Computers in Biology and Medicine 179 (2024) 108925

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Multi-granularity prior networks for uncertainty-informed patient-specific quality assurance

Xiaoyang Zeng [a], Qizhen Zhu [b], Awais Ahmed [a], Muhammad Hanif [a], Mengshu Hou [a,c,*], Qiu Jie [b], Rui Xi [a], Syed Attique Shah [d]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China - UESTC, Sichuan, 611731, China
[b] Department of Radiation Oncology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[c] School of Big Data and Artificial Intelligence, Chengdu Technological University, Sichuan, 611730, China
[d] School of Computing and Digital Technology, Birmingham City University, STEAMhouse, B4 7RQ, Birmingham, United Kingdom

ARTICLE INFO

Keywords:
Multi-granularity prior networks
Deep learning-based PSQA
Dose difference prediction
Gamma passing rate prediction
Dose plan verification

ABSTRACT

Deep Learning Automated Patient-Specific Quality Assurance (PSQA) aims to reduce clinical resource requirements. It is vital to ensure the safety and effectiveness of radiation therapy by predicting the dose difference metric (Gamma passing rate) and its distribution. However, current research overlooks uncertainty quantification in model predictions, limiting their trustworthiness in real clinical environments. This paper proposes a Multi-granularity Uncertainty Quantification (MGUQ) framework. A Bayesian framework that quantifies uncertainties at multiple granularities for multi-task PSQA, specifically Gamma Passing Rate (GPR) prediction and Dose Difference Prediction (DDP), integrates visualization-based interactive components. Using Bayesian theory, we derive a comprehensive multi-granularity loss function that comprises granularity-specific loss and coherence loss components. Additionally, we proposed Multi-granularity Prior Networks, a dual-stream network architecture, to infer the distributions of DDP (modeled as t-distributions) and GPR (modeled as Gaussian distributions) under specific statistical assumptions. Comprehensive evaluations are conducted on a dataset from "Peeking Union Medical College Hospital", and results show that our proposed method achieves a minimum MAE loss of 0.864 with a 2%/3 mm criterion and realizes the uncertainty visualization of dose difference. Further, it also achieves 100% Clinical Accuracy (CA) with a workload of 67.2%. Experiments demonstrate that the proposed framework can enhance the trustworthiness of deep learning applications in PSQA.

1. Introduction

Cancer is one of the leading causes of death worldwide, and radiation therapy is one of the primary methods of cancer treatment. Many studies have applied deep learning to radiotherapy [1–5]. Patient-Specific Quality Assurance (PSQA) in radiation therapy refers to the comprehensive evaluation and verification processes implemented to ensure an individual patient's treatment plan's accuracy, safety, and effectiveness. This quality assurance is crucial in techniques like Intensity-Modulated Radiation Therapy (IMRT) [6], where the radiation dose delivery is precisely modulated to conform to the tumor's shape. Common metrics include Gamma Passing Rate (GPR) [7] and Dose Difference Prediction (DDP) [8]. The GPR is the percentage of points in a dose distribution that meet predefined criteria for dose difference and spatial distance, which provides an overall assessment of the plan's reliability and accuracy. However, in clinical practice, relying solely on the GPR is insufficient to determine a treatment plan's effectiveness because it does not provide high-granularity spatial information [9,10], thereby limiting the assessment of whether the treatment dosage optimally targets the lesion while sparing healthy, yet vulnerable, organs [9,10]. In contrast, the DDP task aims to forecast the dose difference distribution matrix, concentrating on granular discrepancies in dose delivery accuracy and capturing the difference matrix between the actual and calculated doses at individual points.

Indeed, a prevalent method is applying the Gamma analysis to evaluate the accuracy of deliveries in IMRT [11,12]. Still, this approach demands physical implementation and is time-consuming and

* Corresponding author.
E-mail addresses: 202011081605@std.uestc.edu.cn (X. Zeng), zhuqizhen@pumch.cn (Q. Zhu), 202014080105@std.uestc.edu.cn (A. Ahmed), 202014080107@std.uestc.edu.cn (M. Hanif), mshou@std.uestc.edu.cn, mshou@uestc.edu.cn (M. Hou), qiujie@pumch.cn (Q. Jie), ruix.ryan@gmail.com (R. Xi), syedattique.shah@bcu.ac.uk (S.A. Shah).

https://doi.org/10.1016/j.compbiomed.2024.108925
Received 27 November 2023; Received in revised form 28 May 2024; Accepted 17 July 2024
Available online 26 July 2024

labor-intensive, leading to an additional burden on clinical treatment processes. To resolve it, several algorithms have been recently proposed to analyze the GPR for ensuring compliance with clinical dose distribution standards [13–18]. As shown in Table 1, these methods can be categorized into three main categories: manually defined complexity metrics, traditional machine learning, and deep learning methods.

The first method category manually defines complexity metrics based on plan characteristics [6,19–21], which characterize plan delivery accuracy based on the treatment plan's physical attributes. Although these complexity metrics correlate with delivery accuracy, they are indirect to dose delivery performance and lack immediate clinical significance regarding GPR [22]. The second method category is machine learning-based [23–26], which directly utilizes the machine learning models and predicts the accuracy of plan delivery from the input complexity metrics or other pertinent information like field fluence. However, this category method cannot manage the dose difference distribution prediction and then loses the spatial indicator, which is important in clinical applications. The third category is deep learning-based [9,27–30] methods, which input the designed dose plan array and then predict the difference with measured realistic dose delivery in implementation or predict GPR. However, due to their low transparency, most deep learning-based methods lack explainability, and their use in clinical systems remains limited. While planners can easily assess the results of AI segmentation tools, understanding the predictive models developed through deep learning can be much more challenging [31]. To address this concern, researchers have explored uncertainty quantification to determine when the model is unsure about a prediction [32]. The study [33] developed an uncertainty-aware dual-task deep learning model that predicts GPR and the total uncertainty of the prediction. High uncertainties above a certain threshold necessitate human intervention, fostering interaction between the physicist and the model [34], and GPR lacks sensitivity to clinically relevant dosimetric errors [35].

Furthermore, deep learning models, especially for regression tasks, lack the confidence measure. Recent studies indicate they often produce overly confident results, decreasing their practical usability and raising concerns about their security and trustworthiness in real-world PSQA applications. Uncertainty quantification (UQ) has emerged as a prominent and active area of research. Recent findings indicate that the overall uncertainty can be categorized into two distinct types: model uncertainty [36] and data uncertainty [32]. Model uncertainty, referred to as epistemic uncertainty, arises from inadequate acquisition of the underlying knowledge about the prediction task. This uncertainty can be mitigated through increased data availability and improved learning. On the other hand, data uncertainty, known as aleatoric uncertainty, is induced by inherent noise present within the dataset [37]. For instance, systematic and random errors may occur in the computed and measured dose distributions during the quantitative analysis, leading to data noise. It is worth noting that this particular uncertainty is an inherent dataset attribute, and augmenting the training set with additional data does not suffice to diminish it.

Most existing UQ approaches primarily focus on quantifying model uncertainty. These methods include deep ensemble learning [38,39] and Monte Carlo dropout, approximation Bayesian inference [36,40]. However, only a limited number of techniques available can effectively measure the total uncertainty in deep learning. Kendall and Gal [41] proposed an innovative power exponential loss function derived from maximizing the likelihood of learning the data uncertainty. Combined with MC Dropout, this approach predicted total uncertainty in the results. The training process for this method poses significant challenges. Building upon this foundation, Gast [42] and Loquercio et al. [43] employed a pre-defined density filtering method to facilitate uncertainty feedforward propagation in neural networks. This method does not apply to GPR prediction as it requires predetermined uncertainty of the input information. Yang et al. [44] first integrate UQ with PSQA. However, they only studied low granularity uncertainty and missed dose delivery spatial information, which is very important in clinical applications. Although the Efforts to incorporate uncertainty estimation techniques aim to address these issues and enhance the reliability of deep learning predictions, Quantifying such uncertainty before the input of clinical data into the automatic PSQA model is a complex undertaking. The integration of uncertainty information within AI-based PSQA models provides several benefits. Firstly, it measures confidence in the predicted outcomes, allowing for a more comprehensive evaluation of treatment plans. Secondly, it identifies and mitigates potential risks, reducing the likelihood of erroneous or misleading treatments. Additionally, it facilitates patient-specific quality assurance by tailoring treatments to individual needs and characteristics.

This study makes the following significant contributions.

- We propose Multi-granularity Prior Networks for multi-task PSQA, namely DDP and GPR prediction, modeling different types of granular uncertainties.
- We utilized uncertainties at multiple granular levels and proposed a multi-granularity interactive Deep Bayesian learning framework specifically for PSQA.
- Our method is the first to be capable of presenting the PSQA dose difference distribution confidence interval visualization, which is more secure than the previous method.
- Our method involves an Uncertainty-guided man-machine integrated method for precise and robust PSQA.
- We conducted a detailed comparative analysis on uncertainty estimations for PSQA and achieved state-of-the-art results.

The rest of the study is structured as follows: Section 2 discusses the methodology in a detailed manner, whereas Section 3 presents a detailed overview of the experimental setting and results. Further Section 4 concludes this study and discusses the current study's limitations and possible future work.

2. Materials and methods

This section provides a comprehensive overview of our study's methodology, including the data collection process, pre-processing, and a detailed discussion of network architecture. The subsequent subsections discuss each in a detailed manner.

2.1. Data collection and preprocessing

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 1394 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.

Following the recommendations in the TG-218 report [45], 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%/1 mm, 2%/2 mm, and 2%/3 mm criteria at a 10% threshold of the maximum dose (only points with doses

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

Table 1
Overview of recent studies on PSQA methods.

