Label fusion and training methods for reliable representation of inter-rater uncertainty

Andreanne Lemay   andreanne.lemay@polymtl.ca
NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada
Mila, Quebec AI Institute, Montreal, QC, Canada

Charley Gros   charley.gros@gmail.com
NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada
Mila, Quebec AI Institute, Montreal, QC, Canada

Enamundram Naga Karthik   naga-karthik.enamundram@polymtl.ca
NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada
Mila, Quebec AI Institute, Montreal, QC, Canada

Julien Cohen-Adad   jcohen@polymtl.ca
NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada
Mila, Quebec AI Institute, Montreal, QC, Canada
Functional Neuroimaging Unit, CRIUGM, Université de Montréal, Montreal, QC, Canada

Abstract

Medical tasks are prone to inter-rater variability due to multiple factors such as image quality, professional experience and training, or guideline clarity. Training deep learning networks with annotations from multiple raters is a common practice that mitigates the model's bias towards a single expert. Reliable models generating calibrated outputs and reflecting the inter-rater disagreement are key to the integration of artificial intelligence in clinical practice. Various methods exist to take into account different expert labels. We focus on comparing three label fusion methods: STAPLE, average of the rater's segmentation, and random sampling of each rater's segmentation during training. Each label fusion method is studied using both the conventional training framework and the recently published SoftSeg framework that limits information loss by treating the segmentation task as a regression. Our results, across 10 data splittings on two public datasets (spinal cord gray matter challenge, and multiple sclerosis brain lesion segmentation), indicate that SoftSeg models, regardless of the ground truth fusion method, had better calibration and preservation of the inter-rater rater variability compared with their conventional counterparts without impacting the segmentation performance. Conventional models, i.e., trained with a Dice loss, with binary inputs, and sigmoid/softmax final activate, were overconfident and underestimated the uncertainty associated with inter-rater variability. Conversely, fusing labels by averaging with the SoftSeg framework led to underconfident outputs and overestimation of the rater disagreement. In terms of segmentation performance, the best label fusion method was different for the two datasets studied, indicating this parameter might be task-dependent. However, SoftSeg had segmentation performance systematically superior or equal to the conventionally trained models and had the best calibration and preservation of the inter-rater variability. SoftSeg has a low computational cost and performed similarly in terms of uncertainty to ensembles which require multiple models and forward passes. Our code is available at `https://ivadomed.org`.

Keywords: Inter-rater variability, Calibration, Segmentation, Deep learning, Soft training, Label fusion.

1. Introduction

Manual annotation of medical images is challenged by ill-defined boundaries between anatomical regions, and hence prone to inter-expert variability. Inter-expert disagreement is widely acknowledged as a key limitation in medical image analysis (Schaekermann et al., 2019) as it hinders the definition of ground truth (GT) annotation (Mirikharaji et al., 2021; Shwartzman et al., 2019; Yu et al., 2020). For instance, the multiple sclerosis (MS) brain dataset annotated by 7 experts reported an inter-expert agreement ranging between experts from 0.66 to 0.76 of median Dice score with the consensus (Commowick et al., 2021). This variability can arise from many factors, including image quality, expert experience, or guidelines clarity (Mirikharaji et al., 2021; Schaekermann et al., 2019). To mitigate this issue along with speeding annotating time and enhancing reproducibility, a large number of automatic annotation algorithms have been proposed (Gabr et al., 2020; Gros et al., 2019; Isensee et al., 2017; Lemay et al., 2021b). However, the annotations provided by these automatic algorithms are likely to reflect the characteristics of the data they are trained on, including the biases they carry such as different expert experience or style (Vincent et al., 2021). Therefore, it is common practice to provide, for each image, annotations from multiple experts (Commowick et al., 2018; Mirikharaji et al., 2021; Prados et al., 2017; Schaekermann et al., 2019). It remains, however, unclear how to properly use these multiple experts' annotations, i.e., to combine them to generate a GT, to preserve the inter-rater variability information while limiting the expert bias encoded in the model (Mirikharaji et al., 2021).

1.1 Study outline

This study compares different methods to aggregate multiple experts' annotations as GT in algorithm training. A common method to use multiple experts' annotations is to fuse them to create a single mask per image. The fusion method can lead to masks with either categorical values (e.g., zeros or ones for a one-class segmentation task) or soft values (e.g., between 0 and 1), hereafter called "hard fusion" and "soft fusion", respectively. Hard fusion methods include "Simultaneous truth and performance level estimation" (STAPLE) (Warfield et al., 2004), majority voting, intersection, or union, and were widely used in the automatic segmentation literature. On the other hand, soft fusion methods, e.g., averaging the experts' annotations, received a modest interest, probably because most segmentation algorithms assume GTs with categorical values. A training pipeline, called SoftSeg, has been recently proposed to favor the propagation of soft labels (i.e., non-categorical values) (Gros et al., 2021a). The comparison between soft and hard fusion methods questions the tradeoff between the precision and the generalization of a "gold standard" as a precise ground-truth (i.e., hard / binary) may not be reflective of the underlying inter-expert uncertainty. Alternatively, one can choose not to fuse the experts' annotations and, instead, to use them independently when training a segmentation method. This approach is hereafter called random sampling method and aims to preserve the raw multi-expert labeling while confronting the algorithm to contradictory annotations (Jensen et al., 2019; Jungo et al., 2018; Mirikharaji et al., 2021).

1.2 Related works

Some studies compared methods to generate GT labels when experts disagree. Jensen et al. demonstrated that hard fusion, i.e., majority voting, led to over-confident models on skin disease predictions (i.e., uncalibrated model) (Jensen et al., 2019). They showed that a "no fusion" approach, i.e., label random sampling, could mitigate this miscalibration in the model's prediction. Jungo et al. compared hard fusion (STAPLE, majority voting, intersection, and union) methods with the random sampling approach in terms of segmentation performance and uncertainty estimation (Jungo et al., 2018). The random sampling method yielded uncertainty that was able to reflect the underlying expert disagreement on synthetic data and on subjects with a Dice score superior to the median of a brain tumor dataset, but no positive impact was noticed for subjects with a low segmentation performance. Conversely, the hard fusion methods led to an under-estimation of uncertainty, suggesting that inter-expert variability needs to be explicitly taken into consideration when training models in order to reliably estimate uncertainty. To the best of our knowledge, there is currently no study that compares soft fusion methods with hard fusion and random sampling approaches.

Increasing attention has been given to uncertainty estimation. Notably, Quantification of Uncertainties in Biomedical Image Quantification Challenge (QUBIQ) aimed at developing uncertainty estimation methods that were evaluated on MRI and CT scan datasets with multiple raters per annotation. The QUBIQ2020 winning team proposed training one model per rater and aggregating the outputs to evaluate the segmentation uncertainty (Ma, 2020). Silva and Oliveira (2021) proposed to use soft labels computed by averaging all labels and had well-calibrated outputs, competitive with other approaches from the QUBIQ2021 challenge.

1.3 Our contribution

In this study, we compare the impact of hard fusion, soft fusion, and label random sampling methods using SoftSeg or a conventional training framework. The inter-rater variability is lost in hard fusion methods (Yu et al., 2020) and the conventional framework, which inputs binarized GT and trains with categorical losses, limiting the learning of expert disagreement. Hence, we hypothesize that soft or random sampling methods and the SoftSeg framework will better reflect the inter-rater variability, will generate more calibrated predictions, and will yield improved segmentation performances than hard fusion and conventional training methods. The label generated by these methods is used to feed a U-Net (Ronneberger et al., 2015), widely considered as the state-of-the-art in automatic image segmentation. The training is performed using both a conventional pipeline and the recently proposed alternative, SoftSeg. Each method, six in total (two training pipelines, each using the three methods to generate the GT, see Table 1), are compared on two MRI open-source datasets, the spinal cord gray matter (SCGM) challenge (Prados et al., 2017) and multiple sclerosis (MS) brain lesion challenge (Commowick et al., 2018), in terms of (i) preservation of the inter-rater variability, (ii) model calibration and, (iii) segmentation performance.

3

2. Method and Material

2.1 Method

2.1.1 Label fusion

Three methods to exploit multiple rater labels were studied: STAPLE (Warfield et al., 2004), average across GTs, and random sampling of one annotation during training without fusion (Jungo et al., 2018). STAPLE is an expectation-maximization algorithm widely used for label fusion in medical imaging (Akkus et al., 2017; Commowick et al., 2018; Maier et al., 2017). This method produces binary GTs. The second label fusion strategy studied, averaging across all annotations, aims to preserve all the inter-rater variability information by outputting soft (i.e., values between 0 and 1) GTs. However, conventional segmentation pipelines usually binarize the GTs which leads to a majority voting when averaging segmentations across raters. To fully exploit this label fusion method, a soft segmentation framework such as SoftSeg (Gros et al., 2021a) is required. The third method does not merge the labels. During each training epoch, one rater segmentation is randomly chosen as GT, eventually exposing the model to all the rater's annotations. This means that all training cases from all raters are shown to the model during the training. Therefore, the random sampling method uses binary segmentations.

