Titre: Title:	Impact of Soft Segmentation Training on Medical Image Segmentation and Uncertainty Representation
Auteur: Author:	Andréanne Lemay
Date:	2022
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Lemay, A. (2022). Impact of Soft Segmentation Training on Medical Image Segmentation and Uncertainty Representation [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/10259/

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Impact of soft segmentation training on medical image segmentation and uncertainty representation

ANDRÉANNE LEMAY
Institut de génie biomédical

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie biomédical
Mars 2022

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Impact of soft segmentation training on medical image segmentation and uncertainty representation

présenté par **Andréanne LEMAY**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Benjamin DE LEENER, président
Julien COHEN-ADAD, membre et directeur de recherche
Farida CHERIET, membre

ACKNOWLEDGEMENTS

I would like to sincerely thank my research director, Julien Cohen-Adad, for his supervision and supporting me in my decisions throughout my Master's. I came a long way and honed my research skills which I attribute greatly to his mentorship. I also want to warmly thank my colleague and friend, Charley Gros, who has been there for me throughout my Master's degree. I found in him not only an incredible mentor but also a thoughtful and listening friend who helped me navigate through the obstacles of graduate studies. I am grateful for all my lab colleagues, Marie-Hélène Bourget, Olivier Vincent, Lucas Rouhier, Alexandre D'Astous, Gaspard Cereza, Naga Karthik, Uzay Macar, Konstantinos Nasiotis, Anthime Bucquet, Yang Ding, Nick Guenther, Ainsleigh Hill, Joshua Newton, and Alexandru Foias that helped me in my research and enhanced my experience in the lab.

I want to thank all the organizations that helped finance my Master's degree, FRQNT, NSERC, centre UNIQUE, Maxime Fortin and the "Fondation et Alumni de Polytechnique Montréal et Polytechnique Montréal International", and Mitacs (Globalink). Their generous donations allowed me to focus on my research and finish my Master's degree. I also want to highlight the contribution of collaborators such as Joseph Paul Cohen and Dr. Yaou Liu and thank them for sharing their knowledge and helping me grow as a better scientist.

During my Master's, I had the chance to do an internship at the Athinoula A. Martinos Center for Biomedical Imaging, affiliated with Harvard. I want to thank my principal investigator (PI), Jayashree Kalpathy-Cramer, and my supervisor and colleague, Katharina Hoebel, for their invaluable mentorship. They inspired me by being great examples of women role models in the scientific field. I also want to acknowledge my colleagues and friends from the QTIM lab, Benjamin Bearce, Albert Kim, Charles Lu, Syed Rakin Ahmed, Jay Patel, Ken Chang, Christopher Bridge, Praveer Singh, Ikbeom Jang, William Liu, and Mishka Gidwani.

On a more personal note, I want to thank my parents, Jacinthe and Christian, for supporting me morally, intellectually, and financially throughout my studies. Finally, a special thank you to my boyfriend, Francis Granger, my brother, Jonathan, and my friends, Madeleine Fol, Carole-Anne Daunais, Antoine Boudreau-Alexandre, Cloé Boisclair-Laberge, Chloé Fadel, Christina Mahut, and Philippe Miranda-Jean, for encouraging me and being there for me.

RÉSUMÉ

Avec l'essor de l'apprentissage profond, une quantité croissante de modèles sont développés pour le domaine médical afin d'automatiser les tâches fastidieuses et de réduire les erreurs médicales causées par l'homme. Cependant, en raison de leur impact potentiel sur la vie et la santé humaine, des préoccupations éthiques grandissantes font jour concernant la fiabilité et la transparence des modèles d'apprentissage profond. Une solution potentielle à ces problèmes est d'entraîner des modèles générant des prédictions calibrées avec une représentation fidèle de l'incertitude. De cette manière, les prédictions les plus susceptibles d'être incorrectes ou de donner lieu à des désaccords entre les experts peuvent être isolées et corrigées. Cependant, les réseaux neuronaux de segmentation modernes sont généralement trop confiants, c'est-à-dire qu'ils expriment une grande certitude même pour les prédictions erronées, et ne tiennent pas compte de considérations importantes en imagerie médicale telles que l'effet de volume partiel, la variabilité inter-expert ou la représentation de l'incertitude en raison de la faible qualité des images ou du manque de données. Ceci est dû à la nature binaire de la segmentation qui est considérée comme une tâche de classification où chaque voxel se voit attribuer une valeur de 0 ou 1.

