Neural Computing and Applications (2023) 35:22071–22085
https://doi.org/10.1007/s00521-022-08016-4

S.I.: DEEP LEARNING IN MULTIMODAL MEDICAL IMAGING FOR CANCER DETECTION

Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation

Hao Li[1,2] · Yang Nan[1] · Javier Del Ser[3,4] · Guang Yang[1,5]

Received: 10 August 2022 / Accepted: 26 October 2022 / Published online: 17 November 2022

Abstract
Despite recent advances in the accuracy of brain tumor segmentation, the results still suffer from low reliability and robustness. Uncertainty estimation is an efficient solution to this problem, as it provides a measure of confidence in the segmentation results. The current uncertainty estimation methods based on quantile regression, Bayesian neural network, ensemble, and Monte Carlo dropout are limited by their high computational cost and inconsistency. In order to overcome these challenges, Evidential Deep Learning (EDL) was developed in recent work but primarily for natural image classification and showed inferior segmentation results. In this paper, we proposed a region-based EDL segmentation framework that can generate reliable uncertainty maps and accurate segmentation results, which is robust to noise and image corruption. We used the Theory of Evidence to interpret the output of a neural network as evidence values gathered from input features. Following Subjective Logic, evidence was parameterized as a Dirichlet distribution, and predicted probabilities were treated as subjective opinions. To evaluate the performance of our model on segmentation and uncertainty estimation, we conducted quantitative and qualitative experiments on the BraTS 2020 dataset. The results demonstrated the top performance of the proposed method in quantifying segmentation uncertainty and robustly segmenting tumors. Furthermore, our proposed new framework maintained the advantages of low computational cost and easy implementation and showed the potential for clinical application.

Keywords Evidential deep learning · Brain tumor segmentation · Uncertainty quantification · Robustness

✉ Guang Yang
g.yang@imperial.ac.uk

Hao Li
hao.li19@imperial.ac.uk

Yang Nan
y.nan20@imperial.ac.uk

Javier Del Ser
javier.delser@tecnalia.com

[1] National Heart and Lung Institute, Faculty of Medicine, Imperial College London, London, UK

[2] Department of Bioengineering, Faculty of Engineering, Imperial College London, London, UK

[3] TECNALIA, Basque Research and Technology Alliance (BRTA), Derio, Spain

[4] University of the Basque Country (UPV/EHU), Bilbao, Spain

[5] Royal Brompton Hospital, London, UK

1 Introduction

Automated brain tumor segmentation promises to provide more reliable measurements for cancer diagnosis and assessment, establishing new possibilities for high-throughput analysis [1]. Segmentation enables clinicians to determine tumor location, extent, and subtype. Additionally, brain tumor segmentation on longitudinal MRI scans can facilitate monitoring tumor growth or shrinkage. In current clinical practice, accurate segmentation of brain tumor regions is usually done manually by experienced radiologists, which is time-consuming and labor-intensive. Furthermore, manual labeling of results may involve human bias, as they rely on the physician's experience and subjective decision-making. On the other hand, automated segmentation techniques can reduce labor and human bias

to provide efficient, objective, and reproducible results for tumor diagnosis and monitoring.

The performance of automated brain tumor segmentation methods has grown rapidly over the past few years. This development is due to the growth of annotated datasets and the advent of deep learning models that can leverage large amounts of data [2]. Most methods are based on fully convolutional neural networks (FCN) [3], like U-Net [4] and its variants [5] for improving the performance of brain tumor segmentation. Recently, Transformers and self-attention, originating from natural language processing (NLP), have also been applied to medical image segmentation [6].

Although the segmentation results of deep neural networks are reported to be close to or comparable to human performance [7], their robustness levels are low, and concerns remain with their clinical acceptability [8]. Possible reasons include the large variability in imaging properties, such as artifacts and magnetic field strength, as well as the inherent heterogeneity of brain tumors, which are beyond the training dataset. Furthermore, human bias in dataset annotations can cause models also to inherit this bias. One possible direction to alleviate the reliability problem of deep neural networks is to use uncertainty estimation. The uncertainty reflects how confident the network predicts the class labels. Confidence studies can help identify areas of data dominated by lack of annotations (epistemic uncertainty) or noisy annotations (aleatoric uncertainty). Additional information from uncertainty estimation can be used to quantify segmentation performance or as a post-processing step to correct automatic segmentation. Clinically, uncertainty estimates can feed back potential error regions to guide or automate corrections or be used for patient-level segmentation failure detection [1]. Therefore, reliably quantifying the uncertainty of segmentation performance is critical in clinical applications.

Popular methods for quantifying uncertainty in neural networks include quantile regression (QR) [9], Bayesian neural network (BNN) [10–12], ensemble-based [13], dropout-based [14–16], and evidential deep learning (EDL) [17–19]. Simply interpreting the confidence scores of softmax/sigmoid outputs as event probabilities in a categorical distribution can lead to overconfident wrong prediction [20, 21]. The classical BNN aims to capture uncertainty by learning the weight distribution of the network and approximates the integral of parameters by variational inference or Laplace approximation to estimate the posterior prediction distribution [18]. However, most BNNs are challenging to implement and train since model parameters have to be explicitly modeled as random variables [22], which lack scalability in both architecture and data size [12]. Hence, subsequent approaches focused on being able to reuse the training pipeline and maintain scalability while providing reasonable uncertainty estimates. To this end, more intuitive and simple methods, such as learning an ensemble of deterministic networks [15, 21] and introducing Monte Carlo dropout [13] are proposed for brain tumor segmentation. On the downside, ensemble-based methods need to train multiple models from scratch, which is computationally expensive, and the introduction of dropouts results in inconsistent outputs [23].

On the other hand, EDL has been gradually developed in recent studies, demonstrating more promising and reliable performance in uncertainty estimation. Based on the Dempster-Shafer Evidence Theory (DST) [24], EDL uses the Dirichlet distribution to model the categorical distribution of the output given the input to the network. This class of methods produces closed-form prediction distributions and outperforms BNNs in adversarial queries and out-of-distribution uncertainty quantification [18]. Compared to ensemble-based and dropout-based methods, EDL showed more robust results with lower computational costs [25]. However, most of the recent works focus on the natural image classification and segmentation problem, making the application of EDL in medical image segmentation to be further studied.

In this paper, we propose a region-based EDL network for reliable brain tumor segmentation, which is robust to noise and corruption of images. The network learned classification distribution by minimizing region-based prediction error under the Dirichlet prior distribution. This enabled the proposed network to provide accurate segmentation results and reliable uncertainty estimate simultaneously, even under noise-corrupted inputs. Our method improved the mean squared error (MSE) loss used for the simple natural image classification [17], making it more suitable for semantic segmentation of medical images. The main contributions of our work can be summarized as follows:

- An EDL framework was adopted for accurate brain tumor segmentation, which can quantify the uncertainty of segmentation results and improve the reliability and robustness of segmentation with less computational complexity compared to ensemble-based and dropout-based methods.
- A novel training loss was developed based on minimizing the region-based prediction error under the Dirichlet prior distribution to improve the segmentation accuracy of EDL. Theoretical properties are fully provided to guarantee the evidential learning of the model.
- A new evaluation metric called soft uncertainty-error overlap (sUEO) was designed for uncertainty

estimation to assess the model's ability to localize segmentation errors more easily.
- The robustness of the segmentation accuracy and uncertainty quantification of the proposed method is comparatively evaluated on the BraTS2020 dataset using image corruption techniques. The effectiveness and efficiency of the novel loss function were verified.

The rest of the paper is structured as follows: Sect. 2 briefly introduces EDL and recent development. Section 3 details our segmentation framework, including EDL and novel loss functions. Section 4 illustrates the experimental setup and evaluation metrics, and the results are analysed and discussed in Sect. 5. The conclusion and future research directions are given in Sect. 6.

2 Related work

Despite many uncertainty estimation methods mentioned, our proposed framework resorts to arguably the most cutting-edge methodology, EDL, for this purpose. The rest of the section presents the principles of EDL (Sect. 2.1) and a brief overview (Sect. 2.2) of the scarce contributions in which EDL has been utilized for tumor segmentation.

2.1 Principles of EDL

Evidence Deep Learning (EDL) is based on Dempster-Shafer Evidence Theory (DST) [24], which is a generalization of Bayesian theory to subjective probability. It assigns belief masses to subsets of a discerning frame, representing a set of exclusive potential states, such as possible class labels for a voxel. A belief mass can be assigned to any subset of the frame. Assigning all belief masses to the entire frame represents the opinion that the truth can be any possible state, e.g. any label is equally likely.