2.1.2 Training framework

In this work, we compare each label fusion method when trained with both SoftSeg and a conventional segmentation training framework. SoftSeg has three differences when compared with the conventional approach: no binarization during the preprocessing and data augmentation, soft final activation function, and training using a regression loss function (Gros et al., 2021a). The final activation and the regression loss function are normalized ReLU and Adaptive Wing loss (Wang et al., 2019b) respectively as defined in Gros et al. (2021a). The final activation was adapted for multi-class predictions. When using the conventional approach, the GTs were binarized after preprocessing and data augmentation, the models were trained with a Dice loss, and sigmoid and softmax final activation functions were used for the single-class and multi-class models respectively.

An additional note about SofSeg: in the original work of SoftSeg, the final activation function used was a normalized ReLU. The ReLU prediction was then normalized by the maximum value to have a segmentation prediction corresponding to a level of confidence from 0 to 1. However, this activation function is not directly applicable to multi-class predictions as the different classes would not have probabilities summing up to 1. To generalize the normalized ReLU, the output of the original normalized ReLU is divided by the sum across all classes including the background class. This way, all predicted classes are mutually exclusive and have probabilities summing to 1.

2.1.3 Training protocol

All candidates were trained on 2D U-Net models. Training parameters for this work were the same as the one described in Gros et al. (2021a) for the SCGM and MS brain lesion datasets. The use of the same model architecture, training parameters, dataset, and training environment (ivadomed) facilitate comparisons between the present study and the previous

Table 1: Candidates' description. The columns indicate the label fusion method while the rows present the training framework.

Training	Label fusion method		
framework	**STAPLE**	**Average**	**Random sampling**
Conventional	Conv-STAPLE	Conv-Average	Conv-RandomSamping
SoftSeg	SoftSeg-STAPLE	SoftSeg-Average	SoftSeg-RandomSamping

study by (Gros et al., 2021a). Future studies could consider undertaking similar analyses with different model architectures, such as 3D models. The processing, training and evaluation pipeline is based on the open-source framework `ivadomed.org` (Gros et al., 2021b). ivadomed is an open-source medical image analysis Python library based on PyTorch that provides tools, e.g., data loader, models, losses, transformations, pre- and post-processing, metrics, to train and use deep learning models for medical tasks such as segmentation.

2.2 Datasets

Two publicly available datasets with multiple raters were used to study label fusion: the SCGM challenge (Prados et al., 2017) and MS brain lesion challenge (Commowick et al., 2018).

2.2.1 GRAY AND WHITE MATTER CHALLENGE

The SCGM dataset contains 80 T2*-weighted MRI of the cervical spinal cord, evenly acquired in four centers with different MR protocols and scanner vendors. Four raters segmented the gray and white matter from the scans using different guidelines and segmentation software which increases the inter-rater variability across centers. While the dataset includes 80 subjects, only 40 had all 4 raters publicly available, hence, this subdataset of 40 scans was retained for this study. A detailed description of the dataset and demographics of the scanned subjects and acquisition parameters can be found in Prados et al. (2017) or a summary can be found in Appendix A.

2.2.2 MS BRAIN LESION CHALLENGE

The MS brain lesion dataset containing MRI scans from 15 subjects was presented during the MICCAI 2016 challenge. MS lesions of each subject were annotated by seven expert raters. The dataset includes MRI scans with five contrasts: T1-weighted, T1-weighted Gadolinium-enhanced, T2-weighted, PD T2-weighted, and FLAIR. For a detailed description of the dataset see Commowick et al. (2018).

2.3 Evaluation

2.3.1 EVALUATION PROTOCOL

Each model was trained with multiple random dataset splittings to limit splitting bias. For the SCGM dataset, 40 models were trained with an even split on the test centers (10

trainings with center 1 as test set, 10 trainings with center 2 as test set, etc.), while for the MS brain lesion dataset, 20 random splittings were performed (60/20/20% for training/validation/test sets). Before assessing the predictions, the outputs were resampled in the native resolution. A non-parametric 2-sided Wilcoxon signed-rank test compared the most commonly used approach, "Conv-STAPLE", with every other approach. Statistical difference was assessed by considering 0.05 as p-value threshold.

2.3.2 UNCERTAINTY DUE TO INTER-RATER VARIABILITY

To evaluate the preservation of the inter-rater variability, we assessed the correspondence between the uncertainty from the prediction and the uncertainty associated with the multiple annotations. The patient uncertainty can be measured with the predictive entropy (Equation 1) (Jungo et al., 2018) which can be directly compared with the entropy associated to the multiple rater segmentation (GT average). Entropy was chosen as it can directly be computed from the model's prediction and does not require multiple forward passes as in other popular methods such as Monte Carlo dropout (Gal and Ghahramani, 2016) or test-test augmentation (Wang et al., 2019a). A high entropy value indicates a high inter-rater variability. For example, if the fused label across raters is close to 0 or 1 in a given voxel, the level of agreement is high (i.e., low entropy), while values near 0.5 indicate high disagreement. A reliable model would generate a prediction reflecting the expert disagreement similarly to the GT average. Therefore, we plotted the entropy of the prediction against the entropy of the GT average and we expect the values to match. The correspondence was assessed by computing the mean absolute error (MAE) between both values for each patient data.

$$H = -\sum_{i=0}^{N_{vox}} p_i log(p_i) \tag{1}$$

where p_i is the model's prediction for voxel i and N_{vox} is the total number of voxels in the image.

In addition, we quantified the voxel-wise similarity of the uncertain regions with voxels associated with high inter-rater variability. The Brier score (Equation 2) enables us to assess the similarity of non-binary data, hence was used to evaluate the similarity between the model's prediction and the average labels from the expert raters. The average label from raters was selected as GT to quantify the performance of the soft prediction since information on inter-rater variability is encoded in this label while it cannot be directly observed from the STAPLE GT. The metric was computed for each segmentation class.

$$Brierscore = \frac{1}{N_{vox}} \sum_{i=0}^{N_{vox}} (y_i - \hat{y}_i)^2 \tag{2}$$

where y is the GT average, $\hat{y}$ is the prediction, and N_{vox} is the total number of voxels in the image.

2.3.3 CALIBRATION

Reliable deep learning models should predict calibrated outputs to truthfully indicate regions more prone to error or inter-expert disagreement. The model's calibration quantifies how much the predicted values of a model truly represents the probability of the outcome, hence is an indicator of the quality of the model's confidence. For instance, a perfectly calibrated model predicting 0.9 is confident at 90% of its prediction and, therefore, should be correct 90% of the time. Reliability diagrams (DeGroot and Fienberg, 1983) and the expected calibration error (ECE) (Naeini et al., 2015) were computed with the code from google-research repository[1] as used in Guo et al. (2017) to assess the calibration of the candidates.

Reliability diagram The reliability diagram helps to visualize the calibration of the model and plots the prediction's accuracy (Equation 3) in relation to the model's confidence (Equation 4). The identity function represents a perfectly calibrated model where the accuracy and the model's confidence are always equal. Any deviation from this line can be translated into over- or underconfidence from the model. The model's confidence was discretized into K=10 bins of size 0.1 ($\frac{1}{K}$) based on the literature (Naeini et al., 2015). Moreover, preliminary results showed that changing the number of bins from K=5 to 20 did not change the calibration order of the approaches studied. We define confidence as the maximal prediction across classes for a given voxel. For a 3-class prediction problem, a model predicting [0.9, 0.06, 0.04] is associated with a confidence of 0.9. The minimum confidence for a 3-class prediction problem is 0.33^+ (i.e., $[0.33^-, 0.33^-, 0.33^+]$), while for a binary prediction the minimum confidence is 0.5^+ (i.e., $[0.5^-, 0.5^+]$). The predicted values are compared to the binarized GT, here, the STAPLE GT. The accuracy is the proportion of voxels from a given bin, B_k, where the predicted class corresponds to the GT. The bin B_m includes all predictions associated with a confidence of $[\frac{k}{K}, \frac{k+1}{K})$ (Guo et al., 2017). This accuracy is then compared to the average prediction in the bin. For instance, for the bin including voxels with values from [0.8 to 0.9), we expect that 85% of the voxels in this bin, assuming uniform distribution of predicted values, are well classified. If the accuracy is greater than the model's confidence, the model is underconfident, while a lower accuracy compared with the model's confidence means the model is overconfident.

$$Accuracy(B_k) = \frac{\sum_{i \in B_k} 1(y_i = \hat{y}_i)}{\#B_k} \tag{3}$$

$$Confidence(B_k) = \frac{\sum_{i \in B_k} \hat{y}_i}{\#B_k} \tag{4}$$

where $\#B_k$ corresponds to the number of elements in the bin B_k.

