Aleatoric Uncertainty-Aware Learning in 3D Multi-Class Brain Hemorrhage Segmentation

Wei Han, Chen
Student Number 230249354
MSc. Artificial Intelligence, QMUL
w.chen@se23.qmul.ac.uk
Project Supervisor: Kit Mills Bransby
k.m.bransby@qmul.ac.uk

***Abstract*—In recent years, convolutional neural networks (CNNs) have dominated medical image segmentation tasks. However, their lack of reliability quantification significantly limits their clinical applicability. This is particularly challenging in intracerebral hemorrhage detection due to the high variability in hemorrhage shapes and the similarities between different classes.**

In this study, we propose a method to quantify aleatoric uncertainty. Aleatoric uncertainty effectively captures the noise in input images and labels which is a crucial indicator of potential segmentation errors. Identifying and highlighting regions with high uncertainty can refer these areas to experts for further analysis.

Additionally, we incorporate the predicted aleatoric uncertainty into an adaptive weighting loss framework, employing a supervised learning approach that integrates uncertainty information during the training phase to guide the model's learning process. This framework aims to enhance the robustness and accuracy of segmentation models in handling complex and variable data, resulting in a 6.5% increase in the average Dice coefficient and an 11% reduction in the average Hausdorff distance, demonstrating significant improvements in segmentation performance.

***Keywords*—aleatoric uncertainty, intracranial hemorrhage segmentation, nnUNet.**

I. INTRODUCTION

Intracranial hemorrhage (ICH), commonly referred to as brain hemorrhage, is a highly fatal disease caused by the rupture of blood vessels in the brain (Rahmani et al., 2018). This can be classified into various subtypes according to the amount of bleeding, anatomical location, and shape, including extradural hemorrhage (EDH), subdural hemorrhage (SDH), subarachnoid hemorrhage (SAH), intraparenchymal hemorrhage (IPH), and intraventricular hemorrhage (IVH). There are many causes of intracranial hemorrhage. In addition to common external trauma, tumors, and chronic diseases (Ariesen et al. 2003) are all potential factors. In clinical management, professional doctors use CT images to detect and locate ICH. Treatment methods vary greatly depending on the type and amount of bleeding. Therefore, accurate positioning and classification are crucial.

Although deep learning has significantly advanced the automation of medical image segmentation, recent research on ICH has primarily focused on single-category segmentation or bleeding-type classification (Patel et al. 2019; Xu et al. 2023). Although achieving an average Dice similarity coefficient (DSC) between 0.82 and 0.9 for most models (Zarei et al. 2024), the identification of small lesions and the segmentation of specific subtypes is still not accurate enough. The average DSC was 0.44 when the diameter of the lesion was <1 mm, increasing to 0.68 when the diameter was >5 mm. In this regard, the segmentation of small lesions, the imbalance of categories, and the high variability in bleeding volume and location present significant challenges for multi-category ICH segmentation, see **Fig. 1**.

Furthermore, the application of deep learning in medical image segmentation remains limited due to concerns about model reliability. The unpredictability of models leads to potential errors in them. This can have severe consequences in contexts where accurate answers are solicited, for example, in medical applications. It is this limitation that has driven the exploration of uncertainty quantification (Kendall & Gal,

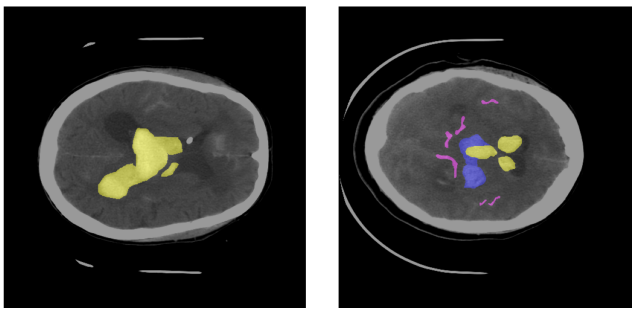

Fig. 1. Comparison of Single-Class and Multi-Class ICH Segmentation. ***Left***: A single-class segmentation of ICH, labeling the hemorrhagic region in yellow. ***Right***: A multi-class segmentation of ICH highlights the increased complexity and variability in features, the difficulty in identifying specific subtypes and small lesions, and the challenges in drawing boundaries where when hemorrhage types coexist. Adapted from 3D BHSD dataset.

2017; Hüllermeier & Waegeman, 2021; Mena, Pujol, & Vitrià, 2021; Valdenegro-Toro & Saromo, 2022). Uncertainty in machine learning can be divided into epistemic and aleatoric uncertainty. In this paper, we focus on aleatoric uncertainty, which reflects the intrinsic variation in the data itself and hence is irreducible by further collection of data or obtaining more training data.

Addressing aleatoric uncertainty is crucial for enhancing the robustness of deep learning models in medical applications. By modeling it, more reliable predictions can be provided, and areas of high uncertainty can be identified for further assessment by clinical practitioners. Recent research has introduced various methods to model aleatoric uncertainty, such as the test-time augmentation (TTA) method introduced by Wang et al. (2018). This method aims to capture the variability in the data, offering confidence measures that are critical for clinical decision-making.

In this study, we employ TTA to quantify aleatoric uncertainty in medical image segmentation tasks. Our approach combines data augmentation techniques with probabilistic statistical theories during inference to achieve robust uncertainty estimation. Through the evaluation of a multi-class ICH dataset, we demonstrated the ability of this method to provide reliable uncertainty quantification, thereby enhancing the trustworthiness of segmentation models in clinical practice. Additionally, we modeled aleatoric uncertainty during training to weight the loss function dynamically, guiding the model's training direction. Incorporating aleatoric uncertainty into the training process enhances the model's robustness by allowing it to adapt its learning based on the inherent uncertainty in the data. This approach not only optimizes the model's performance in challenging areas but also ensures that the model is better equipped to generalize across a wide range of inputs.

The main contributions of this study are as follows:

- Quantification of Aleatoric Uncertainty:
  We proposed a robust method to quantify aleatoric uncertainty, effectively capturing inherent noise and variability in medical images, thus enhancing the model's reliability in clinical applications.

