Deep Evidential Learning for Dose Prediction

H.S.Tan[1], Kuancheng Wang[2], and Rafe Mcbeth[1]

[1] *University of Pennsylvania, Perelman School of Medicine, Department of Radiation Oncology, Philadelphia, USA*
[2] *Georgia Institute of Technology, Atlanta, GA, USA*

Abstract

In this work, we present a novel application of an uncertainty-quantification framework called Deep Evidential Learning in the domain of radiotherapy dose prediction. Using medical images of the Open Knowledge-Based Planning Challenge dataset, we found that this model can be effectively harnessed to yield uncertainty estimates that inherited correlations with prediction errors upon completion of network training. This was achieved only after reformulating the original loss function for a stable implementation. We found that (i)epistemic uncertainty was highly correlated with prediction errors, with various association indices comparable or stronger than those for Monte-Carlo Dropout and Deep Ensemble methods, (ii)the median error varied with uncertainty threshold much more linearly for epistemic uncertainty in Deep Evidential Learning relative to these other two conventional frameworks, indicative of a more uniformly calibrated sensitivity to model errors, (iii)relative to epistemic uncertainty, aleatoric uncertainty demonstrated a more significant shift in its distribution in response to Gaussian noise added to CT intensity, compatible with its interpretation as reflecting data noise. Collectively, our results suggest that Deep Evidential Learning is a promising approach that can endow deep-learning models in radiotherapy dose prediction with statistical robustness. Towards enhancing its clinical relevance, we demonstrate how we can use such a model to construct the predicted Dose-Volume-Histograms' confidence intervals.

Contents

1 Introduction

Radiotherapy treatment planning is a highly intricate process requiring collaborative interactions among medical physicists, dosimetrists and radiation oncologists in formulating the optimal treatment plan (see e.g. [1, 2]) for each individual patient. It is a procedure that is characterized by a potentially large degree of user-variability arising from different inter-institutional guidelines, the fundamental origin of which lies in the subjective nature of tradeoffs like those between tumor control and sparing of normal tissues, etc. [1, 3, 4] Automation by machine learning methods can help to homogenize these variations among the medical professionals, while attaining improvements in the consistency of overall plan quality (see e.g. [5]). The adoption of A.I.-related techniques led to what is known as knowledge-based planning (KBP) [6], which leverages knowledge inherent in past clinical treatment plans to generate new ones with minimal intervention from human experts.

A complete KBP method can normally be regarded as a two-stage pipeline [7] : (i)prediction of the dose distribution that should be delivered to patient (ii)conversion of the prediction into a deliverable treatment plan via optimization. Recent dose prediction models cover quite a range of

sites and modalities [5, 8], including prostate intensity-modulated-radiation-therapy (IMRT) [9, 10], prostate volumetric modulated arc therapy (VMAT) [11], lung IMRT [12] and head and neck VMAT [13]. These models were designed to predict volumetric dose distributions from which dose-volume histogram (DVH) and other dose constraint-related statistics can be deduced. Within the literature, it has been noted [6] that the vast majority of published works was performed using large private datasets which made comparison of model quality challenging. Towards alleviating this issue, the Open Knowledge-Based Planning (OpenKBP) Grand Challenge was organized [7, 14] to enable an international effort for the comparison of dose prediction models on a single open dataset involving 340 patients of head and neck cancer treated by IMRT. To date, there has been many excellent dose prediction models trained on the OpenKBP dataset as reviewed in [5]. Notable examples include *DeepDoseNet* of [15] which combined features of *ResNet* and *DenseNet*, the *MtAA-NET* of [16] which is based on a generative adversarial network, and *TrDosePred* of [17] and *Swin UNETR++* of [18] which are Transformer-based frameworks.[i]. Like the above-mentioned projects, our work here leverages upon the OpenKBP dataset and attempts to furnish dose prediction. However, unlike the majority of these papers, the major focus of our work here lies in developing a dose prediction model that also encapsulates an *uncertainty quantification* framework.

As we navigate towards integrating deep learning methods in the real clinic, a safety-related concern lies in whether and how the model can express its own uncertainty when making predictions based on real-life data distributions beyond those that they were trained on. Uncertainty quantification has been a major theme in the broader context of machine learning [19], and has recently gathered increasing attention in the domain of medical image analysis [20]. However, to our knowledge, there is a scarcity of works devoted specifically to uncertainty quantification in dose prediction models. As already noted in papers dedicated to medical image analysis [20, 21, 22, 23, 24], uncertainty estimates equip deep learning model with statistical robustness and a measure of reliability that can be useful in discovering regions of the model's ignorance, thereby alerting the human expert to potentially erroneous predictions. In the domain of dose prediction, uncertainty estimates can be used to characterize the reliability of a model. In the KBP pipeline, dose prediction models are themselves inputs to some dose mimicking model [7] which generates deliverable treatments via optimization of a set of objective functions [25, 26]. Apart from using reliability as a selection criterion for various dose prediction models, it was pointed out in [7, 27] that probabilistic dose distributions can serve as robust inputs to the objective functions. At the time of writing, we are only aware of [8] and [28] being the only publications that examined the role of uncertainty quantification frameworks for dose prediction. The work in [8] studied how Monte-Carlo Dropout (MC Dropout) and a Deep Ensemble-based bagging method can furnish useful uncertainty estimates for a U-Net-based dose prediction model, whereas [28] proposed a Gaussian mixture model in which the standard deviation of each Gaussian mixture is part of the network's outputs and can be used to reflect the tradeoff implicit in data from two different treatment protocols.

In this work, we examined the applicability of a relatively new uncertainty quantification framework known as *Deep Evidential Learning* [29, 30] for dose prediction. Its fundamental principle lies in asserting a higher-order Bayesian prior over the probabilistic neural network output, embedding it within the loss function and subsequently generating uncertainty estimates together with the completion of model training. There are two major variants of it as proposed in the seminal works of Sensoy et al. in [30] ['*Deep Evidential Classification*'] and Amini et al. in [29] ['*Deep Evidential Regression*']. We apply the formalism of [29] to construct a dose prediction model with 3D U-Net [31] as the backbone model architecture. In the process, we found that this approach required additional crucial refinements for an effective and stable implementation. These refinements are

[i]For a more complete list, see Table 5 of [5].

related to the form of the final model layer which connects inputs and weights to the parameters of the higher-order Bayesian prior distribution. These parameters are formulated as the model's outputs, and are thus naturally obtained when training is completed. They are then used to furnish estimates of model uncertainty. This is in contrast to two major categories of uncertainty quantification frameworks in the current literature[ii]: (i) MC Dropout (ii) Deep Ensemble method. For MC Dropout [33], one typically inserts stochastic dropout variables at various points of the neural networks, and after training, a number of feedforward passes are performed to evaluate the mean and variance of the output. The latter is then regarded as the model uncertainty. For Deep Ensemble method [34, 35], one typically collects a family of models sharing the main network structure but differing in hyperparameters or simply the initial random weight distributions. From the ensemble set of outputs, the model prediction is taken as the average with uncertainty being the variance.

A prominent feature of Deep Evidential Learning is that the higher-order distribution parameters can be used to compute two different types of uncertainties: the aleatoric and epistemic uncertainties. The former probes the level of noise in data while the latter characterizes the uncertainty intrinsic to the model (see e.g. [36] for a detailed explanation). Distinguishing between these two classes of uncertainties is useful since aleatoric uncertainty is largely indicative of noise level in data whereas epistemic uncertainty points to model insufficiency to generalize beyond the training dataset, be it a problem of data insufficiency itself or an unsuitable order of model complexity for the task, etc. In our work, we demonstrated that these two uncertainty types can indeed be differentiated by a noise-sensitivity test and the degree to which each was associated with prediction errors.

Like in [8, 15, 16, 17], our study is based on the OpenKBP dataset, and involved constructing a dose prediction model as part of the KBP pipeline. As emphasized in [7], although the mean absolute error (MAE) is an indicator of the model's feasibility as a dose prediction system, one should bear in mind that dose prediction models enact an intermediate role in the complete KBP pipeline. Ultimately, it is the deliverable treatment plan most compatible with various clinical criteria that we wish to generate. For the OpenKBP Challenge, while prediction models with better MAE dose score generally led to treatment plans with higher criteria satisfaction, the best KBP treatment plan in [7] turned out to be associated with the one that ranked 16^{th} on the dose score chart with a MAE of ~ 3.19 Gy. In comparison, our Deep Evidential model's MAE score was 3.09 Gy. We also found that for a stable and effective implementation of the Deep Evidential model, much fine-tuning of model hyperparameters was required. Thus, we picked a vanilla 3D U-Net as our backbone architecture of which learning curve approximately flattened within a couple of hours' training. This enabled us to perform extensive ablation experiments for completing a reformulation of the original theory in [29], so that it can be efficiently adapted for dose prediction with excellent uncertainty-error correlation. The primary focus of our work here is to elucidate the extent to which uncertainty estimates are correlated with model prediction errors, as they can be further harnessed to characterize reliability of a model.

Our paper is organized as follows. In Section 2, we furnish a brief exposition of the theoretical basis underpinning Deep Evidential Learning [29, 30], our reformulation of the loss function and some brief comments on related works in [8, 28]. In Section 3, we outline details of our model structure and implementation. This is followed by a presentation of various results in Section 4. We end with a summary of key findings, main limitations and significance of our work in Sections 5 and 6.