The belief distribution of DST in the discerning framework can be formalized as a Dirichlet distribution by Subjective Logic (SL) [26]. For a voxel i, the Dirichlet distribution $Dir(\boldsymbol{\alpha}_i)$ is parameterized by a vector of Dirichlet parameters α_{ij} for K classes, where j denotes the j-th class. (The denotations of subscripts i and j hold for the entire paper.) The neural network collects evidence e_{ij} from the input data, a measure of support that facilitates classifying samples into the class j. The belief mass distribution, i.e. subjective opinion, in [17] corresponds to a Dirichlet distribution with parameter $\alpha_{ij} = e_{ij} + 1$.

As a result, it is equivalent to placing a Dirichlet distribution on the predicted categorical distribution, allowing a single network to output different predictions. The output layer of an EDL-based neural network parameterizes a simplex distribution representing the probability distribution of class assignments. The softmax/sigmoid classification layer is replaced with a ReLU activation layer that outputs non-negative continuous values, resulting in e_{ij}. The vector of predicted classification probabilities can be computed by:

$$\hat{p}_{ij} = \frac{\alpha_{ij}}{S_i}, \tag{1}$$

where $S_i = \sum_{j=1}^{K} \alpha_{ij}$ is called the Dirichlet strength. The class probability vector for voxel i given by $\mathbf{p}_i$ is modeled as a random vector drawn from the Dirichlet distribution [18].

Let $\mathbf{y}_i$ be the one-hot encoded labels with $y_{ik} = 1$ and $y_{ij} = 0$ for all $j \neq k$. Treating the Dirichlet distribution $Dir(\mathbf{p}_i \mid \boldsymbol{\alpha}_i)$ as a prior on the multinomial likelihood $Mult(\mathbf{y}_i \mid \mathbf{p}_i)$, one can minimize the negative logarithm of the marginal likelihood:

$$\begin{aligned}\mathcal{L}_{\mathrm{ML,i}} &= -\log\left(\int \prod_{j=1}^{K} p_{ij}^{\,y_{ij}} \frac{1}{\mathcal{B}(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} p_{ij}^{\,\alpha_{ij}-1} d\mathbf{p}_i\right) \\ &= \sum_{j=1}^{K} y_{ij}\big(\log(S_i) - \log(\alpha_{ij})\big),\end{aligned} \tag{2}$$

where $\mathcal{B}$ is the multinomial beta function [27]. Alternatively, one can minimize the Bayes risk of the cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{CE,i}} &= \int \left[-\sum_{j=1}^{K} y_{ij} \log(p_{ij})\right] \frac{1}{\mathcal{B}(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} p_{ij}^{\,\alpha_{ij}-1} d\mathbf{p}_i \\ &= \sum_{j=1}^{K} y_{ij}\big(\psi(S_i) - \psi(\alpha_{ij})\big),\end{aligned} \tag{3}$$

or the mean squared error:

$$\begin{aligned}\mathcal{L}_{\mathrm{MSE,i}} &= \int \|\mathbf{y}_i - \mathbf{p}_i\|^2 \frac{1}{\mathcal{B}(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} p_{ij}^{\,\alpha_{ij}-1} d\mathbf{p}_i \\ &= \sum_{j=1}^{K} \big(y_{ij} - \hat{p}_{ij}\big)^2 + \frac{\hat{p}_{ij}\big(1-\hat{p}_{ij}\big)}{(S_i+1)},\end{aligned} \tag{4}$$

where ψ refers to the digamma function [28]. Sensoy et al. [17] observed that $\mathcal{L}_{\mathrm{ML,i}}$ and $\mathcal{L}_{\mathrm{CE,i}}$ produced excessively high belief masses and were less stable than $\mathcal{L}_{\mathrm{MSE,i}}$. This

can be attributed to the fact that these two loss functions encourage maximizing the correct likelihood.

2.2 Related work of EDL

Sensoy et al. [17] used the MSE loss for natural image classification. They showed that the loss decreases as the correct class parameter grows and decreases when the largest incorrect parameter decays. Furthermore, they integrated the KL divergence loss to narrow the error class parameters further. However, the properties of the aggregated loss function were not shown, and the behavior of the loss was not studied for all parameters. Also, for the image classification problem, [18] improved the square-norm of MSE loss to max-norm and achieved higher performance. Because max-norm minimizes the highest class prediction error, and square-norm minimizes the total sum of squares, which is more susceptible to outliers. However, this situation may not be applicable for tumor segmentation with severe class imbalance. The TBraTS network [25] attempted to apply EDL's CE loss to brain tumor segmentation. In order to improve the segmentation accuracy, the network output was additionally passed through the softmax layer to calculate the soft Dice loss, which was added with the CE loss. However, this increases training costs and complexity, and an incomplete deployment of the EDL framework may cause the network to fail to produce true evidence values.

Different from these methods that employed MSE or CE loss and show inferior segmentation results, our approach minimized region-based prediction error (soft Dice loss) under the Dirichlet prior distribution, which significantly facilitated the segmentation performance of the EDL framework. The improvement of EDL in segmentation was statistically verified in the medical image dataset, paving the way for the clinical application of the EDL-based segmentation and uncertainty estimation framework.

3 Method

This section details our approach, a novel region-based EDL framework for 3D brain tumor segmentation (Sect. 3.1) and describes how we quantify the uncertainty (Sect. 3.2).

3.1 Region-based evidential deep learning

For semantic segmentation of medical images, it is important to consider the accuracy of segmented regions in addition to standard classification errors. Hence, we proposed a region-based objective to minimize the expected prediction error in the EDL framework while maintaining high segmentation accuracy. Unlike Zou et al. [25] who added the soft Dice (sDice) loss based on the result of softmax activation to $\mathcal{L}_{\mathrm{CE,i}}$, we directly minimized the Bayes risk of sDice loss:

$$\mathrm{sDice} = \frac{1}{K}\sum_{j=1}^{K} 1 - \frac{2\sum_i y_{ij}p_{ij}}{\sum_i {y_{ij}}^2 + {p_{ij}}^2}, \tag{5}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{DICE}} &= \int [\mathrm{sDice}]\frac{1}{\mathcal{B}(\boldsymbol{\alpha}_i)}\prod_{j=1}^{K} {p_{ij}}^{\alpha_{ij}-1} d\mathbf{p}_i \\ &= \frac{1}{K}\sum_{j=1}^{K}\mathbb{E}\left[1 - \frac{2\sum_i y_{ij}p_{ij}}{\sum_i {y_{ij}}^2 + {p_{ij}}^2}\right] \\ &= 1 - \frac{2}{K}\sum_{j=1}^{K}\frac{\sum_i y_{ij}\mathbb{E}[p_{ij}]}{\sum_i {y_{ij}}^2 + \mathbb{E}[{p_{ij}}^2]}.\end{aligned} \tag{6}$$

By using the identity:

$$\mathbb{E}[{p_{ij}}^2] = \mathbb{E}[p_{ij}]^2 + \mathrm{Var}(p_{ij}), \tag{7}$$

the equation can be formulated in an easily interpretable form:

$$\begin{aligned}\mathcal{L}_{\mathrm{DICE}} &= 1 - \frac{2}{K}\sum_{j=1}^{K}\frac{\sum_i y_{ij}\hat{p}_{ij}}{\underbrace{\sum_i {y_{ij}}^2 + \hat{p}_{ij}^2}_{\mathrm{sDiceDen}} + \underbrace{\frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{(S_i+1)}}_{\mathrm{Var}}} \\ &= 1 - \frac{2}{K}\sum_{j=1}^{K}\frac{\sum_i y_{ij}\frac{\alpha_{ij}}{S_i}}{\sum_i {y_{ij}}^2 + \left(\frac{\alpha_{ij}}{S_i}\right)^2 + \frac{\alpha_{ij}(S_i-\alpha_{ij})}{{S_i}^2(S_i+1)}}.\end{aligned} \tag{8}$$

By factoring out the denominator of sDice (sDiceDen) and variance (Var), the loss aims to achieve the joint goal of minimizing the region-based prediction error and variance of the Dirichlet experiments generated by the neural network for the training set.

In order to ensure an effective EDL framework that allows the network to learn to generate subjective opinions from evidence correctly, the loss function needs to have the following properties.