- Adaptive Weighting Loss Function:
  We developed a loss function that dynamically weights based on uncertainty, improving the model's ability to focus on challenging regions during training and increasing overall segmentation accuracy.

- Enhanced Model Robustness:
  We demonstrated significant improvements in model robustness, with a 6.5% increase in average DSC, by integrating uncertainty into the training process.

The remainder of this paper is structured as follows: The section II reviews related work in the dataset, aleatoric uncertainty quantification, and how uncertainty enhances the model performance. Section III details our methodology, Section IV presents the experimental setup and results, and finally concludes the paper in Section V, and discusses our potential directions for future works in Section VI.

II. RELATED WORKS

A. *ICH Segmentation*

ICH segmentation involves automatically detecting and delineating bleeding areas within the brain through CT images. This is crucial for clinical applications in medical diagnostics, as it is a life-threatening disease that requires immediate attention. Automated segmentation allows for faster diagnosis compared to manual segmentation by radiologists. Elsheikh et al. (2024) explore the use of CNN for ICH segmentation and volume measurement on small data sets. This study shows that CNN can achieve accurate ICH segmentation even on small data sets, which is of great significance for the treatment of acute ICH. Kothala and Guntur (2022) use a hybrid model of EfficientNet-B7 and UNet. Through EfficientNet-B7's compound scaling concepts and fewer model parameters to achieve higher accuracy. Paper (Wang, Tang & Hu 2023) designed a cluster capsule network called GroupCapsNet for ICH segmentation. This model is superior to traditional U-Net in terms of the Dice coefficient and effectively reduces the computational cost and memory consumption of capsule networks.

B. *Uncertainty in medical image segmentation*

Uncertainty has played a crucial role in medical image segmentation tasks and has been widely studied. For example, Whitbread and Jenkinson (2022) explored various types of uncertainty in medical image segmentation and their source diversity. Ng et al. (2023) conducted a systematic evaluation of multiple methods for uncertainty estimation in cardiac MRI segmentation. Their study compared the performance of different uncertainty quantification methods, including Bayes by Backprop, Monte Carlo (MC) Dropout, Deep Ensembles, and Stochastic Segmentation Networks in terms of segmentation accuracy and the ability to measure uncertainty in distorted images. Another study (Judge et al. 2022) introduced a method named CRISP (ContRastive Image Segmentation for Uncertainty Prediction), which utilizes contrastive learning to robustly quantify uncertainty in datasets.

In addition to methods for quantifying uncertainty, Zhovannik et al. (2022) studied the importance of considering uncertainty in improving model robustness and accuracy. Furthermore, other studies (Hershkovitch & Riklin-Raviv 2018; Krygier et al. 2022) discussed the potential value of uncertainty in medical applications such as image analysis and clinical decision-making.

C. *Uncertainty-Aware Learning*

Uncertainty-aware learning and difficulty-aware learning share the same core concept: we expect regions with high uncertainty to be associated with higher learning difficulty. This association has been shown to improve model performance effectively. (Liu, Zhang & Barnes 2022) proposed a confidence estimation network and dynamic supervision framework, which significantly improved the accuracy and robustness of camouflaged target detection (Zhao et al. 2024). Reduce the impact of pseudo-label inaccuracies on model training by designing uncertainty-

aware cross-teaching strategies. In (He & Li 2024), uncertainty was integrated across regional evidence prediction results, achieving excellent results in supervised learning for various medical image segmentation tasks.

D. Test-time augmentation

Data augmentation is a method that applies different transformations including flipping, cropping, and rotating training data to enhance the model's robustness. Goceri (2023) conducted a comprehensive review of various data augmentation techniques applied to different medical image modalities and confirmed the effectiveness of different enhancement methods. Basaran et al proposed a new method called LesionMix (Basaran et al. 2024) for lesion-level data enhancement in medical image segmentation. Unlike traditional image enhancement methods that are applied to the entire data, this method achieves better performance than other hybrid enhancement methods by enhancing specific lesion areas.

Several studies have found that combining data augmentation on test images can help improve performance. (Test-Time Augmentation) TTA can reduce the impact of noise by enhancing input diversity to produce more reliable final predictions. For example, Shanmugam et al. (2020) explored the factors influencing TTA performance, including dataset characteristics, model architecture, and augmentation strategies. They identified a drawback of TTA: even though TTA produces a net improvement in accuracy, it can change many correct predictions into incorrect ones. Another study (Pokhrel et al. 2024) investigates using test-time augmentation techniques to enhance the performance of out-of-distribution detection (anomaly detection) in gastrointestinal images.

Recent studies have increasingly applied TTA for uncertainty quantification and model output calibration. Conde & Barros (2023) introduced two TTA-based variants aimed at improving uncertainty calibration in deep models for image classification, emphasizing their higher stability and overall performance, particularly under distribution shift scenarios. Similarly, Hekler et al. (2023) utilized TTA in the context of skin cancer classification, demonstrating that it offers more reliable uncertainty estimates in real-world applications.

III. METHOD

Our work can be divided into two stages. In the first stage, we use TTA to estimate the aleatoric uncertainty of the data set. This stage is formulated in an image augmentation model and an uncertainty estimation method. The input image is represented using this image augmentation model in Section III.B. In Section III.C, we compile the varied prediction outcomes from the test-time augmentation, which are then used to estimate the aleatoric uncertainty associated with image transformation. In the second stage, we apply aleatoric uncertainty into the training stage to guide the learning of the model. This is detailed in Section III.D.

A. Setting

The intracerebral hemorrhage multi-class segmentation task considers an input image $I \in \mathbb{R}^{D\times H\times W\times C}$ (CT scans), with the goal of classifying each pixel in the image to predict which of the six different classes it belongs to (5 hemorrhage types and background). For each pixel, model produces a tuple of six softmax probabilities $O_i(x,y,z) = (p_1, p_2, \ldots, p_6)$, where p_i represents the probability that the pixel belongs to class j. The final predicted class $c_i(x,y,z)$ is determined by taking the argmax of these probabilities, selecting the class with the highest probability. This setup frames the segmentation task as a multi-class classification problem, where the model learns to predict a multi-class softmax probability distribution for each pixel during training.