[ii] A general survey can be found in [19] whereas comprehensive recent reviews in the field of medical imaging can be found in [20] and [32].

2 Preliminaries

2.1 Model uncertainties from a Bayesian prior probability distribution

In the standard supervised learning formalism, typically the model is trained via minimizing a loss function such as the mean squared error so that the neural networks' weights converge to at least a local minimum point, yielding a good accuracy defined in terms of the error term implied by the loss function. This standard approach does not directly give any estimates of model or data noise uncertainties. In a more refined approach, one can attempt to provide an estimate of uncertainty by the following assertion:

- The target outputs can be modeled as being drawn from a probability distribution. For example, in our context of dose prediction, we let the predicted dose y_k in voxel V_k be described by a Gaussian probability distribution function (PDF) with mean and variance parameters (μ_k, σ_k).
- We then attempt to train the model to infer the set of (μ_k, σ_k) (for every voxel k) by the principle of maximum likelihood estimation. Typically, this is performed by minimizing the negative log-likelihood function $\mathcal{L}$ defined as

$$f(y_k|\mu_k,\sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(y_k-\mu_k)^2}{2\sigma_k^2}}, \quad \mathcal{L}(\vec{\mu},\vec{\sigma}) = -\log\left(\Pi_k f(y_k|\mu_k,\sigma_k)\right). \tag{1}$$

  Consider an ensemble of models, not necessarily identical in architecture, being trained upon the same dataset to yield estimates of the parameters (μ_k, σ_k), we would then obtain a spectrum of differing values of (μ_k, σ_k), the distribution being stochastic in nature due to the choice of model hyperparameters and their sensitivity to the random noise present in the training dataset.

The Deep Evidential Learning framework of [30] and [29] posits that there is another PDF which describes the distribution of the Gaussian mean and variance parameters. This can be regarded as a 'higher-order' PDF that in turn describes the distribution of (μ_k, σ_k) of the 'first-order' PDF $f(y|\mu,\sigma)$ in eqn. (1). In Bayesian theory, there is a natural mathematical entity that enacts this role – the prior probability distribution that treats (μ_k, σ_k) as random rather than deterministic variables. In [29], various model training and simulations were performed by taking the prior distribution to be a product of a Gaussian distribution (for μ_k) and an inverse-gamma distribution (for σ_k). Each of them is separately a two-parameter distribution. Suppressing the voxel index k, we have

$$\begin{aligned} &\mu \sim \mathcal{N}(\gamma, \sigma^2/\nu), \;\; p(\mu|\gamma,\sigma^2/\nu) = \frac{\sqrt{\nu}}{\sqrt{2\pi\sigma^2}} e^{-\frac{\nu(y-\mu)^2}{2\sigma^2}}, \\ &\sigma^2 \sim \Gamma^{-1}(\alpha,\beta), \;\; p(\sigma|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)\sigma^{2(\alpha+1)}} e^{-\frac{\beta}{\sigma^2}}, \end{aligned} \tag{2}$$

where $\Gamma^{-1}(x)$ is the inverse gamma distribution. The overall prior distribution is a four-parameter *normal-inverse-gamma* distribution defined as the product of the two PDFs above.

$$\mathcal{P}(\mu,\sigma|\alpha,\beta,\nu,\gamma) = p(\mu|\gamma,\sigma^2/\nu)p(\sigma|\alpha,\beta) = \frac{\beta^\alpha\sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{2\beta+\nu(\gamma-\mu)^2}{2\sigma^2}\right]. \tag{3}$$

Given this prior distribution, we can take expectation values and other statistical moments with respect to it to compute mean and variances. In the simplest form, Deep Evidential Learning implies attaching a final feedforward layer (e.g. of the convolutional type) to an existing model architecture such that we have $\{\alpha, \beta, \nu, \gamma\}$ as the eventual model outputs.

The mean prediction, epistemic and aleatoric uncertainties are then defined with respect to $\mathcal{P}$ as follows.

$$\mathbb{E}_{\mathcal{P}}[\mu] = \gamma, \quad U_a = \mathbb{E}_{\mathcal{P}}[\mathrm{Var}(f)] = \mathbb{E}_{\mathcal{P}}[\sigma^2] = \frac{\beta}{\alpha - 1}, \quad U_e = \mathrm{Var}_{\mathcal{P}}[\mathbb{E}(f)] = \mathrm{Var}_{\mathcal{P}}[\mu] = \frac{\beta}{\nu(\alpha - 1)}, \quad (4)$$

where f is the 'first-order' PDF of eqn. (1). It is in this sense that $\mathcal{P}$ is the 'higher-order' PDF providing the description of μ, σ of f as random variables. We note that the aleatoric and epistemic uncertainties are obtained after integrating over μ, σ using eqn. (3). The factorized form of $\mathcal{P}$ enables an analytic expressions for each of these uncertainties. In deep evidential learning, the model output parameters are $\{\alpha, \beta, \nu, \gamma\}$. From these parameters, we can then obtain the dose prediction (given by γ) as well as the uncertainty estimates U_a, U_e for each voxel.

2.2 Deep Evidential regression for a dose prediction model: a refined loss function

We now apply such an uncertainty quantification framework in the context of a dose prediction model which estimates the dose delivered to each CT voxel based on an input set of CT images and various binary-valued masks representing the radiotherapy's target regions and the organs-at-risk. The OpenKBP dataset of [7, 14] was used as our source, and a vanilla 3D U-Net was employed as the backbone architecture.

In [29], the first-order PDF pertaining to the model output y was assumed to be the normal distribution $\mathcal{N}(y|\mu, \sigma)$, with a 4-parameter normal-inverse-gamma distribution (eqn. (3)) adopted to be the higher-order evidential distribution. For our case, a crucial caveat is that the radiation dose value is definitively constrained within a finite interval, and not unbounded. To incorporate this physical nature of the radiation dose, we first define a dimensionless dose parameter $y = y(D_p) = \frac{0.9D_p+10}{100}$ where D_p is the physical dose in units of Gy. We then pass y to its logit representation $\log\left(\frac{y}{1-y}\right)$ which we take as the form of output for the neural network (rather than the physical dose D_p). Ablation experiments using the validation dataset were used to determine the coefficients in the linear map $y = y(D_p)$ such that the logit representation yielded the optimal model performance. Our choice of the linear map $y = y(D_p)$ also avoided the singular points $\{0, 1\}$ of the logit function by construction.

Formally, our choice of the output data representation implies that the first-order PDF for each voxel is related to the *logit-normal* distribution more precisely instead of a Gaussian. Its full form reads

$$f\left(y|\mu, \sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{y(1-y)} \exp\left(-\frac{1}{2\sigma^2}\left[L(y) - \mu\right]^2\right), \quad L(y) \equiv \log\left(\frac{y}{1-y}\right). \quad (5)$$

This is a close cousin of the normal distribution, and has a bounded domain $y \in (0, 1)$ as the support. The parameters μ, σ are now dimensionless parameters defined as the Gaussian mean and standard deviation of the logit of the (dimensionless) dose y. For their higher-order evidential distributions, we can still follow the procedure of [29], i.e. we adopt a normal distribution for μ and an inverse-gamma distribution for the variance to yield a normal-inverse-gamma distribution for (5). Also, similar to the approach in [29], we can derive the aleatoric and epistemic uncertainties.

In their original forms as defined in eqn. (4), they are dimensionless variances associated with the logit representation. To convert them to the corresponding dimensionful ones in units of Gy^2, we first invoke the standard approximation technique of relating the uncertainty in the $L(y)$ (logit of y) to that in y itself, before using the linear map $y = y(D_p)$ to deduce the physical aleatoric and epistemic uncertainties in units of Gy^2. For clarity, we refer to them separately as U_{alea}, U_{epis} which are related to U_a, U_e in eqn. (4) via $U_{alea,epis} = \left(\frac{100}{0.9} y(1-y)\right)^2 U_{a,e}$.

Following [29], the primary loss function can be taken to be the maximum likelihood loss defined as the marginal likelihood obtained when we integrate over the first-order mean μ, and variance σ^2. After some algebra, we find that we obtain the likelihood function

$$\begin{aligned}
\mathcal{L}_{pri} &= \int_0^\infty d\sigma^2 \int_{-\infty}^\infty d\mu \; f(y|\mu,\sigma)\, \mathcal{P}(\mu,\sigma|\alpha,\beta,\nu,\gamma) \\
&= \frac{1}{y(1-y)} \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \sqrt{\frac{\nu}{\pi}} \left[2\beta(1+\nu)\right]^\alpha \left[\nu\left(\log\left(\frac{y}{1-y}\right)-\gamma\right)^2 + 2\beta(1+\nu)\right]^{-(\alpha+\frac{1}{2})}, \quad (6)
\end{aligned}$$

where the expression lying on the right of the term $1/(y(1-y))$ can be identified as the Student's t distribution $\mathrm{St}\left(\log(\frac{y}{1-y}); \gamma, \frac{\beta(1+\nu)}{\nu\alpha}, 2\alpha\right)$. In [29], the model's primary loss function was then derived from eqn. (6) by taking the negative logarithm of it and adding a suitable regularization loss term.