Categorization	Study ref.	Methods and description	Advantages	Limitations
Complexity Metrics	[6]	PSQAIDC using TPS-generated plans and treatment delivery log files.	The feasibility of substituting PSQAM with PSQAIDC, interpretation using NTCP/TCP models, time savings, adaptive proton therapy facilitation, and sensitivity to delivery errors are discussed.	• Indirect characterization: Complexity metrics indirectly characterize plan delivery accuracy. While they may be correlated with delivery accuracy, they do not directly measure clinical significance or patient outcomes.
	[19]	Beam complexity metrics based on aperture analysis (mean aperture area, extent of aperture shape irregularity, degree of beam modulation)	It quantitatively assesses beam complexity in IMRT plans, characterizing different aspects of beam complexity across various disease sites and treatment modalities, while also correlating beam complexity metrics with quality assurance (QA) results. Additionally, it identifies factors that influence beam complexity.	• Complexity metrics have a limited scope and applicability, focusing on specific plan characteristics, and may not capture all factors influencing delivery accuracy across different treatment centers, techniques, or disease sites.
	[20]	Modulation index (MI), Fluence map complexity (FMC)	Successfully distinguished deliverable treatment plans from plans that did not pass quality assurance (QA) tests with a low false negative rate. Provides information on the complexity of a given plan relative to the achievable complexity of treatments for the same site.	• Incomplete understanding: While complexity metrics can provide insights into the relationship between plan characteristics and delivery accuracy, they may not fully capture the underlying mechanisms or specific causes of discrepancies between planned and delivered dose distributions.
	[21]	Volumetric modulated arc therapy (VMAT), Local (γ) index analysis (3%, 3 mm, and 2%, 2 mm), Leaf travel and modulation complexity score (MCS)	Decreased probability of intrafraction errors, Enables comparison between delivered and calculated dose and Provides a quantitative measure of dosimetric accuracy, enables comparison of dosimetric accuracy between different plan parameters.	• Limited validation against clinical endpoints: Complexity metrics often lack validation against clinical endpoints or patient outcomes. The association between complexity metrics and delivery accuracy does not guarantee improved treatment outcomes or increased clinical efficacy.
Machine Learning	[23]	Quality assurance-based optimization (QAO) framework utilizing machine learning (SVM algorithm)	Directly maximizes predicted quality assurance (QA) outcomes without compromising the dosimetric quality of plans designed with knowledge-based planning (KBP). The feasibility of QAO was explored as a proof-of-concept. Potential for future incorporation of QA endpoints into plan optimization.	• Inability to predict clinical significance: Machine learning methods that rely on complexity metrics or other input parameters may struggle to predict the clinical significance of plan delivery accuracy directly. These methods may not capture the complexity of patient-specific factors influencing treatment outcomes.
	[24]	Radiomics-based ML, Elastic net (EN) and extra trees (ET), Feature selection and hyperparameter tuning, GPR, DD, and DTA metrics, Radiomic features as plan complexity indicators	High performance for predicting dose-evaluation metrics, especially in high dose region, Accurate prediction of patient-specific QA results for VMAT, Quantitative assessment of dosimetric accuracy, Enable prediction based on VMAT plan parameters, Insights into dose distribution and QA relationship	• Reliance on GPR alone is insufficient. While GPR can provide valuable insights, it may not be comprehensive enough for accurate plan delivery accuracy assessment. Incorporating additional information, such as spatial data, is crucial to capturing the full complexity of patient-specific quality assurance (PSQA).
	[25]	Gamma analysis with 2%2mm, 2%3mm, and 3%3mm, Utilization of PTW OCTAVIUS 4D (4DOCT) and Arc Check detectors	Investigates the influence of calibration field size on GPR in PSQA. Demonstrates variations in GPR across different calibration field sizes. Shows increasing GPR trend below $10x10cm^2$ calibration field sizes and decreasing trend above $10x10cm^2$. There is a strong correlation between 4DOCT and Arc Check.	• Data limitations and bias: Machine learning methods heavily rely on training data quality, quantity, and representativeness. Insufficient or biased training data can lead to suboptimal performance, inaccurate predictions, or biased results.
	[26]	Gradient Boosting Decision Tree (GBDT), Random Forest (RF), Poisson Lasso (PL)	ML model based on GBDT algorithm showed a certain degree of accuracy for predicting gamma passing rates (GPRs) in dual-layered MLC linac. Models after reweighting the data showed up to 30% improvement in the rare region. Complexity metrics were studied to identify important features for PSQA prediction.	• Dependency on complexity metrics overlooks critical factors affecting plan delivery accuracy, such as patient-specific variations and uncertainties, compromising model accuracy and generalizability.

(*continued on next page*)

Table 1 (*continued*).

Deep Learning	[27]	Complexity metrics (CM) and binarization of PSQA, C# script used for integration into the TPS, Evaluation of predictive performance using AUROC, sensitivity, and specificity	Optimizes workload of PSQA, model exhibited an AUROC of 0.88, Sensitivity and specificity exceeding 50% and 90%, Estimated yearly savings of about 140 h with the use of the script	• Lack of explainability: Deep learning methods are black-box and lack transparency. This lack of explainability can hinder trust and understanding in the decision-making process. On the other hand, it also hinders the integration of prior knowledge with the deep learning model.
	[9]	Dual Training Approach (Cycle-GAN and c-GAN), Error Detection Model, Deep Learning Approaches (ML, CNN, UNet++, GAN).	Automation of patient-specific quality assurance (PSQA) process. Improved prediction accuracy of both cycle-GAN and c-GAN models. Accurate generation of synthetic measured fluence for VMAT PSQA and identification of errors within the synthetic fluence.	• Limited uncertainty estimation: While deep learning methods can provide effective predictions, uncertainty estimation is often limited, leading to a lack of security.
	[28]	TransQA (Deep Hybrid Transformer Network). It incorporates a self-attention mechanism-based transformer for modeling long-range dependencies and combines modified U-Net architecture with volumetric dose prediction.	Achieves comparable or better performance in MDose prediction and reduces the measurement work of prePSQA, providing a clinical guidance tool.	• Computation Resource Consumption: Deep learning methods demand more computation resources than other methods. Therefore, it is crucial to analyze the method computation availability for practical medical application
	[29]	UNet++, Support Vector Machines (SVMs), Valdes, Lam, Li models	Provides accurate patient-specific QA for radiotherapy and classifies failed or pass fields also Predicts gamma passing rates for different criteria.	
	[30]	Multilayer Perceptron, Neural Network	Considers dynamic delivery process of VMAT plans during PSQA. Incorporates regression and ranking loss for improved accuracy.	
UQ Deep Learning	[33]	Uncertainty-aware dual-task, deep learning (UDDL) model	Simultaneously predicts GPR and corresponding uncertainty, which balances effectiveness and safety of automatic PSQA model. Identifies samples with substantial prediction errors and failed samples with low-label GPR.	• Dismiss the dose difference distribution prediction task and lose the dose delivery high-granularity spatial information, which is crucial in clinical practice.

Table 2
Five samples of variations of GPR values.

Particular	1%/1 mm	2%/2 mm	3%/2 mm
/0001/Calc/Y1.dcm	91.8006	96.5789	98.3126
/0002/Calc/Y5.dcm	98.1378	99.5006	99.7908
/0003/Calc/Y3.dcm	94.8047	96.9754	98.0618
/0004/Calc/Y3.dcm	94.8458	97.9322	98.9743
/0005/Calc/Y1.dcm	93.3672	96.89	98.5501

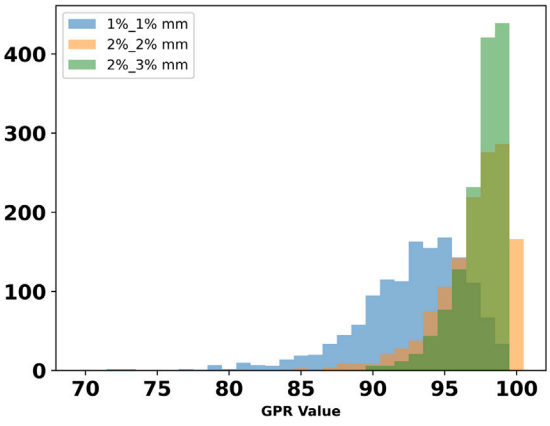


Fig. 1. Variations in GPR Values.

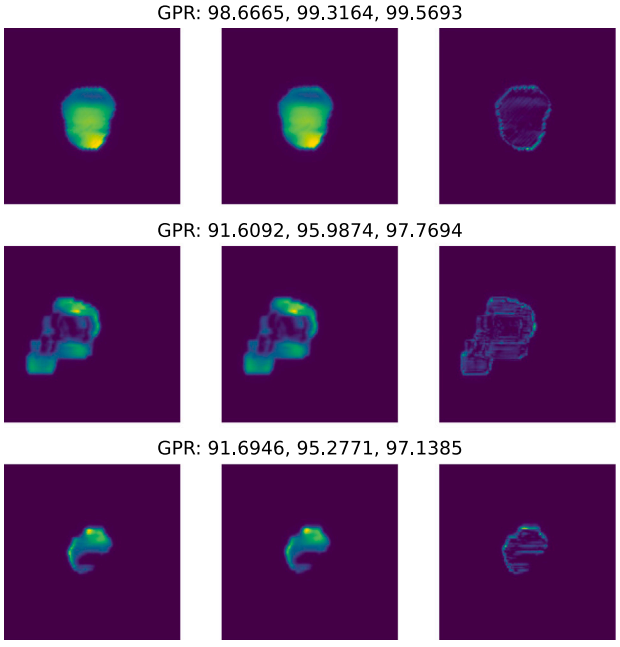


Fig. 2. Three IMRT samples of PSQA. In the figure, the left column is the calculation plan, the Middle is a real measurement, and the right is the dose difference. The pixel intensity denotes the delivery dose intensity in the radiation therapy plan (IMRT).

greater than 10% of the global maximum dose per beam were analyzed) as presented in Table 2 for reference. The gamma analyses were performed in absolute dose mode, and global normalization was applied to the results. The Treatment Planning System (TPS) calculated

Algorithm 1 Multi granularity Prior Networks High-Level Pseudo code

Input: dataset
Output: L_{total}
Function `Train(`*dataset*`)`:
  $L_{total} \leftarrow 0$
  for *X*, Y_{gpr}, Y_{ddp} *in dataset* **do**
    X← Preprocess(X)
    FeaturesList← encoder(X)
    $Prior \leftarrow \mathcal{NW}(m, L, \kappa, \nu) \leftarrow$ regression_head(FeaturesList[-1]) //Calculate GPR prior distribution parameters following the assumption of Normal Wishart distribution
    $P(Y_{ddp}) \leftarrow \mathcal{N}(\mu, \Lambda) \leftarrow$ decoder(FeaturesList) // Calculate DDP distribution parameters following the assumption of Normal distribution
    $L_1 \leftarrow -lnP(Y_{gpr}) - ln\mathcal{T}(y|m, \frac{\kappa+1}{\kappa(\nu-K+1)}L^{-1}, \nu-K+1)$ //Loss of GPR task, namely the loglikelihood of marginal distribution according to the $\mathcal{NW}$ prior
    $L_2 \leftarrow E_{p(\mu,\Lambda|x,\theta)}[-ln\mathcal{N}(Y_{ddp}|\mu, \Lambda)]$//Loss of DDP task, namely the loglikelihood of DDP Normal distribution
    $L_3 \leftarrow KL[p(\mu, \Lambda|x, \theta)||\mathcal{NW}(\mu, \Lambda|m, L, \kappa, \nu)]$ //Coherence loss between two granularity task
    $L_{total} \leftarrow L_1 + L_2 + \alpha L_3$
  end
  return L_{total}
Function `Eval(`*dataset*`)`:
  $L_{total} \leftarrow 0$
  for *X in dataset* **do**
    X← Preprocess(X)
    FeaturesList← encoder(X)
    $Prior \leftarrow \mathcal{NW}(m, L, \kappa, \nu) \leftarrow$ regression_head(FeaturesList[-1])
    $P(Y_{gpr}) \leftarrow \mathcal{T}(y|m, \frac{\kappa+1}{\kappa(\nu-K+1)}L^{-1}, \nu-K+1)$
    $P(Y_{ddp}) \leftarrow \mathcal{N}(\mu, \Lambda) \leftarrow$ decoder(FeaturesList)
  end
  return $H(Y_{ddp}), H(Y_{gpr}), m, \mu$ //H is the Entropy Calculation Function for uncertainty quantification, and m, μ is the maximum estimation of the GPR prediction and DDP.

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 1394 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 1 mm × 1 mm and then cropped to 512 × 512 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. Table 2 offers a glimpse of the dataset samples, highlighting the complexities and potential of our dataset. At the same time, the accompanying Fig. 1 visually underscores the need for meticulous data balancing due to evident imbalances in certain measures' 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. 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 calculation plan (the left column), real measurement (the middle column), and dose difference (the right column).

2.2. Multi-granularity regression prior networks

This section thoroughly explores the Multi-granularity Regression Prior to Networks, delving into a series of mathematical derivations that underpin the framework's foundation. With the help of the following mathematical derivation, our proposed Multi-granularity Prior Networks are architecturally detailed, emphasizing their ability to model diverse granular uncertainties. The network architecture is optimized for multi-task learning, allowing simultaneous training for DDP and GPR prediction tasks. Integrating granular uncertainty information enhances the network's adaptability to various PSQA scenarios. The equations mentioned below are common settings to denote prior networks-related work, and they are considered as prior hypotheses [46,47]; in such cases, we followed these equations to form a base for our proposed multi-granularity regression prior networks.

$$P(y|x^*, D) = \iint p(y|\mu)p(\mu|x^*, \phi)p(\phi|D), d\mu, d\phi \tag{1}$$

In Eq. (1), $P(y|x^*, D)$ represents the probability of the regress variable y given input (x^*) and the training data D. It denotes the predictive uncertainty. Whereas $p(y|\mu)$ is the likelihood of the y given the model parameter μ. It quantifies the uncertainty associated with the model parameters. $p(\mu|x^*, \phi)$ is the posterior distribution of the model parameter μ given the input x^* and the hyperparameter ϕ. It represents the uncertainty in the model parameters conditioned on the input and hyperparameters. $p(\phi|D)$ is the posterior distribution of the hyperparameter ϕ given the training data D.