B. Image augmentation model

Aleatoric uncertainty is caused by the inherent variability of the data itself, such as observation noise or other random factors. This image augmentation model aims to simulate these inherent variabilities by performing random spatial transformations on the image.

$$\mathrm{X} = \mathrm{T_n} \cdot \mathrm{T_{n-1}} \cdot \ldots \cdot \mathrm{T_1}(\mathrm{X_0}) \tag{1}$$

where X_0 represents the original image, and n is the number of transformations performed on X_0 . Each T is a transformation operator applied to X_0 which is randomly selected from rotation, scaling, and flipping, as follows:

$$T_i = \begin{cases} R(\theta_i, v_i), & rotating \\ F(v_i), & flipping \\ S(s_i, v_i), & scaling \end{cases} \tag{2}$$

where $\theta_i \in [-10°, 10°]$ is the random rotation angle, $s_i \in [0.9,1.1]$ is the random scaling factor, and v_i is the unit vector defining the axis of transformation.

The generated images X cover the distribution of X_0, while also simulating the prior distribution of some random noise in the real world. Let Y and Y_0 be the discrete label prediction of X and X_0 , respectively, and they are equivariant with the spatial transformation. Therefore, we can obtain the original prediction Y_0 through the inverse transformation corresponding to (Eq.1) :

$$Y_0 = T_1^{-1} \cdot T_2^{-1} \cdot \ldots \cdot T_n^{-1}(Y) \tag{3}$$

C. Aleatoric uncertainty quantification

Aleatoric uncertainty can estimated by measuring the spread of predictions for a given image. However, simulating real-world data distribution accurately is computationally expensive because the parameters of the transformation operators (Eq.2) are continuous. Therefore, we implement the Monte Carlo simulation. This approach allows us to approximate the distribution of possible outcomes by performing multiple inferences with different random transformations applied to the input image. By aggregating the results of these multiple runs, we can effectively estimate the aleatoric uncertainty with its variance.

For segmentation task, we need to estimate pixel-level uncertainty maps. Let M represent the total number of simulations runs, Y_i denote the predicted label for the i-th pixel. The aleatoric uncertainty on i-th pixel U_i can be denoted as:

$$U_i = \frac{1}{M}\sum_{m=1}^{M}(Y_i^m - \bar{Y}_i)^2 \quad (4)$$

where Y_i^m is the softmax prediction results for i-th pixel in m simulations, and $\bar{Y}_i$ is the average value of the i-th pixel in M simulations.

At test time, we apply random spatial transformations to the input data (Eq.1) and then reverse these transformations on the inference results (Eq.3). By utilizing Monte Carlo Simulation with these processes M times, we can effectively quantify the aleatoric uncertainty for each pixel with its variance (Eq.4). The whole algorithm is shown in Alg. 1.

Algorithm 1 Aleatoric Uncertainty Quantification

Input:

(1) spatial transformation counts N

(2) Monte Carlo Simulation counts M

(3) training data X_0

(4) model function $f(\cdot)$

Output:

uncertainty map of X_0

1: for i ← 1 to M do
2: for i ← 1 to N do
3: random transformation T on X_0, get $X = T(X_0)$
4: get model prediction $Y = f(X)$
5: for i ← 1 to N do
6: inverse transformation T^{-1} on Y, get $Y^0 = T^{-1}(Y)$
7: Calculate uncertainty map from Eq.4

D. *Uncertainty-aware learning*

Multi-class ICH segmentation task has very different learning difficulties in different areas. Due to its pathological characteristics, the model needs to focus more on the areas of small lesions and the junction of physiological structures.

Therefore, in this paper, we adopt an uncertainty-aware approach to enhance the robustness and accuracy of the segmentation model. Specifically, we estimate uncertainty during the training phase and incorporate these estimates into the loss function to guide the learning process. We integrate the uncertainty in both Cross-Entropy Loss (L_{ce}) and Dice Coefficient Loss (L_{dc}), which are commonly used in segmentation tasks. The modified loss functions are defined as follows:

$$\text{Loss} = \mathrm{L_{ce}}(1 + \lambda \mathrm{U}) + L_{dc}(1 + \lambda \mathrm{U}) \quad (5)$$

where λ is a hyper-parameter controlling the degree of attention given to uncertain pixels. A higher λ results in greater emphasis on uncertain pixels, while a λ of 0 treats all pixels equally. This method aims to leverage the correlation between uncertainty and error segmentation by assigning higher learning weights to regions with high uncertainty, guiding the model to focus on areas that are likely to be misclassified.

IV. Experiments And Results

We validated our work using a 3D ICH dataset (Wu et al. 2023), with detailed implementation provided in Section IV.A. In Section IV.B, we present the aleatoric uncertainty obtained using Equation 4. Section IV.C compares the segmentation accuracy achieved through our uncertainty-aware learning approach with the baseline model.

A. *Dataset and training details:*

Dataset:

All of our experiments have been carried on a Brain Hemorrhage Segmentation Dataset (BHSD) (Wu et al. 2023) since this dataset covers all types and is well annotated for ICH segmentation tasks. This dataset consists of 192 high-resolution 3D CT scans of the brain that were annotated on pixel level by medical imaging professionals.

This dataset presents some challenges including significant class imbalance and the difficulty in accurately identifying due to their complex and overlapping characteristics. These challenges lead to more unstable model predictions and give greater significance to quantifying aleatoric uncertainty, making it an ideal choice for our research.

Training details:

We utilized the nnU-Net (Isensee et al. 2021) model and modified both the training and inference phases to apply our uncertainty-aware learning and quantify aleatoric uncertainty. Model nnU-Net is a state-of-the-art framework for biomedical image segmentation that automatically adapts its architecture and training pipeline to the specific dataset. Its core structure consists of symmetric encoder-decoder and skip connection technique, enabling it to capture global context while maintaining precise localization, making it particularly well-suited for high-precision medical image segmentation tasks.