Unfortunately, we found that in this form, the loss function did not lead to stable learning curves that could generate uncertainty heatmaps that correlated with prediction errors. After some experimentation, we found that making the following refinements to eqn. (6) significantly enhanced the effectiveness of the model:

- removal of the factor $1/y(1-y)$ from eqn. (6),
- passing the negative log-likelihood function through a regularizing positive-definite activation function (our choice: standard logistic function),
- adding a mean-squared-error loss function term.

These refinements led to the eventual form of our loss function being as follows.

$$\begin{aligned}
L_{pri} &= f_s\left(\frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \sqrt{\frac{\nu}{\pi}} \left[2\beta(1+\nu)\right]^\alpha \left[\nu\left(\log\left(\frac{y}{1-y}\right)-\gamma\right)^2 + 2\beta(1+\nu)\right]^{-(\alpha+\frac{1}{2})}\right), \\
\mathcal{L}_f &= \mathcal{L}_{pri} + \lambda_{KL}|y-\gamma|(2\nu+\alpha) + \lambda_{mse}\left(\log\left(\frac{y}{1-y}\right)-\gamma\right)^2, \;\; f_s(g) \equiv \frac{1}{1+g}, \quad (7)
\end{aligned}$$

where $\lambda_{KL}, \lambda_{mse}$ are crucial hyperparameters of which optimal values for our purpose will be discussed in Sec. 3.2. The first regularization term $\mathcal{L}_{reg} = |y-\gamma|(2\nu+\alpha)$ preceded by λ_{KL} was proposed in [29] to penalize statistical evidence for incorrectly labeled terms. We note in passing that in [29], it was argued that the prior (the function $\mathcal{P}(\mu,\sigma|\alpha,\beta,\nu,\gamma)$ in eqn. (3)) has a singular limit when we take parameter values corresponding to 'zero evidence', and that softening this limit by introducing finite yet small values for $\alpha = 1+\epsilon$, $\nu = \epsilon$ didn't appear to be effective. We found this to be the case as well. Another possibility that we explored is using Jeffrey's prior (which in this case involves an improper uniform prior for the mean of the model output). But this turned out to be similarly ineffective relative to the choice adopted in [29].

2.3 Related Work

Among the many papers devoted to the subject of dose prediction models (see e.g. [5] for an extensive review and compilation), to our knowledge, there has been only two which directly delved into the notion of uncertainty quantification frameworks: [8] and [28]. These papers contained highly interesting results which we would like to briefly discuss in relation to our work.

In [8], the authors studied the frameworks of MC Dropout and bootstrap aggregation which is essentially a Deep Ensemble method where each model is trained on only a portion of the entire training dataset. Like in this work, they used the OpenKBP Challenge dataset of [14] for validating the performance of each uncertainty quantification framework, concluding that Deep Ensemble yielded a lower mean absolute error (MAE) while showing better correlation between uncertainty and prediction error. In particular, the authors proposed that with the aid of an additional scaling factor, the models generated uncertainty heatmaps and DVH confidence intervals which appeared reasonable. This scaling factor was defined differently for each region of interest (ROI) and was argued to be necessary to bring the uncertainty values into a more 'interpretable scale' as the raw ones were only providing 'relative measures'. For each set of points belonging to a ROI in the validation dataset, this ROI-dependent factor was defined simply as the standard deviation of the ratio between the prediction error and the uncertainty value (see eqn. (3),(4) of [8]). In our work, we did not incorporate such an empirical scaling factor in our definitions of various uncertainties, including our final uncertainty heatmaps and the DVH confidence intervals. For Deep Evidential Learning where the predictive variance can be expressed precisely in terms of learned output variables of the model, in principle, it is not clear to us how an additional empirical scaling factor should be justified. By construction, this rescaling factor naturally enhances the correlation between uncertainty and error since the uncertainty values themselves are part of the factor's definition, assuming that it generalizes well for unknown (e.g. testing) data. Conversely, we took the resulting average value of such a ratio (across all voxels) to assess whether the uncertainty distribution generated by the model was reasonable in terms of the overall scale of magnitude. Through the numerical values (see Table 1) and the DVH certainty intervals in Fig. 8, we found that this was indeed the case.

The work in [28] proposes the use of Gaussian mixture models of which parameters are the neural network's outputs, and with U-Net as the backbone architecture. Such a mixture density network is very similar in form to our Deep Evidential Learning model since the output variables of our model are also parameters of a PDF. In our case, the PDF is a higher-order normal-inverse-gamma distribution whereas in [28], the parameters are the means and variances of the component normal distributions (see eqn. (1)) and the coefficients defining their linear combination. Each component of the Gaussian mixture pertained to a treatment protocol representing specific priorities (e.g. organs at risk (OAR) sparing over target coverage, etc.), and the relative values of the variances quantified the trade-offs between these protocols. The cross-entropy loss function associated with the cumulative distribution function (CDF) of the predicted dose was then taken as part of the objective function for generating a dose mimicking model. In [28], deliverable treatment plans were the endpoints of their dose prediction algorithm, with the optimization problem solved by RayStation's native sequential quadratic programming solver and final dose computed via a collapsed cone algorithm [28]. Although we noted that there was no scrutiny of the correlation between uncertainty and error, the authors of [28] (and its follow-up work [27]) notably pointed out that probabilistic dose prediction models such as [8] and ours could play a crucial role in their pipeline since our model would yield dose CDFs that can be used to define the cross-entropy loss function as part of the dose-mimicking model's objective function. It would thus be interesting to

extend results of our work here by integrating our probabilistic dose predictions into a pipeline like [28] where deliverable treatment plans with specified machine parameter settings can be anticipated as final outcomes.

3 Methodology

3.1 On the OpenKBP Challenge Datset

The OpenKBP Challenge dataset of [7, 14] is devoted towards establishing an open framework for the development of plan optimization models for knowledge-based planning (KBP) in radiotherapy. The data was obtained for 340 patients with head-and-neck cancer. The dataset split proposed for training deep learning models was: 200 (training), 40 (validation), 100 (testing). In our work, we followed this decomposition of data.

The radiotherapy plans were delivered using nine equidistant coplanar beams at various angles with a 6 MV step-and-shoot intensity-modulated-radiation-therapy (IMRT) in 35 fractions. The organizational team of the Challenge used the IMRTP library from *A Computational Environment for Radiotherapy Research* implemented in MATLAB [37] to generate the dose deposited at each voxel. These dose distributions were from fluence-based treatment plans with similar degrees of complexity [14, 38].

We used a $(128, 128, 128, 11)$-dimensional input representation consisting of (i)CT in Hounsfield units(clipped to be within [0, 4095] before being normalized to [0,1]), (ii)structure masks of three planning target volumes (PTV) and seven organs-at-risk (OAR) regions : each is a Boolean tensor labeling any voxel contained within the respective structure. The OARs are brainstem, spinal cord, right and left parotids, larynx, mandible and esophagus. The three PTV regions are: PTV56, PTV63, PTV70 which are targets that should receive 56 Gy, 63 Gy and 70 Gy of radiation dose respectively.

3.2 On model architecture and hyperparameters

Since our primary goal is to examine the uncertainty estimation aspects of the model, in this work, we adopt a simple vanilla 3D U-Net as the backbone architecture upon which we insert two additional layers so that they are compatible with the framework of Deep Evidential Learning.

Our U-Net has the following structure (see Fig. 1):

- it takes in a 11-channel input of size $128 \times 128 \times 128$ voxels,
- each downsampling level consists of two convolutional layers each with a $(3 \times 3 \times 3)$ kernel, ReLU activation and equipped with a dropout unit, followed by maximum pooling with kernel size $(2 \times 2 \times 2)$,
- the downsampling operation is performed 4 times, with the dropout rates increasing consecutively as $\{0.10, 0.15, 0.20, 0.25\}$, and the number of convolutional filters for each layer being $\{16, 32, 64, 128\}$ respectively,
- the bottleneck layer has 256 filters and dropout rate of 0.30,

- at each of the four upsampling levels, features from the contraction path are concatenated with the corresponding upsampled features, and the dropout rates decrease similarly in the reverse order.

- a 8-channel pointwise convolution is then applied followed by a final 4-channel pointwise convolution that yields the four parameters $\{\alpha, \beta, \nu, \gamma\}$ of the Bayesian prior distribution $\mathcal{P}(\mu, \sigma|\alpha, \beta, \nu, \gamma)$ of eqn. (3).

This model carries about 6×10^6 free weight parameters. Various hyperparameters such as the dropout rate at each level, etc. were eventually adopted after running ablation experiments to determine their optimal values.

Imposing some choice of numerical bounds to $\{\alpha, \beta, \nu, \gamma\}$ in the final layer is a crucial aspect of the hyperparameter tuning process unique to this framework. The parameter β sets the overall (dimensionless) scale to both uncertainties of which ratio is set by $\nu = U_{alea}/U_{epis}$ which has to be positive. For this dataset, we found good uncertainty-error correlation after imposing $\beta \geq 10^{-3}$. In [29], the authors imposed the constraint $\alpha > 1$ which we also followed here, as for this range of values, the aleatoric and epistemic uncertainties can be conveniently described by simple analytic expressions. More broadly speaking, $\alpha > 0$ is the larger admissible range for α as a parameter of the inverse-gamma distribution. We also found it useful to map the zero dose to a small positive value $\epsilon = 0.1$ in the (dimensionless) normalized dose D_N defined to be $D_N = \frac{0.9D_p+10}{100}$ where D_p is the physical dose in Gy. In the loss function, we work in the logit-representation of D_N, with the predicted dose $\gamma = \log\left(\frac{D_N}{1-D_N}\right)$. Since this logistic function is monotonic, γ is bounded from below as $\gamma \geq \log\left(\frac{\epsilon}{1-\epsilon}\right) = -\log 9$ with our choice of $\epsilon = 0.1$.