Fundamental Assumption: The fundamental assumption involves modeling the lower granularity task, DDP, as a Gaussian distribution. Considering that GPR aggregates points from DDP above a threshold, we view the probability distribution of the higher granularity GPR task as an average representation of the uncertainty in DDP. Furthermore, within the Bayesian inference framework, the probabilities used for inference can be regarded as a probabilistic average of the likelihood probabilities. Therefore, it is reasonable to consider the GPR distribution as the prior for DDP. The Normal-Wishart (NW) distribution is commonly employed as the prior when dealing with a Gaussian

Table 3
Mathematical symbol table

Symbol	Description
Y_{ddp}	Specific target variable of DDP task
Y_{gpr}	Specific target variable of GPR prediction task
K	The dimension of GPR, namely 3 in our implementation, since in clinical application there are 3 common types of GPR criteria
KL	Kullback–Leibler divergence
$\mathcal{N}(Y_{ddp}\|\mu, \Lambda)$	Normal distribution to model the DDP task. μ is the mean and Λ is the variance matrix.
$\mathcal{N}(\mu\|m, \kappa\Lambda)$	Normal distribution. The normal distribution is the common prior distribution with an unknown mean, where κ is the mean, and Λ is the variance matrix
$\mathcal{W}(\Lambda\|L, \nu)$	Wishart distribution. The Wishart distribution is the common prior distribution to normal distribution with unknown variance, where the L is scale matrix, and ν is freedom degree
$\mathcal{NW}(\mu, \Lambda\|m, L, \kappa, \nu)$	Normal-Wishart distribution, the Normal distribution is the common joint prior distribution with unknown mean and variance. The specific form is in the Eq. (4)
$\mathcal{T}$	Student's t-distribution. The parameters of the Student's t-distribution are orderly parameterized as follows: first, the mean, followed by the scale matrix, and finally, the degrees of freedom
lnP	The log-likelihood of distribution P
H_p	The entropy of distribution p
V_p	The variance of distribution p
α	The coefficient of coherence loss

distribution with unknown mean and variance.

$$\begin{aligned}\int p(y|\mu) &= [\int p(\mu|x^*, \phi)p(\phi|D)d\phi]d\mu \\ &= [\int p(y|\mu)p(\mu|x^*, D)]d\mu\end{aligned} \tag{2}$$

$$p(y|x, \theta) = \mathcal{N}(y|\mu, \Lambda), \{\mu, \Lambda\} = f(x; \theta) \tag{3}$$

After stating the fundamental assumption, Table 3 defines the mathematical symbols employed in this study. Eq. (3) represents the probabilistic regression model, where $p(y|x, \theta)$ is the target variable y distribution given input x and parameters θ. The distribution is assumed to be a multivariate normal distribution with mean μ and precision matrix Λ, where $\{\mu, \Lambda\}$ is obtained from the function $f(x; \theta)$.

$$\mathcal{NW}(\mu, \Lambda|m, L, \kappa, \nu) = \mathcal{N}(\mu|m, \kappa\Lambda)\mathcal{W}(\Lambda|L, \nu) \tag{4}$$

Eq. (4) defines the Normal-Wishart distribution, denoted as $\mathcal{NW}(\mu, \Lambda|m, L, \kappa, \nu)$. It is a compound distribution consisting of a conditional normal distribution over the mean μ and a Wishart distribution over the precision matrix Λ.

$$\begin{aligned}&p(\mu, \Lambda|x, \theta) = \mathcal{NW}(\mu, \Lambda|m, L, \kappa, \nu), \\ &\{m, L, \kappa, \nu\} = \Omega = f(x; \theta)\end{aligned} \tag{5}$$

Eq. (5) represents the parameterization of the Normal-Wishart distribution over the mean μ and precision matrix Λ given input x and parameters θ. The distribution parameters $\{m, L, \kappa, \nu\}$ are obtained from the function $f(x; \theta)$.

$$\begin{aligned}p(y|x, \theta) &= \mathbb{E}_{p(\mu,\Lambda|x,\theta)}[p(y|\mu, \Lambda)] \\ &= \mathcal{T}(y|m, \frac{\kappa+1}{\kappa(\nu-K+1)}L^{-1}, \nu-K+1)\end{aligned} \tag{6}$$

Eq. (6) calculates the predictive distribution $p(y|x, \theta)$ by taking the expected value over the parameter distribution $p(\mu, \Lambda|x, \theta)$. The resulting distribution is denoted as $\mathcal{T}(y|m, \kappa+1/\kappa(\nu-K+1), L^{-1}, \nu-K+1)$ [46]. Meanwhile, the T distribution converges to a normal distribution if υ is infinity.

$$\begin{aligned}H_{\text{ModelUnc}} =& H\left[\mathbb{E}_{p(\mu,\Lambda|x,\theta)}\left[p(y|\mu, \Lambda)\right]\right] \\ &- \mathbb{E}_{p(\mu,\Lambda|x,\theta)}\left[H\left[p(y|\mu, \Lambda)\right]\right]\end{aligned} \tag{7}$$

As discussed in Dirichlet Prior Networks [48], Eq. (7) represents the model uncertainty, denoted as H_{ModelUnc}. It is calculated based on the determinant of the precision matrix Λ. Whereas, $H\left[\mathbb{E}_{p(\mu,\Lambda|x,\theta)}[p(y|\mu, \lambda)]\right]$ represents the entropy of predictive probability distribution $p(y|\mu, \Lambda)$ over μ and Λ which denotes the total uncertainty and $\mathbb{E}_{p(\mu,\Lambda|x,\theta)}[H\left[p(y|\mu, \Lambda)\right]]$ denotes the data uncertainty.

The model uncertainty, total uncertainty, and expected model uncertainty in the Prior Networks can also be expressed in variance form:

$$\begin{aligned}V_{p(\mu,\Lambda|x,\theta)}[\mu] =& V_{p(y|x,\theta)}[y] \\ &- \mathbb{E}_{p(\mu,\Lambda|x,\theta)}\left[\Lambda^{-1}\right]\end{aligned} \tag{8}$$

In Eq. (8), this $V_{p(\mu,\Lambda|x,\theta)}[\mu]$ term represents model uncertainty, further $V_{p(y|x,\theta)}[y]$ presents total uncertainty, whereas $\mathbb{E}_{p(\mu,\Lambda|x,\theta)}[\Lambda^{-1}]$ showcases expected data uncertainty.

$$p(\mu, \Lambda|\Omega^{\hat{i}}) \propto p(y^{(i)}|\mu, \Lambda) \cdot p(\mu, \Lambda|m, L, \kappa, \nu) \tag{9}$$

The Eq. (9) specifies the relation for inducing precision in Normal-Wishart prior distribution for uncertainties, where $p(y(i)|\mu, \Lambda)$ is a normal distribution

Based on the fundamental assumption, the predictive distribution of the high-granularity task, GPR, serves as the prior for the low-granularity task, DDP. Therefore, following Bayesian theory, we construct a variational distribution, denoted as $p(\mu, \Lambda|x, \theta)$, based on a deep neural network to approximate the posterior distribution $p(\mu, \Lambda|\Omega)$. This approximation is used to obtain the evidence lower bound (ELBO) [47,49] as the loss function for training DDP. The equation is as follows:

$$\begin{aligned}L_{ddp} =& KL[p(\mu, \Lambda|x, \theta)\|p(\mu, \Lambda|\Omega)] \\ =& E_{p(\mu,\Lambda|x,\theta)}[-ln\mathcal{N}(y|\mu, \Lambda)] \\ &+ KL[p(\mu, \Lambda|x, \theta)\|\mathcal{NW}(\mu, \Lambda|m, L, \kappa, \nu)] + Z\end{aligned} \tag{10}$$

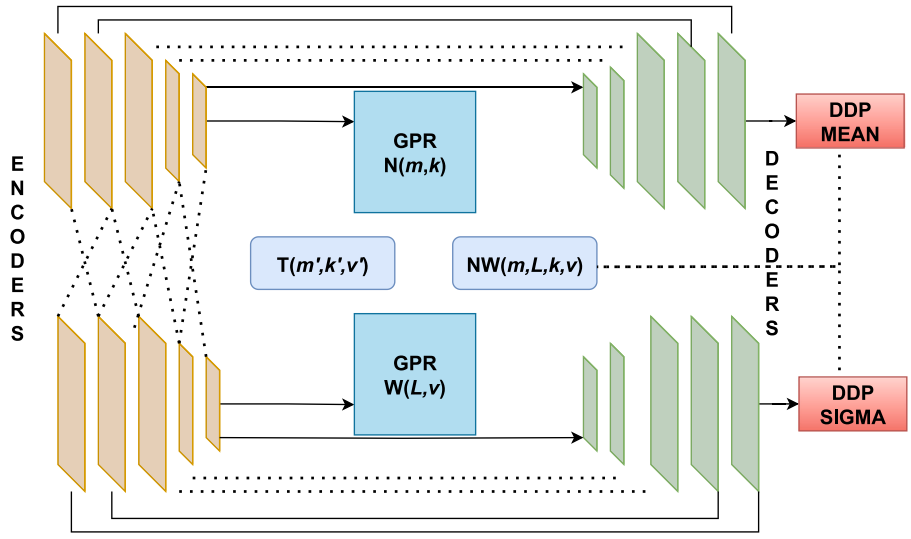


Fig. 3. Overview of proposed multi-granularity prior networks model architecture. The diagram illustrates the intricate components and relationships within the model, showcasing the integration of various encoders and decoders for capturing uncertainties and generating predictions.

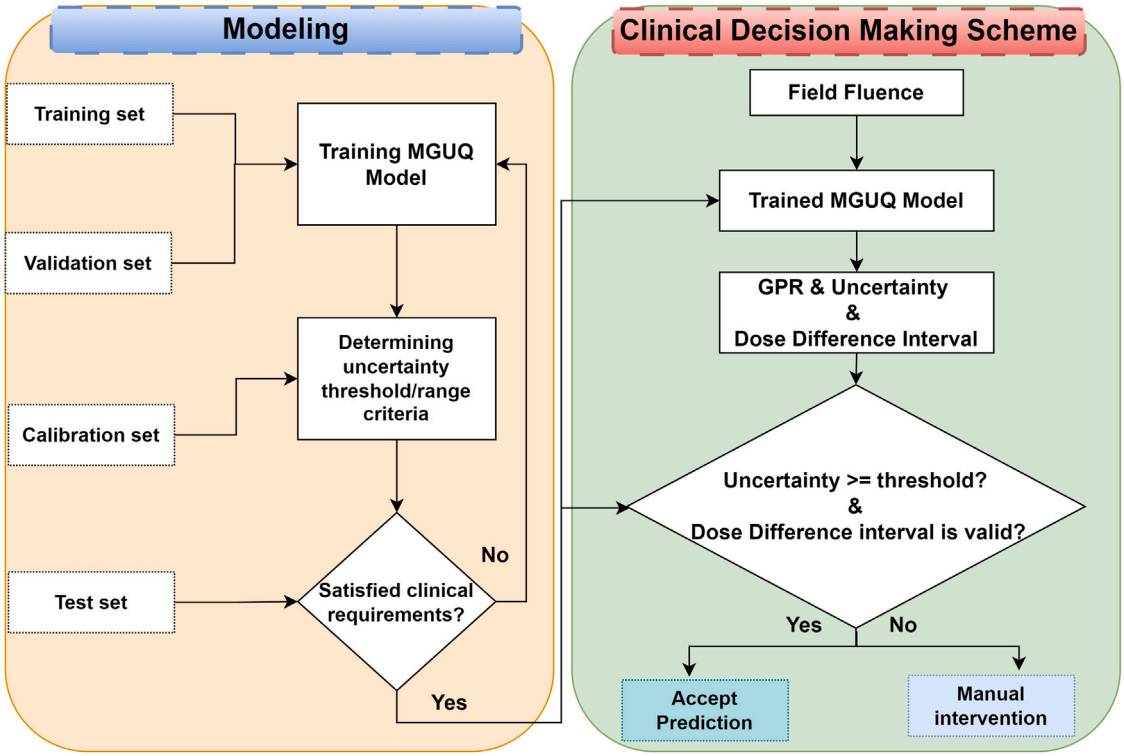


Fig. 4. Modeling clinical decision making scheme: A flowchart for proposed multi-granularity uncertainty quantification.