For training, we employed the "nnUNetResEncUNet-MPlans" configuration, which features a residual encoder U-Net model. We also adjust the number of training epochs to 300. This adjustment was made to highlight how uncertainty guides the learning of the model during the early stage of training. All training and inference were performed on a single Nvidia A100 80G GPU.

B. *Aleatoric Uncertainty Quantification:*

To obtain robust uncertainty quantification results, we set *n=5* to apply random data augmentation in the inference stage (Eq.1) and implement Monte Carlo simulation where *M=20* (Eq.4).

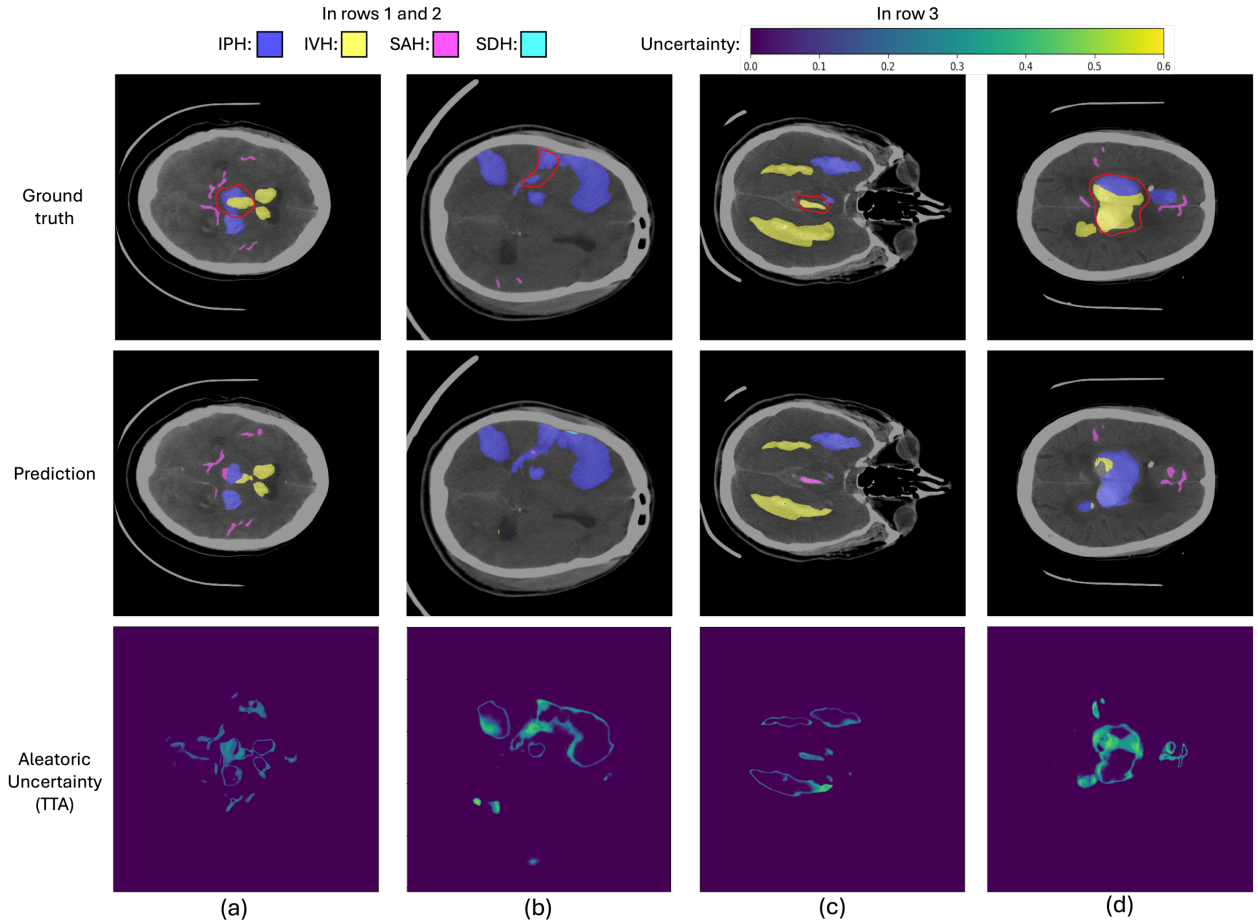

Fig. 2. Visual comparison of uncertainties and their corresponding segmentations. The uncertainty maps in the third row are based on Monte Carlo simulation with M = 20 and encoded by the color bar in the top right corner (low uncertainty shown in purple and high uncertainty shown in yellow).

Fig. 2 shows the visual comparison of uncertainties and their corresponding segmentations of multi-type intracranial hemorrhage. In each column, the first and second rows present the ground truth and segmentation results with the corresponding image, respectively. In column (a), the red circle in the ground truth image marks a multi-class intersection area. Such overlapping regions of different lesions are often difficult to segment correctly. Although our model fails to segment this region correctly in prediction. The uncertainty map captures this information effectively, it indicates higher uncertainty in the middle of mis-segmented areas. Similar to (a), the red circle in column (b) marks an area where the model is over-segmented, and this phenomenon is also reflected in the uncertainty map we obtained. From the uncertainty maps of columns (a) and (b), we can find that the uncertainty is mostly concentrated at the edges of the segmented areas and places where segmentation may be wrong, which indicates the model's learning difficulty in these areas.

In column (c), we can find that our model incorrectly predicts the IVH lesion area within the red circle as an SAH lesion. Normally, the model only outputs the prediction result after argmax, which is incorrect in this case. By incorporating the output of the uncertainty map, we can focus on areas where the model has lower confidence, providing more information for clinical judgment.

However, column (d) also presents an error case where the model incorrectly segments the IVH region as an ICH region. The uncertainty is higher around the edges and in the middle of this region, while the lower part shows quite high confidence. This may be due to the high similarity of features, causing the model to have overly high confidence in the incorrect prediction in some cases.