Hypothesis 1 When the network optimises, the loss function prioritizes data fitting over variance estimation.

Hypothesis 2 The loss function has a tendency to fit the data.

Hypothesis 3 The loss function avoids generating evidence for all observations it cannot explain.

These properties of the proposed DICE loss ($\mathcal{L}_{\mathrm{DICE}}$) can be guaranteed by the following theorems, each numbered one-to-one with the hypothesis. The proofs of all Theorems are presented in Appendix 1.

Theorem 1 *For any* $\alpha_{ij} \geq 1$, *the inequality sDiceDen > Var is satisfied.*

Theorem 2 *For a given voxel p with the correct label c,* $\mathcal{L}_{\text{DICE}}$ *decreases when new evidence is added to* α_{pc} *and increases when evidence is removed from* α_{pc}.

Theorem 3 *For a given voxel p with the correct label c,* $\mathcal{L}_{\text{DICE}}$ *decreases when evidence is removed from all incorrect Dirichlet parameters* α_{pw} *for all* $w \neq c$.

To summarise, Theorems 1 to 3 demonstrate that the proposed loss function can optimize the neural network to provide more evidence for the correct class of each voxel while avoiding misclassification by discarding misleading evidence. By accumulating evidence, the loss also tends to reduce the variance of its predictions on the training set, but only if the additional evidence leads to a better fit to the data.

Furthermore, to further minimize the contribution of parameters associated with incorrect classes, a KL divergence loss function is introduced to shrink their evidence to 0 as follows:

$$\begin{aligned}\mathcal{L}_{\text{KL,i}} = &\log\left(\frac{\Gamma\left(\sum_{j=1}^{K}\widetilde{\alpha}_{ij}\right)}{\Gamma(K)\prod_{j=1}^{K}\Gamma\left(\widetilde{\alpha}_{ij}\right)}\right)\\ &+\sum_{j=1}^{K}\left(\widetilde{\alpha}_{ij}-1\right)\left[\psi\left(\widetilde{\alpha}_{ij}\right)-\psi\left(\sum_{j=1}^{K}\widetilde{\alpha}_{ij}\right)\right],\end{aligned} \tag{9}$$

where $\Gamma(\cdot)$ is the gamma function [28] and $\widetilde{\boldsymbol{\alpha}}_i = \mathbf{y}_i + (\mathbf{1} - \mathbf{y}_i) \odot \boldsymbol{\alpha}_i$ is the Dirichlet parameters after removal of the non-misleading evidence. The following theorem shows a desirable monotonicity property of this regularization loss as a supplementary to [17].

Theorem 4 *For a voxel i with the correct label c,* $\mathcal{L}_{\text{KL,i}}$ *increases in* α_{iw} *for all* $w \neq c$.

Theorems 3 and 4 show that the strength of parameters associated with misleading results is expected to decrease during training. Since the parameters are all expected to be minimized, the preferred behavior of the proposed loss function results in a higher uncertainty of misclassification.

The final loss function is defined as:

$$\mathcal{L}_{\text{EDL}} = \mathcal{L}_{\text{DICE}} + \lambda \mathcal{L}_{\text{KL,mean}}, \tag{10}$$

where $\mathcal{L}_{\text{KL,mean}}$ is the mean KL divergence loss over all voxels and λ is an annealing coefficient. The KL divergence loss is gradually introduced by λ for a stable training due to its strong regularization effect. The annealing scheme is set to reach a maximum $\frac{1}{10}$ as: $\lambda = \frac{1}{10}\min\left(1, \frac{n}{100}\right)^2$ where n is the current epoch.

In addition, the weighted sDice loss, $\mathcal{L}_{\text{wDICE}}$, is also proposed to ease the class imbalance between tumor and background voxels. The weight for each segmentation class is one minus the ratio of foreground voxels to background voxels. Since the weights are all positive and class-wise, all theoretical properties of the loss function still hold.

Furthermore, the parameter of Dirichlet distribution in our framework is re-defined as:

$$\alpha_{ij} = \left(e_{ij} + 1\right)^2. \tag{11}$$

Unlike [17] defined the Dirichlet $\alpha_{ij} = e_{ij} + 1$, the alternative formula allows the network to output high Dirichlet parameters more easily. This avoids the defect that it is almost impossible for the network to express a high degree of uncertainty for a particular outcome since each outcome gives a minimal proof of one, i.e. $\alpha_{ij} \geq 1$.

3.2 Uncertainty quantification

Calculating the predictive entropy (PE) is a common way to quantify uncertainty. Based on the information theory, PE uses confidence scores of predictions to calculate the total uncertainty for a voxel i, which is defined as:

$$\mathcal{H}(\mathbf{p}_i) = -\sum_{j=1}^{K} p_{ij}\log(p_{ij}), \tag{12}$$

where $\mathbf{p}_i$ is the confidence score vector [29]. In order to better compare different methods, we normalized the PE by its maximum possible value as:

$$\mathcal{H}(\mathbf{p}_i) = -\frac{1}{\log(K)}\sum_{j=1}^{K} p_{ij}\log(p_{ij}). \tag{13}$$

As a result, the value range of normalized predictive entropy (NPE) is normalized to [0, 1], where 1 implies the maximum uncertainty and 0 implies the absolutely confident prediction.

4 Experiment setup

Experiments on the standard benchmark (BraTS 2020) were conducted to compare different techniques for uncertainty quantification and evaluate qualitatively the produced segmentation along with the uncertainty associated with each voxel. We first present the implementation details (Sect. 4.1) and then introduce the models (Sect. 4.2) and evaluation metrics (Sect. 4.3) for comparative experiments.

4.1 Data acquisition and processing

The BraTS 2020 [7, 30, 31] dataset comprises brain MRI images of various scanners and protocols. The ground truth (GT) label includes the GD-enhancing tumor (ET),

peritumoral edema (ED), and necrotic and non-enhancing tumor core (NCR + NET). The segmentation masks were evaluated on three tumor subregions: the ET, the tumor core (TC = ET + NCR + NET), and the whole tumor (WT = ET + NCR + NET + ED). Four MRI modalities of T1, T1ce, T2, and T2-FLAIR were co-registered with a size of 240 × 240 × 155. They were then interpolated to $1mm^3$ and skull-stripped. Since GT labels are only available for the training set (369 cases), we split the original training set into a new training set of 236 cases, a validation set of 59 cases, and a test set of 74 cases.

All images are cropped to 160 × 192 × 128 to reduce computational waste in the background and are then pre-processed by intensity normalization. During the training, various data augmentation techniques were applied on-the-fly as in [32] to artificially increase the dataset size and minimize the risk of overfitting.

4.2 Model Training and Optimization

We chose the well-validated nnU-Net [33] as our Base network model and configured as in our previous work [32, 34]. All softmax/sigmoid layers in the Base network were replaced with ReLU activation layers as described in the previous section. For comparison, we used different loss functions to optimize the network: $\mathcal{L}_{\text{CE,i}}$, $\mathcal{L}_{\text{MSE,i}}$, $\mathcal{L}_{\text{DICE}}$, and $\mathcal{L}_{\text{wDICE}}$. Since the evaluation would be based on more meaningful tumor subregions, the network was trained to segment each overlapping subregion separately. However, we also trained the network for multi-class segmentation of the basic labels using $\mathcal{L}_{\text{DICE}}$ of $K = 4$ for ablation study.

In addition, we also employed training strategies of Ensemble [15], Dropout [14], and TBraTS [25], which all used soft Dice based loss function for fair comparisons. For Ensemble, we trained five networks with different initialized weights, which has proven to be sufficient in practice [35]. Dropout layers (factor of 0.5) were added to the deepest three layers of the Base network to handle high-level features, which is the most efficient [16].These layers were also active during inference, and the same images were passed 10 times to quantify the prediction uncertainty [14]. Previous research has found that a sampling rate of 10 is adequate for reasonable uncertainty estimation [14]. Moreover, we used the strategy of TBraTS, which combined existing losses for multi-label segmentation.

The adaptive moment estimator (Adam) optimizer was used to optimize all networks in 200 epochs, with a batch size of 1 and an initial learning rate of 0.0003. Experiments were implemented using PyTorch 1.10 on NVIDIA GeForce RTX 3090 GPUs.