Thus, we can get:

$$L_1 = -lnP(Y_{gpr}) = -ln\mathcal{T}(Y_{gpr}|m, \frac{\kappa+1}{\kappa(\nu-K+1)}L^{-1}, \nu-K+1) \tag{11}$$

$$L_2 = E_{p(\mu,\Lambda|x,\theta)}[-ln\mathcal{N}(Y_{ddp}|\mu,\Lambda)] \tag{12}$$

$$L_3 = KL[p(\mu,\Lambda|x,\theta)\|\mathcal{NW}(\mu,\Lambda|m,L,\kappa,\nu)] \tag{13}$$

$$L_{total} = L_1 + L_2 + \alpha L_3 \tag{14}$$

The presented Eq. (10) illustrates an intricate relationship between lower and higher granularity levels. Here, L_1 is the maximum likelihood of GPR distribution, and then the GPR distribution is used as the prior distribution to DDP. By optimizing the posterior distribution of DDP, we get L_2 and L_3, L_2 can be regarded as the L_{ddp} loss, which means if we only concern the DDP task, only this loss exists. L_3 models the coherence of two granularity tasks, which is hereafter called coherence loss. We can achieve a more comprehensive and accurate uncertainty estimation by applying a coherence loss across different granularities of uncertainty types. This supervised information allows the neural network to learn more reasonable and informative representations within limited data.

Table 4
Parameter settings.

Parameter	Description	Value
Resized Dose Plan Array Sample Shape	To organize these dose plan array samples into a batch, we do the ResizeWithPadOrCrop preprocessing. And here is the resized shape.	[512,512]
Encoder Architecture	The pretrained backbone network is used as the encoder. And the structure of the dual network of the original and uncertain tasks is identical.	Resnet18(Imagenet)
Regression head Architecture	Architecture of GPR prediction output, the output dimension is 3 since there we select three GPR criteria [1%/1 mm,2%/2 mm,2%/3 mm]	AdaptiveAvgPool2d[1,1] Flatten() Dropout(p=0.1) Linear[512,3]
Decoder Architecture	Architecture of DDP output is Unet decoder, and here is input channel size.	[768, 384, 192, 128,32]
Dropout Rate	The dropout rate implemented in network	0.1
α	The coefficient of coherence loss	2.5
Optimizer	Type of optimizer used	Adam
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 (EXP-A and EXP-B)	Experiment Method dataset split ratio, namely Training Dataset Size: Validation Dataset Size: Calibration Dataset size: Test Dataset Size which is used in Experiment A and B, Calibration Dataset is used to determine the uncertainty score threshold. For the method without UQ, the calibration dataset is not used	7:1:1:1

2.3. Network architecture

The given model architecture diagram, as depicted in Fig. 3 offers detailed insights into the components and relationships. The first component is a set of encoders consisting of GPR prediction with Student-t distribution derived by predicted NW distribution parameter. Student-t distribution handles uncertainties, employing parameters like Location Factor (LF), Scaled Factor (SF), and Degree of Freedom (DF). LF represents the mean value in the Gaussian distribution, determining the central or average value of the distribution. SF controls the variance of the Student-t distribution. DF influences the shape and tails of Student-t distribution. The next component is the decoder, which produces output for DDP and processes the encoded input to generate output predictions. Where (Mean) indicates the mean value of the distribution and Sigma represents the standard deviation of output distribution, signifying the variability in the model's predictions. This architecture integrates diverse encoders to capture different uncertainties in the PSQA data.

Furthermore, Fig. 3 showcases the utilization of Unet with ResNet [50], a classical and efficient structure that has been widely used in many medical image tasks [51,52]. Unet and ResNet suggested considering the backbone we employed as an example, as this model architecture is flexible enough to use others. When a neural network reaches a certain depth, increasing the number of layers cannot bring improvements due to the gradient vanishing and network degradation. ResNet proposed the idea of residual learning to alleviate the above problems by fitting a residual function F(x)=H(x)-x (wherein x and H(x) are the input and output of the residual block, respectively). Here, we employed ResNet18 as the feature extraction backbone in our network, which mainly consists of four stages, each including two cascaded Residual Blocks. The Residual Blocks comprise Convolution, Instance Normalization, and ReLU activation layers to learn the residual mapping. The backbone can be adjusted to other pre-trained networks as needed.

The specific implementation setting is shown in Table 4. Our network consists of two identical backbone networks. We extract features from each layer of the backbone networks and then fuse the features from the two networks through interaction links. The output of the last layer is then individually fed into a regression head to predict the prior distribution parameters for the mean (Normal distribution) and variance (Wishart distribution). These parameters then composite the overall prior distribution, a Normal-Wishart distribution. The Student-t distribution of the GPR is then calculated according to this prior and Gaussian likelihood. On the other hand, we feed the feature list into a U-Net decoder to output the mean and the variance of the Gaussian distribution for the DDP.

2.4. Flow of MGUQ method

The flowchart presented in Fig. 4 outlines the Multi-granularity Uncertainty Quantification (MGUQ) in the Clinical Decision Making Scheme, demonstrating the procedure for assessing uncertainty and intervention thresholds. The flow starts with the modeling phase, wherein several datasets, including the training, validation, and calibration sets, are employed. The MGUQ model, which has been trained, is then used to establish criteria for uncertainty thresholds or ranges. This stage entails evaluating and measuring the level of ambiguity linked to making clinical decisions. After determining the uncertainty level, it is compared to a preset threshold. When uncertainty is above the threshold, it signifies an elevated degree of uncertainty in decision-making. In such instances, it may be necessary to intervene manually to arrive at a more informed and dependable choice.

Conversely, if the level of uncertainty falls below the specified threshold and the GPR prediction and DDP is qualified, the choice is deemed acceptable and does not require manual intervention. This indicates that the decision-making process has a reduced amount of ambiguity and fulfills the predetermined clinical criteria. In summary, the flowchart illustrates a systematic process for measuring uncertainty in clinical decision-making, enabling improved comprehension and control of uncertainty by implementing suitable intervention thresholds.

3. Evaluation and results

This section comprehensively analyzes our proposed multi-granularity prior networks for achieving uncertainty-informed robust and efficient PSQA.

Table 5
Mean Absolute Error (MAE) comparison of dose difference for various Uncertainty Quantification (UQ) criterion. The best results are shown with bold text.

Methods category	Method	Criteria	GPR MAE (%)				
			All	95-100	90-95	85-90	<85
Machine Learning	GBDT [26]	1%/1 mm	2.054	1.867	1.914	2.157	6.402
		2%/2 mm	1.333	1.173	1.723	4.441	–
		2%/3 mm	0.950	0.896	2.108	–	–
	RF [26]	1%/1 mm	2.153	1.969	1.994	2.527	5.589
		2%/2 mm	1.386	1.262	1.712	3.730	–
		2%/3 mm	1.058	1.007	1.973	–	–
	PL [26]	1%/1 mm	2.928	2.952	2.894	2.769	4.126
		2%/2 mm	2.258	2.202	2.367	3.630	–
		2%/3 mm	1.763	1.737	2.277	–	–
Deep Learning	UNET++ [29]	1%/1 mm	2.317	1.507	2.049	3.622	6.415
		2%/2 mm	1.278	0.881	2.189	3.519	–
		2%/3 mm	1.047	0.975	1.500	–	–
	TransQA [28]	1%/1 mm	2.081	1.924	1.612	2.605	5.605
		2%/2 mm	1.545	1.399	1.656	3.808	–
		2%/3 mm	1.003	0.912	1.716	–	–
	CycleGAN [9]	1%/1 mm	2.213	1.520	1.460	4.013	8.167
		2%/2 mm	1.324	0.837	2.373	4.625	–
		2%/3 mm	0.945	0.744	2.661	–	–
UQ Deep Learning	MCBN_LAP [53]	1%/1 mm	8.889	2.700	5.805	27.919	25.262
		2%/2 mm	3.369	2.202	6.807	7.219	–
		2%/3 mm	1.880	1.127	7.856	–	–
	MCBN_Gaussian [53]	1%/1 mm	2.332	1.238	2.197	3.958	6.224
		2%/2 mm	1.400	0.936	2.523	4.254	–
		2%/3 mm	1.085	0.937	2.257	–	–
	MCDO_LAP [33]	1%/1 mm	2.164	1.211	1.732	4.090	6.719
		2%/2 mm	2.554	1.480	5.345	8.096	–
		2%/3 mm	0.982	0.822	2.252	–	–
	MCDO_GAUSS [33]	1%/1 mm	2.116	1.489	1.893	3.262	4.930
		2%/2 mm	1.474	0.901	2.796	5.305	–
		2%/3 mm	1.013	0.880	2.069	–	–
	MCDO_Alpha_1	1%/1 mm	2.858	1.117	2.164	6.046	11.229
		2%/2 mm	2.044	1.011	3.551	13.636	–
		2%/3 mm	1.176	0.909	3.301	–	–
	MCDO_Alpha_2	1%/1 mm	2.176	1.353	1.549	3.903	7.858
		2%/2 mm	1.713	1.213	3.254	3.004	–
		2%/3 mm	1.081	0.945	2.165	–	–
	MGUQ	1%/1 mm	**2.024**	1.342	1.798	2.722	4.243
		2%/2 mm	**1.246**	0.897	1.960	4.082	–
		2%/3 mm	**0.864**	0.750	1.766	–	–

Evaluation Metrics The objective of our multi-granularity prior networks to automate PSQA in radiation therapy presents a multifaceted challenge involving classification and regression tasks. Here, we choose six pivotal metrics to assess robustness: (i) Clinical Accuracy (CA), (ii) Manual Measurement Fraction (MMF), (iii) Proportion of Correct-certain (PCC) [44], and (iv) Mean Absolute Error (MAE). Additionally, we maintain sensitivity (TPR) and specificity (TNR).

Manual Measurement Fraction (MMF)

$$MMF = \frac{N_{unc}}{N_{total}} \tag{15}$$

Proportion of Correct-certain (PCC)

$$PCC = P(correct|certain) = \frac{N_{CC}}{N_{CC} + N_{IC}} \tag{16}$$

Clinical Accuracy (CA)

$$CA = 1 - P(incorrect|certain) = \frac{N_{CC} + N_{unc}}{N_{total}} \tag{17}$$

Where the N_{unc} is the number of samples below the uncertain threshold, N_{CC} is the number of certain correct samples, and N_{IC} is the number of certain incorrect samples, which indicates the real fault in clinical application. MMF is the uncertain sample ratio, and $1-MMF$ denotes a certain sample ratio, namely the workload.

3.1. Baseline methods for PSQA

This section presents the baseline PSQA methods regarded as comparative measures with the proposed MGUQ method.