As shown in **Fig. 3**, we analyzed the relationship between aleatoric uncertainty and error rate. Due to the imbalance between background and foreground in the 3D CT dataset, most areas are background regions with very low uncertainty. Therefore, we observe that when uncertainty approaches 0, the error rate also approaches 0. In other regions, we can see a

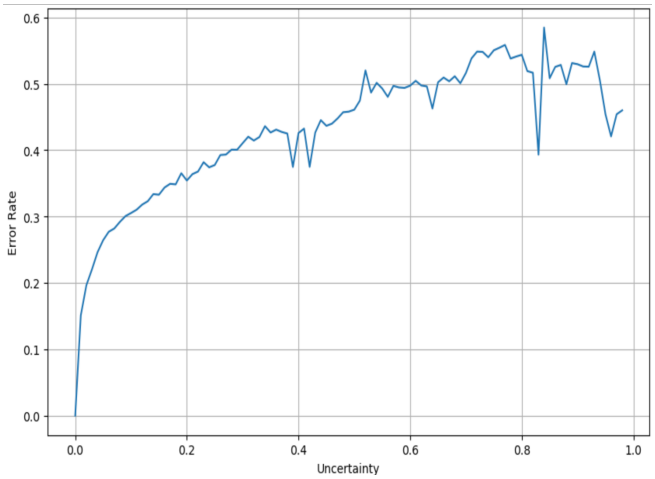

Fig. 3. The relationship between aleatoric uncertainty and error rate.

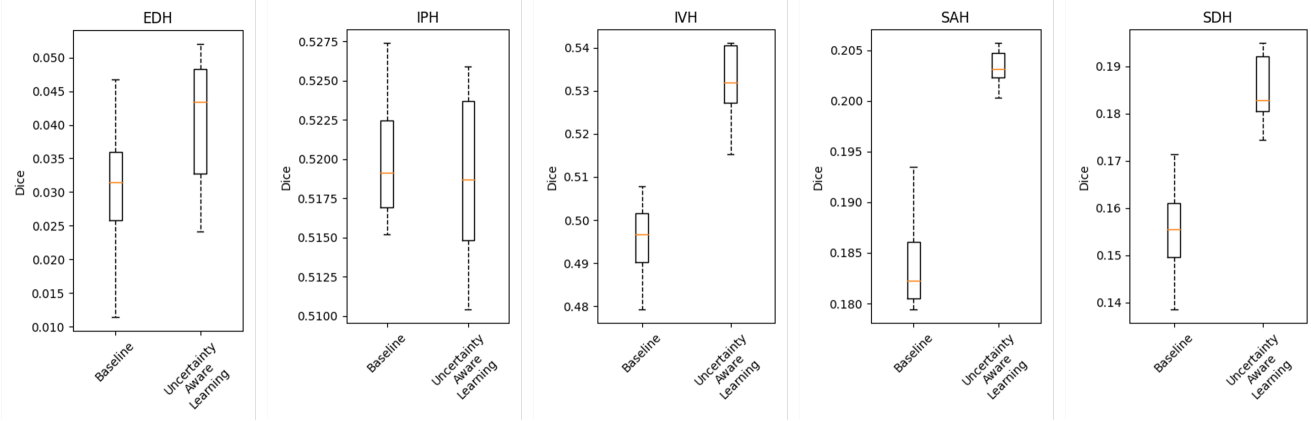

Fig. 4. Comparison of Dice coefficient distributions between Baseline and Uncertainty-Aware Learning Using nnUNet on BHSD Dataset.

Table 1. Comparison of Hausdorff Distance (HD) and Average Symmetric Surface Distance (ASSD).

		EDH	IPH	IVH	SAH	SDH	Mean
HD	*Baseline*	123.37 ± 6.33	49.78 ± 2.53	51.35 ± 1.78	87.03 ± 1.76	106.12 ± 2.59	83.53
	Ours	**86.01 ± 6.42**	**49.67 ± 3.01**	**49.48 ± 2.99**	**85.16 ± 3.33**	**101.17 ± 2.65**	**74.30**
ASSD	*Baseline*	46.35 ± 2.68	**8.32 ± 0.31**	7.13 ± 1.02	23.62 ± 3.91	25.17 ± 2.02	22.11
	Ours	**25.31 ± 2.81**	8.35 ± 0.35	**6.78 ± 1.05**	**20.26 ± 2.12**	**19.15 ± 2.39**	**15.97**

strong positive correlation between uncertainty and error rate. Apart from a few dips caused by insufficient sample size, the error rate consistently increases with rising uncertainty. It can be seen that Aleatoric uncertainty not only indicates the uncertainty and noise of the data itself but also has a strong correlation with the error rate of prediction.

C. Segmentation Result with uncertainty:

Quantitative evaluation:

To compare the performance of our approach, we trained nnUNet on the BHSD dataset (Wu et al. 2023) for 350 epochs. We applied the default nnUNet model and our modified uncertainty-aware learning model during the training process.

Fig. 4 presents the Dice coefficients distribution of the baseline model and our method across five categories. The results show that the Uncertainty-Aware learning method's Dice coefficients for the five different categories (EDH, IPH, IVH, SAH, SDH) are 0.0401, 0.5187, 0.5312, 0.2032, and 0.1849, respectively, compared to the baseline values of 0.0303, 0.5202, 0.4951, 0.1844, and 0.1552. Our method exhibits noticeable improvements in four categories (EDH, IVH, SAH, SDH), especially in categories with relatively low baseline values, such as EDH and SDH, with increases of 32.34% and 19.13%, respectively. However, in the IPH category, our model's performance decreased by 0.29% and showed a broader distribution, indicating lower stability and accuracy for this category.

For a more detailed evaluation, we also used other evaluation metrics including Hausdorff Distance (HD) and Average Symmetric Surface Distance (ASSD). HD is used to measure the maximum deviation between the segmentation results and the ground-truth annotations, reflecting the model's performance at the most extreme errors, while ASSD calculates the average deviation between segmentation results and ground-truth annotations, giving a more comprehensive performance evaluation. In **Tab. 1**, our model achieved results similar to the Dice coefficient in HD and ASSD. The model improved in other categories except IPH. The overall average dropped by 9.23 and 6.14 in HD and ASSD metrics respectively. With these two additional metrics, we can more accurately evaluate the extreme errors of the model and the overall performance.