4.3 Evaluation metrics

Our method was evaluated using the independent test set of BraTS 2020 (74 cases). The segmentation performance was evaluated using the Dice score, which is defined as:

$$\text{Dice} = \frac{2|\mathcal{X} \cap \mathcal{Y}|}{|\mathcal{X}| + |\mathcal{Y}|}, \tag{14}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are sets of GT and prediction. The Dice score measures spatial overlap between the GT and segmentation results, where a score of 1 indicate a complete overlap.

In addition, the following metrics were utilized to evaluate uncertainty estimation: expected calibration error (ECE) [1], soft uncertainty-error overlap (sUEO), and BraTS score (BraS) [36]. ECE is defined by the absolute calibration error between the confidence interval and the accuracy interval (c_m and a_m, where m is the m-th bin defined in the interval [0, 1]), weighted by the number of voxels (n_m) in the interval. ECE is given by

$$\text{ECE} = \sum_{m=1}^{M} \frac{n_m}{N} |c_m - a_m|, \tag{15}$$

where N and M are the total numbers of voxels and bins, and the confidence is calculated by one minus the uncertainty. ECE ranges from 0 to 1, where lower values indicate better calibration. To reduce the effect of the large, confident, and accurate extracranial regions typically found in brain tumor MRI, we only considered voxels within the brain. Improved on the uncertainty-error overlap (UEO) [1], we proposed the soft uncertainty-error overlap (sUEO) that directly uses the uncertainty quantities (u_i) to measure the overlap:

$$\text{sUEO} = \frac{2\sum_i y_i u_i}{\sum_i {y_i}^2 + {u_i}^2}. \tag{16}$$

sUEO does not require thresholding the uncertainty map, which can save time optimizing the threshold over the validation set. It shows whether a model can precisely localize segmentation errors. Moreover, the comprehensive BraS is defined by:

$$\begin{aligned} \text{BraS} = \frac{1}{3}\big[&\text{AUC}_{\text{Dice}} + (1 - \text{AUC}_{\text{FTP}}) \\ &+ (1 - \text{AUC}_{\text{FTN}})\big], \end{aligned} \tag{17}$$

where AUC_{Dice}, AUC_{FTP}, and AUC_{FTN} are area under three curves: 1) Dice vs. confidence threshold, 2) ratio of filtered True Positives (FTP) vs. confidence threshold, and 3) ratio of filtered True Negatives (FTN) vs. confidence threshold. The curves are plotted against the segmentation filtered by different confidence levels, which only voxels with confidence greater than the threshold retain. This metric rewards uncertainty estimates that yield high confidence for correct segmentations or assigns a low confidence level to incorrect segmentations and penalizes uncertainty measures that result in a higher percentage of under-confident correct segmentations.

5 Results and discussion

This section first evaluates the performance of the novel region-based EDL framework for brain tumor segmentation and uncertainty quantification through experiments on the original dataset (Sect. 5.1). It then examines its robustness by applying various image processing techniques to the test image data (Sect. 5.2).

5.1 Segmentation and uncertainty estimation

Our method generated comparable segmentation results with the GT labels, as visualized in Fig. 1. The quantitative results of our methods averaged over three tumor subregions are compared in Table 1. Although the proposed DICE and wDICE loss functions achieved the highest Dice scores (0.791 and 0.793) among all EDL-based methods, Ensemble and Dropout methods performed slightly more accurately in segmentation (0.807 and 0.804). The success of Ensemble and Dropout was attributed to the variance reduction by combining predictions prone to various errors. However, the dominance of the proposed region-based losses in all EDL frameworks still proved their effectiveness in improving EDL in segmentation performance.

Table 1 Quantitative comparisons of different uncertainty estimation methods on the BraTS 2020 test set

Method	Dice↑	ECE↓	sUEO↑	BraS↑
Ensemble	**0.807** †‡	0.010†‡	0.409†‡	0.873†
Dropout	0.804†‡	**0.009** †‡	0.412‡	0.869†‡
EDL (TBraTS)	0.790‡	0.019‡	0.383‡	0.862†‡
EDL (CE)	0.783†‡	0.038†‡	0.325†‡	0.829†‡
EDL (MSE)	0.783†‡	0.038†‡	0.325†‡	0.829†‡
EDL (DICE)	0.791	0.017	0.414‡	**0.876**
EDL (wDICE)	0.793	0.016	**0.420** †	0.874
EDL (DICE-M)	0.771†‡	0.035†‡	0.283†‡	0.860†‡

†: p-value < 0.05 compared with EDL (DICE) by paired t-test. ‡: p-value < 0.05 compared with EDL (wDICE) by paired t-test. Bold numbers are the best results

Compared to CE-based or MSE-based losses, the DICE-based losses significantlly improved the Dice score by 0.01.

As for the uncertainty estimation, Ensemble and Dropout obtained the lowest ECE metrics of 0.009 and 0,010, which indicated they were well-calibrated. However, our methods achieved the highest sUEO and BraS of 0.420 and 0.875, showing their ability to more precisely locate errors and estimate uncertainty. The EDL model optimized by the proposed wDice loss generated the most accurate uncertainty map to indicate the potential false predictions. On the other hand, the proposed EDL (DICE) model made the most reliable uncertainty estimation, maintaining the lowest error while thresholding along the uncertainty dimension. The advantages of our region-based EDL methods are also shown in Fig. 2. The proposed methods had the most precise uncertainty map consistent with the error map. Ideally, a learned model only give high uncertainty for a possible erroneous prediction. Despite the high segmentation accuracy, both Ensemble and Dropout methods generated more uncertainty around mask boundaries and other correct regions, leading to inferior uncertainty estimation performance in terms of sUEO and BraS. It is also worth mentioning that EDL models trained to segment each

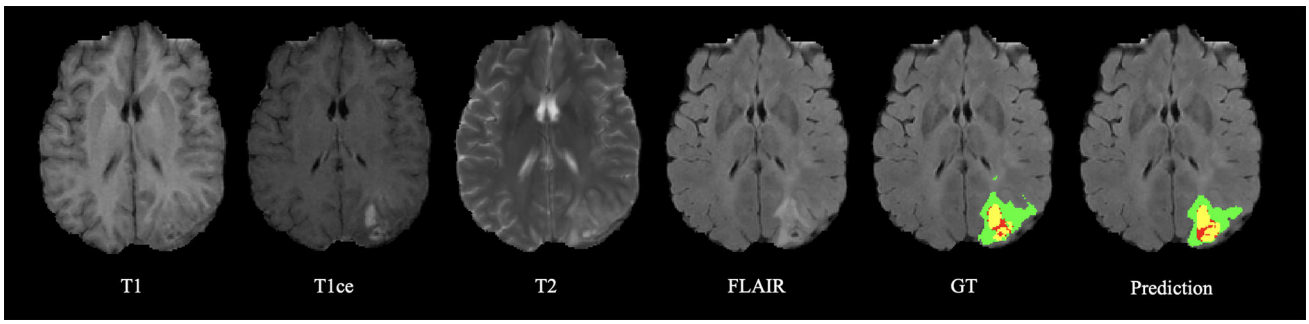
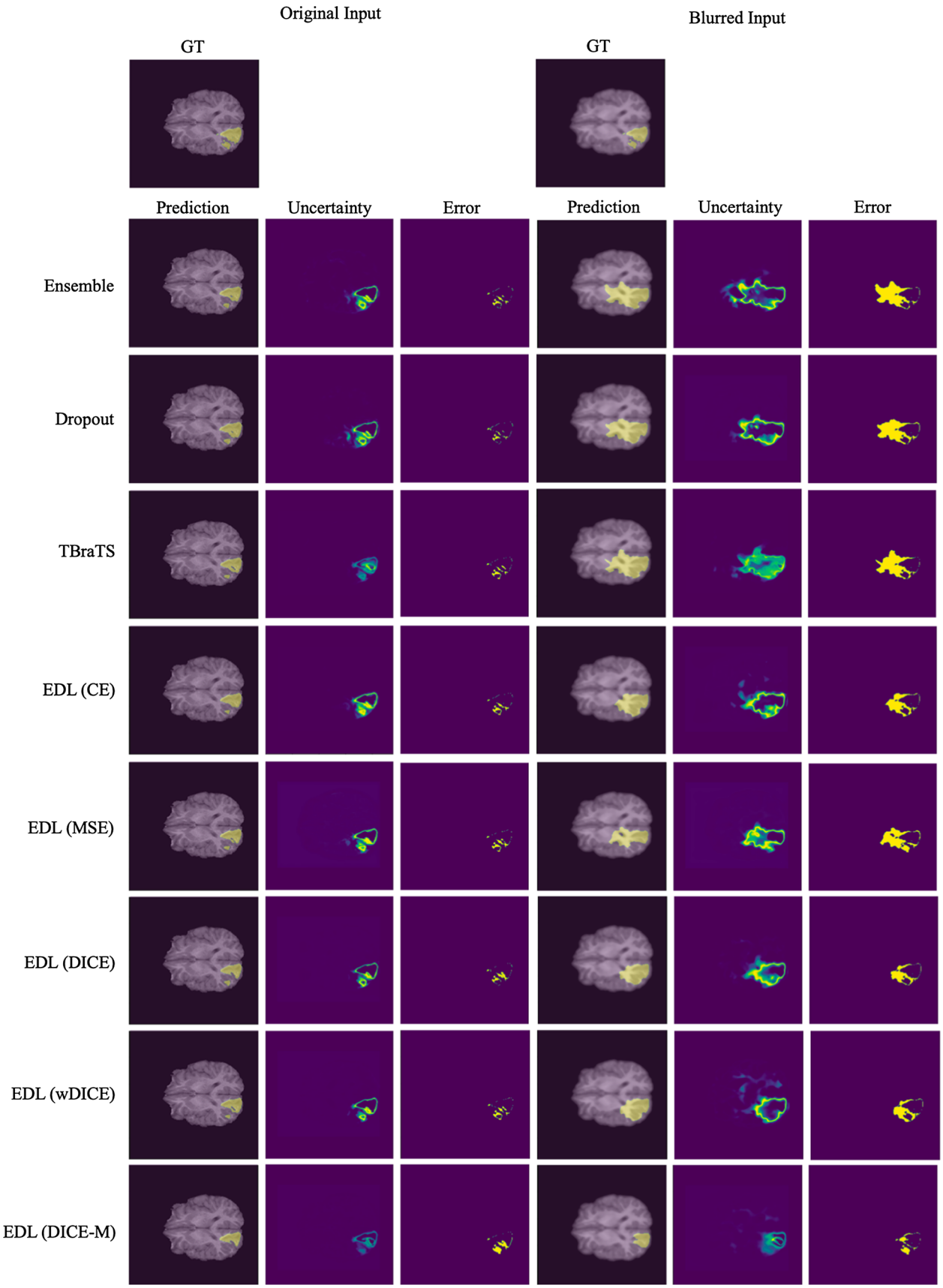