For the loss function originally proposed in [29], we found that unfortunately, it led to training curves with frequent oscillations and poor eventual outcomes. As explained in detail in Sec. 2.2, we reformulated the loss function for a logistic representation of the dose variable such that one could obtain a stable implementation. Our refined loss function in eqn. (7) is characterized by the hyperparameters $\lambda_{mse}, \lambda_{KL}$ which describe couplings of the loss function to a mean-squared-error (MSE) term and a KL-divergence-like regularization term respectively. Results from ablation experiments with validation dataset indicated that optimal values of these hyperparameters lie within the intervals $\lambda_{mse} \in (0.01, 0.1)$, $\lambda_{KL} \lesssim 0.01$. In particular, the finite interval of optimal λ_{mse} arose from our finding that enhancing MSE-type performance too much would compromise the correlation between uncertainty and prediction error. For the results reported here for the rest of our paper, we took $\lambda_{mse} = 0.05$, $\lambda_{KL} = 0.01$. In Fig. 2, we plot the learning curves pertaining to the original and our refined loss functions for comparison.

Towards comparing the Deep Evidential model against more common approaches of uncertainty quantification in literature, we implemented a Deep Ensemble model and Monte-Carlo Dropout model for the same dataset. Each of the 5 neural networks used in the ensemble model share the same 3D U-Net architecture with all hyperparameters preserved. Random weight initialization was all performed via the *He*-uniform initializer [39]. Similarly, for the Monte-Carlo Dropout model, the same base model was used for passing 30 forward passes with the dropout layers activated for model prediction. All models were trained for 200 epochs with a learning rate of 10^{-4} using Adam optimizer. Our focus in this work lies in studying aspects of uncertainty estimation, and thus our choice of a relatively simple backbone architecture equipped with only $\sim 6 \times 10^6$ weight parameters. Compared to other more complicated setups, each of these models converged quickly in about a couple of hours while attaining a mean-absolute-error of dose prediction that laid within the top 20 of the score chart of the OpenKBP Challenge [14].

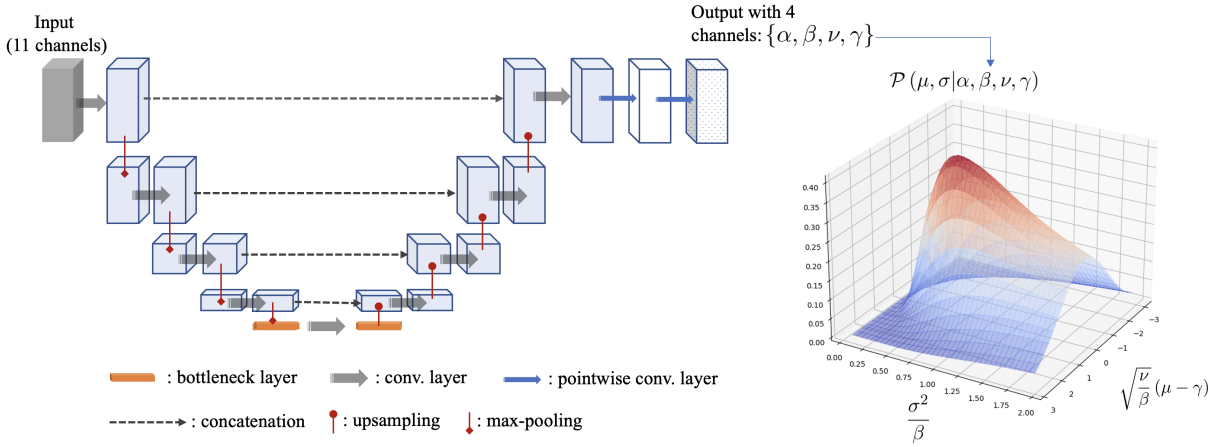

Figure 1: A sketch of the Deep Evidential model with 3D U-Net backbone architecture. The 4-channel outputs of the model are the parameters of a normal-inverse-gamma distribution schematically plotted on the right. Details of the convolutional and other layer operations in the backbone segment are described earlier in Sec. 3.2 and largely identical to the original 3D U-Net of [31]. We adjoined the U-Net structure to the Deep Evidential framework by passing the output obtained after the final upsampling layer to two consecutive pointwise convolution layers with number of channels = 8, 4 respectively. The final output has dimensions $(128, 128, 128, 4)$.

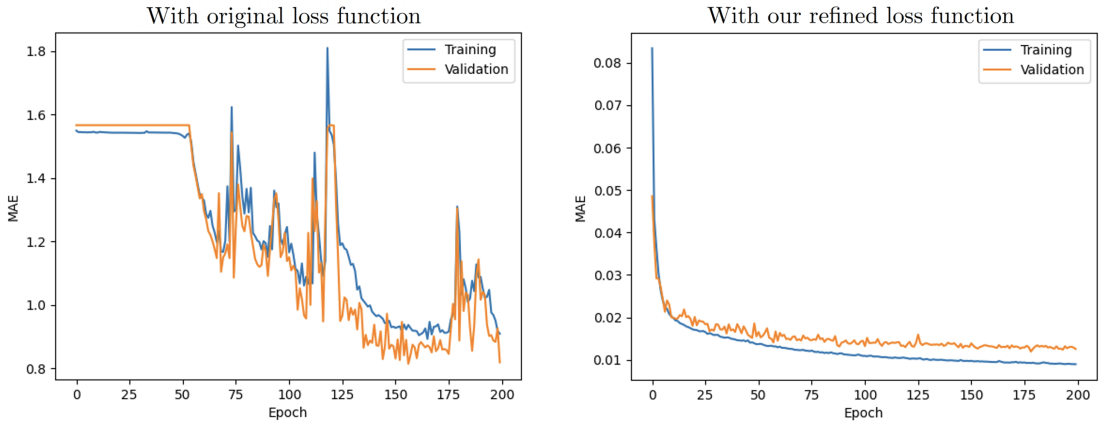

Figure 2: Plots of the validation and training mean-absolute-error (MAE) for the model equipped with the original (left) and our refined (right) loss functions.

4 Results

We found that our Deep Evidential Learning model yielded uncertainty estimates which demonstrated strong correlation with prediction errors, while achieving a similar level of accuracy relative to the methods of Monte-Carlo Dropout and Deep Ensemble. The metrics that we used as measures of any putative uncertainty-error associations were (i)the Spearman's rank correlation coefficient

between the uncertainty and error values across the testing dataset and their patient-averaged distributions, (ii)the mutual information between them.

Model	MAE (Gy)	U_{avg} (Gy^2)	$r_s(\overline{U}, \overline{D_e})$	$r_s(U, D_e)$	$M.I.(U, D_e)$
Deep Evidential (U_e, U_a)	3.09	(2.80, 0.20)	(0.83, 0.69)	(0.69, 0.63)	(0.67, 0.61)
Monte-Carlo Dropout	3.28	2.12	0.73	0.62	0.60
Deep Ensemble	3.10	3.42	0.86	0.66	0.56

Table 1: Table collecting various measures of uncertainty-error correlations. *Abbreviations*: MAE = mean absolute error, U_{avg} = mean uncertainty value, $r_s(\overline{U}, \overline{D_e})$ = Spearman's coefficient between patient-averaged uncertainty and error distributions, $r_s(U, D_e)$ = Spearman's coefficient between uncertainty and error distributions, $M.I.(U, D_e)$ = mutual information between uncertainty and error distributions, D_e = prediction error, (U_e, U_a) = (epistemic uncertainty, aleatoric uncertainy).

As summarized in Table 1, one observes that across the various correlation indices, the epistemic uncertainty's degree of association with error was higher than Monte-Carlo Dropout and Deep Ensemble, with the only exception being its $r_s(\overline{U}, \overline{D_e})$ slightly lower than that of Deep Ensemble. For all three models, the correlation between patient-averaged uncertainties and errors was always higher, suggesting that there is a certain degree of stochasticity characterizing this relationship within each patient's dataset. Relative to epistemic uncertainty U_e, aleatoric uncertainty U_a in the Deep Evidential Learning model consistently demonstrated lower correlations with error, and its r_s value was less affected by averaging within individual patient (compared to other models). We also note that the order-of-magnitudes of the various averaged uncertainty values U_{avg} were such that $\text{MAE}/\sqrt{U_{avg}} \sim 2$, the exception being U_a which was generally an order-of-magnitude below U_e and uncertainties of the other two models.

4.1 Variation of prediction error with uncertainty threshold

The results in Table 1 indicated that, at least compared to Monte-Carlo Dropout and Deep Ensemble methods, Deep Evidential Learning model appeared to yield spectra of uncertainty values which were strongly correlated with prediction error, indicating that it could enact a feasible uncertainty estimation framework that helps to discover and highlight potential regions of ignorance within the deep learning model.