Dans la première partie de ce travail, nous proposons une méthode appelée SoftSeg qui traite la segmentation comme une tâche de régression afin d'encourager la représentation des informations sur les volumes partiels, la variabilité inter-expert et la représentation de l'incertitude. L'approche de segmentation non-binaire vise à réduire la confiance excessive du modèle. Trois caractéristiques principales définissent SoftSeg par rapport aux modèles de segmentation conventionnels : (i) la préservation du caractère non-binaire, i.e., entre 0 et 1, des segmentations utilisées pour l'entraînement après le traitement et l'augmentation des données, (ii) une fonction d'activation finale linéaire normalisée pour éviter la perte d'information contrairement aux fonctions sigmoïde ou softmax non linéaires, et (iii) l'utilisation d'une fonction de perte de régression plutôt que de classification comme Dice ou d'entropie croisée. Nous avons exploré ces nouvelles fonctionnalités et évalué l'impact de chacune d'entre elles lors d'une étude d'ablation. La combinaison de ces trois nouvelles caractéristiques a permis d'obtenir de meilleures performances de segmentation sur trois ensembles de données de segmentation publiquement accessibles : matière grise de la moelle épinière, lésions de sclérose en plaques du cerveau et tumeur du cerveau.

Dans un deuxième article, trois méthodes de fusion d'annotations d'expert, soit STAPLE, moyennage et l'échantillonnage aléatoire (c'est-à-dire sans fusion), pairées à SoftSeg ou à

un entraînement conventionnel, ont été comparées. Les approches ont été étudiées sur deux ensembles de données avec respectivement quatre et sept annotations d'évaluateurs pour chaque image : segmentation de la matière grise et blanche de la moelle épinière et lésions de la sclérose en plaques du cerveau. La préservation de l'incertitude due au désaccord entre les évaluateurs, la calibration des prédictions, la qualité visuelle et les performances de segmentation ont été évaluées. Bien qu'il n'y ait pas eu de consensus entre les ensembles de données en ce qui concerne la meilleure méthode de fusion des annotations d'expert, les résultats étaient équivoques en ce qui concerne le type d'entraînement. Nos résultats indiquent que SoftSeg a produit une prédiction avec une meilleure calibration ainsi qu'une préservation de la variabilité inter-expert accrue, et ce, avec une performance de segmentation améliorée, ou minimalement équivalente.

Toutes les approches étudiées dans ce travail ont été répétées 10 à 40 fois avec des séparations aléatoires des données pour éviter un biais au niveau des données de test et garantir des différences statistiques (valeur $p < 0{,}05$). Toutes les recherches effectuées dans ce projet ont été réalisées et rendues accessibles via le projet en libre accès ivadomed (`https://ivadomed.org`).

ABSTRACT

With the rise of deep learning, an increasing amount of models are being developed and researched for the medical field to automate tedious tasks and mitigate medical errors. However, due to the sensitive nature of medical tasks and their impact patient's health, increasing ethical concerns are arising regarding reliability and transparency of deep learning models. A potential avenue to address these issues is to have calibrated predictions with truthful uncertainty representation. Reliable uncertainty representation help identify predictions that are prone to model failure or inter-rater disagreement. However, modern segmentation neural networks are usually overconfident, i.e., express a high certainty even for incorrect predictions, and disregard important considerations in medical imaging such as partial volume effect, inter-rater variability, or uncertainty representation due to low image quality or lack of data. This is partly due to the inherent binary nature of segmentation that is considered a classification task where each voxel is attributed a value of 0 or 1.

In the first part of this work, we propose a method named SoftSeg that treats segmentation as a regression task to encourage the representation of partial volume information, inter-rater variability, and uncertainty. The soft segmentation approach aims at mitigating overconfidence. Three main features characterize SoftSeg compared with the conventional segmentation models: (i) preservation of soft input labels following data processing and augmentation, (ii) a normalized linear final activation to avoid information loss instead of the non-linear sigmoid or softmax, and (iii) the use of a regression loss function rather than the classification Dice or cross-entropy loss. We explored these new features and evaluated the impact of each feature through an ablation study. The combination of these three new features resulted in better segmentation performance on three publicly available segmentation datasets: spinal cord gray matter, brain multiple sclerosis lesions, and brain tumor.

In a second article, three label fusion methods, STAPLE, average, and random sampling (no fusion), paired with SoftSeg or a conventional training framework, were compared. The approaches were studied on two datasets with respectively four and seven rater annotations for each image: spinal cord gray and white matter segmentation and brain multiple sclerosis lesions. The uncertainty preservation due to inter-rater disagreement, the calibration of predictions, the visual predictions, and the segmentation performance were evaluated. While there was no consensus between datasets in terms of the best label fusion method, results were equivocal regarding the training framework. Our results indicate that SoftSeg yielded prediction with better calibration and inter-rater variability preservation with higher, or

minimally equivalent, segmentation performance.