Fig. 1 Representative visual segmentation results of the proposed region-based EDL method on the BraTS 2020 test set. The labels are enhancing tumor (yellow), edema (green), and necrotic and non-enhancing tumor (red)


Original Input
Blurred Input
GT
GT
Prediction
Uncertainty
Error
Prediction
Uncertainty
Error
Ensemble
Dropout
TBraTS
EDL (CE)
EDL (MSE)
EDL (DICE)
EDL (wDICE)
EDL (DICE-M)

◀**Fig. 2** Representative visual results of the whole tumor (WT) produced by different uncertainty estimation methods on the BraTS 2020 test set. The right half of the figure was evaluated on the test images blurred by a Gaussian filter of sigma = 1.5

Table 2 Inference runtimes of different uncertainty estimation methods for one image

Method	Runtime (sec)
Ensemble	6.94 ± 0.05
Dropout	63.68 ± 0.17
EDL (TBraTS)	3.32 ± 0.05
EDL (CE)	3.39 ± 0.06
EDL (MSE)	3.41 ± 0.05
EDL (DICE)	3.38 ± 0.03
EDL (wDICE)	3.40 ± 0.02
EDL (DICE-M)	3.23 ± 0.04

tumor subregion separately outperformed the ones trained with multi-class labels (DICE-M).

In addition, the inference runtimes of the uncertainty estimation methods on one sample are reported in Table 2. Runtimes of all EDL-based methods are lower than the others. This is because both Ensemble and Dropout use multiple sampling mechanisms at inference time to obtain uncertainty estimates.

5.2 Robustness experiment

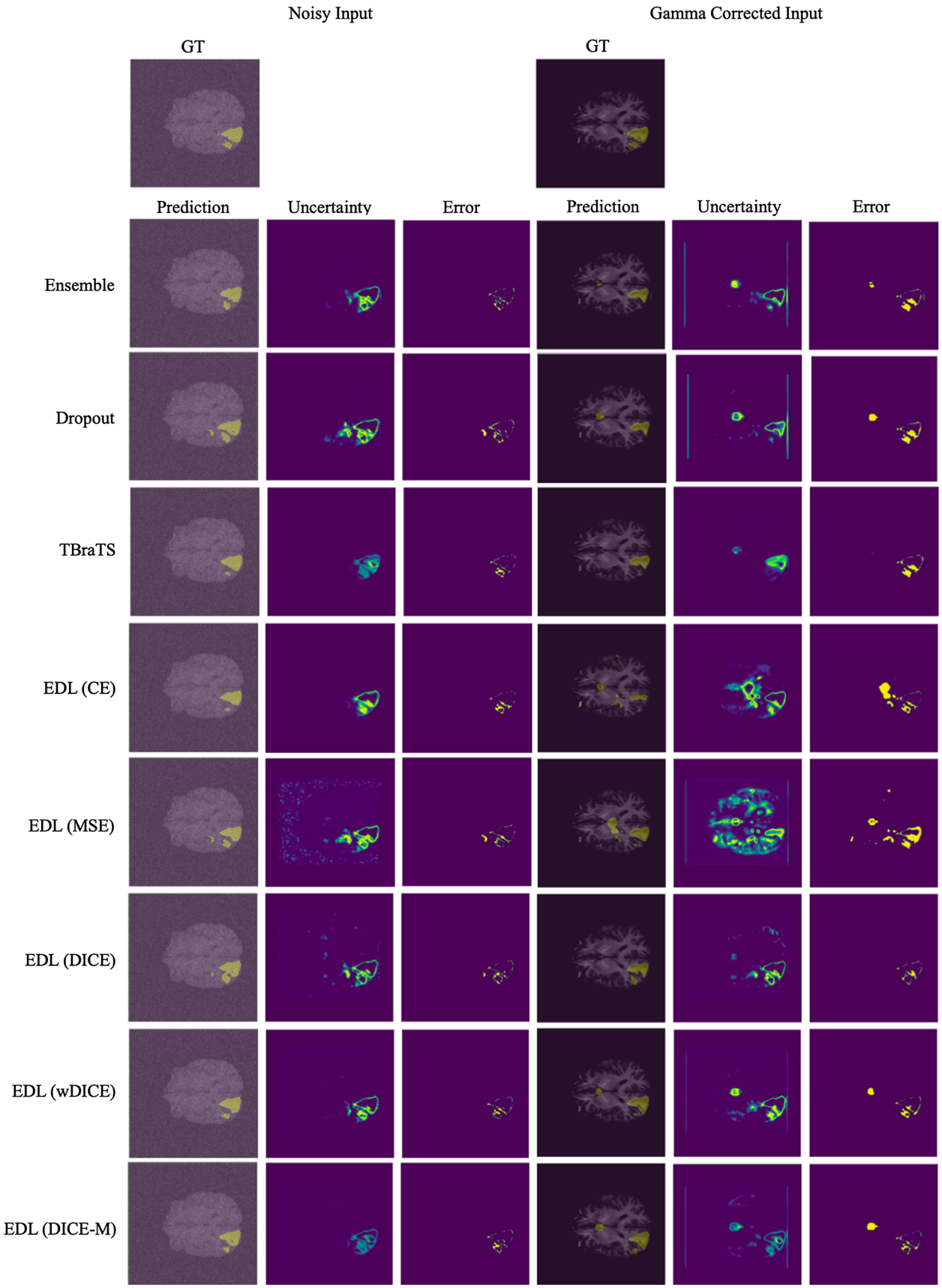
To verify the robustness of the segmentation model, we applied several image processing techniques to simulate the low-quality acquisition that usually happens in actual practice. We first blurred the four modalities of the MRI images using a Gaussian filter with sigma = 1.5. Subsequently, we re-evaluated the performance of all methods for segmentation and uncertainty estimation, as shown in Table 3. We can observe that with the addition of Gaussian blur, the segmentation performance of all methods dropped significantly, especially Ensemble and Dropout. Our method leaped to the highest Dice metric of 0.572 for blurry images, demonstrating its robustness. By comparing the segmentation results with the original input and high-noise input in Fig. 2, it can be seen that the EDL using our loss function segmented the WT region more accurately than all other methods. This is due to the evidence extracted from the data that produced these subjective opinions.