We selected the gradient boosting decision tree (GBDT), random forest (RF), and Poisson Lasso (PL) models as traditional machine learning-based baseline methods. These techniques incorporate manually designed complexity metrics as features for predicting the Gamma Passing Rate (GPR). The implementation details of these machine learning-based methods are elaborated in [26].

For deep learning-based baseline methods, the study by [29] utilizes a ResNet-based UNet++ architecture to predict GPR and dose differences. Beyond conventional CNN architectures, the research presented in [9] introduces a transformer-based methodology, TranQA, for predicting GPR and dose differences. Alongside feedforward neural network approaches, study [9] utilizes CycleGAN [54] to model the Dose Difference Prediction (DDP) task, hereafter referred to by the abbreviation CycleGAN.

3.2. Baseline methods for UQ

To validate the superiority of our method over other existing uncertainty quantification deep learning methods in PSQA uncertainty-based performance metrics and thus demonstrate its security, we also selected some uncertainty deep learning algorithms for comparison. Uncertainty estimation algorithms can be classified into (i) Bayesian methods, (ii) Deterministic methods, and (iii) Ensemble-based methods.

Bayesian methods introduce uncertainty by integrating probabilistic variance into parameters. MCDO (Monte Carlo Dropout) [55] derived a method to use dropout as an uncertainty measure. Dropout can be seen as modeling some neurons as a binomial distribution as the variational distribution. Similarly, MCBN (Monte Carlo Batch Normalization) [53]

derived a method to consider the mean and variance used for normalization in batch normalization as a Gaussian distribution, thereby completing uncertainty quantification.

In contrast, deterministic methods typically model neural network output as probability distribution parameters and then train based on maximum likelihood loss. For example, [41,44] utilize Gaussian distributions, [56] use Laplace distributions, or even [57] use uniform distribution intervals to model the output distribution. OCs [58] also utilize uniform distribution intervals while applying a different form of loss function called simultaneous quantile regression. Note that Bayesian methods and deterministic methods are not contradictory in practice and are often used together to quantify uncertainty simultaneously, so in our setup, naming such as MCDO_Gaussian appears, representing the use of both MCDO and Gaussian distribution to model the output distribution. This is also the original setting of MCDO [41,44]. We also provide experimental results of combining different alpha values with MCDO. MCDO_Alpha_N denotes the MGUQ method with an alpha value of N. For example, MCDO_Alpha_2 denotes the MGUQ method with an alpha value of 2. MGUQ shows the best performance with an alpha value of 2.5.

The idea of ensemble-based methods is simple. They use the variance of sub-model predictions in the ensemble model as the uncertainty output. Deep Ensemble [38] integrates identical sub-deep learning models trained on the overall training dataset for the same epochs, each of which, similar to [41], treats the predicted values as samples from a Gaussian distribution with mean and variance. Deep Ensemble models the ensemble result as a Gaussian distribution to approximate the uniform mixture of Gaussian distributions predicted by the sub-models. The ensemble mean and variance are approximated by

$$\mu_* = M^{-1}\sum_{m=1}^{M}\mu_m, \sigma_*^2 = M^{-1}\sum_{m=1}^{M}\left(\sigma_m^2+\mu_m^2\right)-\mu_*^2$$

respectively. The original paper also proposed adversarial sample training as an optional additional training step. However, it was not utilized in regression experiments on real-world datasets because even subtle perturbations can affect the ground truth of regression tasks. We followed this setting. However, ensemble-based methods usually achieve good effectiveness, but the model's resource consumption is also larger. SWAG [59] proposes directly integrating the last few training epoch versions of the same model to optimize training time. If not otherwise stated, all hyperparameters, such as the backbone network, are set according to the settings in Table 4.

3.3. EXP-A: Comparative experiment with PSQA methods

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

Table 5 lists the comparative results of the Mean Absolute Error (MAE) for different criteria in the context of MCBN, MCDO, and the proposed MGUQ method for dose differences over various dose percentage ranges. 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 "GPR Range (%)" includes precise dose ranges, such as "90–95", "95–100", and "85–90".

The MAE values are given for each combination of criteria and GPR range. The table also categorizes the method into several main categories. The machine learning category records three traditional methods: GBDT, RF, and PL. Whereas deep learning records three further methods, including UNET++, TransQA, and CycleGAN. Lastly, uncertainty quantification-based deep learning methods are presented; the proposed MGUQ method and several others are included.

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, for uncertainty, quantification-based deep learning methods show stable performance across all methods with some variations among each. As an illustration, in the row that corresponds to the criteria of "1%/1 mm", the example MAE values for the dose percentage ranges are as follows:

- The MAE for the dose range of 95%–100% is 1.342 for MGUQ.
- The MAE for the dose range of 90%–95% is 1.798 for MGUQ.
- The MAE for the dose range of 85%–90% is 2.722 for MGUQ.

The best MAE values recorded with All ranges are 2.024, 1.246, and 0.864 for (1%/1 mm, 2%/2 mm, and 2%/3 mm) respectively. Smaller MAE values imply stronger concordance between the estimated and reference dose, indicating enhanced dose calculation or prediction procedure precision.

Table 6
Evaluation metrics for different uncertainty criteria and alpha variable setting.

Method	MMF	TPR	TNR	PCC	CA
MCBN_Gaussian	0.0	0.948	0.529	0.901	0.901
	0.1	0.944	0.538	0.906	0.914
	0.2	0.973	0.455	0.927	0.941
	0.3	0.971	0.500	0.936	0.954
	0.4	0.989	0.571	0.958	0.974
MCDO_Gaussian	0.0	0.926	0.235	0.849	0.849
	0.1	0.932	0.167	0.899	0.908
	0.2	0.925	0.250	0.903	0.921
	0.3	0.943	0.250	0.918	0.941
	0.4	0.935	0.333	0.917	0.947
MCDO_Alpha_1	0.0	0.941	0.471	0.888	0.888
	0.1	0.937	0.500	0.899	0.908
	0.2	0.956	0.455	0.911	0.928
	0.3	0.960	0.455	0.909	0.934
	0.4	0.966	0.625	0.938	0.961
MCDO_Alpha_2	0.0	0.956	0.647	0.921	0.921
	0.1	0.967	0.600	0.928	0.934
	0.2	0.973	0.545	0.935	0.947
	0.3	0.981	0.571	0.955	0.967
	0.4	0.978	0.250	0.948	0.967
MCDO_Alpha_3	0.0	0.970	0.529	0.921	0.921
	0.1	1	0.417	0.949	0.954
	0.2	1	0.333	0.952	0.961
	0.3	1	0.286	0.955	0.967
	0.4	1	0.2	0.958	0.974
Interval Network [57]	0.0	0.956	0.647	0.921	0.921
	0.1	0.967	0.600	0.928	0.934
	0.2	0.973	0.545	0.935	0.947
	0.3	0.981	0.571	0.955	0.967
	0.4	0.978	0.250	0.948	0.967
OCs [58]	0.0	0.993	0.294	0.914	0.914
	0.1	0.992	0.333	0.935	0.941
	0.2	0.991	0.333	0.927	0.941
	0.3	0.990	0.364	0.927	0.947
	0.4	0.988	0.400	0.927	0.954
SWAG (3×) [59]	0.0	0.985	0.353	0.914	0.914
	0.1	0.992	0.417	0.942	0.947
	0.2	0.991	0.556	0.960	0.967
	0.3	0.990	0.571	0.964	0.974
	0.4	1	0.6	0.979	0.987
Deep Ensemble(3×) [38]	0.0	0.926	0.647	0.895	0.895
	0.1	0.959	0.667	0.928	0.934
	0.2	0.955	0.714	0.927	0.941
	0.3	0.949	0.636	0.918	0.941
	0.4	0.954	0.667	0.927	0.954
MGUQ	0.0	**0.970**	0.529	0.921	0.921
	0.1	**1**	0.285	0.956	0.961
	0.2	**1**	0.233	0.975	0.980
	0.3	**1**	0.000	0.981	0.986
	0.328	**1**	**1**	**1**	**1**

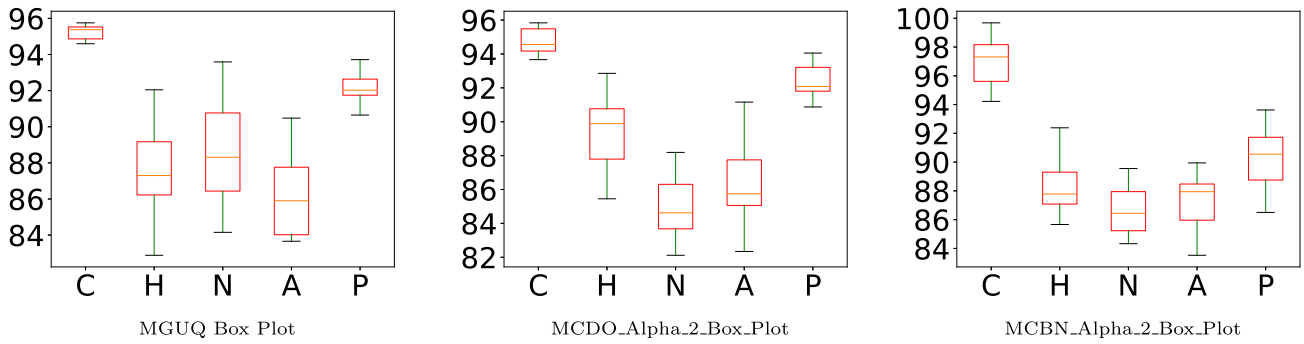


Fig. 5. Various box plots illustrating Monte Carlo variations such as MCDO (Monte Carlo dropout) and MCBN (Monte Carlo batch normalization). The labels denote body regions: C for chest, H for head, *N* for neck, A for abdominal, and P for pelvic. Among the methods, MGUQ demonstrates the most significant differences in uncertainty qualifications.

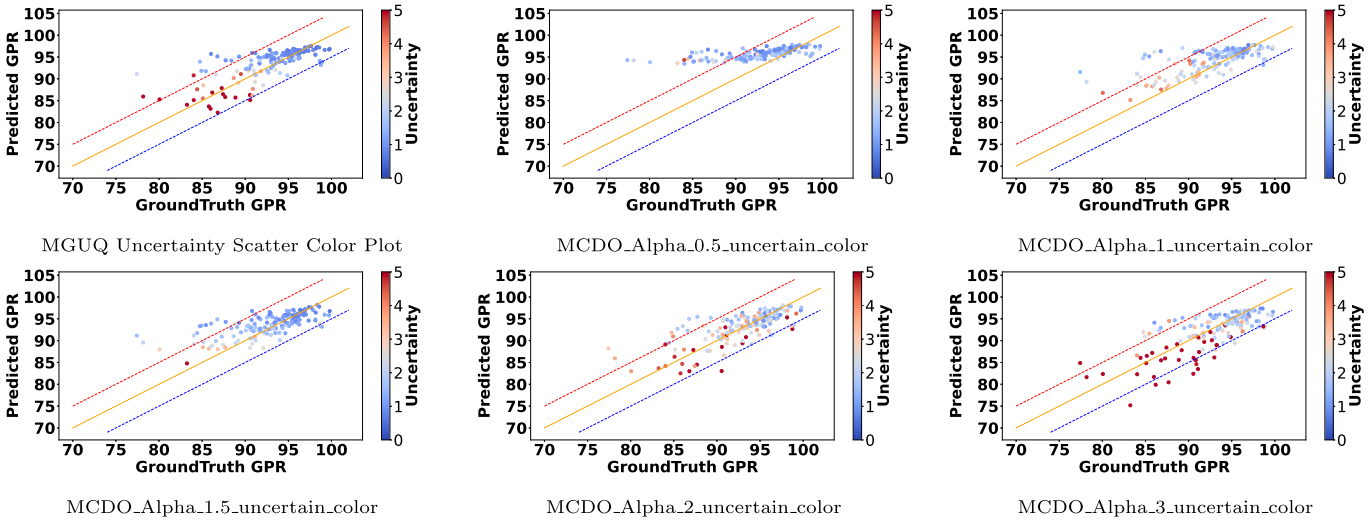


Fig. 6. Various color scatter plots with Monte Carlo variation such as MCDO (Monte Carlo Dropout) and MCBN (Monte Carlo Batch Normalization). MGUQ indicates the most uncertainty difference between the extreme and norm samples. Meanwhile, other plots are used to compare our proposed method.