Fig. 5 shows the segmentation results of the baseline model and the uncertainty-aware learning model across four different cases, along with their corresponding ground truths and uncertainty maps. In column (a), a red circle marks an IVH area, which also shows high uncertainty in the corresponding uncertainty map. Our method achieved more accurate segmentation results than the baseline model by weighting these areas in the loss function. Similarly, column (b) highlights the intersection of an IPH and IVH lesion. The baseline model failed to segment this boundary accurately, but our model overcame this challenge.

In column (c), the red circle indicates an IPH area that the baseline model failed to identify correctly, whereas the uncertainty-aware learning model achieved better results. Finally, in row (d), the marked IPH area was identified as a multi-class region (IPH, IVH, SAH) by the baseline model. The corresponding uncertainty map also showed high uncertainty in these areas. Our model not only accurately segmented the IPH region but also reduced the misclassified SAH area.

With these cases, we can see how uncertainty-aware learning guides model learning through uncertainty weighting. However, the final learning results still depend greatly on the data itself and the complexity of the model.

Training Time Comparison:

Our uncertainty-aware learning approach involves performing multiple forward passes during training to

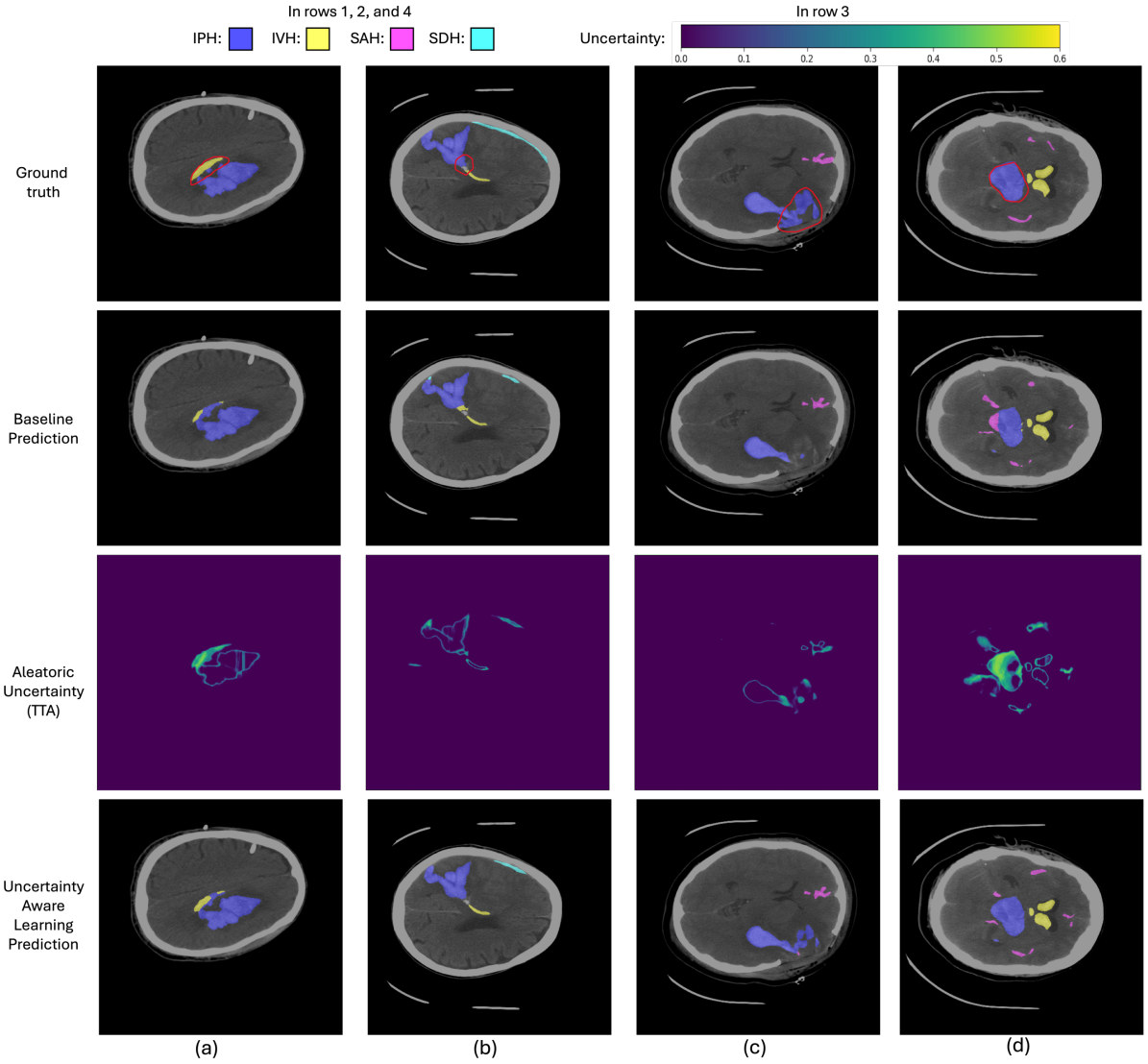

Fig. 5. Visual comparison of uncertainties, the baseline prediction, and the prediction after applying uncertainty-aware learning. The red circle in the first row marks our area of interest and shows how uncertainty-aware learning improves the model's output.

compute the current uncertainty map, which naturally increases the computational load during the loss backpropagation process. While the default nnUNet takes 30 seconds per epoch for training, our method extends this to 90 seconds per epoch.

However, despite the longer training time, inference time and memory usage remain similar. This represents a favorable trade-off, particularly in a clinical setting where accurate and reliable segmentation is critical, and training is typically a one-time process.