To further scrutinize the difference among the various frameworks, we examined how error distributions varied with changing threshold values of uncertainty. Since the error distributions were generically found to be quite skewed, we used the median of the error distribution as its characterizing parameter. Fig. 3 and 4 below reveal how median error changed with increasing levels of uncertainty thresholds for the different models. Assuming that uncertainty values correlate positively with prediction error implies that one expects a monotonically increasing curve in such a plot, a trend that was indeed manifest for these models as shown in Fig. 3 and 4.

In particular, from Fig. 4, the visibly more evident linearity of the curve for Deep Evidential Learning indicated a sensitivity of epistemic uncertainty to prediction errors that was relatively more uniformly calibrated compared to Monte-Carlo Dropout and Deep Ensemble methods.

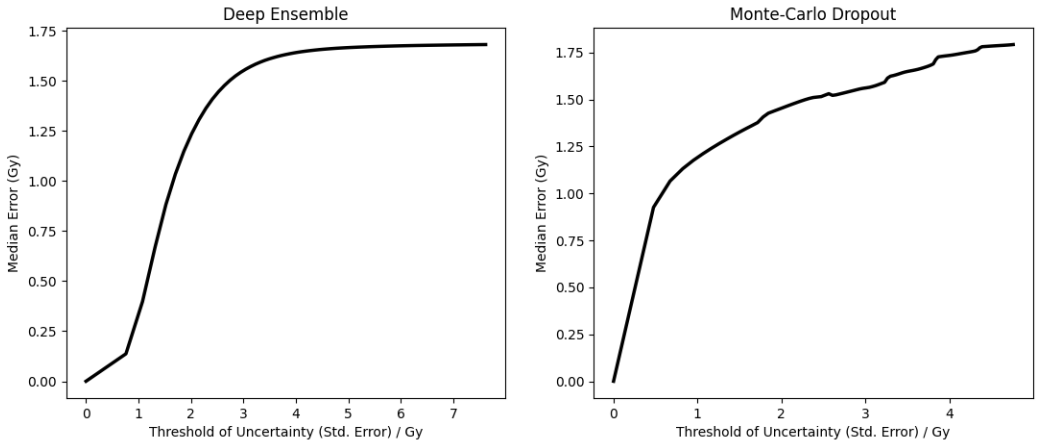

Figure 3: The median error plotted is that of the subset of the testing dataset which remains after imposing an upper threshold value for the uncertainty. We note that each model generated different ranges of uncertainty values.

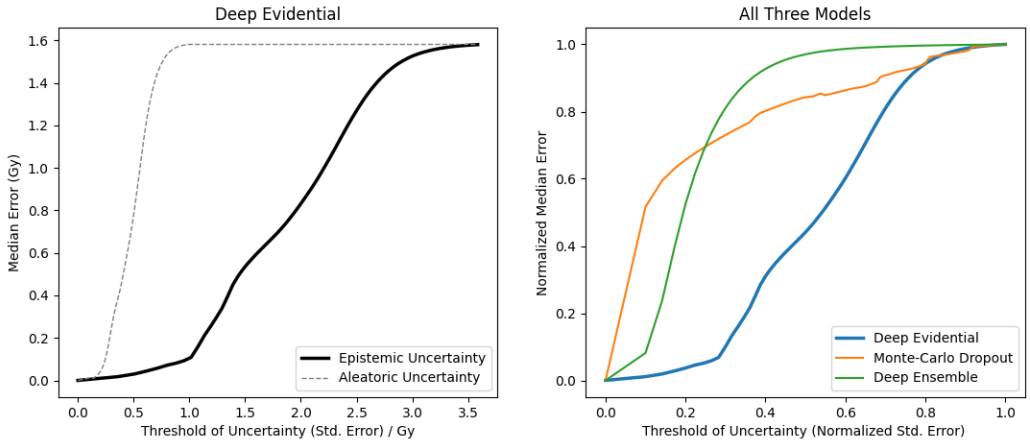

Figure 4: The left diagram pertained to Deep Evidential Learning where we included both the curves for aleatoric and epistemic uncertainties. On the right, we furnished a comparative plot where both median error and uncertainty threshold were normalized using their maximum values. This allows one to differentiate among the curves based on a common overall scaling.

4.2 Probing noise-sensitivity of uncertainty measures

A prominent feature of Deep Evidential Learning is a clear principled approach towards distinguishing between aleatoric and epistemic uncertainties. They can be defined precisely using the higher-order PDF $\mathcal{P}$ as expressed in eqn. (4). Traditionally, the definition of aleatoric uncertainty is commonly taken to imply that it can be understood to capture any inherent random noise in the data, in contrast to epistemic uncertainty which presumably captures uncertainty originating from model-suitability and data-sufficiency (see e.g. [32, 36]).

We would like to probe the level of sensitivity of various uncertainty measures to noise perturbations of the input data. In particular, we would like to identify if there would be any difference

in noise-sensitivity between U_{alea} and U_{epis}. Our choice of perturbation is the addition of Gaussian noise [40] of mean zero and standard deviation 0.5 to each CT voxel (of which intensity was pre-processed to lie within $[0, 1]$). To enable a more holistic description of response to noise, we plot the empirical cumulative distribution function (eCDF) of various uncertainty measures preceding and following the noise addition.

In Fig. 5, the plot of various eCDFs collectively illustrated the differences among the uncertainty measures in their responses towards the Gaussian perturbation. Since different measures yielded distinct overall scale of uncertainty magnitudes, we normalized each with respect to the maximum value for each uncertainty type. In the limit of the Gaussian noise completely dominating over any pre-existing noisiness, and assuming that the uncertainty value of each voxel is correlated only with the noise PDF (identical for every voxel), we would expect the uncertainty eCDF to approach a symmetric distribution with mean (normalized) uncertainty at 0.5 in this limiting scenario. As a simple reference distribution, in Fig. 5, we included the line corresponding to the uniform distribution. From just visual inspection, aleatoric uncertainty distribution and that of Deep Ensemble changed more significantly relative to others. The noise-induced fractional changes in the relative entropy (measuring the Kullback-Leibler divergence between each eCDF and the uniform distribution) turned out to be ~ 0.1 for aleatoric uncertainty and Deep Ensemble, about ten times larger than those for epistemic uncertainty and Monte-Carlo dropout.

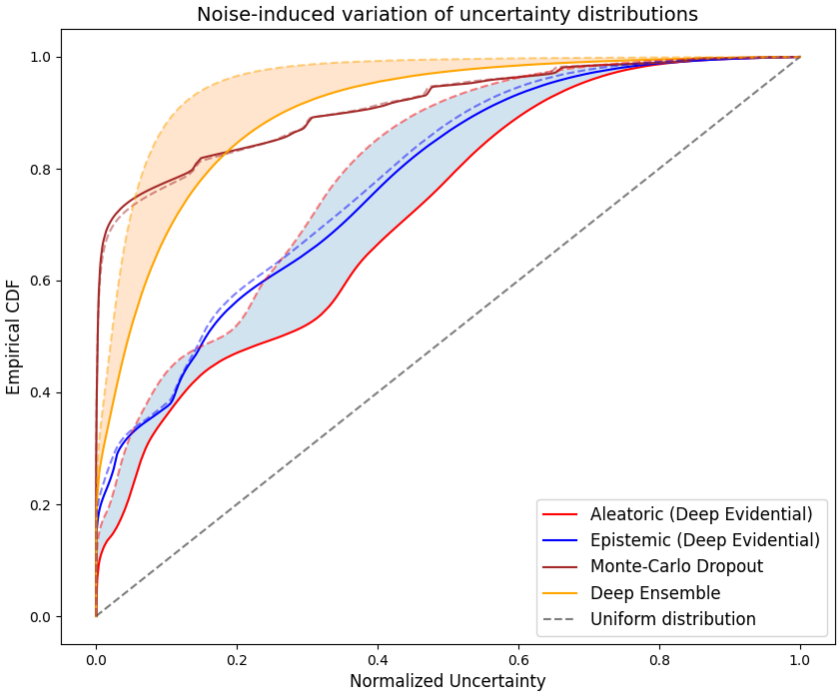

Figure 5: Dashed lines are the uncertainty eCDF in the absence of noise, while the corresponding solid curves show distributions in the presence of an added Gaussian noise with zero mean and standard deviation of 0.5 for the CT intensity value in each voxel. Shaded regions accentuate the more significant shift associated with aleatoric uncertainty (of the Deep Evidential Model) and Deep Ensemble's uncertainty distributions.

The various curves appeared to suggest the following.

- The relative responsiveness of aleatoric (stronger) and epistemic (weaker) uncertainties are indeed compatible with their conventional interpretations as reflecting inherent data noise and model-related uncertainties respectively.
- Monte-Carlo Dropout yielded an uncertainty distribution that, like epistemic uncertainty, was not sensitive to the addition of the Gaussian noise. This favored its interpretation as an epistemic uncertainty in nature.
- Deep Ensemble method yielded an uncertainty distribution that was responsive to the addition of noise. Like the aleatoric uncertainty in Deep Evidential Learning, the eCDF shifted to being less long-tailed and leaning towards being more uniform.

For the Deep Ensemble method, it was explained in [34, 41] that one can often use a Gaussian mixture model as an effective description of the ensemble prediction. By Eve's law of total variance [41], the predictive variance admits a natural decomposition into aleatoric and epistemic components. This appears to be consistent with the noise sensitivity displayed by Deep Ensemble method in Fig. 5.