All the approaches studied in this work were repeated 10 to 40 times with varying random seeds to avoid data splitting biases and ensure statistical differences (p-value < 0.05). All the research done in this project was developed and made accessible via the open-source project ivadomed (`https://ivadomed.org`).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

AVD	Absolute volume difference
BraTS	Brain tumor segmentation
CNN	Convolutional neural network
DL	Deep learning
ECE	Expected calibration error
GT	Ground truth
H	Entropy
LFDR	Lesion false detection rate
LTPR	Lesion true positive rate
MAE	Mean absolute error
MRI	Magnetic resonance imaging
MS	Multiple sclerosis
MSE	Mean squared error
PVE	Partial volume effect
ReLU	Rectified linear unit
RPI	Right-to-left, posterior-to-anterior, inferior-to-superior
RVD	Relative volume difference
SCGM	Spinal cord gray matter
STAPLE	Simultaneous truth and performance level estimation

CHAPTER 1 INTRODUCTION

1.1 Problem statement

Deep learning (DL) in healthcare has witnessed unprecedented attention and progression in the past years [1–3] suggesting its potential to automate tedious medical tasks or reduce human-related errors [4, 5]. However, DL faces numerous ethical concerns due to, among others, its bias and lack of transparency [6]. Indeed, DL is often criticized for being a black box as it is hard to isolate the reasons motivating its final prediction the way humans would do it. Moreover, DL models carry the characteristics of the labels, i.e., bias, it was trained on, which is problematic as medical tasks are prone to inter-rater disagreement [7–9]. Using multiple expert annotations is a common method to mitigate the models' bias. However, there is no clear consensus on the best way to combine expert labels to represent the inter-rater variability. To ensure safe integration of DL models into clinical settings and gain the public's trust, these concerns must be addressed along ensuring state-of-the-art performance. Proper uncertainty representation highlights predictions that are more likely to be misclassified or challenging to experts. Flagging uncertain predictions limit the frequency of silent-failures and makes the model more transparent since it indicates to the user which output are prone to error.

Image segmentation, notably of pathological tissues, allows morphometric quantification of structures of interest helping with longitudinal monitoring or treatment planning [10, 11]. While segmentation is mostly treated as a classification task where each voxel of an image is assigned a binary value to highlight a given structure, this approach does not truthfully reflect the complexity of the task. Conventional DL segmentation models overlook partial volume effect, inter-rater variability, or ill-defined boundaries by outputting overconfident predictions [12] that are mostly binary. Reliable DL models should yield calibrated output truthfully characterizing these phenomena with proper uncertainty representation.

1.2 Research objectives and hypothesis

This work aims at:

1. Proposing a new approach to generate predictions reflecting partial volume effect, inter-rater variability, and ill-defined boundaries through proper uncertainty representation.

2. Validating the output quality of this new method on medical datasets in terms of

segmentation performance, calibration, and uncertainty representation.

3. Implementing and giving open access to the research tools and results.

Ultimately, these research objectives intend to propose and evaluate a method to have more reliable and transparent models while yielding better, or minimally equivalent, segmentation performances to the conventional training framework. A secondary objective was to implement and give open access to this new approach and the experiments to validate it.

We hypothesize that limiting information loss during training by modifying the data processing, the final activation layer, and the loss function will improve segmentation performance and uncertainty representation. Tackling segmentation as a regression task should help yield valuable information for decision-making derived from partial volume effect, inter-rater variability, or limited image quality.

1.3 Thesis outline

This work starts with an overview of the background and related works to this Master's project in Chapter 2. The purpose of each article presented in this work and their role in addressing the research objectives are described in Chapter 3. Chapter 4 presents SoftSeg [13] which fulfills the first research objective (see Section 1.2) of this Master's project. SoftSeg treats segmentation as a regression task and takes advantage of soft labels by (i) removing all binarization steps during preprocessing or data augmentation, (ii) using a normalized ReLU as final activation function, and (iii) training the model using a regression loss function. To accomplish the second research objective (see Section 1.2), Chapter 5 focuses on assessing inter-rater preservation and output calibration of SoftSeg while comparing label fusion methods [14]. A discussion on the results presented in the two articles and their clinical usefulness is presented in Chapter 6. Finally, Chapter 7 concludes and gives recommendations based on this research project.