Hyper-parameter analysis:

The hyper-parameter λ controls the uncertainty guidance in the modified loss function (Eq. 5). **Tab. 2** shows the impact of selecting different λ values. Since our dataset has a significantly larger background compared to the foreground, and most of the background has near-zero uncertainty, we require a relatively large λ to effectively guide the model's learning process. The results indicate that setting λ = 50 achieves better performance. Therefore, in this study, we defined a fixed λ = 50 for the entire training phase.

Table 2. Comparison of different λ in Dice, HD, and ASSD.

	Dice ↑	HD ↓	ASSD ↓
λ=5	0.2895	78.0848	16.4188
λ=50	**0.2956**	**74.3002**	**15.9725**
λ=100	0.2873	80.9452	17.6182

V. Discussion and Conclusion

The uncertainty-aware learning pipeline for 3D multi-class segmentation of brain hemorrhages developed in this work shows the capability of handling challenges associated with huge variability of the shape of hemorrhages and similarity between different classes. We further showed that quantifying aleatoric uncertainty through test-time data augmentation and integrating it into the training process via a modified loss function leads to significant improvements in segmentation robustness and accuracy.

Compared with large imaging datasets such as ImageNet, which have millions of images, the ICH dataset we use only has 192 CT scans each containing 24 to 40 slices. This situation is particularly common in medical imaging datasets. First, there are difficulties in collecting clinical data, and

secondly, labeling pixel-level annotations is not only time-consuming but also requires professional doctors to perform. Additionally, relatively small data sets are more suitable for data augmentation, which is more consistent with our approach of using multiple data augmentation in the training and inference process. Image segmentation under limited data has always been an important issue, and Aleatoric uncertainty plays an important role in this challenge. Our research shows that understanding the inherent variability and noise in the data itself not only reveals uncertainties in model predictions but also helps the model perform more robustly in the face of new data.

The method we used for quantifying uncertainty only implements spatial transformations to model aleatoric uncertainty. However, TTA is a simple and intuitive method that can be easily extended to include artificially added noise or spatial transformations in specific regions to model more robust uncertainty output. In our study, we fix $M = 20$ as the parameter of Monte Carlo Simulation, which allows us to obtain relatively robust uncertainty results on the dataset used. However, the value of M depends greatly on the differences in the data set itself. An excessively large M value will cause unnecessary sacrifices in computational efficiency.

Section III.B shows the positive relationship between confidence level and error rate, making it an effective indicator of potential prediction error. However, in Fig.2. (d) we also provide an error case to illustrate the model It is still possible to achieve high confidence in areas that were incorrectly predicted. Therefore, the significance of uncertainty as an auxiliary judgment and providing additional information in clinical applications is still greater than its actual ability to predict the correct type of expression.

We have also demonstrated how uncertainty guides the learning process in the model. While there was an improvement in all three different evaluation metrics with our uncertainty-weighted loss function, the associated increase in computational cost and extra time were important factors, particularly with large datasets where these effects could be more pronounced.

In summary, first, we applied a test-time augmentation technique to quantify uncertainty on a multi-category intracerebral hemorrhage dataset. And analyze its interaction with data variability and error rate. Second, through real-time uncertainty assessment, we used a weighted loss function to guide the model's learning in high-uncertainty areas during the model training phase and analyzed the model's performance.

VI. Future Work

While our research has demonstrated advancements in 3D multi-class brain hemorrhage segmentation through the application of uncertainty-aware learning, several avenues remain open for future exploration.

One is dynamic hyper-parameter adjustment. While we used a fixed λ in the experiment, developing how to dynamically adjust such hyper-parameters during training could lead to even further performance gains in the model's adaptivity. This would help the model appropriately handle the different levels of uncertainty that arise during different stages of the training process.

Another research direction is to incorporate additional information while training the uncertainty maps. For example, in our samples, some areas of wrong predictions still hold high confidence, which misguides our uncertainty-aware learning. Combining error prediction targets with the uncertainty learning method may make guidance more effective in training the model.

Moreover, although our method improves accuracy, it also increases the computational cost. Future work can focus on optimizing the algorithm to reduce training and inference time without affecting performance. For example, under the original method, the number of spatial transformations and Monte Carlo simulations is reduced to reduce the calculation time of backpropagation in the loss function. However, this requires more experiments to strike a balance between uncertainty quantification accuracy and computational efficiency. This optimization is particularly important for practical applications in clinical settings where computational resources and time are limited.

We also aim to extend our current model to other medical conditions. While our research focuses on intracranial hemorrhage, future studies could apply this approach to other medical conditions such as brain tumors, ischemic strokes, or pathologies specific to other organs. This would help evaluate the generalizability of our method. Furthermore, the dataset used in this study has relatively low uncertainty at baseline; however, our method may be limited in datasets that already have high accuracy. We hope to test and analyze our approach on other datasets to further assess its performance.

By exploring these directions, we can continue to improve the effectiveness and applicability of uncertainty-aware learning in medical image segmentation, ultimately improving clinical outcomes.

REFERENCES LIST

Ariesen, M.J., Claus, S.P., Rinkel, G.J.E., and Algra, A., 2003, "Risk Factors for Intracerebral Hemorrhage in the General Population: A Systematic Review." Available at: https://doi.org/10.1161/01.STR.0000080678.09344.8D.

Basaran, B.D., Zhang, W., Qiao, M., Kainz, B., Matthews, P.M., Bai, W., 2024, "LesionMix: A Lesion-Level Data Augmentation Method for Medical Image Segmentation." In: Xue, Y., Chen, C., Chen, C., Zuo, L., Liu, Y., 2024, "Data Augmentation, Labelling, and Imperfections." Available at: https://doi.org/10.1007/978-3-031-58171-7_8

Conde, B., and Barros, T., 2023, "Approaching Test Time Augmentation in the Context of Uncertainty Calibration for Deep Neural Networks." Available at: https://doi.org/10.48550/arXiv.2304.05104

Elsheikh, S., Elbaz, A., Rau, A. et al., 2024, "Accuracy of Automated Segmentation and Volumetry of Acute Intracerebral Hemorrhage Following Minimally Invasive Surgery Using a Patch-Based Convolutional Neural Network in a Small Dataset." Available at: https://doi.org/10.1007/s00234-024-03311-4.