Expected calibration error The reliability diagram does not display the information about the quantity of voxels in each bin. The ECE (Equation 5) is a metric extracted from the reliability diagram that takes into account the occurrence of voxels in each bin. The ECE corresponds to the sum of the absolute difference between the confidence of the model

---

1. https://github.com/google-research/google-research/blob/master/uncertainties/sources/
   postprocessing/metrics.py

and the accuracy (i.e., the miscalibration) weighted by the number of voxels. The ECE was measured on all predictions from a model and averaged across models with different random splittings.

$$ECE = \sum_{k=0}^{K} \frac{\#B_k}{N_{vox}} \big| Accuracy(B_k) - Confidence(B_k) \big| \tag{5}$$

where N_{vox} is the total number of voxels in the image.

2.3.4 Segmentation accuracy

Metrics for binarized predictions To evaluate the quality of the segmentation the following metrics were used: Dice score, precision, recall, relative volume difference between the GT and prediction divided by the GT volume (RVD), and absolute volume difference (AVD) which is the absolute value of RVD. Due to the binary nature of these metrics, the predictions of the model were binarized. For example, a prediction of 0.5 with a GT of 0.5 obtained by averaging labels results in a Dice score of 0.5 even though both values are the same and should reach a maximal score. For this same reason, the STAPLE annotations were used as GTs for these metrics. Approaches using STAPLE during training (SoftSeg-STAPLE and Conv-STAPLE) were positively biased in that the evaluation metrics were computed with STAPLE as GT. Despite this limitation, STAPLE was the best option available for binary GT as it takes into account all the raters' opinions at once. For the MS dataset which has two classes (i.e., lesion or background), the binarization threshold was found by searching for the optimal value (between 0 and 1 with an increment of 0.05) in terms of Dice score as done in Gros et al. (2021a). The threshold was optimized for each model individually, regardless of the training method (i.e., conventional or SoftSeg models). For the SCGM dataset, there are three classes: gray matter, white matter and background. The predicted class is obtained by selecting the maximum prediction across the three classes.

Composite score To represent the overall segmentation accuracy performance, a composite score is computed by aggregating the above metrics. Firstly, z-scores for each metric are derived by standardizing the results across candidates (i.e., zero mean and unit standard deviation). Secondly, the z-scores are linearly aggregated to compute the composite score, with equal absolute weights across metrics. A weight of 1 was used for the Dice, precision, and recall (because they need to be maximized), and a weight of -1 was used for the AVD (because it needs to be minimized).

## 3. Results

3.1 Inter-rater uncertainty

The predicted segmentation should ideally reflect the uncertainty associated with the disagreement between experts. Figure 1 illustrates the agreement between the entropy generated from the multiple expert ratings and the predicted segmentation's entropy. Similar observations can be drawn for both SCGM and MS brain segmentation. The SoftSeg models showed better correspondence between the predicted and true entropy which can be seen

Table 2: Quantitative assessment of the inter-rater variability preservation on the SCGM and MS brain datasets (MEAN $\pm$ STD). Brier score is reported by segmentation class while the MAE is computed on the total entropy of the entire image. Each row represents a candidate. The best averaged result for each metric and tissue is displayed in bold. Statistical differences are computed between "Conv-STAPLE" (ref) and each other candidate (**: $p < 0.05$). Abbreviations: Opt.: optimal; MAE: mean absolute error; GM: gray matter; WM: white matter.

		SCGM			MS brain	
		Brier score ($\times 10^{-4}$) *Opt. value: 0*		**MAE** ($\times 10^3$) *Opt. value: 0*	**Brier score** ($\times 10^{-4}$) *Opt. value: 0*	**MAE** ($\times 10^4$) *Opt. value: 0*
		GM	**WM**	**Entire image**	**MS lesions**	**Entire image**
Conventional	**STAPLE (ref)**	1.05 ± 0.04	3.08 ± 0.11	2.30 ± 0.02	7.78 ± 5.84	5.32 ± 2.86
	Average	$1.18 \pm 0.07*$	$3.02 \pm 0.11*$	$2.33 \pm 0.01*$	7.91 ± 5.05	5.43 ± 3.07
	Random sampling	$0.98 \pm 0.05*$	$2.81 \pm 0.16*$	$2.37 \pm 0.01*$	8.21 ± 3.84	4.86 ± 2.53
SoftSeg	**STAPLE**	$0.96 \pm 0.06*$	$2.92 \pm 0.10*$	$\mathbf{0.98 \pm 0.06}^*$	$5.71 \pm 2.74*$	$\mathbf{2.97 \pm 2.59}^*$
	Average	$1.14 \pm 0.07*$	$3.00 \pm 0.10*$	$1.13 \pm 0.09*$	$5.55 \pm 2.91*$	5.19 ± 3.20
	Random sampling	$\mathbf{0.90 \pm 0.03}^*$	$\mathbf{2.68 \pm 0.03}^*$	$1.08 \pm 0.08^*$	$\mathbf{5.01 \pm 2.41}^*$	$3.40 \pm 3.96^*$

by data points lying near the identity line (perfect agreement) and smaller MSE values. All models trained with the conventional framework showed a tendency to generate less entropy which can be interpreted as overconfidence and an underestimation of the uncertainty. Random sampling and STAPLE have similar patterns with the SoftSeg framework and reflect the more truthfully the entropy linked to multiple raters. "SoftSeg-Average" showed a slight tendency to overestimate the uncertainty. Clusters can be observed in Figure 1a and are associated with the different data centers of the SCGM dataset.

Table 2 summarizes the metrics associated with the preservation of the inter-rater variability. A general trend that can be observed is that SoftSeg candidates performed better than their conventional counterparts. When performing pairwise comparisons of each candidate using SoftSeg vs. the conventional framework, SoftSeg systematically yielded the best average metric. More precisely, for all metrics, "SoftSeg-RandomSampling" and "SoftSeg-STAPLE" were always the top two performing candidates. For both dataset and on all classes, "SoftSeg-RandomSampling" yielded the lowest Brier score indicating the greater resemblance with the segmentation from the averaged labels. "SoftSeg-STAPLE" obtained the best correspondence, i.e., lowest MAE, between the predicted uncertainty and the inter-rater variability. The MAE, which should be minimized, of conventional models was on average 46% and 45% higher compared with SoftSeg models for the SCGM and MS brain datasets respectively.
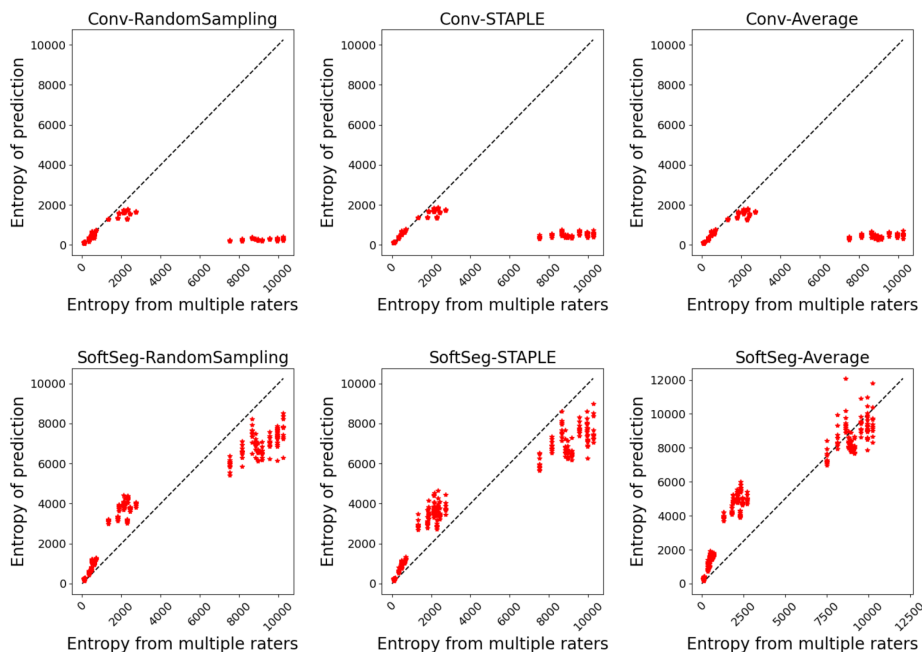
(a) SCGM
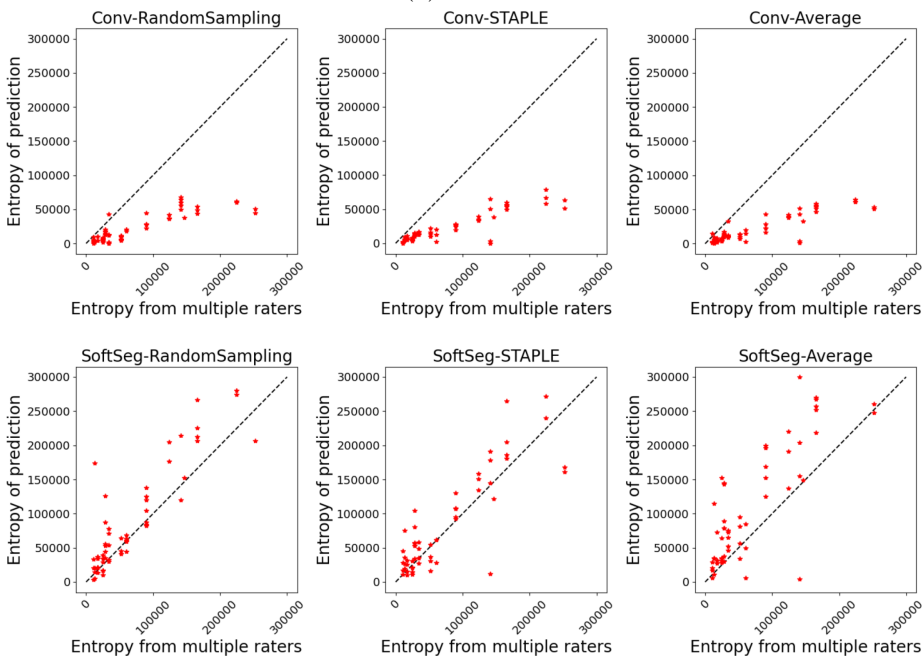
(b) MS brain