4.3 Applications

4.3.1 Uncertainty heatmaps

Heatmaps of the dose uncertainty distributions can be used through direct visual inspection to discover regions of potentially acute errors made by the dose prediction model. This capability hinges on the strength of the correlation between uncertainty and error distributions. In Table 1, epistemic uncertainty in the Deep Evidential model appeared to perform comparatively well, if not better, than the two more conventional uncertainty quantification frameworks as measured by mutual information and Spearman's correlations. These statistical indices were defined at the level of the entire (testing) dataset. For each of the seven OAR and three PTV regions, one could in principle compute the uncertainty distribution localized within its interior. Thus, in Table 2 we collect for various ROIs their mean uncertainty (U_{alea}, U_{epis}) values together with the errors in predicted dose D_e. The Spearman's coefficients between each uncertainty type and the corresponding dose prediction error $\overline{D}_e$ are also indicated (all with p-values $p < .01$). Among all the ROIs, the L,R-parotid regions exhibited the highest correlations between $\overline{D}_e$ and epistemic uncertainty, while the larynx region showed the lowest measures of correlation for both types of uncertainties. The mandible region carried the highest aleatoric uncertainty and was associated with the highest $\overline{D}_e$. Across the spectrum of ROIs in Table 2, we found each mean epistemic uncertainty to be strongly correlated with $\overline{D}_e$), with Spearman's correlation coefficient of $0.80\,(p = 0.005)$, whereas that for the aleatoric uncertainty was much less significant at $0.57\,(p = 0.09)$.

At the level of each ROI, apart from the larynx, $r_s(U_{epis}, D_e) \gtrapprox 0.7$ (1 s.f.) for all other OAR and PTV regions. Coupled with the error-uncertainty threshold study in Sec. 4.1 and the noise-sensitivity test in Sec. 4.2, this tapestry of results equipped us with the basis for interpreting elevated heatmap regions of epistemic uncertainty as indicators of potential regions of prediction errors and those of aleatoric uncertainty as correlated with inherent data noise.

In Fig. 6, we display an illustrative set of axial CT images portrayed alongside their corresponding uncertainty heatmaps. In principle, the aleatoric uncertainty probes the level of inherent data

ROI	$\overline{U}_{alea}$ (Gy^2)	$\overline{U}_{epis}$ (Gy^2)	$r_s(U_{alea}, D_e)$	$r_s(U_{epis}, D_e)$	$\overline{D}_e = \overline{\lvert D_p - D_{GT} \rvert}$ (Gy)
PTV70	0.21	2.63	0.56	0.68	2.23
PTV63	0.21	2.65	0.65	0.74	2.16
PTV56	0.20	2.36	0.68	0.78	1.87
Brainstem	0.10	0.94	0.69	0.69	0.65
Spinal Cord	0.27	2.49	0.79	0.82	1.59
R-Parotid	0.13	2.43	0.80	0.84	1.38
L-Parotid	0.10	2.78	0.80	0.84	1.45
Esophagus	0.06	1.18	0.71	0.71	0.98
Larynx	0.17	1.95	0.43	0.48	2.46
Mandible	0.21	3.77	0.83	0.79	4.02

Table 2: Table collecting various ROIs, their associated uncertainty (U_{alea}, U_{epis}) values together with the errors in predicted dose D_e. These values are averaged over the entire testing dataset. $r_s(U_{alea}, D_e)$ and $r_s(U_{epis}, D_e)$ denote the Spearman's coefficients between each uncertainty type and the corresponding dose prediction error D_e.

noise, while the epistemic uncertainty is more strongly correlated with the elevated regions in the (leftmost) dose prediction error heatmap. Thus, we expect the latter to be more heterogeneous which was indeed evident in our survey of uncertainty heatmaps across the testing dataset.

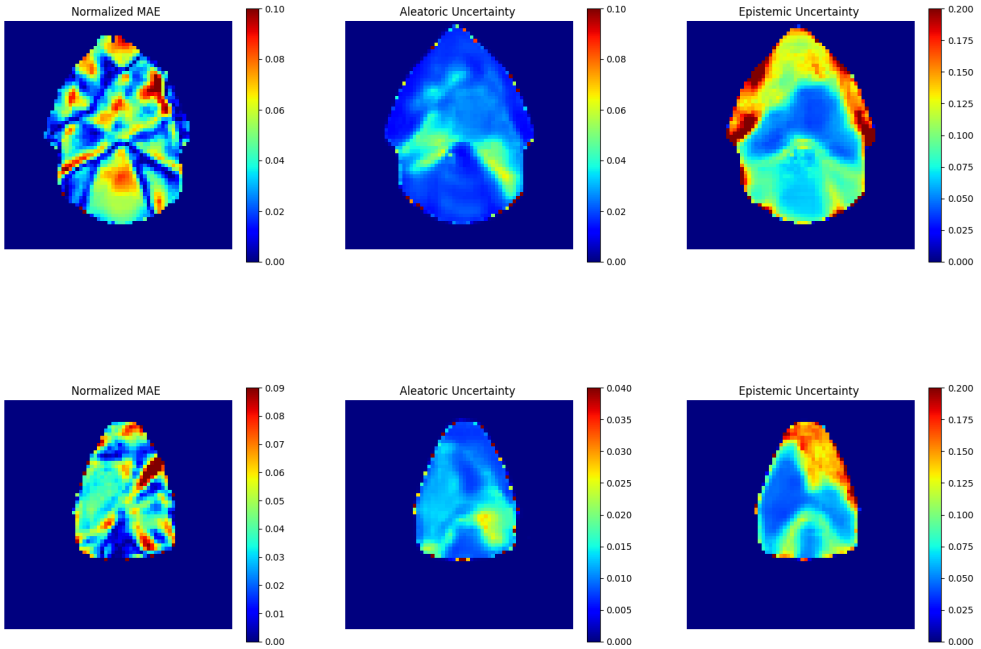

Figure 6: Axial CT slices with their model prediction error and their corresponding uncertainty heatmaps. In each row, the leftmost image shows the mean absolute dose prediction error (MAE) heatmap, the center and rightmost images showing the aleatoric and epistemic uncertainty distributions respectively.

4.3.2 DVH with confidence bands

For each patient, we can combine the aleatoric and epistemic uncertainties learned by the neural network to construct confidence intervals for the Dose-Volume-Histogram for each of the region-of-interest. Denoting the predictive variance of the dose in each voxel by δD_p^2, and the predicted dose variable by μ_p, one can invoke the general Eve's law of total variance to obtain

$$\delta D_p^2 \equiv \mathrm{Var}[\mu_p] = \mathbb{E}\left[\sigma^2\right] + \mathrm{Var}\left[\mu\right], \tag{8}$$

where μ, σ are the mean and variance defined in (2). To see this, we recall that in general, Eve's law of total variance (see e.g. [42]) states the following relation for random variables Y and X:

$$\mathrm{Var}[Y] = \mathbb{E}\left[\mathrm{Var}\left[Y|X\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[Y|X\right]\right], \tag{9}$$

where conditional expectations such as $\mathbb{E}\left[Y|X\right]$ are random variables themselves. To apply (9) appropriately to our context, we can identify $Y \sim \mu_p$, and X to collectively represent all model-dependent and data-dependent variables. The relation (8) then follows directly from (9) by virtue of the definitions of σ^2, μ. In the following, we furnish a contextual proof of this result taking into account our framework explicitly.[iii] We begin with the definition

$$\mathrm{Var}\left[\mu_p\right] = \mathbb{E}\left[\mu_p^2\right] - \overline{\mu}_p^2, \;\; \overline{\mu}_p \equiv \iint d\mu d\sigma \;\mathcal{P}(\mu,\sigma|\vec{\alpha})\;\mu, \quad \vec{\alpha} \equiv \{\alpha,\beta,\nu,\gamma\}. \tag{10}$$

Recall that in the Deep Evidential learning framework, a higher-order prior distribution $\mathcal{P}$ describes the randomness of σ, μ. Also, taking into account the definition of σ, μ, we can then write

$$\iint d\mu d\sigma\, \mathcal{P}(\mu,\sigma|\vec{\alpha})\,\sigma^2 = \iint d\mu d\sigma\, \mathcal{P}(\mu,\sigma|\vec{\alpha})\;\left(\mu_p^2 - \mu^2\right). \tag{11}$$

Substituting (11) into (10), we then have

$$\begin{aligned} \delta D_p^2 \equiv \mathrm{Var}[\mu_p] &= \iint d\mu d\sigma\, \mathcal{P}(\mu,\sigma|\vec{\alpha})\;\left(\sigma^2+\mu^2\right) - \overline{\mu}_p^2 \\ &= \mathbb{E}[\sigma^2] + \iint d\mu d\sigma\, \mathcal{P}(\mu,\sigma|\vec{\alpha})\;\left(\mu^2 - \overline{\mu}_p^2\right) = \mathbb{E}[\sigma^2] + \mathrm{Var}[\mu], \end{aligned} \tag{12}$$

hence recovering eqn. (8) expected from more general arguments. Thus, we see that the predictive variance is the sum of the aleatoric and epistemic uncertainties. Our derivation is consistent with a similar result expressed in [36], and provides the theoretical basis for an operational definition of δD_p from knowledge of aleatoric and epistemic uncertainties.

