: deep hybrid transformer network for measurement-guided volumetric dose prediction of pre-treatment patient specific quality assurance. Physics in Medicine & Biology 68(20), 205010 (2023). https://doi.org/10.1088/1361-6560/acfa5e
21. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: Asurvey. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(4), 4396–4415 (2022)
22. Zhu, H., Zhu, Q., Wang, Z., Yang, B., Zhang, W., Qiu, J.: Patient specific quality assurance prediction models based on machine learning for novel dual-layered mlc linac. Medical Physics 50(2), 1205–1214 (2023). https://doi.org/https://doi.org/10.1002/mp.16091
23. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image trans lation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017).
24. Ahmed, A., Xi, R., Hou, M., Shah, S. A., & Hameed, S. (2023). Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts. IEEE Access.