Furthermore, our method exhibited the most reliable uncertainty quantification on blurred images compared to other uncertainty estimation methods. Unlike the uncertainty of the Ensemble and Dropout methods, which only came from the variance of the prediction, the uncertainty of the EDL method represented whether the prediction was supported by sufficient evidence. Therefore, the uncertainty estimates of EDL methods can more correctly indicate possible prediction errors or provide a better rationale for erroneous predictions, such as learning the wrong evidence or failing to identify the correct features. As shown in Table 3, all uncertainty evaluation metrics decreased, except for sUEO for all EDL-based methods. This might benefit from the robust evidence captured by the EDL segmentation framework, where the advantage is more noticeable with larger error regions. The proposed EDL (DICE) method achieved top performance, especially for generating reliable uncertainty maps. Besides showing the robustness of our method, this again demonstrated the importance of region-based loss for locating semantic segmentation errors. As shown in the right half of Fig. 2,

Table 3 Quantitative comparisons of different uncertainty estimation methods on preprocessed BraTS 2020 test set. (Bold numbers: best results)

Method	Blurred				Noisy				Gamma corrected			
	Dice ↑	ECE ↓	sUEO ↑	BraS ↑	Dice ↑	ECE ↓	sUEO ↑	BraS ↑	Dice ↑	ECE ↓	sUEO ↑	BraS ↑
Ensemble	0.561†‡	0.025	0.408†‡	0.772†‡	0.751†‡	0.024	0.407†‡	0.780†‡	0.702†‡	0.035	0.388†‡	0.730†‡
Dropout	0.544†‡	0.026	0.401†‡	0.772†‡	0.729†‡	0.027†‡	0.405†‡	0.774†‡	0.693†‡	0.040†	0.392†‡	0.724†‡
EDL (TBraTS)	0.546†‡	0.025	0.384†‡	0.636†‡	0.756†‡	0.024	0.387†‡	0.654†‡	0.702†‡	0.044†‡	0.345†‡	0.680†‡
EDL (CE)	0.562†‡	0.028†‡	0.416†‡	0.775†‡	0.756†‡	0.027†‡	0.440†‡	0.785†‡	0.659†‡	0.053†‡	0.460†‡	0.751†‡
EDL (MSE)	0.559†‡	0.038†‡	0.419†‡	0.770†‡	0.739†‡	0.040†‡	0.391†‡	0.773†‡	0.632†‡	0.071†‡	0.413†‡	0.700†‡
EDL (DICE)	0.571	**0.024**	**0.458** ‡	**0.796**	0.769	**0.022**	0.447	**0.803**	**0.711**	**0.033**	**0.480** ‡	**0.768**
EDL (wDICE)	**0.572**	0.025	0.442†	0.793	**0.771**	0.023	**0.449**	0.799	0.707	0.036	0.448†	0.763
EDL (DICE-M)	0.438†‡	0.037†‡	0.343†‡	0.682†‡	0.758†‡	0.027†‡	0.398†‡	0.723†‡	0.704†	0.037	0.350†‡	0.626†‡

†: p-value < 0.05 compared with EDL (DICE) by paired t-test. ‡: p-value < 0.05 compared with EDL (wDICE) by paired t-test. Bold numbers are the best results


Noisy Input
Gamma Corrected Input
GT
GT
Prediction
Uncertainty
Error
Prediction
Uncertainty
Error
Ensemble
Dropout
TBraTS
EDL (CE)
EDL (MSE)
EDL (DICE)
EDL (wDICE)
EDL (DICE-M)

◀**Fig. 3** Representative visual results of the whole tumor (WT) produced by different uncertainty estimation methods on the BraTS 2020 test set. The left half of the figure was evaluated on the test images added with Gaussian noise of variance = 1.5. The right half was evaluated on the test images after Gamma correction of gamma = 5

the uncertainty map generated by the EDL (DICE) method is the most relevant to the error map. Unlike other methods that only generate high uncertainty at the edges of the predicted mask, the proposed method can indicate potential error regions inside masks. It is true that regional boundaries should a priori undergo higher uncertainty, but this high uncertainty should not be assumed to be exclusive of these regions. This is what is favored by our proposed methods, as they declare high uncertainty in delimiting and non-delimiting regions of the segmented image. This demonstrates the potential of our proposed method for clinical application. Potential error regions fed back by the model can assist in automatic correction or quality quantification of tumor segmentation.

To enrich the robustness experiment, we further applied Gaussian noise and Gamma correction to the original input to simulate the noise and the contrast variability introduced by imaging or enhancement technique. With the Gaussian noise of variance = 1.5, the segmentation performance of all methods was superior to burred ones, as shown in Table 3. The proposed EDL (wDICE) remained the top in The proposed EDL (wDICE) method remained the top in terms of the Dice metric (0.771), followed by the proposed EDL (DICE) with Dice 0.769. The metrics for uncertainty estimation resembled blurred input, while the sUEO of EDL (wDICE) became 0.002 higher than that of EDL (DICE). However, as for the result after Gamma correction with gamma = 5, the proposed EDL (DICE) method excelled in all metrics, which showed its robustness to unexpected contrast variance.

These observations can also be visually inspected in Fig. 3. Compared to other methods, our methods provided more precise uncertainty maps. Ensemble and Dropout models had trouble handling the boundaries, especially for Gamma corrected input. The extracranial area is no longer zero after Gamma correction, which might cause problems when applying zero-padding. Moreover, their overconfident prediction using softmax/sigmoid is shown for Gamma corrected input where the main error regions were not indicated in the uncertainty map. Non-region-based EDL methods also showed inaccurate uncertainty maps. The shortcoming of using MSE loss in EDL to quantify the uncertainty of medical image segmentation can be seen in Fig. 3, which was significantly biased by the interference.

6 Conclusions and future research directions

In this paper, we proposed a region-based EDL framework to segment brain tumors and quantify their uncertainty reliably and robustly. We demonstrated that the proposed region-based loss could generate reliable prediction confidence by gathering evidence in the output image by demonstrating four theoretical properties. Our method produced voxel-level uncertainty maps for brain tumor segmentation, which provided additional information on segmentation confidence for cancer diagnosis. Extensive experiments showed that the proposed method is more robust than previous methods on the BraTS 2020 dataset and achieves the best performance in segmentation uncertainty estimation. Furthermore, the novel framework maintained the low computational cost properties of EDL and can be easily integrated into any neural network.

Unfortunately, the performance of our method was currently slightly inferior to Ensemble and Dropout methods in terms of ECE and Dice when segmenting raw images. Calibration methods such as temperature scaling can be applied to improve the ECE, while EDL frameworks with higher segmentation accuracy are worthy of further study. Moreover, tuning and optimizing the parameters of EDL to achieve faster inference is a known problem, especially the suitability of the Dirichlet prior, that will be addressed in follow-up studies. In addition, since the predictive uncertainty can be separated into epistemic and aleatoric uncertainty, future work can also focus on the inherent value for automated diagnosis that uncertainty estimation brings when differentiating between the two sources of uncertainty. The fourth direction is validating this framework in other diagnostic applications, possibly favoring the fusion of more multimodal information sources. Then, we can assess whether the fusion of different information modes permits a decrease in the overall uncertainty of the model estimated by our EDL segmentation framework.

Appendix A: Proofs of Theorems

This section provides full proofs for Theorems 1 to 4.