In Fig. 7, we plot examples of individual patient's DVH for various PTV and OAR regions, each equipped with a 95% confidence interval. A visible feature of these bands is that they tend to contain the more extreme deviations of each groundtruth DVH from its corresponding predicted one along their peripheral edges. This is consistent with the order-of-magnitude estimate of $MAE/\delta D_p \sim 1.7$ as drawn from numerical values of the mean error and uncertainties in Table 1. Thickness and various features of these DVH confidence bands for the OARs and PTV regions can be used to characterize the extent of reliability of various dose prediction models, differentiating among them in this aspect.

[iii] Our style of proof is similar in spirit to that presented in [41] for a class of Deep Ensemble models that admit interpretations as Gaussian mixture models, and of which outputs are engineered to be μ, σ.

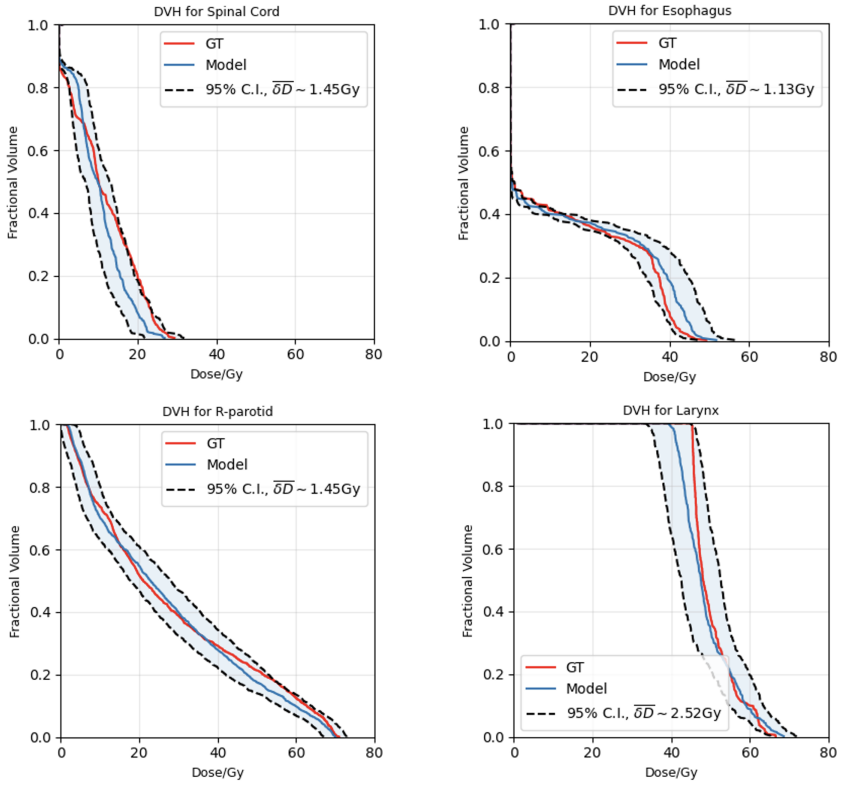

Figure 7: A set of examples of DVH for spinal cord, esophagus, right parotid and larynx each equipped with confidence bands. Dashed curves are the bounds for the 95% intervals defined by $\pm 1.96\,\delta D_p \equiv \pm 1.96\sqrt{U^2_{alea} + U^2_{epis}}$. Red and blue curves pertain to the ground truth and model prediction respectively. In each diagram, $\overline{\delta D}$ refers to the standard error in dose in units of Gy obtained by averaging δD_p over all voxels of the ROI in the patient.

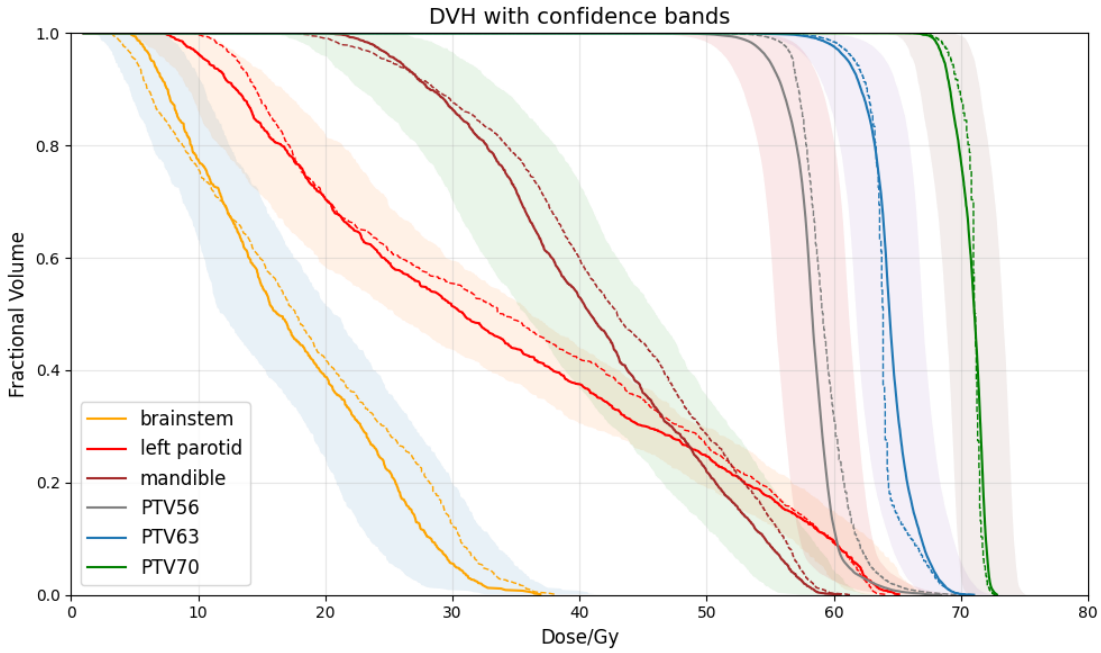

Figure 8: An example of the DVH of a patient in the Open-KBP dataset equipped with 95% confidence intervals indicated as shaded regions. Dashed and solid curves pertain to the ground truth and model prediction respectively. The overall DVH score as defined in [14] was 2.07 which was within the top 17 of the DVH score chart in [14].

5 Discussion

We have refined and applied Deep Evidential Learning framework to radiotherapy dose prediction, our most prominent finding being that the uncertainty estimates obtained from model training inherited strong correlations with prediction errors. In the following, we conclude with a summary list of key findings accompanied by a discussion of various limitations of our work.

5.1 Summary list of key findings

(a) The original loss function proposed in [29] did not work in this particular context of dose prediction. We found that regularizing the final layer with a sigmoid function and adding a mean-squared-error term led to a stable and effective implementation of Deep Evidential Learning.

(b) Epistemic uncertainty (U_{epis}) in Deep Evidential Learning was highly correlated with prediction errors. Correlation indices (mutual information, Spearman's coefficients) were comparable or higher than those for MC Dropout and Deep Ensemble methods (see Table 1).

(c) The median error varied with uncertainty threshold much more linearly for U_{epis} as compared to MC Dropout and Deep Ensemble, indicative of a more uniformly calibrated sensitivity to model errors (see Fig. 3, 4).

(d) Aleatoric uncertainty (U_{alea}) demonstrated a more significant shift in its empirical CDF upon addition of Gaussian noise to CT intensity distribution compared to U_{epis} (see Fig. 5). It also displayed visibly weaker correlation with prediction errors relative to U_{epis}. These traits of U_{alea}, U_{epis} appeared to be supportive of their conventional interpretations as reflecting data noise and model-related uncertainties respectively.

(e) We demonstrated how Eve's law of total variance enables one to express the predictive variance in terms of U_{alea}, U_{epis}, and used this result to construct confidence intervals for DVH (see Fig. 8) as an application that can be further used to characterize reliability of the model.

5.2 Limitations and Future Directions

In the following, we outline several limitations of our work together with corresponding suggestions for future directions.

(i) Although the simple 3D U-Net backbone architecture led to efficient and fast convergence, it would be interesting to study how more complicated network structures like the transformers-based models of [17, 18] perform after modifying their final layers so that there are four output heads corresponding to the parameters of $\mathcal{P}$ in Deep Evidential Learning. This will provide a more extensive study of the degree to which a low MAE can be attained while not compromising the uncertainty-error correlation.

(ii) The MC Dropout and Deep Ensemble methods we studied for comparison purpose admit more complex variants which, in principle, may also yield aleatoric and epistemic uncertainties [36, 41, 43]. It would be interesting to perform a similar analysis for them, although they would bring with them additional challenges (e.g. the use of Laplacian priors for MC Dropout

setting as described in [36]; computational cost in picking the optimal ensemble parameters as explained in [43] where a novel automated approach for Deep Ensemble was proposed).

(iii) We had used the same regularization term $\mathcal{L}_{reg} = |y-\gamma|(2\nu+\alpha)$ in the loss function following [29], which is independent of the β parameter of the prior distribution $\mathcal{P}$ in eqn. (3). It would be interesting to explore if there are other alternatives which express full dependence on all four parameters of $\mathcal{P}$. In an analogous framework for classification, Sensoy et al. in [30] showed that the regularization term measuring the Kullback-Leibler divergence from an uninformative prior (e.g. uniform distribution) worked well for classification problems, and notably, the authors of [29] had attempted to find the corresponding version for regression yet without success.