Figure 1: Comparison of entropy generated by inter-rater variability and entropy from the model's prediction for the SCGM (a) and MS brain lesions (b) datasets. Each red dot corresponds to a participant. The dashed line represents the identity line where data points from an ideal model should lie.
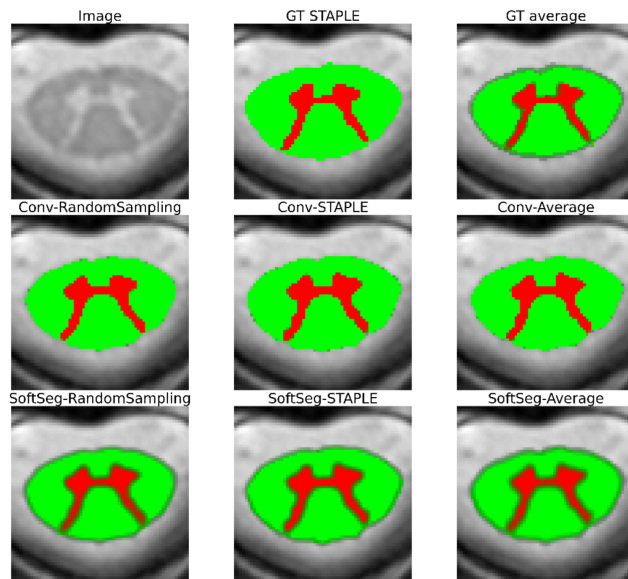
Figure 2: Visual assessment of STAPLE and average GTs and predictions from the six candidates on spinal gray and white matter segmentation. Abbreviations: GT: ground truth.

3.2 Visual assessment

Figure 2 and Figure 3 contain the segmentations from the STAPLE and GT average and from the predictions of the six candidates. Regardless of the label fusion method and the dataset, predictions using conventional models have sharp edges between tissue types (similar to the GT STAPLE) and underestimated the inter-rater variability. All SoftSeg candidates display smoother boundaries (similar to the GT average). When comparing the SoftSeg models, "SoftSeg-Average" presents the softest edges followed by "SoftSeg-RandomSampling", then "SoftSeg-STAPLE". These differences are especially noticeable in Figure 2 at the extremity of the dorsal horns and near the central canal (black arrows) and in Figure 3 on the lesion aggregate on the top-left (yellow arrows). An ideal prediction should reflect the inter-rater variability similarly to the GT average. Hence, predictions should not be too sharp or too smooth compared to the GT average.

3.3 Calibration

Figure 4 presents the reliability diagrams generated from the predictions on SCGM and on MS brain lesion datasets. The conventional approach is overconfident with most of its predictions for all the datasets and label fusion methods. This overconfidence results in high ECE: 16.2% for SCGM and 20.4% for MS lesions on average. In contrast, the SoftSeg is on average better calibrated with an ECE of 2.9% for SCGM and 2.4% for MS lesions. SoftSeg candidates mostly present slight underconfidence with the exception of "SoftSeg-STAPLE" on MS lesions which is overall well calibrated with minimal overconfidence. "SoftSeg-STAPLE" and "SoftSeg-RandomSampling" are the candidates presenting the best calibration. "SoftSeg-Average" presents more underconfidence compared to the
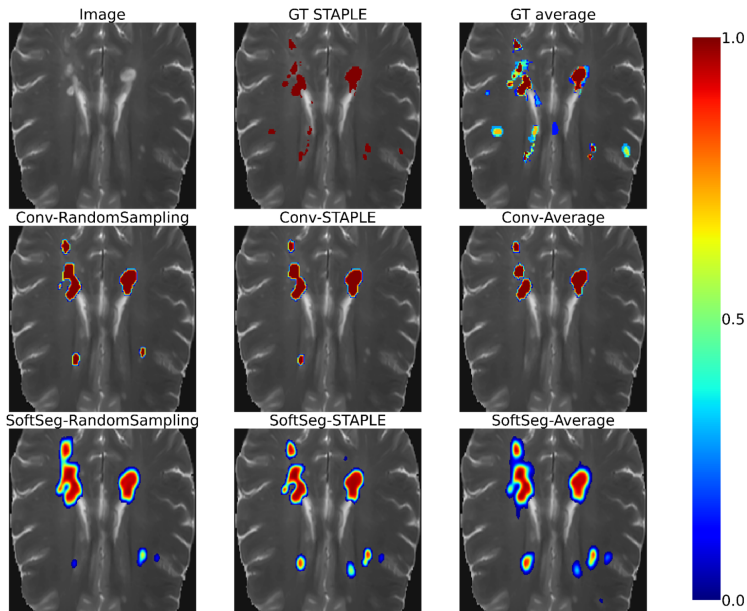
Figure 3: Visual assessment of STAPLE and GT average and predictions from the six candidates on MS brain segmentation. Red: Gray matter. Green: White matter. Voxels at tissue boundaries represent values between 0 and 1. Abbreviations: GT: ground truth.

other SoftSeg candidates due to the overly soft predictions encouraged by the non-binary GT.

3.4 Segmentation accuracy

Table 3a presents the quantitative results of the segmentation accuracy assessment on the SCGM dataset. When comparing the binarized predictions to the STAPLE GT, "SoftSeg-STAPLE" yielded the best Dice, recall, AVD, and RVD score for both white and gray matter segmentation and significantly outperformed the "Conv-STAPLE" method ($p <$ 0.05). Figure 5a summarizes the metrics presented in Table 3a using a composite score. The averaged composite score indicates that, regardless of the training pipeline (i.e., conventional and SoftSeg), the best label fusion method is STAPLE (see Figure 5a). The composite score of SoftSeg was consistently higher compared with the conventional framework for a given label fusion method. All composite scores were statistically different from the "Conv-STAPLE" candidate.