Proof of Theorem 1 Since $\frac{(S_i-\alpha_{ij})}{(S_i+1)}<1$ and $\frac{\alpha_{ij}}{S_i^2}\le\left(\frac{\alpha_{ij}}{S_i}\right)^2$,

$$\left(\frac{\alpha_{ij}}{S_i}\right)^2>\frac{\alpha_{ij}(S_i-\alpha_{ij})}{S_i^2(S_i+1)}. \tag{A1}$$

As $y_{ij}\in 0,\ 1$,

$${y_{ij}}^2+\left(\frac{\alpha_{ij}}{S_i}\right)^2>\frac{\alpha_{ij}(S_i-\alpha_{ij})}{S_i^2(S_i+1)}. \tag{A2}$$

Proof of Theorem 2 The loss function becomes:

$$\mathcal{L}_{DICE}=1-\frac{2}{K}\left[\underbrace{\frac{\frac{\alpha_{pc}+\varepsilon}{S_p+\varepsilon}+\overbrace{\sum_{i\neq p}\frac{\alpha_{ic}}{S_i}}^{L_{n,c}}}{1+\left(\frac{\alpha_{pc}+\varepsilon}{S_p+\varepsilon}\right)^2+\underbrace{\frac{(\alpha_{pc}+\varepsilon)(S_p-\alpha_{pc})}{(S_p+\varepsilon)^2(S_p+1+\varepsilon)}}_{Var_{pc}}+\underbrace{\sum_{i\neq p}1+\left(\frac{\alpha_{ic}}{S_i}\right)^2+\frac{\alpha_{ic}(S_i-\alpha_{ic})}{S_i^2(S_i+1)}}_{L_{d,c}}}}_{C_c}\right.$$
$$\left.+\underbrace{\sum_{j\neq c}^{K}\frac{\sum_{i\neq p}\frac{\alpha_{ij}}{S_i}}{\left(\frac{\alpha_{pj}}{S_p+\varepsilon}\right)^2+\frac{\alpha_{pj}(S_p-\alpha_{pj})}{(S_p+\varepsilon)^2(S_p+1+\varepsilon)}+\sum_i\left(\frac{\alpha_{ij}}{S_i}\right)^2+\frac{\alpha_{ij}(S_i-\alpha_{ij})}{S_i^2(S_i+1)}}}_{D_c}\right], \tag{A3}$$

where ε is the change in α_{pc}. For $\varepsilon>0$, $\frac{\alpha_{pc}+\varepsilon}{S_p+\varepsilon}>\frac{\alpha_{pc}}{S_p}$, so $\hat{p}_{pc}=\frac{\alpha_{pc}}{S_p}$ increases as α_{pc} increases. Taking the partial derivative of C_c, we get:

$$\frac{\partial C_c}{\partial\hat{p}_{pc}}=\frac{1+Var_{pc}+L_{d,c}-2\hat{p}_{pc}L_{n,c}-\hat{p}_{pc}^2}{\left(1+Var_{pc}+L_{d,c}+\hat{p}_{pc}^2\right)^2}. \tag{A4}$$

The numerator can be organized as:

$$numerator=\left(1-\hat{p}_{pc}^2\right)+\left(L_{d,c}-2\hat{p}_{pc}L_{n,c}\right)+Var_{pc}. \tag{A5}$$

Since $0<\hat{p}_{pc}<1$, $\frac{L_{n,c}}{L_{d,c}}\le\frac{1}{2}$, $Var_{pc}\ge 0$, and the denominator is non-negative,

$$\frac{\partial C_c}{\partial\hat{p}_{pc}}>0. \tag{A6}$$

Thus, C_c increases as $\hat{p}_{pc}$ increases, equally as α_{pc} increases. As for D_c, when $\varepsilon>0$, $\left(\frac{\alpha_{pj}}{S_p+\varepsilon}\right)^2<\left(\frac{\alpha_{pj}}{S_p}\right)^2$ and

$$\frac{\alpha_{pj}(S_p-\alpha_{pj})}{(S_p+\varepsilon)^2(S_p+1+\varepsilon)}<\frac{\alpha_{pj}(S_p-\alpha_{pj})}{S_p^2(S_p+1)}.$$

So D_c also increases as α_{pc} increases. As a result, $\mathcal{L}_{DICE}=1-\frac{2}{K}[C_c+D_c]$ decreases as α_{pc} increases. The reasoning is the same for $\varepsilon<0$.

Proof of Theorem 3 The loss function becomes:

$$\mathcal{L}_{DICE} = 1 - \frac{2}{K}\left[\underbrace{\frac{\frac{\alpha_{pc}}{S_p+\varepsilon_0} + L_{n,c}}{1 + \left(\frac{\alpha_{pc}}{S_p+\varepsilon_0}\right)^2 + \frac{\alpha_{pc}\left(S_p-\alpha_{pc}\right)}{\left(S_p+\varepsilon_0\right)^2\left(S_p+1+\varepsilon_0\right)} + L_{d,c}}}_{C_w} + \underbrace{\sum_{j\neq c}^{K} \frac{\overbrace{\sum_{i\neq p} \frac{\alpha_{ij}}{S_i}}^{L_{n,j}}}{\left(\frac{\alpha_{pj}+\varepsilon_j}{S_p+\varepsilon_j}\right)^2 + \frac{\left(\alpha_{pj}+\varepsilon_j\right)\left(S_p-\alpha_{pj}\right)}{\left(S_p+\varepsilon_j\right)^2\left(S_p+1+\varepsilon_j\right)} + \underbrace{\sum_i \left(\frac{\alpha_{ij}}{S_i}\right)^2 + \frac{\alpha_{ij}\left(S_i-\alpha_{ij}\right)}{S_i^2(S_i+1)}}_{L_{d,j}}}}_{W_w}\right], \tag{A7}$$

where ε_j is the change in α_{pj} and $\varepsilon_0 = \sum_{j\neq c}^{K} \varepsilon_j$. For $\varepsilon_0 < 0$ and $j \neq c$, $\frac{\alpha_{pc}}{S_p+\varepsilon_0} > \frac{\alpha_{pc}}{S_p}$, so $\hat{p}_{pc}$ increases as α_{pj} increases. Taking the partial derivative of C_w, we get:

$$\frac{\partial C_w}{\partial \hat{p}_{pc}} = \frac{\left(1-\hat{p}_{pc}^2\right) + \left(L_{d,c} - 2\hat{p}_{pc}L_{n,c}\right) + Var_{pc}}{\left(1 + Var_{pc} + L_{d,c} + \hat{p}_{pc}^2\right)^2}. \tag{A8}$$

Since $0 < \hat{p}_{pc} < 1$, $\frac{L_{n,c}}{L_{d,c}} \leq \frac{1}{2}$, $Var_{pc} \geq 0$, and the denominator is non-negative,

$$\frac{\partial C_w}{\partial \hat{p}_{pc}} > 0. \tag{A9}$$

Thus, C_w increases as $\hat{p}_{pc}$ increases, equally as α_{pj} decreases for all $j \neq c$. Similarly, when $\varepsilon_j < 0$ and $j \neq c$, $\left(\frac{\alpha_{pj}+\varepsilon_j}{S_p+\varepsilon_j}\right)^2 < \left(\frac{\alpha_{pj}}{S_p}\right)^2$, $\hat{p}_{pj} = \frac{\alpha_{pj}}{S_p}$ decreases as α_{pj} decreases. The partial derivative of W_w is:

$$\frac{\partial W_w}{\partial \hat{p}_{pj}} = \frac{-2\hat{p}_{pj}L_{n,j}}{\left(L_{d,j} + \hat{p}_{pj}^2\right)^2}. \tag{A10}$$

Since $0 < \hat{p}_{pj} < 1$, $L_{n,j} > 0$, and the denominator is non-negative,

$$\frac{\partial W_w}{\partial \hat{p}_{pj}} < 0. \tag{A11}$$

Thus, W_w increases as α_{pj} decreases. As a result, $\mathcal{L}_{DICE} = 1 - \frac{2}{K}[C_w + W_w]$ decreases as α_{pj} decreases for all $j \neq c$. The reasoning is the same for $\varepsilon > 0$.

Proof of Theorem 4 The loss function becomes:

$$\begin{aligned}\mathcal{L}_{KL,i} = &\underbrace{\log\left(\frac{\Gamma\left(\varepsilon_w + \sum_{j=1}^{K} \widetilde{\alpha}_{ij}\right)}{\Gamma(K)\Gamma(\widetilde{\alpha}_{iw} + \varepsilon_w)\prod_{j\neq w}^{K}\Gamma(\widetilde{\alpha}_{ij})}\right)}_{A} \\ &+ \underbrace{(\widetilde{\alpha}_{iw} - 1 + \varepsilon_w)\left[\psi(\widetilde{\alpha}_{iw} + \varepsilon_w) - \psi\left(\varepsilon_w + \sum_{j=1}^{K}\widetilde{\alpha}_{ij}\right)\right]}_{B} \\ &+ \underbrace{\sum_{j=1}^{K}(\widetilde{\alpha}_{ij} - 1)\left[\psi(\widetilde{\alpha}_{ij}) - \psi\left(\varepsilon_w + \sum_{j=1}^{K}\widetilde{\alpha}_{ij}\right)\right]}_{C},\end{aligned} \tag{A12}$$

where ε_w is the change in $\widetilde{\alpha}_{iw}/\alpha_{iw}$. As for A, when $\varepsilon_w > 0$,