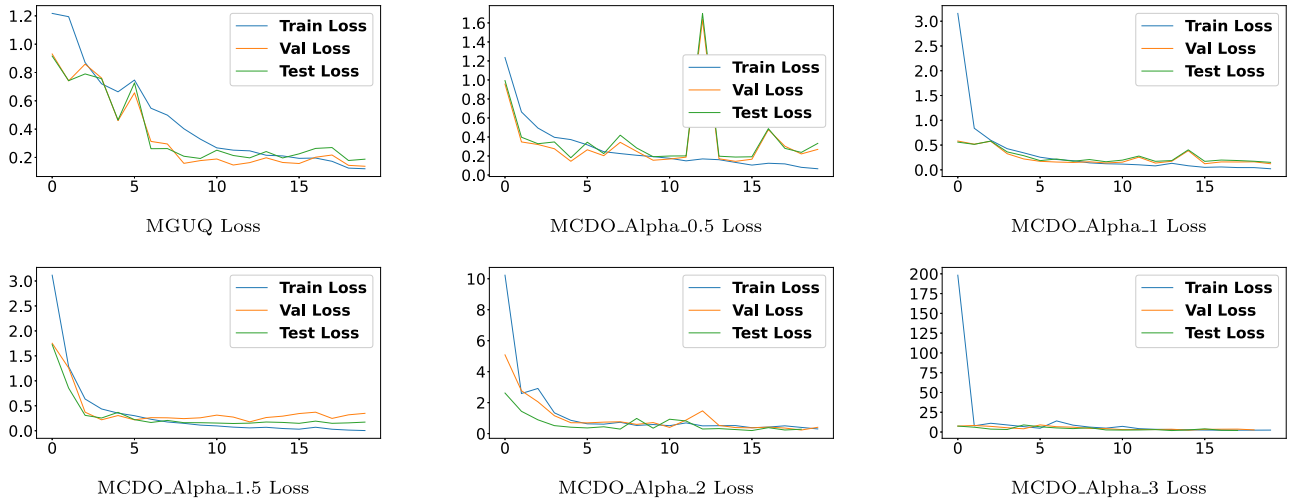


Fig. 7. Various loss plots with Monte Carlo variation such MCDO (Monte Carlo Dropout) and MCBN (Monte Carlo Batch Normalization). The first MGUQ Loss plot is our proposed method's loss plot, while other loss plots are presented for comparison with our framework. Results suggest that the proposed method MGUQ achieves significant performance.

3.4. EXP-B: Comparative experiment with UQ deep learning methods

In EXP-B, we analyze the UQ-based metrics to prove the MGUQ's superiority in security and effectiveness compared with other UQ methods.

Table 6 compares our proposed method MGUQ with common uncertainty deep learning methods. MMF is the ratio of the samples with an uncertainty score larger than the uncertainty threshold, which is regarded as uncertain. Bigger MMF means a more strict uncertainty threshold. So, we show performance in five MMFs: 0.0, 0.1, 0.2, 0.3,

and 0.4. We then present an individual comprehensive evaluation of each metric.

For the MCDO_Alpha_1 setting, we observe an improvement in TPR as the MMF increases from 0 to 0.4, indicating better identification of positive and negative instances. PCC steadily increases, showing an enhanced correlation between predicted and ground truth values. The decrease in MMF implies increased prediction stability. CA also improves, indicating better overall classification accuracy and practical application accuracy. On the other hand, TNR does not steadily improve with MMF because, in this task, many uncertain samples are negative samples due to data distribution. Therefore, as the uncertainty threshold worsens, many negative samples are excluded beyond the threshold. Similar trends are observed for the MGUQ and MCDO_Alpha_2 methods. With the decrease in MMF, TPR, TNR, PCC, and CA overall improve. For the "MCBN Gaussian" and "MCDO Gaussian", TPR and TNR values vary significantly but generally maintain a high level, such as TPR from 0.948 to 0.989, while TNR oscillates around 0.5. Lastly, consistent PCC values indicate a good correlation between predicted and ground truth values.

MGUQ converged to 1 for TPR, TNR, PCC, and CA when MMF reached 0.328 due to its good uncertainty metric performance. TNR becomes 0.0 at MMF 0.3 because the MGUQ uncertainty metric is more sensitive to rare sample patterns. Combined with PCC reaching a good value of 0.981, this indicates that most extreme samples are detected as uncertain at the MMF uncertainty threshold. In addition, when MMF reached 0.328, all the negative samples were detected as uncertain, showcasing MGUQ's security advantage.

Fig. 8 illustrates our work as the first to visualize the confidence interval of dose difference. In clinical practice, solely relying on the GPR is insufficient to reflect the rationality of a plan, as it disregards spatial information and may result in overall compliance while locally violating medical criteria. Therefore, visualizing dose differences is also a crucial task. However, as demonstrated earlier, deep learning models may have limited performance on extreme data. For data the model is uncertain about, predicting its confidence interval boundaries rather than a single value is secure and more valid. Our implementation's dose difference prediction task output is modeled as a multivariate independent Gaussian distribution. Therefore, we visualize the predicted dose difference by representing the $[\mu - 1.96 \cdot diag(\Lambda), \mu + 1.96 \cdot diag(\Lambda)]$, which covers a 95% confidence interval. These visualizations offer a nuanced understanding of the model's explainability, enhancing clinical practice's deep learning security and trustworthiness.

Tables 5 and 6 demonstrate that MGUQ performs better than other uncertainty quantification techniques regarding different criteria and settings. MGUQ substantially reduces MAE for dose variations, demonstrating accurate prediction and adherence to reference doses. Furthermore, MGUQ consistently attains high scores for CA, PCC, TPR, and overall PSQA metrics, highlighting its effectiveness in precisely capturing dose distributions and identifying specific dose regions. These findings demonstrate that MGUQ is strong and dependable in increasing the accuracy and efficiency of PSQA in radiation therapy, making it a potential method for strengthening the quality assurance process in clinical settings.

Furthermore, Figs. 5 to 7 shows various loss plots, color scatter plots with uncertainty, and box plots showcasing Monte Carlo variations such as MCDO and MCBN. The presented plots are divided into three portions, each of which includes several sub-plots for showcasing performances with Monte Carlo variation on Alpha α methods such as MCDO_Alpha_0.5, MCDO_Alpha_1, ⋯ and MCDO_Alpha_3. This illustration provides a detailed visual representation of the performance and uncertainty features of the Monte Carlo variations in the studied models. As shown, MGUQ measured uncertainty differs the most for the out-of-distribution and extreme samples. The figures provided thoroughly compare Monte Carlo variations, such as MGUQ, MCDO, and MCBN, using box, scatter, and loss plots. Fig. 5 displays box plots illustrating the varied uncertainty qualification levels in different body regions. Among them, MGUQ exhibits the most notable variances. Fig. 6 displays scatter plots demonstrating MGUQ's ability to detect substantial uncertainty differences between extreme and normal samples. It also offers a comparative study with other approaches. The loss graphs in Fig. 7 highlight the superior performance of MGUQ since it continuously outperforms other approaches at various α levels. Together, these statistics provide useful knowledge about the effectiveness of MGUQ in measuring uncertainty and its prospective uses in other fields.

3.5. EXP-C: Ablation experiment with Alpha

This section explores ablation investigations, concentrating on optimizing the parameter α in our experimental paradigm.

Fig. 9 shows Alpha plots at different training dataset sizes starting from 500 to 1200 samples with α values from 0 to 2. When α equals 0, it is a single-level approach, whereas > 0 is active multi-granularity coherence loss. The larger the size of the training dataset, the lower the MAE will be across all values. This effect is more pronounced for uncommon data (< 90). If the amount of data reaches a certain level, it is advantageous to choose a larger α for predicting common values; otherwise, it will have the opposite effect; regardless of whether the amount of data reaches a certain level, α is advantageous for predicting uncommon values. A larger α is more advantageous for predicting uncommon values when the data is small; these analyses validate the initial observations of the dataset as depicted in Fig. 1. This analysis emphasizes that the coarse-grained extreme values in samples may not be extreme on a fine-grained level. Therefore, modeling uncertainties at multiple levels can potentially improve PSQA.

3.6. EXP-D: Medical application generalization

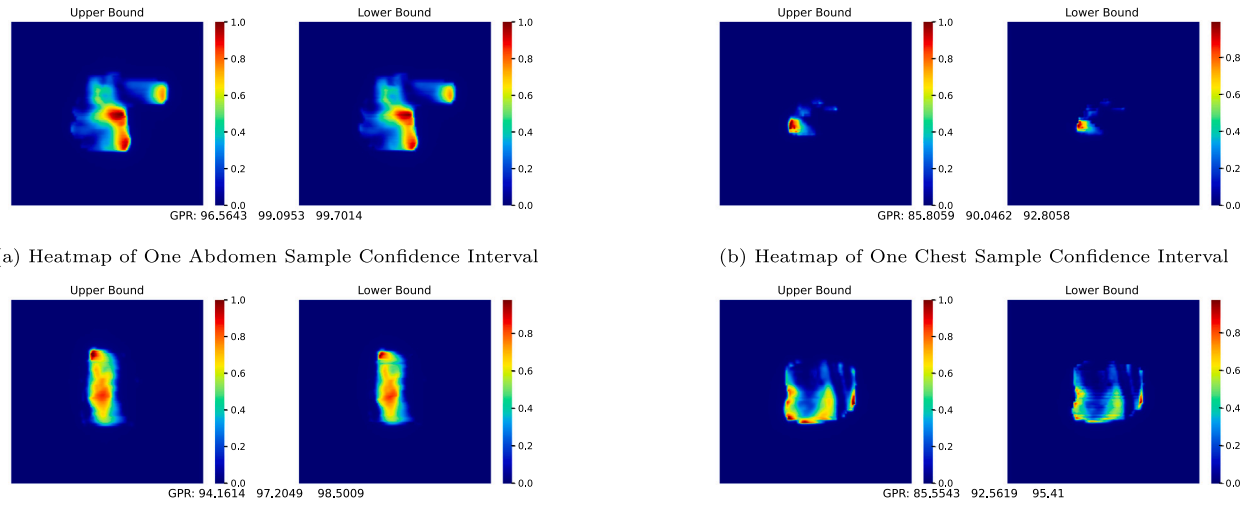
In EXP-D, we discuss the generalization of medical applications leveraging MGUQ to evaluate the generalization capability of the proposed method.

Although the MGUQ method is specifically designed for PSQA and has shown promising results, it would be more encouraging to prove whether the MGUQ framework can be generalized to other medical tasks. On the other hand, due to the relatively small proprietary dataset, further evidence is needed to demonstrate that MGUQ performance improvements are not incidental. Therefore, we tested several medical tasks that can be applied to this framework. However, PSQA is a dual-granularity regression task, which is rare in the medical deep-learning domain, and dual-granularity classification tasks are more common. Following the method derived in the previous section, we can provide the loss formulation for dual granularity classification tasks. Dual-granularity classification can model some medical segmentation scenarios and medical classification tasks, such as the major category and sub-category.