Table 3b introduces the MS brain segmentation performance metrics. Most metrics on the binary prediction demonstrated no significant difference compared with the "Conv-STAPLE" candidate. Only the "Conv-RandomSampling" candidate had a significantly lower Dice score compared to the "Conv-STAPLE". Figure 5b summarizes the metrics presented in Table 3b using a composite score. "SoftSeg-Average" provided the best composite score, followed by "Conv-STAPLE". When comparing the composite scores of the candidates with "Conv-STAPLE", no significant difference was found, except "Conv-RandomSampling" and "SoftSeg-RandomSampling" which led to significantly lower results.
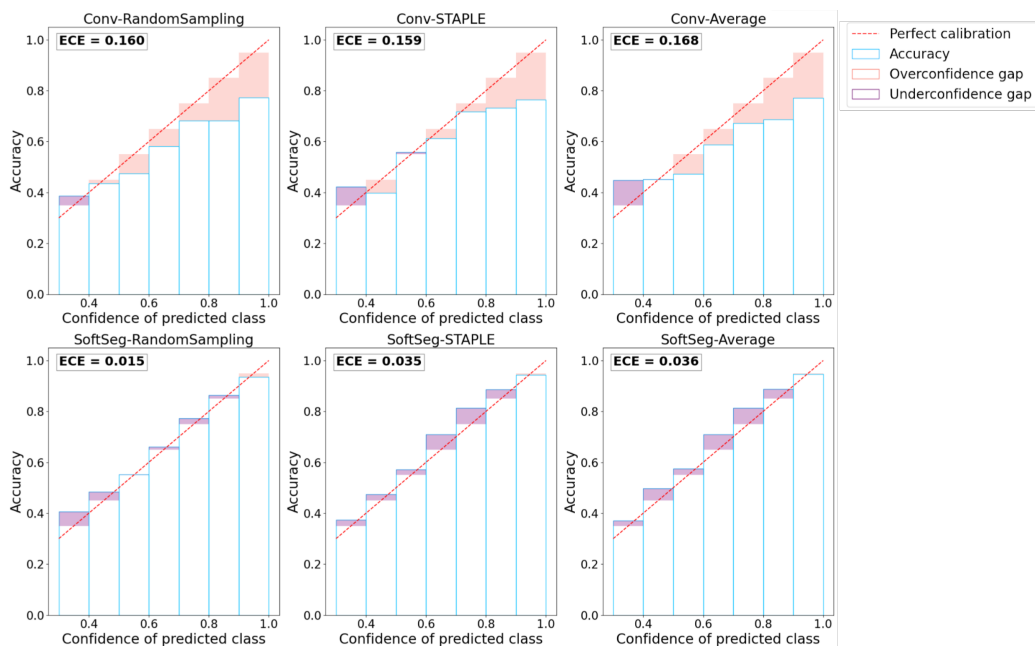
(a) SCGM
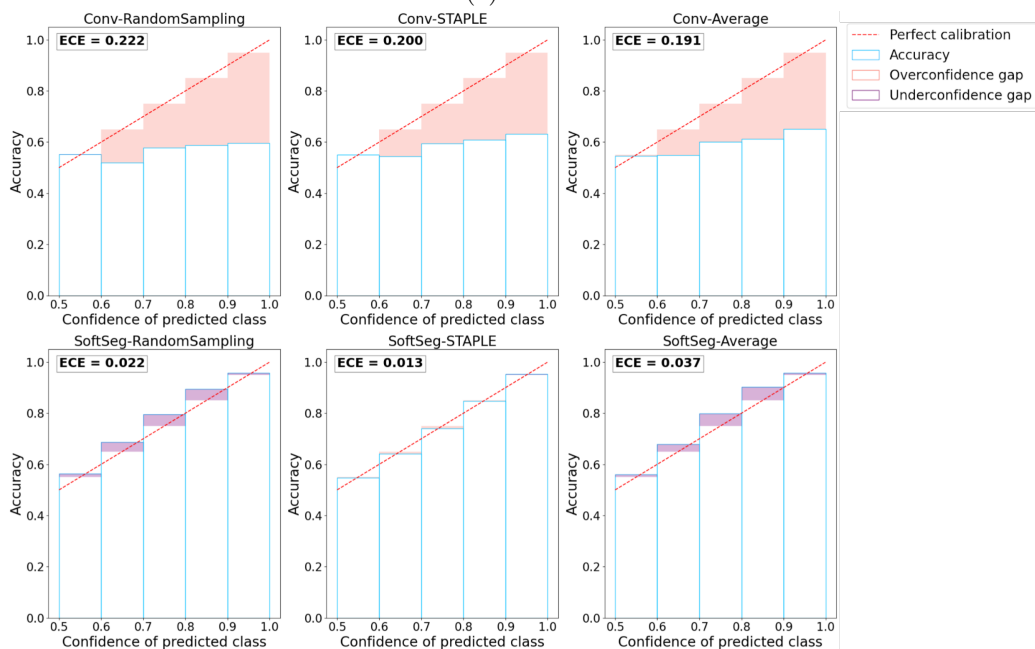
(b) MS brain

Figure 4: Reliability diagram for all candidates on SCGM (a) and MS brain lesions (b) datasets. The red identity line illustrates a perfect calibration. Orange bands represent overconfidence while the purple ones indicate underconfidence.

Table 3: Quantitative assessment of the segmentation performance on the SCGM and brain MS lesions datasets. For SCGM, each value represents the average over 40 models and intervals are the standard deviation over these. Mean and standard deviation are reported for both gray and white matters. For brain MS lesions, each value represents the average and standard deviation over 20 models and intervals are the standard deviation over these. Each row represents a candidate. The best averaged result for each metric and tissue is displayed in bold. All metrics are computed on binarized predictions against the "STAPLE GT". Statistical differences are computed between "Conv-STAPLE" (ref) and each other candidate (*: $p < 0.05$). Abbreviations: Opt.: optimal; GM: gray matter; WM: white matter.

(a) SCGM

		Dice [%] *Opt. value: 100*		Precision [%] *Opt. value: 100*		Recall [%] *Opt. value: 100*		AVD [%] *Opt. value: 0*		RVD [%] *Opt. value: 0*	
		GM	WM	GM	WM	GM	WM	GM	WM	GM	WM
Conventional	**STAPLE** (ref)	86.6 ±2.8	93.1 ±2.0	89.0 ±6.5	94.0 ±1.0	85.1 ±3.1	92.5 ±3.5	10.1 ±3.8	**5.2** **±1.8**	3.6 ±9.9	1.5 ±3.5
	Average	83.6 ±2.1*	90.7 ±2.6*	91.6 ±6.7*	94.5 ±0.7*	77.9 ±5.1*	87.5 ±4.8*	17.4 ±6.9*	7.9 ±4.6*	14.1 ±12.1*	7.4 ±5.1*
	Random sampling	84.2 ±4.4*	90.1 ±4.2*	**93.6** **±5.4***	**95.5** **±0.4***	77.8 ±7.4*	86.0 ±6.6*	18.1 ±9.4*	10.1 ±6.7*	16.3 ±11.9*	10.0 ±6.8*
SoftSeg	**STAPLE**	**87.1** **±3.4***	**93.3** **±2.1***	88.7 ±7.3*	93.7 ±1.0*	**86.4** **±2.5***	**93.2** **±3.7***	**9.9** **±4.5***	5.2 **±2.0***	**1.6** **±10.7***	**0.5** **±3.7***
	Average	83.7 ±3.6*	91.0 ±2.5*	91.1 ±7.5*	95.0 ±0.8*	78.8 ±6.1*	87.5 ±4.4*	16.8 ±7.5*	8.2 ±4.5*	12.4 ±13.6*	7.9 ±4.6*
	Random sampling	84.8 ±4.0	90.4 ±3.6*	93.1 ±5.6*	**95.5** **±0.8***	79.1 ±6.3*	86.6 ±5.8*	16.4 ±8.1*	9.4 ±6.0*	14.4 ±11.0*	9.2 ±6.2*

(b) MS brain

		Dice [%] *Opt. value: 100*	Precision [%] *Opt. value: 100*	Recall [%] *Opt. value: 100*	AVD [%] *Opt. value: 0*	RVD [%] *Opt. value: 0*
Conventional	**STAPLE** (ref)	54.9 ± 12.8	56.3 ± 14.0	**57.0 ± 12.4**	31.8 ± 18.2	-7.5 ± 26.0
	Average	54.3 ± 12.8	58.0 ± 14.8	$55.6 \pm 13.0^*$	42.0 ± 57.1	-11.8 ± 65.5
	Random sampling	$50.6 \pm 11.8^*$	61.4 ± 13.4	52.9 ± 12.6	$54.8 \pm 54.8^*$	-9.9 ± 64.3
SoftSeg	**STAPLE**	55.0 ± 11.1	**60.4 ± 13.2**	56.6 ± 11.1	51.2 ± 56.2	-16.6 ± 63.2
	Average	**56.0 ± 10.7**	59.8 ± 13.0	55.5 ± 11.5	**30.6 ± 17.7**	**-1.7 ± 26.0**
	Random sampling	53.7 ± 10.3	58.5 ± 13.1	$52.8 \pm 9.4^*$	39.7 ± 50.2	-8.3 ± 59.2
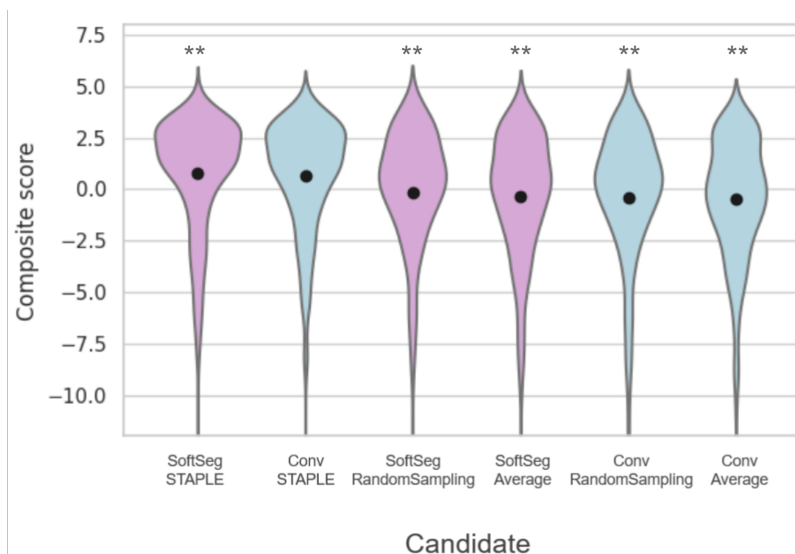
(a) SCGM
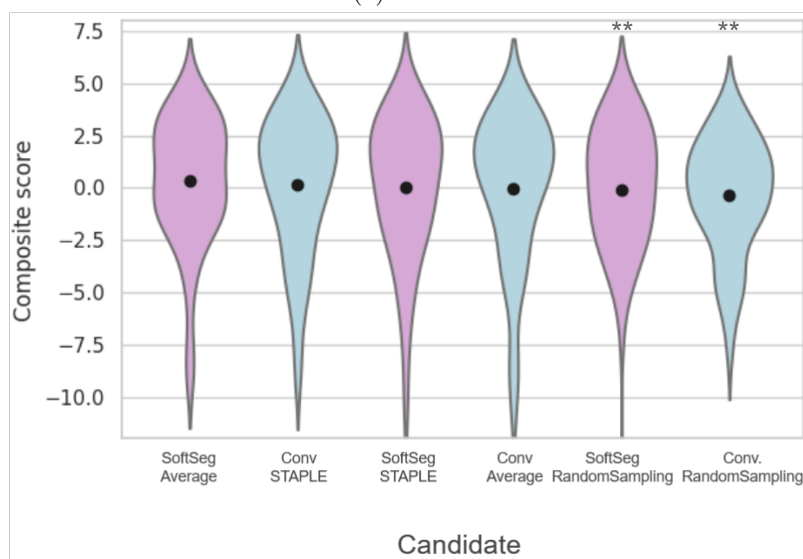
(b) MS brain

Figure 5: Composite score across candidates on the SCGM (a) and MS brain (b) datasets. The composite score aggregates the following segmentation metrics: Dice, Precision, Recall, and AVD. For each candidate, each violin plot represents the distribution of composite scores across testing patients and random splittings. They are sorted from the best to the worst averaged composite score (black dot). Statistical differences are computed between "Conv-STAPLE" (ref) and each other candidate (**: $p < 0.05$).