Goceri, E., 2023, "Medical Image Data Augmentation: Techniques, Comparisons and Interpretations." Available at: https://doi.org/10.1007/s10462-023-10453-z

Hüllermeier, E., Waegeman, W., 2023, "Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods." Available at: https://doi.org/10.1007/s10994-021-05946-3

Hekler, A., Brinker, T.J. and Buettner, F., 2023, "Test Time Augmentation Meets Post-hoc Calibration: Uncertainty Quantification under Real-World Conditions." Available at: https://doi.org/10.1609/aaai.v37i12.26735

Isensee, F., Jaeger, P.F., Kohl, S.A.A., et al., 2021, "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation." Available at: https://doi.org/10.1038/s41592-020-01008-z

Judge, T. et al., 2022, "CRISP - Reliable Uncertainty Estimation for Medical Image Segmentation." Available at: https://doi.org/10.48550/arXiv.2206.07664

Kendall, A., Gal, Y., 2017, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" Available at: https://doi.org/10.48550/arXiv.1703.04977

Krygier, M.C., LaBonte, T., Martinez, C., et al., 2022, "Quantifying the Unknown Impact of Segmentation Uncertainty on Image-Based Simulations." Available at: https://doi.org/10.1038/s41467-021-25493-8

Liu, J., Zhang, J., and Barnes, N., 2022, "Modeling Aleatoric Uncertainty for Camouflaged Object Detection." Available at: https://doi.org/10.1109/WACV51458.2022.00267.

Mena, J., Pujol, O., and Vitrià, J., 2021, "A Survey on Uncertainty Estimation in Deep Learning Classification Systems from a Bayesian Perspective." Available at: https://doi.org/10.1145/3477140

Ng, M. et al., 2023, "Estimating Uncertainty in Neural Networks for Cardiac MRI Segmentation: A Benchmark Study." Available at: https://doi.org/10.48550/arXiv.2012.15772

Patel, A., et al., 2019, "Intracerebral Haemorrhage Segmentation in Non-contrast CT." Available at: https://doi.org/10.1038/s41598-019-54491-6

Rahmani, F., Rikhtegar, R., Ala, A., Farkhad-Rasooli, A. and Ebrahimi-Bakhtavar, H. (2018) 'Predicting 30-day mortality in patients with primary intracerebral hemorrhage: Evaluation of the value of intracerebral hemorrhage and modified new intracerebral hemorrhage scores', Iranian Journal of Neurology, 17(1), pp. 47-52. PMID: 30186559; PMCID: PMC6121209.

Sandesh Pokhrel et al., 2024, "TTA-OOD: Test-time Augmentation for Improving Out-of-Distribution Detection in Gastrointestinal Vision." Available at: https://doi.org/10.48550/arXiv.2407.14024

Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J., 2020, "When and Why Test-Time Augmentation Works" Available at: https://www.researchgate.net/publication/346143761_When_and_Why_Test-Time_Augmentation_Works

Valdenegro-Toro, M., Saromo, D., 2022, "A Deeper Look into Aleatoric and Epistemic Uncertainty Disentanglement." Available at: https://doi.org/10.48550/arXiv.2204.09308

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2018, "Aleatoric Uncertainty Estimation with Test-Time Augmentation for Medical Image Segmentation with Convolutional Neural Networks." Available at: https://doi.org/10.1016/j.neucom.2019.01.103

Wang, L., Tang, M., Hu, X., 2023, "Evaluation of Grouped Capsule Network for Intracranial Hemorrhage Segmentation in CT Scans." Available at: https://doi.org/10.1038/s41598-023-30581-4

Whitbread, L., Jenkinson, M. (2022). "Uncertainty Categories in Medical Image Segmentation: A Study of Source-Related Diversity." In: "Sudre, C.H., et al. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging." UNSURE 2022. Lecture Notes in Computer Science, vol 13563. Springer, Cham. Available at: https://doi.org/10.1007/978-3-031-16749-2_3

Xu, B., Fan, Y., Liu, J., Zhang, G., Wang, Z., Li, Z., Guo, W., Tang, X., 2023, "CHSNet: Automatic Lesion Segmentation Network Guided by CT Image Features for Acute Cerebral Hemorrhage" Available at: https://doi.org/10.1016/j.compbiomed.2023.107334

Zhao, X. et al., 2024, "SAM-Driven Weakly Supervised Nodule Segmentation with Uncertainty-Aware Cross Teaching." Available at: https://doi.org/10.48550/arXiv.2407.13553

He, Y., and Li, L., 2024, "Uncertainty-Aware Evidential Fusion-Based Learning for Semi-Supervised Medical Image Segmentation." Available at: https://doi.org/10.48550/arXiv.2404.06177

Zhovannik, I. et al., 2022, "Segmentation Uncertainty Estimation as a Sanity Check for Image Biomarker Studies." Available at: https://doi.org/10.3390/cancers14051288

Zarei, D., Issaiy, M., Kolahi, S., and Liebeskind, D.S., 2024, "Do Deep Learning Algorithms Accurately Segment Intracerebral Hemorrhages on Noncontrast Computed Tomography? A Systematic Review and Meta-Analysis" Available at: https://doi.org/10.1161/SVIN.123.001314

Hershkovitch, T., Riklin-Raviv, T., 2018, "Model-Dependent Uncertainty Estimation of Medical Image Segmentation." Available at: https://doi.org/10.1109/ISBI.2018.8363827.

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993, "Comparing Images Using the Hausdorff Distance." Available at: https://doi.org/10.1109/34.232073

Kothala, L.P., Guntur, S.R., 2022, "Segmentation of Intracranial Hemorrhage Through an EfficientNetB7-Based UNET Model." Available at:

Wu, B. et al., 2023, "BHSD: A 3D Multi-Class Brain Hemorrhage Segmentation Dataset" Available at: https://doi.org/10.1109/SMARTGENCON56628.2022.10083730