Assuming we have two tasks with different granularities, where Task 2 has a higher granularity compared to Task 1. y_1 is the prediction for Task 1, with a Dirichlet prior Dir(a), and y_2 is the prediction for Task 2, with a Dirichlet prior Dir(b). The loss formulation is as follows:

$$L = E[\ln P(y_1|a)] + E_{Dir(b|x,\theta)}[-\ln P(Y_2|b)] + KL(Dir(b|x,\theta) \| Dir(a))$$

As shown in Table 7, Breakhis and Ham10000 are two common datasets for cancer classification. They both have a label structure consisting of a major class (benign and malignant cases) and subclasses representing specific types of cancer. This aligns well with our assumption. We adopt a dataset split ratio of 7:1:2, and further other parameters we follow as discussed in Table 4. We evaluate three popular neural network architectures: ResNet, EfficientNet, and MobileViT. We tested the accuracy using our proposed framework (w/) and without the proposed framework (w/o). Adopting MGUQ's coherence loss primarily benefits the major categories but does not



(a) Heatmap of One Abdomen Sample Confidence Interval

(b) Heatmap of One Chest Sample Confidence Interval

(c) Heatmap of One Head Sample Confidence Interval

(d) Heatmap of One Pelvic Sample Confidence Interval

Fig. 8. DDP confidence interval bound heatmap plots for various lesioned.

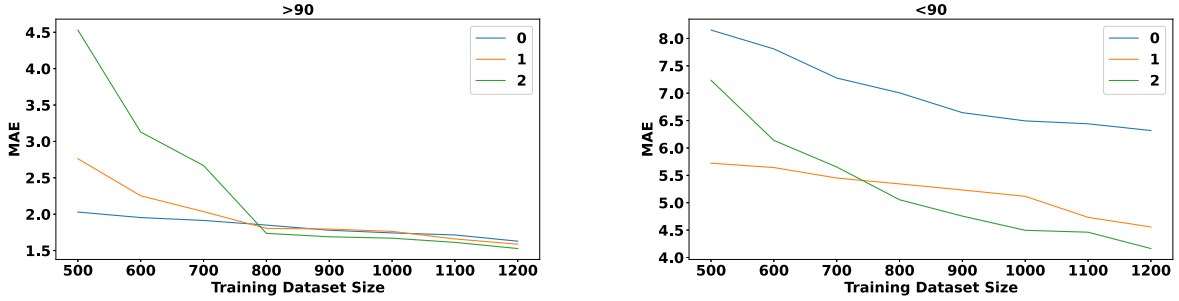


Fig. 9. Comparative analysis of various aplha plots at different training dataset size.

Table 7
Performance comparison.

Task	Method	Performance	
		Major category	Sub-category
Breakhis [60]	Resnet18(w/o)	0.91	0.48
	Resnet18(w/)	0.93	0.49
	Efficientnet(w/o)	0.95	0.51
	Efficientnet(w/)	0.96	0.51
	MobileVit(w/o)	0.93	0.49
	MobileVit(w/)	0.95	0.49
Ham10000 [61]	Resnet18(w/o)	0.92	0.87
	Resnet18(w/)	0.93	0.89
	Efficientnet(w/o)	0.95	0.81
	Efficientnet(w/)	0.96	0.81
	MobileVit(w/o)	0.97	0.83
	MobileVit(w/)	0.98	0.82

show clear advantages for the sub-categories. This is likely because coherence loss between two granularities captures label structure information about mappings between major and sub-categories, which is advantageous for the major categories but, to some extent, redundant for the sub-categories.

3.7. Computational resources utilization

This section showcases the resource consumption of the tested method and proves MGUQ availability in practice.

Table 8
Comparative analysis of computational resources utilization by baseline and the proposed MGUQ for uncertainty quantification methods in deep learning.

Method name	Flops (GMac)	Param size(M)
Unet++	63.7	15.9
TranQA	240.34	215.38
Cyclegan	42.7	28.6
Deep Ensemble	63.9	43.0
SWAG	63.9	43.0
MGUQ	**42.6**	**28.6**

A comparison of computational resource use between baseline approaches and the proposed MGUQ method is shown in Table 8. The table shows the method name, Flops (GMac) (billion multiply-accumulate operations), and Parameter Size (millions). All methods are deep learning models. Unet++, TranQA and Cyclegan are baseline PSQA approaches with different Flops and Parameter Size computational requirements. Deep Ensemble and SWAG are UQ deep learning baseline approaches that use the same computational resources. However, we skipped some UQ deep learning methods due to their similar resource consumption because they do not specify the architecture. The suggested MGUQ technique uses fewer computational resources than other baseline methods to reach this performance, as seen by its reduced Flops and Parameter Size values. The proposed MGUQ method quantifies uncertainty in deep learning models more efficiently, making it a suitable candidate for applications with limited computational resources or optimization needs.

4. Conclusion

This study categorizes and summarizes the current PSQA work into three main categories and analyzes the limitations. Only deep learning methods can fully meet the clinical demand by predicting two granularities of tasks, namely GPR and DDP. However, the lack of transparency in deep learning methods presents significant challenges to the explainability and security of algorithms, greatly limiting the clinical significance of deep learning in PSQA applications and restricting our integration of prior knowledge with deep learning. In this work, we adopt a Bayesian framework based on prior knowledge of PSQA dual-granularity regression tasks, analyze and derive probabilistic models, and thereby obtain a loss function that can express the inherent consistency information and probabilistic nature of multi-granularity tasks. We then correspondingly design the dual-stream network and training evaluation algorithm to form the MGUQ framework to improve deep learning's effectiveness, security, and trustworthiness in PSQA.

Experiments show that our approach achieves state-of-the-art performance compared to traditional and benchmark methods in PSQA, and it is the first work to visualize confidence interval bounds for dose difference distribution tasks, which is highly relevant to clinical needs. In terms of PSQA uncertainty measurement, MGUQ demonstrates that the correlation of multi-granularity tasks can optimize the performance of uncertainty quantification, thereby enhancing security and interpretability. This highlights its potential for real-world clinical implementation. The study advances the field of DL-based PSQA and provides valuable insights for safer and more reliable radiation therapy practices. Consequently, clinical staff can make targeted planning adjustments accordingly.

However, the MGUQ framework still has some limitations. Although this study extends MGUQ to dual-granularity classification tasks, its limitations remain in its limited prior assumption and application domain. Theoretically, the probabilistic modeling process requires that the low-granularity task be approximately regarded as a probabilistic sum of high-granularity tasks. Future research can further explore improvements and extend our method to more complex scenarios.

Dataset acknowledgment

We sincerely thank the "Peeking Union Medical College Hospital" for granting us access to their invaluable dataset.

Funding acknowledgment

We acknowledge that the paper received financial support from Birmingham City University, United Kingdom.

CRediT authorship contribution statement

Xiaoyang Zeng: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Qizhen Zhu:** Resources, Project administration, Investigation, Formal analysis, Data curation, Conceptualization. **Awais Ahmed:** Writing – review & editing, Writing – original draft, Visualization. **Muhammad Hanif:** Writing – review & editing, Writing – original draft, Visualization. **Mengshu Hou:** Writing – review & editing. **Qiu Jie:** Resources, Data curation. **Rui Xi:** Writing – review & editing. **Syed Attique Shah:** Writing – review & editing, Funding acquisition.

Declaration of competing interest

None Declared.

References

[1] L. Deng, Y. Ji, S. Huang, X. Yang, J. Wang, Synthetic CT generation from CBCT using double-chain-CycleGAN, Comput. Biol. Med. 161 (2023) 106889, http://dx.doi.org/10.1016/j.compbiomed.2023.106889.

[2] A. Jabbarpour, S.R. Mahdavi, A.V. Sadr, G. Esmaili, I. Shiri, H. Zaidi, Unsupervised pseudo CT generation using heterogenous multicentric CT/MR images and cyclegan: Dosimetric assessment for 3D conformal radiotherapy, Comput. Biol. Med. 143 (2022) 105277, http://dx.doi.org/10.1016/j.compbiomed.2022.105277.

[3] J. Sheng, S. Lam, J. Zhang, Y. Zhang, J. Cai, Multi-omics fusion with soft labeling for enhanced prediction of distant metastasis in nasopharyngeal carcinoma patients after radiotherapy, Comput. Biol. Med. (2023) 107684, http://dx.doi.org/10.1016/j.compbiomed.2023.107684.

[4] X. Zhou, W. Cai, J. Cai, F. Xiao, M. Qi, J. Liu, L. Zhou, Y. Li, T. Song, Multimodality MRI synchronous construction based deep learning framework for MRI-guided radiotherapy synthetic CT generation, Comput. Biol. Med. 162 (2023) 107054, http://dx.doi.org/10.1016/j.compbiomed.2023.107054.

[5] Y. Wang, Y. Sun, K. Gan, J. Yuan, H. Xu, H. Gao, X. Zhang, Bone marrow sparing oriented multi-model image registration in cervical cancer radiotherapy, Comput. Biol. Med. 166 (2023) 107581, http://dx.doi.org/10.1016/j.compbiomed.2023.107581.

[6] A. Meijers, G.G. Marmitt, K.N.W. Siang, A. van der Schaaf, A.C. Knopf, J.A. Langendijk, S. Both, Feasibility of patient specific quality assurance for proton therapy based on independent dose calculation and predicted outcomes, Radiother. Oncol. 150 (2020) 136–141, http://dx.doi.org/10.1016/j.radonc.2020.06.027.

[7] J.M. Park, J.-i. Kim, S.-Y. Park, D.H. Oh, S.-T. Kim, Reliability of the gamma index analysis as a verification method of volumetric modulated arc therapy plans, Radiation Oncol. 13 (1) (2018) 1–14, http://dx.doi.org/10.1186/s13014-018-1123-x.

[8] Y. Huang, Y. Pi, K. Ma, X. Miao, S. Fu, H. Chen, H. Wang, H. Gu, Y. Shao, Y. Duan, et al., Image-based features in machine learning to identify delivery errors and predict error magnitude for patient-specific IMRT quality assurance, Strahlentherapie und Onkol. 199 (5) (2023) 498–510, http://dx.doi.org/10.1007/s00066-023-02076-8.

[9] S. Yoganathan, S. Ahmed, S. Paloor, T. Torfeh, S. Aouadi, N. Al-Hammadi, R. Hammoud, Virtual pretreatment patient-specific quality assurance of volumetric modulated arc therapy using deep learning, Med. Phys. 50 (12) (2023) 7891–7903, http://dx.doi.org/10.1002/mp.16567.

[10] C. Wu, D. Nguyen, J. Schuemann, A. Mairani, Y. Pu, S. Jiang, Applications of artificial intelligence in particle radiotherapy, 2021, http://dx.doi.org/10.48550/arXiv.2102.03061, arXiv preprint arXiv:2102.03061.

[11] 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, http://dx.doi.org/10.1016/j.radonc.2020.09.008.

[12] V. Liesbeth, C. Michal, M.D. Anna, L.B. Charlotte, V.E. Wouter, Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance, Radiother. Oncol. (2020) http://dx.doi.org/10.1016/j.radonc.2020.09.008.

[13] 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, http://dx.doi.org/10.1002/mp.13669.

[14] D.A. Granville, J.G. Sutherland, J.G. 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, http://dx.doi.org/10.1088/1361-6560/ab142e.

[15] G. Valdes, M.F. Chan, S.B. Lim, R. Scheuermann, J.O. Deasy, T.D. Solberg, IMRT QA using machine learning: a multi-institutional validation, J. Appl.clinical Med. Phys. 18 (5) (2017) 279–284, http://dx.doi.org/10.1002/acm2.12161.

[16] 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, http://dx.doi.org/10.1002/mp.13112.

[17] Y. Interian, V. Rideout, V.P. Kearney, E. Gennatas, O. Morin, J. Cheung, T. Solberg, G. Valdes, Deep nets vs expert designed features in med. phys.: an IMRT qa case study, Med. Phys. 45 (6) (2018) 2672–2680, http://dx.doi.org/10.1002/mp.12890.