$$\frac{\Gamma\left(\varepsilon_w + \sum_{j=1}^{K}\widetilde{\alpha}_{ij}\right)}{\Gamma(\widetilde{\alpha}_{iw} + \varepsilon_w)} > \frac{\Gamma\left(\sum_{j=1}^{K}\widetilde{\alpha}_{ij}\right)}{\Gamma(\widetilde{\alpha}_{iw})}, \tag{A13}$$

since $\sum_{j=1}^{K}\widetilde{\alpha}_{ij} > \widetilde{\alpha}_{iw}$. Hence, A increases as $\widetilde{\alpha}_{iw}/\alpha_{iw}$ increases. Similarly for B and C, when $\varepsilon_w > 0$,

$$\psi(\widetilde{\alpha}_{iw} + \varepsilon_w) - \psi\left(\varepsilon_w + \sum_{j=1}^{K}\widetilde{\alpha}_{ij}\right) > \psi(\widetilde{\alpha}_{iw}) - \psi\left(\sum_{j=1}^{K}\widetilde{\alpha}_{ij}\right), \tag{A14}$$

$$\psi(\widetilde{\alpha}_{ij}) - \psi\left(\varepsilon_w + \sum_{j=1}^{K}\widetilde{\alpha}_{ij}\right) > \psi(\widetilde{\alpha}_{ij}) - \psi\left(\sum_{j=1}^{K}\widetilde{\alpha}_{ij}\right). \tag{A15}$$

As a result, $\mathcal{L}_{KL,i}$ increases as $\widetilde{\alpha}_{iw}/\alpha_{iw}$ increases.

Acknowledgements This study was supported in part by the BHF (TG/18/5/34111, PG/16/78/32402), the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC/NSFC/211235), the Imperial College Undergraduate Research Opportunities Programme (UROP), the NVIDIA Academic Hardware Grant Program, the SABER project supported by Boehringer Ingelheim Ltd, NIHR Imperial Biomedical Research Centre (RDA01), and the UKRI Future Leaders Fellowship (MR/V023799/1). J. Del Ser acknowledges funding support from the Basque Government through grant number IT1456-22 (MATHMODE), as well as the Spanish Centro para el Desarrollo Tecnologico Industrial (CDTI, Ministry of Science and Innovation) through the "Red Cervera" Programme (AI4ES project).

Author contributions Hao Li: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft. **Yang Nan**: Conceptualization, Methodology, Writing - review and editing, Supervision. **J. Del Ser**: Conceptualization, Writing - review and editing, Supervision. **Guang Yang**: Conceptualization, Methodology, Writing - review and editing, Supervision, Funding acquisition.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Jungo A, Balsiger F, Reyes M (2020) Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Front Neurosci 14:282. https://doi.org/10.3389/fnins.2020.00282
2. Muhammad K, Khan S, Ser JD, de Albuquerque VHC (2021) Deep learning for multigrade brain tumor classification in smart healthcare systems: a prospective survey. IEEE Trans Neural Netw Learn Syst 32(2):507–522. https://doi.org/10.1109/TNNLS.2020.2995800
3. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR). Boston, MA, USA: IEEE. pp 3431–3440
4. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical image computing and computer-assisted intervention – MICCAI 2015, vol 9351. Springer International Publishing, Cham, pp 234–241
5. Dong H, Yang G, Liu F, Mo Y, Guo Y (2017) Automatic brain tumor detection and segmentation using u-net based fully convolutional networks. In: Valdés Hernández M, González-Castro V (eds) Medical image understanding and analysis, vol 723. Springer International Publishing, Cham, pp 506–517
6. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. (2021) TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306 [cs]
7. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, et al. (2019) Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv:1811.02629 [cs, stat]
8. Müller S, Weickert J, Graf N (2020) Robustness of brain tumor segmentation. J Med Imaging 7(6):064006
9. Das K, Krzywinski M, Altman N (2019) Quantile regression. Nat Methods 16(6):451–452. https://doi.org/10.1038/s41592-019-0406-y
10. Hinton GE, van Camp D (1993) Keeping the neural networks simple by minimizing the description length of the weights. In: proceedings of the sixth annual conference on computational learning theory - COLT '93. Santa Cruz, California, USA: ACM Press. pp 5–13
11. MacKay DJC (1992) A practical bayesian framework for backpropagation networks. Neural Comput 4(3):448–472. https://doi.org/10.1162/neco.1992.4.3.448
12. Hernandez-Lobato JM, Adams R (2015) Probabilistic backpropagation for scalable learning of Bayesian neural networks. In: Bach F, Blei D (eds). proceedings of the 32nd international conference on machine learning. vol. 37 of proceedings of machine learning research. Lille, France: PMLR. pp 1861–1869
13. Nair T, Precup D, Arnold DL, Arbel T (2020) Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med Image Anal 59:101557. https://doi.org/10.1016/j.media.2019.101557
14. Gal Y, Ghahramani Z (2016) Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: International conference on machine learning, PMLR. pp 1050-1059
15. Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inform Proess Syst, 30.
16. Kendall A, Badrinarayanan V, Cipolla R (2017) Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In: Procedings of the British machine vision conference 2017. London, UK: British Machine Vision Association. p 57
17. Sensoy M, Kaplan L, Kandemir M (2018) Evidential Deep Learning to Quantify Classification Uncertainty. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc
18. Tsiligkaridis T (2021) Information aware max-norm dirichlet networks for predictive uncertainty estimation. Neural Netw 135:105–114. https://doi.org/10.1016/j.neunet.2020.12.011
19. Tong Z, Xu P, Denœux T (2021) Evidential fully convolutional network for semantic segmentation. Appl Intell 51(9):6376–6399. https://doi.org/10.1007/s10489-021-02327-0arXiv:2103.13544. [cs]
20. Guo C, Pleiss G, Sun Y, Weinberger KQ. (2017) On calibration of modern neural networks. In: International conference on machine learning

21. Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T (2020) Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans Med Imaging 39(12):3868–3878. https://doi.org/10.1109/TMI.2020.3006437
22. Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, et al. (2022) A survey of uncertainty in deep neural networks. arXiv
23. Kohl SAA, Romera-Paredes B, Meyer C, De Fauw J, Ledsam JR, Maier-Hein KH, et al. (2019) A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems, 31.
24. Dempster AP (2008) A generalization of Bayesian inference. In: Yager RR, Liu L (eds) Classic works of the dempster-shafer theory of belief functions. Springer, Berlin, Heidelberg, pp 73–104
25. Zou K, Yuan X, Shen X, Wang M, Fu H (2022) TBraTS: Trusted brain tumor segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, Cham.
26. Jøsang A (2016) Subjective logic. Artificial intelligence: foundations, theory, and algorithms. Springer International Publishing, Cham
27. Kotz S, Balakrishnan N, Johnson NL (2005) Continuous multivariate distributions, vol 1. Wiley, Hoboken
28. Morales M (2008) Construction of the digamma function by derivative definition. International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 503–513
29. Malinin A, Gales M (2019) Reverse KL-divergence training of prior networks: improved uncertainty and adversarial robustness. Advances in Neural Information Processing Systems
30. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS et al (2017) Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data 4(1):170117. https://doi.org/10.1038/sdata.2017.117
31. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J et al (2015) The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 34(10):1993–2024. https://doi.org/10.1109/TMI.2014.2377694
32. Li H, Nan Y, Yang G (2022) LKAU-Net: 3D large-kernel attention-based u-net for automatic MRI brain tumor segmentation. In: Yang G, Aviles-Rivero A, Roberts M, Schönlieb CB (eds) Medical image understanding and analysis, vol 13413. Springer International Publishing, Cham, pp 313–327
33. Isensee F, Jäger PF, Full PM, Vollmuth P, Maier-Hein KH (2021) nnU-net for brain tumor segmentation. In: Crimi A, Bakas S (eds) Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries, vol 12659. Springer International Publishing, Cham, pp 118–132
34. Li H, Nan Y, Del Ser J, Yang G (2022) Large-kernel attention for 3D medical image segmentation. arXiv
35. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. (2019) Can you trust your models uncertainty? evaluating predictive uncertainty under dataset shift. In: Wallach H, Larochelle H, Beygelzimer A, dAlché-Buc F, Fox E, Garnett R, (eds). Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc
36. Mehta R, Filos A, Baid U, Sako C, McKinley R, Rebsamen M, et al. (2021) QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation – Analysis of Ranking Metrics and Benchmarking Results. arXiv