4. Comparison with loss ensembles

In this section, we compare our method to loss ensembles Ma (2020) - winner of the QUBIQ 2020 challenge. Loss ensembles are inspired by the observation that the manual labels are typically generated by several expert radiologists with varying expertise. Therefore, during the training phase, this method requires training of multiple independent models, each with labels corresponding to a different rater. During inference, the test input was passed through each model and the output was averaged.

We used the MS brain lesion dataset Commowick et al. (2016) for this experiment, containing GT labels from 7 expert raters. Therefore, we trained 7 separated 2D UNet models, each corresponding to a different rater. The final test output was the averaged across the predictions for all the raters. We call this method "Conv-STAPLE-QUBIQ". Figure 6 shows the reliability diagrams for our method and the loss ensembles. We observed that our method SoftSeg-STAPLE is relatively less over-confident compared to loss ensembles Ma (2020) and also achieves lower ECE.
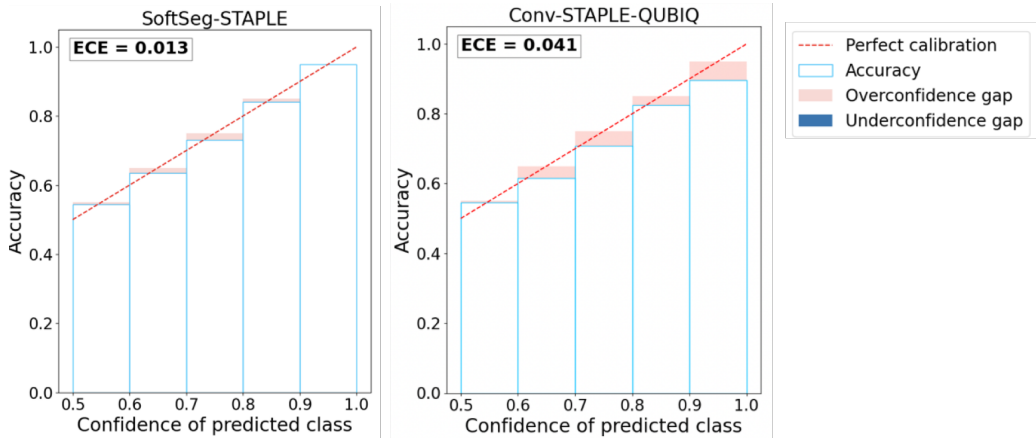

Figure 6: Reliability diagrams comparing our method SoftSeg-STAPLE (left) with the Deep Ensembles method (right). The red identity line illustrates a perfect calibration and the orrange bands represent overconfidence.

We also computed the Brier Score and the MAE of these two methods to better quantify the uncertainty arising from inter-rater variability (see Table 4). It can be seen that SoftSeg-STAPLE performs better on both the uncertainty quantification metrics (while also being computationally efficient).

Table 4: Comparison based on Inter-rater Uncertainty Quantification Metrics (lower the better)

Method	Brier Score $(\times 10^{-4})$	MAE $(\times 10^4)$
Loss Ensembles Ma (2020)	5.39 ± 1.94	1.08 ± 0.59
SoftSeg-STAPLE	$\mathbf{5.28 \pm 1.83}$	$\mathbf{1.01 \pm 0.50}$

5. Discussion

Data labeling is prone to inter-rater variability, and it is still unclear how to best preserve this valuable information when training a deep learning model. In this study, we compared three label fusion methods, using both SoftSeg or a conventional training framework. Overall, SoftSeg models were shown to provide a more reliable representation of the inter-rater variability than using the conventional approach, in terms of (i) correspondence between the predicted and true uncertainty, (ii) visual assessment, (iii) calibration, and (iv) segmentation accuracy. This study suggests that the conventional framework has a tendency to be overconfident and to underestimate the uncertainty, regardless of the label fusion method used. When using SoftSeg, random sampling and STAPLE label fusion methods showed a more reliable inter-rater uncertainty and calibration than the average label fusion method. In this section, we further discuss avenues to preserve information from all raters, via label fusion and/or training pipeline, then we discuss the need to go beyond the Dice score for model evaluations, particularly in the context of multiple raters and soft predictions. Finally, we discuss the importance of repeatability in medical deep learning research.

5.1 The preservation of the inter-rater variability

Encoding the inter-rater variability in the model training is important as it helps tailor models that reflect the experts' disagreement through the predictions. In this study, we investigated two avenues to preserve the inter-rater variability when training a deep learning model: how the raters' labels are fused, and how the labels are processed by the training framework. Overall, we found that the way the labels are used by the training framework is important to preserve the inter-rater variability, while the results of the label fusion methods' comparison were less univocal.

5.1.1 WHEN FUSING THE LABELS

Label fusion is a critical step in many image segmentation frameworks as it is often used to condense a collection of labels from multiple raters into a single estimate of the underlying segmentation. Although the GT generated by averaging the raters' labels is intrinsically a more truthful representation of the inter-rater disagreement than STAPLE (see Figure 2 and Figure 3), training a deep learning model with GT average showed less promising results in this study. The models trained using the averaged GT were underconfident and tended to overestimate the uncertainty, which can be seen on the extended soft edges around the segmented structures (Figure 2 and Figure 3), the larger underconfidence gaps on reliability diagrams and the associated higher ECE (Figure 4), and uncertainty correspondence plots (Figure 1). The models trained with labels from individual raters, i.e., random sampling, were less overconfident than when using consensus labels, which is in line with previous studies (Jensen et al., 2019; Ji et al., 2021). The best calibration results were obtained when using STAPLE as label fusion method, for the MS lesion dataset, and random sampling for SCGM dataset. However, both STAPLE and random sampling had similar reliability diagrams and ECE values (Figure 4), suggesting only small differences between "SoftSeg-STAPLE" and "SoftSeg-RandomSampling" candidates in terms of calibration. A similar trend can be observed for the MAE on the uncertainty correspondence plots (Figure 1). For

both datasets, the Brier score between the GT average and predictions was the best when using random sampling, which is in line with the results obtained by Jungo et al. (2018). In terms of segmentation performance, no clear consensus was reached between the two datasets. "SoftSeg-Average" achieved the best performance for MS lesion segmentation, while "SoftSeg-STAPLE" was the best candidate for SCGM. This could be explained by the fact that MS lesion segmentation is more subject to inter-rater disagreement than spinal cord segmentation. MS lesion segmentation models might benefit more from being explicitly exposed to the rater inter-rater variability than the spinal cord segmentation models. Unlike Jensen et al. (2019); Jungo et al. (2018); Mirikharaji et al. (2021), no equivocal conclusion can be drawn in terms of the best label fusion method. It would be interesting to extend the study to more datasets to confirm our observations that the more appropriate label fusion method might be dataset-specific.

Ideally, all deep learning models would be trained with GTs derived from multiple raters to account for inter-rater variability. However, the availability of datasets with multiple experts segmenting each image is rare in medical settings, because manual segmentation is time-consuming and expensive. In future works, the impact of having one to N raters in the GT annotations on uncertainty representation should be explored to ensure our results stand with a varying number of raters.

In this work, we focused on three label fusion methods to limit the number of model comparisons. However, more fusion methods exist, such as majority voting, soft STAPLE (Kats et al., 2019), Bayesian fusion (Audelan et al., 2022), SIMPLE (Langerak et al., 2010), inter-rater variability sampling scheme (Jensen et al., 2019), and others. Fusing labels with majority voting is a simpler approach compared to STAPLE and can yield similar performance depending on the number of raters and/or the type of task (McGurk et al., 2013). While the present study did not lead to an unequivocal label fusion method to recommend, some works (Audelan et al., 2022; Langerak et al., 2010) suggest methods for improving the STAPLE algorithm, and hence could outperform the fusion methods studied in this work. Future studies could also consider other label fusion methods, such as recently proposed deep-learning approaches to explicitly model the consensus process (Nichyporuk et al., 2021; Yu et al., 2020).