[18] G. Valdes, R. Scheuermann, C. Hung, A. Olszanski, M. Bellerive, T. Solberg, A mathematical framework for virtual IMRT QA using machine learning, Med. Phys. 43 (7) (2016) 4323–4334, http://dx.doi.org/10.1118/1.4953835.

[19] W. Du, S.H. Cho, X. Zhang, K.E. Hoffman, R.J. Kudchadker, Quantification of beam complexity in intensity-modulated radiation therapy treatment plans, Med. Phys. 41 (2) (2014) 021716, http://dx.doi.org/10.1118/1.4861821.

[20] S. Crowe, T. Kairn, J. Kenny, R. Knight, B. Hill, C.M. Langton, J. Trapp, Treatment plan complexity metrics for predicting IMRT pre-treatment quality assurance results, Australasian Phys. Eng. Sci. Med. 37 (3) (2014) 475–482, http://dx.doi.org/10.1007/s13246-014-0274-9.

[21] L. Masi, R. Doro, V. Favuzza, S. Cipressi, L. Livi, Impact of plan parameters on the dosimetric accuracy of volumetric modulated arc therapy, Med. Phys. 40 (7) (2013) 071718, http://dx.doi.org/10.1118/1.4810969.

[22] M.C. Glenn, V. Hernandez, J. Saez, D.S. Followill, R.M. Howell, J.M. Pollard-Larkin, S. Zhou, S.F. Kry, Treatment plan complexity does not predict IROC houston anthropomorphic head and neck phantom performance, Phys. Med. Biol. 63 (20) (2018) 205015, http://dx.doi.org/10.1088/1361-6560/aae29e.

[23] P.D. Wall, J.D. Fontenot, Quality assurance-based optimization (QAO): Towards improving patient-specific quality assurance in volumetric modulated arc therapy plans using machine learning, Phys. Medica 87 (2021) 136–143, http://dx.doi.org/10.1016/j.ejmp.2021.03.017.

[24] N. Ishizaka, T. Kinoshita, M. Sakai, S. Tanabe, H. Nakano, S. Tanabe, S. Nakamura, K. Mayumi, S. Akamatsu, T. Nishikata, et al., Prediction of patient-specific quality assurance for volumetric modulated arc therapy using radiomics-based machine learning with dose distribution, J. Appl. Clinical Med. Phys. 25 (1) (2024) e14215, http://dx.doi.org/10.1002/acm2.14215.

[25] S. Raj, V. Sundaram, H.F. Godson, R. John, Wobbling nature of gamma passing rate as a function of calibration field sizes in patient-specific quality assurance, J. Radiother. Pract. 23 (2024) e2, http://dx.doi.org/10.1017/S1460396923000444.

[26] H. Zhu, Q. Zhu, Z. Wang, B. Yang, W. Zhang, J. Qiu, Patient-specific quality assurance prediction models based on machine learning for novel dual-layered MLC linac, Med. Phys. 50 (2) (2023) 1205–1214, http://dx.doi.org/10.1002/mp.16091.

[27] C. Noblet, M. Maunet, M. Duthy, F. Coste, M. Moreau, A TPS integrated machine learning tool for predicting patient-specific quality assurance outcomes in volumetric-modulated arc therapy, Phys. Medica 118 (2024) 103208, http://dx.doi.org/10.1016/j.ejmp.2024.103208.

[28] L. Zeng, M. Zhang, Y. Zhang, Z. Zou, Y. Guan, B. Huang, X. Yu, S. Ding, Q. Liu, C. Gong, TransQA: deep hybrid transformer network for measurement-guided volumetric dose prediction of pre-treatment patient-specific quality assurance, Phys. Med. Biol. 68 (20) (2023) 205010, http://dx.doi.org/10.1088/1361-6560/acfa5e.

[29] Y. Huang, Y. Pi, K. Ma, X. Miao, S. Fu, H. Chen, et al., Virtual patient-specific quality assurance of IMRT using unet++: classification, gamma passing rates prediction, and dose difference prediction., Front Oncol 11 (2021) 700343, http://dx.doi.org/10.3389/fonc.2021.700343.

[30] W. Liu, L. Zhang, L. Xie, T. Hu, G. Li, S. Bai, Z. Yi, Multilayer perceptron neural network with regression and ranking loss for patient-specific quality assurance, Knowl.-Based Syst. 271 (2023) 110549, http://dx.doi.org/10.1016/j.knosys.2023.110549.

[31] G. Li, X. Wu, X. Ma, Artificial intelligence in radiotherapy, in: Seminars in Cancer Biology, 2022, http://dx.doi.org/10.1016/j.rpor.2020.03.015.

[32] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Inf. Fusion 76 (2021) 243–297, http://dx.doi.org/10.1016/j.inffus.2021.05.008.

[33] X. Yang, S. Li, Q. Shao, Y. Cao, Z. Yang, Y.-q. Zhao, Uncertainty-guided man–machine integrated patient-specific quality assurance, Radiother. Oncol. 173 (2022) 1–9, http://dx.doi.org/10.1016/j.radonc.2022.05.016.

[34] M. Stasi, S. Bresciani, A. Miranti, A. Maggio, V. Sapino, P. Gabriele, Pretreatment patient-specific IMRT quality assurance: a correlation study between gamma index and patient clinical dose volume histogram, Med. Phys. 39 (12) (2012) 7626–7634, http://dx.doi.org/10.1118/1.4767763.

[35] S.F. Kry, A. Molineu, J.R. Kerns, A.M. Faught, J.Y. Huang, K.B. Pulliam, J. Tonigan, P. Alvarez, F. Stingo, D.S. Followill, Institutional patient-specific IMRT QA does not predict unacceptable plan delivery, Int. J. Radiat. Oncol. Biol. Phys. 90 (5) (2014) 1195–1201, http://dx.doi.org/10.1016/j.ijrobp.2014.08.334.

[36] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059, https://proceedings.mlr.press/v48/gal16.html?trk=public_post_comment-text.

[37] J. Mukhoti, A. Kirsch, J. van Amersfoort, P.H. Torr, Y. Gal, Deterministic neural networks with inductive biases capture epistemic and aleatoric uncertainty, 2021, arXiv preprint arXiv:2102.11582.

[38] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Adv. Neural Inf. Process. Syst. 30 (2017) http://dx.doi.org/10.48550/arXiv.1612.01474.

[39] G. Huang, Y. Li, G. Pleiss, Z. Liu, J.E. Hopcroft, K.Q. Weinberger, Snapshot ensembles: Train 1, get m for free, 2017, http://dx.doi.org/10.48550/arXiv.1704.00109, arXiv preprint arXiv:1704.00109.

[40] A. Mobiny, P. Yuan, S.K. Moulik, N. Garg, C.C. Wu, H. Van Nguyen, Dropconnect is effective in modeling uncertainty of bayesian deep networks, Sci. Rep. 11 (1) (2021) 5458, http://dx.doi.org/10.1038/s41598-021-84854-x.

[41] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30 (2017) https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html.

[42] J. Gast, S. Roth, Lightweight probabilistic deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3369–3378, http://dx.doi.org/10.1109/CVPR.2018.00355.

[43] A. Loquercio, M. Segu, D. Scaramuzza, A general framework for uncertainty estimation in deep learning, IEEE Robot. Autom. Lett. 5 (2) (2020) 3153–3160, http://dx.doi.org/10.1109/LRA.2020.2974682.

[44] X. Yang, S. Li, Q. Shao, Y. Cao, Z. Yang, Y.Q. Zhao, Uncertainty-guided man-machine integrated patient-specific quality assurance, Radiother. Oncol. J. Eur. Soc. Therapeutic Radiol. Oncol. 173 (2022) 1–9, http://dx.doi.org/10.1016/j.radonc.2022.05.016.

[45] M. Miften, A. Olch, D. Mihailidis, TG 218: Tolerance limits and methodologies for IMRT measurement-based verification QA: recommendations of AAPM task group no. 218, Med. Phys. 45 (2018) http://dx.doi.org/10.1002/mp.12810.

[46] K.P. Murphy, Machine learning: a probabilistic perspective, MIT Press, 2012, https://books.google.co.jp/books?hl=en&lr=&id=RC43AgAAQBAJ&oi=fnd&pg=PR7&dq=Murphy,+Machine+learning:+a+probabilistic+perspective&ots=ungvfAPs-c&sig=SkmNtFtMmZ69FtPA3W4lSYNKIW4&redir_esc=y#v=onepage&q=Murphy.

[47] C.M. Bishop, N.M. Nasrabadi, Pattern recognition and machine learning, vol. 4, (no. 4) Springer, 2006, http://cs.haifa.ac.il/~rita/uml_course/lectures/prml8_update_p1.pdf.

[48] A. Malinin, M. Gales, Predictive uncertainty estimation via prior networks, Adv. Neural Inf. Process Syst 31 (2018) https://proceedings.neurips.cc/paper/2018/hash/3ea2db50e62ceefceaf70a9d9a56a6f4-Abstract.html.

[49] D.P. Kingma, M. Welling, Auto-encoding variational bayes, 2013, http://dx.doi.org/10.48550/arXiv.1312.6114, arXiv preprint arXiv:1312.6114.

[50] 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, http://dx.doi.org/10.1109/CVPR.2016.90.

[51] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu, Ce-net: Context encoder network for 2d medical image segmentation, IEEE Trans. Med. Imaging 38 (10) (2019) 2281–2292, http://dx.doi.org/10.1109/TMI.2019.2903562.

[52] A.A. Ardakani, A.R. Kanafi, U.R. Acharya, N. Khadem, A. Mohammadi, Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks, Comput. Biol. Med. 121 (2020) 103795, http://dx.doi.org/10.1016/j.compbiomed.2020.103795.

[53] M. Teye, H. Azizpour, K. Smith, Bayesian uncertainty estimation for batch normalized deep networks, in: International Conference on Machine Learning, PMLR, 2018, pp. 4907–4916, https://proceedings.mlr.press/v80/teye18a.html?fbclid=IwAR1j0fZGbKLNdpSX2GEeT95X8PYtCHZv-4-Is9Y8I_dgifnzDp01mkDW_C4.

[54] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232, http://dx.doi.org/10.48550/arXiv.1703.10593.

[55] Y. Gal, Z. Ghahramani, Bayesian convolutional neural networks with Bernoulli approximate variational inference, 2015, http://dx.doi.org/10.48550/arXiv.1506.02158, arXiv preprint arXiv:1506.02158.

[56] T. Kawashima, Q. Yu, A. Asai, D. Ikami, K. Aizawa, The aleatoric uncertainty estimation using a separate formulation with virtual residuals, in: 2020 25th International Conference on Pattern Recognition, ICPR, IEEE, 2021, pp. 1438–1445, http://dx.doi.org/10.1109/ICPR48806.2021.9412324.

[57] L. Oala, C. Heiß, J. Macdonald, M. März, W. Samek, G. Kutyniok, Interval neural networks: Uncertainty scores, 2020, http://dx.doi.org/10.48550/arXiv.2003.11566, arXiv preprint arXiv:2003.11566.

[58] N. Tagasovska, D. Lopez-Paz, Single-model uncertainties for deep learning, Adv. Neural Inf. Process. Syst. 32 (2019) http://dx.doi.org/10.48550/arXiv.1811.00908.

[59] W.J. Maddox, P. Izmailov, T. Garipov, D.P. Vetrov, A.G. Wilson, A simple baseline for bayesian uncertainty in deep learning, Adv. Neural Inf. Process. Syst. 32 (2019) http://dx.doi.org/10.48550/arXiv.1902.02476.

[60] F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, A dataset for breast cancer histopathological image classification, Ieee Trans. Biomed. Eng. 63 (7) (2015) 1455–1462, http://dx.doi.org/10.1109/TBME.2015.2496264.

[61] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (1) (2018) 1–9, http://dx.doi.org/10.1038/sdata.2018.161.