6 Conclusion

In this work, we have presented a novel application of Deep Evidential Learning in the domain of radiotherapy dose prediction. Using medical images of the OpenKBP Challenge dataset, we found that this model can be effectively harnessed to yield uncertainty estimates upon completion of network training. This was achieved only after reformulating the original loss function of [29] for a stable implementation. Since epistemic uncertainty was found to be highly correlated with prediction errors, its distribution could be used to discover and highlight areas of potential inaccuracies of the neural network, apart from being a diagnostic indicator of model reliability. Towards enhancing its clinical relevance, we demonstrated how to construct the predicted Dose-Volume-Histograms' confidence intervals. We hope that this work has furnished the crucial preliminary steps towards realizing Deep Evidential Learning for dose prediction models, paving another path towards quantifying the reliability of treatment plans in the context of knowledge-based-planning.

References

[1] Woody NM, Gregory M M Videtic MDCMF, Vassil AD. Handbook of Treatment Planning in Radiation Oncology. Springer Publishing Company; 2014.

[2] Fraass B, Doppke K, Hunt M, Kutcher G, Starkschall G, Stern R, et al. American Association of Physicists in Medicine Radiation Therapy Committee Task Group 53: Quality assurance for clinical radiotherapy treatment planning. Medical Physics. 1998;25(10):1773-829.

[3] Nelms BE, Robinson G, Markham J, Velasco K, Boyd S, Narayan S, et al. Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems. Pract Radiat Oncol. 2012 Oct-Dec;2(4):296-305.

[4] Craft DL, Hong TS, Shih HA, Bortfeld TR. Improved planning time and plan quality through multicriteria optimization for intensity-modulated radiotherapy. Int J Radiat Oncol Biol Phys. 2012 Jan;82(1):e83-90.

[5] Kui X, Liu F, Yang M, Wang H, Liu C, Huang D, et al. A review of dose prediction methods for tumor radiation therapy. Meta-Radiology. 2024;2(1):100057.

[6] Wu B, Ricchetti F, Sanguineti G, Kazhdan M, Simari P, Chuang M, et al. Patient geometry-driven information retrieval for IMRT treatment plan quality control. Medical Physics. 2009;36(12):5497-505.

[7] Babier A, Mahmood R, Zhang B, Alves VGL, Barragán-Montero AM, Beaudry J, et al. OpenKBP-Opt: an international and reproducible evaluation of 76 knowledge-based planning pipelines. Phys Med Biol. 2022 Sep;67(18).

[8] Nguyen D, Barkousaraie AS, Bohara G, Balagopal A, McBeth R, Lin MH, et al. A comparison of Monte Carlo dropout and bootstrap aggregation on the performance and uncertainty estimation in radiation therapy dose prediction with deep learning neural networks. Physics in Medicine and Biology. 2021;66(5):054002.

[9] Nguyen D, Long T, Jia X, Lu W, Gu X, Iqbal Z, et al. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. Sci Rep. 2019 Jan;9(1):1076.

[10] Kearney V, Chan JW, Haaf S, Descovich M, Solberg TD. DoseNet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks. Physics in Medicine and Biology. 2018 dec;63(23):235022. Available from: `https://dx.doi.org/10.1088/1361-6560/aaef74`.

[11] Shiraishi S, Moore KL. Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy. Medical Physics. 2016;43(1):378-87.

[12] Barragán-Montero AM, Nguyen D, Lu W, Lin MH, Norouzi-Kandalan R, Geets X, et al. Three-dimensional dose prediction for lung IMRT patients with deep neural networks: robust learning from heterogeneous beam configurations. Medical Physics. 2019;46(8):3679-91.

[13] Nguyen D, Jia X, Sher D, Lin MH, Iqbal Z, Liu H, et al. 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. Physics in Medicine and Biology. 2019 mar;64(6):065020.

[14] Babier A, Zhang B, Mahmood R, Moore KL, Purdie TG, McNiven AL, et al. OpenKBP: The open-access knowledge-based planning grand challenge and dataset. Medical Physics. 2021 Jun;48(9):5549–5561. Available from: http://dx.doi.org/10.1002/mp.14845.

[15] Soomro MH, Alves VGL, Nourzadeh H, Siebers JV. DeepDoseNet: a deep learning model for 3D dose prediction in radiation therapy. arXiv preprint arXiv:211100077. 2021.

[16] Li H, Peng X, Zeng J, Xiao J, Nie D, Zu C, et al. Explainable attention guided adversarial deep network for 3D radiotherapy dose distribution prediction. Knowledge-Based Systems. 2022;241:108324.

[17] Hu C, Wang H, Zhang W, Xie Y, Jiao L, Cui S. TrDosePred: A deep learning dose prediction algorithm based on transformers for head and neck cancer radiotherapy. J Appl Clin Med Phys. 2023 Jul;24(7):e13942.

[18] Wang K, Tan HS, Mcbeth R. Swin UNETR++: Advancing Transformer-Based Dense Dose Prediction Towards Fully Automated Radiation Oncology Treatments; 2024.

[19] Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion. 2021 dec;76:243-97. Available from: https://doi.org/10.1016%2Fj.inffus.2021.05.008.

[20] Lambert B, Forbes F, Doyle S, Dehaene H, Dojat M. Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. Artificial Intelligence in Medicine. 2024;150:102830.

[21] Ghesu FC, Georgescu B, Mansoor A, Yoo Y, Gibson E, Vishwanath RS, et al. Quantifying and leveraging predictive uncertainty for medical image assessment. Medical Image Analysis. 2021;68:101855.

[22] Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'19/IAAI'19/EAAI'19. AAAI Press; 2019. Available from: https://doi.org/10.1609/aaai.v33i01.3301590.

[23] Zou K, Yuan X, Shen X, Wang M, Fu H. TBraTS: Trusted Brain Tumor Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII. Berlin, Heidelberg: Springer-Verlag; 2022. p. 503–513. Available from: https://doi.org/10.1007/978-3-031-16452-1_48.

[24] Jones CK, Wang G, Yedavalli V, Sair H. Direct quantification of epistemic and aleatoric uncertainty in 3D U-net segmentation. J Med Imaging (Bellingham). 2022 May;9(3):034002.

[25] Benson HP. Existence of efficient solutions for vector maximization problems. Journal of Optimization Theory and Applications. 1978;26(4):569-80.

[26] Chan TCY, Craig T, Lee T, Sharpe MB. Generalized Inverse Multiobjective Optimization with Application to Cancer Therapy. Operations Research. 2014;62(3):680-95.

[27] Zhang T, Bokrantz R, Olsson J. Probabilistic feature extraction, dose statistic prediction and dose mimicking for automated radiation therapy treatment planning. Med Phys. 2021 Sep;48(9):4730-42.

[28] Nilsson V, Gruselius H, Zhang T, De Kerf G, Claessens M. Probabilistic dose prediction using mixture density networks for automated radiation therapy treatment planning. Phys Med Biol. 2021 Feb;66(5):055003.

[29] Amini A, Schwarting W, Soleimany A, Rus D. Deep evidential regression. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20. Red Hook, NY, USA: Curran Associates Inc.; 2020. .

[30] Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems. 2018;31.

[31] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham: Springer International Publishing; 2016. p. 424-32.

[32] Zou K, Chen Z, Yuan X, Shen X, Wang M, Fu H. A review of uncertainty estimation and its application in medical imaging. Meta-Radiology. 2023;1(1):100003.

[33] Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Balcan MF, Weinberger KQ, editors. Proceedings of The 33rd International Conference on Machine Learning. vol. 48 of Proceedings of Machine Learning Research. New York, New York, USA: PMLR; 2016. p. 1050-9. Available from: https://proceedings.mlr.press/v48/gal16.html.

[34] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6405–6416.

[35] Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence. 2022 Oct;115:105151. Available from: http://dx.doi.org/10.1016/j.engappai.2022.105151.

[36] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 5580–5590.

[37] Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med Phys. 2003 May;30(5):979-85.

[38] Craft D, Süss P, Bortfeld T. The tradeoff between treatment plan quality and required number of monitor units in intensity-modulated radiotherapy. Int J Radiat Oncol Biol Phys. 2007 Apr;67(5):1596-605.

[39] He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: 2015 IEEE International Conference on Computer Vision (ICCV); 2015. p. 1026-34.

[40] Gravel P, Beaudoin G, De Guise JA. A method for modeling noise in medical images. IEEE Trans Med Imaging. 2004 Oct;23(10):1221-32.

[41] Valdenegro-Toro M, Saromo D. A Deeper Look into Aleatoric and Epistemic Uncertainty Disentanglement. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:1508-16. Available from: https://api.semanticscholar.org/CorpusID:248266412.

[42] Weiss NA, Holmes PT, Hardy M. A Course in Probability. Pearson Addison Wesley; 2006. Available from: https://books.google.com/books?id=Be9fJwAACAAJ.

[43] Egele R, Maulik R, Raghavan K, Lusch B, Guyon I, Balaprakash P. AutoDEUQ: Automated Deep Ensemble with Uncertainty Quantification. In: 2022 26th International Conference on Pattern Recognition (ICPR). Los Alamitos, CA, USA: IEEE Computer Society; 2022. p. 1908-14. Available from: https://doi.ieeecomputersociety.org/10.1109/ICPR56361.2022.9956231.