5.1.2 When using the labels through the training pipeline

The way the labels are processed to train a model has important implications on the preservation of the inter-rater variability in this study. SoftSeg training framework led to a more reliable inter-rater uncertainty and models better calibrated than when using a more conventional training approach. This increased ability to encode the inter-rater variability is probably due to the fact that SoftSeg facilitates the propagation of soft labels throughout the training scheme: (1) no binarization of the input labels, (2) a loss function which does not penalize uncertain predictions, and (3) an activation function which does not enforce binary outputs. SoftSeg has a computational advantage over other uncertainty quantification methods such as the ensembling proposed by Ma (2020). An interesting avenue would be to combine SoftSeg and ensemble, thereby creating an ensemble of SoftSeg models to potentially improve calibration and uncertainty representation. Considered with equivalent expertise in this work, future studies could account for the different expertise across raters,

for instance by modulating the training scheme with FiLM layers (Lemay et al., 2021a), or by the use of expertise-aware inferring module (Ji et al., 2021).

5.1.3 When comparing with loss ensembles

Based on the experiment described in Section 4, on comparison with an ensembling approach that aimed at better preserving the inter-rater variability and model uncertainties, we observed that SoftSeg-STAPLE performed better in terms of calibrated predictions and uncertainty quantification metrics. This is an important result because not only it suggests that our method is more computationally efficient, i.e., does not require multiple forward passes through each model at test time, unlike loss ensembles Ma (2020)). It is also better at preserving inter-rater variability with a single forward pass.

5.2 A multifaceted evaluation with model training repetitions

While it is common to select the best model solely based on segmentation accuracy considerations (Commowick et al., 2018; Isensee et al., 2017; Prados et al., 2017), we argue that a more exhaustive evaluation is needed, e.g., by including model calibration and uncertainty assessments. For instance, "Conv-STAPLE" is among the best approaches in terms of segmentation accuracy on the MS dataset (see Figure 5), but is not properly calibrated as it showed an important overconfident gap (see Figure 4). A multifaceted evaluation scheme has the potential to facilitate model acceptance and integration in the clinical routine, which still remains limited (Jungo et al., 2018). Some avenues are discussed below.

5.2.1 The ongoing research around uncertainty and calibration estimation

In the same way that there are numerous segmentation accuracy metrics (Yeghiazaryan and Voiculescu, 2015), there are many ways to assess the model uncertainty and calibration. For instance, recent studies suggested the voxel-wise aleatoric (Wang et al., 2019a) and epistemic (Xia et al., 2020) uncertainties, or the structure-wise uncertainty (Roy et al., 2019), just to name a few uncertainty evaluation methods. The medical image analysis community has only recently started to report measures of model uncertainty and model calibration, and the best practices on how to estimate them are yet to be determined (Abdar et al., 2021; Gal and Ghahramani, 2016; Guo et al., 2017). We acknowledge the exhaustive comparison performed by Jungo et al. (2020) across different uncertainty estimation methods. Their study showed the limits of voxel-wise uncertainty measures in terms of subject-level calibration and recommended the use of subject-wise uncertainty estimates. We followed their recommendations in the present study. Uncertainty was computed directly from the model prediction, rather than from Monte Carlo iterations or deep ensembles, which does not require more computational power and can be measured during inference. Calibration was qualitatively assessed with reliability diagrams and quantitatively analyzed with the ECE as done by Guo et al. (2017). While multiple studies suggest post-hoc strategies to improve calibration (Guo et al., 2017; Kuleshov et al., 2018; Zhang et al., 2020), we suggest a training strategy that directly generates calibrated outputs without the need of extra computation or hyperparameters.

5.2.2 When using the labels through the training pipeline

With the increased number of evaluation criteria often comes the complexity to select a model as the preferred one. The prioritization of one criterion over the others can be application- or user- specific. Alternatively, in this study, we introduce the use of a composite score to represent the overall segmentation accuracy performance by aggregating multiple scores. This approach assumes equal weights for each evaluation criteria, which can be modified depending on the model user's needs. In lesion segmentation tasks, the overall score proposed by Carass et al. (2017) can also be used in lieu of the composite score used here. Another avenue would be to represent the performance across the different criteria using a radar visualization, e.g., used by Placidi et al. (2021).

5.2.3 The importance of training repetitions

Common in studies using machine learning approaches, we observe that experience repetition (e.g., cross-validation, random dataset splittings) is not often performed by studies using deep learning approaches. This is likely due to the long training time required by deep learning model training (often several days). However, our experiments showed that a large variability can be observed across the dataset splittings, especially when data is limited which is often the case in medical settings. For instance, the standard deviation of Dice across the 40 "Conv-STAPLE" models was 12.8%. In the present study, we performed 40 random dataset splittings for the experiments on the SCGM dataset, and 10 on the MS brain dataset. We encourage future deep learning studies to implement experience repetitions in their evaluation scheme.

5.3 Limitations

Since SoftSeg generates softer segmentations, this method is more sensitive to the choice of binarization threshold compared with conventional models where predictions are mostly binary. Hence, thoughtful postprocessing is advised when using SoftSeg. Validation loss progression had different behaviors for conventional and SoftSeg models. While in the previous (Gros et al., 2021a) and current work, all SoftSeg models converged, the Adaptive Wing loss stagnated for the first 25 epochs on the MS lesion models while the Dice loss gradually decreased during all training (see Appendix A).

6. Conclusion

In this study, we evaluated three methods to combine labels from multiple raters using a conventional training framework and SoftSeg, aiming to preserve the inter-rater variability. Our study highlights overconfidence and inter-rater variability underestimation of the conventional framework while SoftSeg models with STAPLE or random sampling were well calibrated and reflected more truthfully the variability due to multiple experts. While fusing annotations using the average encodes the disagreement between experts, predictions were underconfident and the rater uncertainty was overestimated. No consistent observation was made throughout datasets to determine an overall best label fusion method. However, Soft-Seg was systematically superior or equal in terms of segmentation performance and had the best calibration and preservation of the inter-rater variability. SoftSeg showed similar re-

sults to an ensemble method in terms of uncertainty. SoftSeg has the advantage of requiring the training of a single model and uses a single forward pass while the ensemble requires training of multiple models, one per rater, and requires one forward pass per model. While these observations should be confirmed on other datasets, using SoftSeg could potentially be an effective strategy to capture inter-rater variability in segmentation tasks.

Acknowledgments

Ethical Standards

All the data used for this study was de-identified. No IRB was necessary for this work.

Conflicts of Interest

The authors declare that they have no conflicts of interest including financial interests or personal relationships that could impact the reported results in this paper.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson. Deep learning for brain mri segmentation: state of the art and future directions. *Journal of digital imaging*, 30(4):449–459, 2017.

Benoît Audelan, Dimitri Hamzaoui, Sarah Montagne, Raphaële Renard-Penna, and Hervé Delingette. Robust bayesian fusion of continuous segmentation maps. *Medical Image Analysis*, 78:102398, 2022.

Olivier Commowick, Frédéric Cervenansky, and Roxana Ameli. Msseg challenge proceedings: multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure. In *Miccai*, 2016.

Olivier Commowick, Audrey Istace, Michael Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, Pascal Girard, Roxana Ameli, Jean-Christophe Ferré, et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific reports*, 8(1):1–17, 2018.

Olivier Commowick, Michaël Kain, Romain Casey, Roxana Ameli, Jean-Christophe Ferré, Anne Kerbrat, Thomas Tourdias, Frédéric Cervenansky, Sorina Camarasu-Pop, Tristan Glatard, et al. Multiple sclerosis lesions segmentation from multiple experts: The miccai 2016 challenge dataset. *NeuroImage*, 244:118589, 2021.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Refaat E Gabr, Ivan Coronado, Melvin Robinson, Sheeba J Sujit, Sushmita Datta, Xiaojun Sun, William J Allen, Fred D Lublin, Jerry S Wolinsky, and Ponnada A Narayana. Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study. *Multiple Sclerosis Journal*, 26(10):1217–1226, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Charley Gros, Benjamin De Leener, Atef Badji, Josefina Maranzano, Dominique Eden, Sara M Dupont, Jason Talbott, Ren Zhuoquiong, Yaou Liu, Tobias Granberg, et al. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. *Neuroimage*, 184:901–915, 2019.

Charley Gros, Andreanne Lemay, and Julien Cohen-Adad. Softseg: Advantages of soft versus binary training for image segmentation. *Medical image analysis*, 71:102038, 2021a.

Charley Gros, Andreanne Lemay, Olivier Vincent, Lucas Rouhier, Marie-Helene Bourget, Anthime Bucquet, Joseph Paul Cohen, and Julien Cohen-Adad. ivadomed: A medical imaging deep learning toolbox. *Journal of Open Source Software*, 6(58):2868, 2021b. doi: 10.21105/joss.02868. URL https://doi.org/10.21105/joss.02868.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In *International MICCAI Brainlesion Workshop*, pages 287–297. Springer, 2017.

Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–548. Springer, 2019.

Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021.

Alain Jungo, Raphael Meier, Ekin Ermis, Marcela Blatti-Moreno, Evelyn Herrmann, Roland Wiest, and Mauricio Reyes. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 682–690. Springer, 2018.

Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14: 282, 2020.

Eytan Kats, Jacob Goldberger, and Hayit Greenspan. A soft staple algorithm combined with anatomical knowledge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 510–517. Springer, 2019.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR, 2018.

Thomas Robin Langerak, Uulke A van der Heide, Alexis NTJ Kotte, Max A Viergever, Marco Van Vulpen, and Josien PW Pluim. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple). *IEEE transactions on medical imaging*, 29(12):2000–2008, 2010.

Andreanne Lemay, Charley Gros, Olivier Vincent, Yaou Liu, Joseph Paul Cohen, and Julien Cohen-Adad. Benefits of linear conditioning for segmentation using metadata. In *Medical Imaging with Deep Learning*, pages 416–430. PMLR, 2021a.

Andreanne Lemay, Charley Gros, Zhizheng Zhuo, Jie Zhang, Yunyun Duan, Julien Cohen-Adad, and Yaou Liu. Automatic multiclass intramedullary spinal cord tumor segmentation on mri with deep learning. *NeuroImage: Clinical*, 31:102766, 2021b.

Jun Ma. Estimating segmentation uncertainties like radiologists. *MICCAI: QUBIQ*, 2020.

Oskar Maier, Bjoern H Menze, Janina von der Gablentz, Levin Häni, Mattias P Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, et al. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Medical image analysis*, 35:250–269, 2017.

Ross J McGurk, James Bowsher, John A Lee, and Shiva K Das. Combining multiple fdg-pet radiotherapy target segmentation methods to reduce the effect of variable performance of individual segmentation methods. *Medical physics*, 40(4):042501, 2013.

Zahra Mirikharaji, Kumar Abhishek, Saeed Izadi, and Ghassan Hamarneh. D-lema: Deep learning ensembles from multiple annotations-application to skin lesion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1837–1846, 2021.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Brennan Nichyporuk, Justin Szeto, Douglas L Arnold, and Tal Arbel. Optimizing operating points for high performance lesion detection and segmentation using lesion size reweighting. *arXiv preprint arXiv:2107.12978*, 2021.

Giuseppe Placidi, Luigi Cinque, Filippo Mignosi, and Matteo Polsinelli. Multiple sclerosis lesions identification/segmentation in magnetic resonance imaging using ensemble cnn and uncertainty classification. *arXiv preprint arXiv:2108.11791*, 2021.

Ferran Prados, John Ashburner, Claudia Blaiotta, Tom Brosch, Julio Carballido-Gamio, Manuel Jorge Cardoso, Benjamin N Conrad, Esha Datta, Gergely Dávid, Benjamin De Leener, et al. Spinal cord grey matter segmentation challenge. *Neuroimage*, 152: 312–329, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22, 2019.

Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. Understanding expert disagreement in medical data analysis through structured adjudication. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

Or Shwartzman, Harel Gazit, Ilan Shelef, and Tammy Riklin-Raviv. The worrisome impact of an inter-rater bias on neural network training. *arXiv preprint arXiv:1906.11872*, 2019.

Joao Lourenço Silva and Arlindo L Oliveira. Using soft labels to model uncertainty in medical image segmentation. *arXiv preprint arXiv:2109.12622*, 2021.

Olivier Vincent, Charley Gros, and Julien Cohen-Adad. Impact of individual rater style on deep learning uncertainty in medical imaging segmentation. *arXiv preprint arXiv:2105.02197*, 2021.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019a.

Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6971–6981, 2019b.

Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

Yingda Xia, Dong Yang, Zhiding Yu, Fengze Liu, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical Image Analysis*, 65:101766, 2020.

Varduhi Yeghiazaryan and Irina Voiculescu. An overview of current evaluation methods used in medical image segmentation. *Department of Computer Science, University of Oxford*, 2015.

Shuang Yu, Hong-Yu Zhou, Kai Ma, Cheng Bian, Chunyan Chu, Hanruo Liu, and Yefeng Zheng. Difficulty-aware glaucoma classification with multi-rater consensus modeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 741–750. Springer, 2020.

Le Zhang, Ryutaro Tanno, Kevin Bronik, Chen Jin, Parashkev Nachev, Frederik Barkhof, Olga Ciccarelli, and Daniel C Alexander. Learning to segment when experts disagree. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 179–190. Springer, 2020.

Appendix A.

Training process   Table 5 enumerates the training parameters used to train the models from this study. All models were trained for a maximum a 200 epochs with an early stopping of 50 epochs ($\epsilon = 0.001$). Figure 7 illustrates the validation losses progression for on seed from each model. The validation loss reached a plateau for at least the last 50 epochs of the training. The small amount of training data and the high level of difficulty of the MS brain lesion segmentation tasks generated more instability in the validation loss.
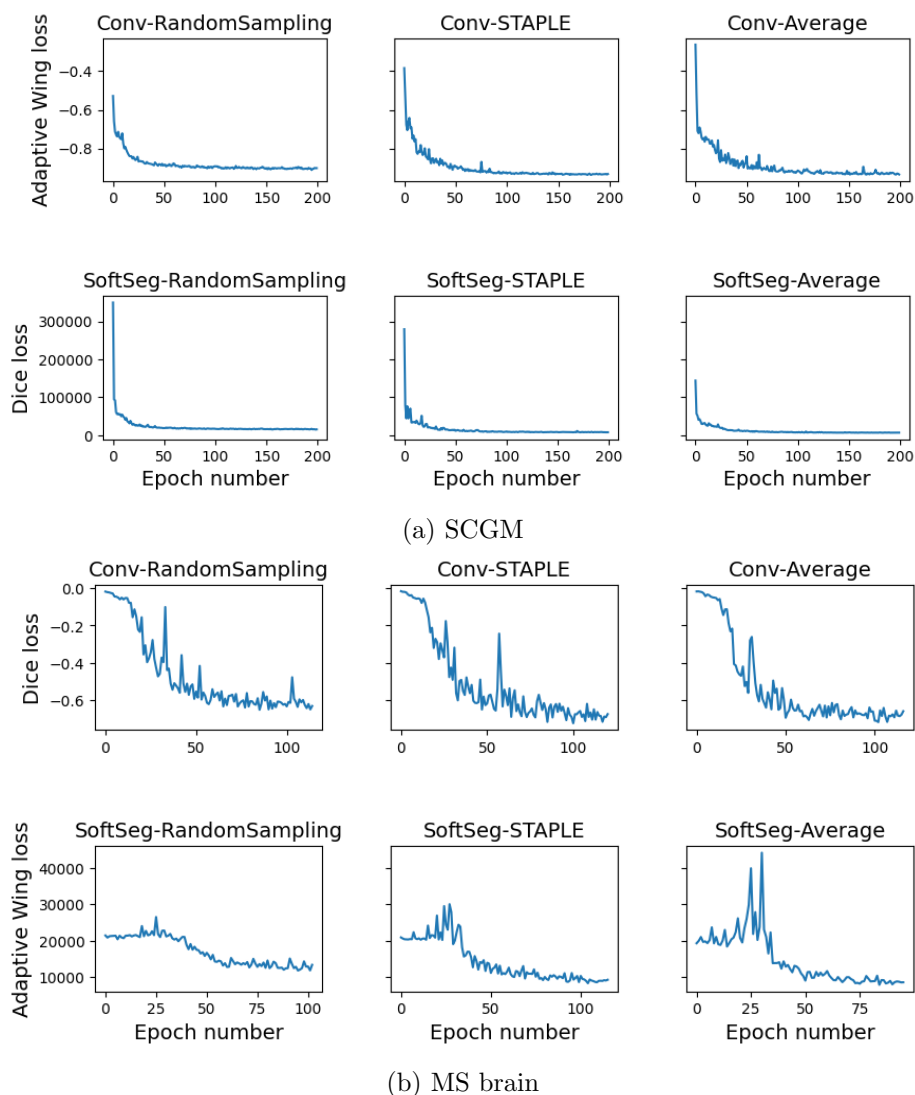
Figure 7: Validation loss progression for a seed from each candidate for the SCGM (a) and MS brain lesions (b) datasets.

Table 5: Training parameters for each dataset. This table was partially extracted from Gros et al. (2021a). Abbreviations: MS: multiple sclerosis; RPI: right-to-left, posterior-to-anterior, inferior-to-superior orientation; SCGM: spinal cord gray matter.

		SCGM dataset	Brain MS lesion dataset
Preprocessing	**Resample**	0.25×0.25 mm^3 (RPI)	1 mm isotropic
	Batch format	2D axial slices	
	Crop	128×128 pixels^2	160 pixels^2
Data augmentation	**Rotation**	± 20 degrees	
	Translation	$\pm 3\%$	
	Scale	$\pm 10\%$	
Batch size		8	24
U-Net depth		3	4
Dropout rate		30%	
Learning Rate	**Initial**	0.001	0.00005
	Scheduler	Cosine Annealing	
Adaptive Wing Loss		$\epsilon = 1; \alpha = 2.1; \Omega = 0.5; \omega = 8$	
Early stopping		Patience: 50 epochs; $\epsilon : 0.001$	
Maximum number of epochs		200	