CHAPTER 2 LITERATURE REVIEW

2.1 Deep learning segmentation models

Medical image segmentation consists of the delineation of an anatomical structure and helps with diagnosis, disease monitoring, and treatment planning [10]. Each voxel of an image is associated with a class representing one or multiple structures of interest. Segmentation on 3D medical images, often composed of hundreds of 2D slices, is tedious and time-consuming. DL segmentation models have been increasingly studied [15–17] to help automate this laborious process. Convolutional neural networks (CNN), especially U-Net [18], stand out as the most popular DL approach to medical segmentation [19–21] due to its state-of-the-art performances [22–24]. CNN extracts features from the image using convolution to identify shapes and textures characterizing the objects of interest. U-Net (Figure 2.1) has an encoder generate, from the extracted features, activation maps that are gradually losing spatial resolution to the profit of high-level abstraction. The second part of the U-Net is the decoder block that generates a high-resolution output based on the features extracted during the encoding phase. U-Net has the particularity to simultaneously consider features at different scales, i.e., high-resolution details and feature-rich semantics, because of its skip connections between the encoder and decoder blocks.

2.1.1 Label processing

Most segmentation models are trained through supervised learning using binary expert annotations. Before training DL models, preprocessing, including resampling to a common resolution, and data augmentation, including affine or elastic transformations to avoid overfitting [25], are applied to the training images and associated GT. These operations require interpolation. To preserve the binary nature of segmentations, the interpolation chosen for GT is often nearest neighbor [26–28] or another interpolation technique followed by a binarization. Restricting values to 0 or 1, while it's optimal when using Dice or cross-entropy losses (see Section 2.1.2), can cause partial volume information loss at tissue boundary.

2.1.2 Loss function

Various loss functions to train segmentation models have been introduced, but the two of the most commonly used in the literature are the cross-entropy [29–33] (Equation 2.1) and soft Dice loss [11, 17, 30–32, 34–39] (Equation 2.2). Both losses are designed for classification

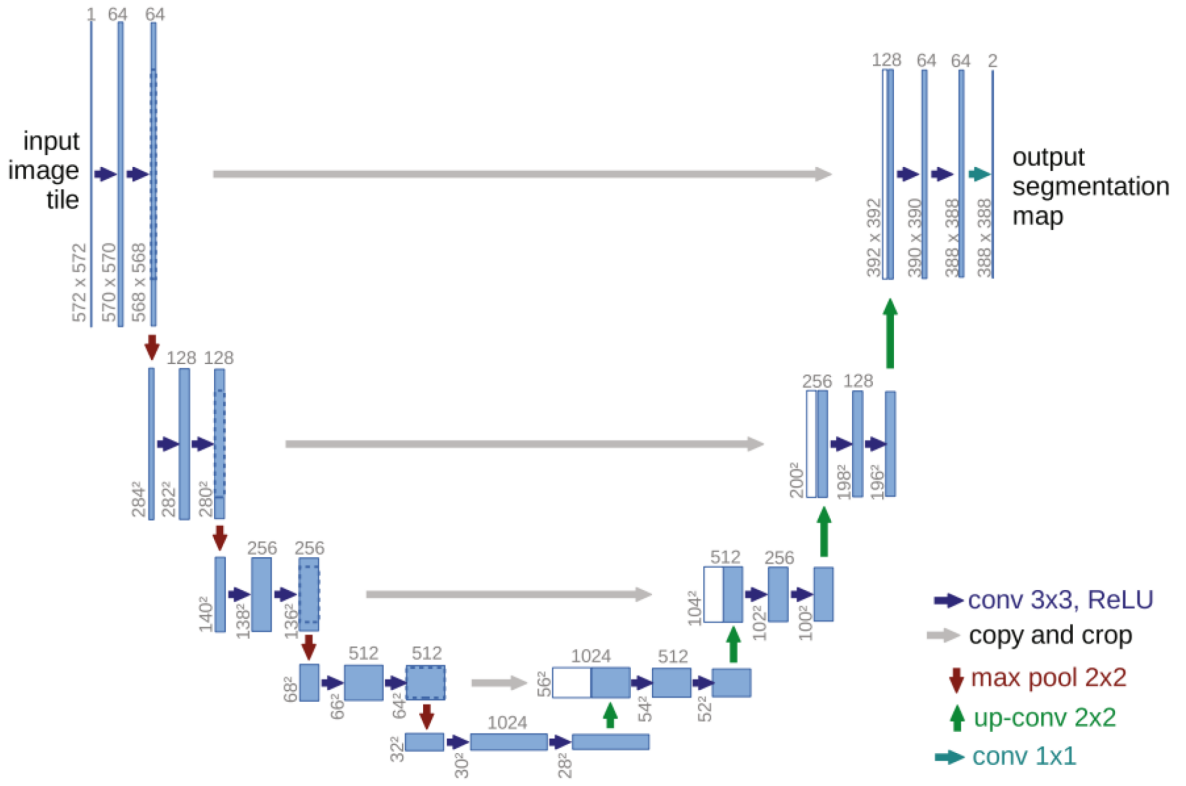


Figure 2.1 U-Net architecture.

An image is input in the neural network (left side) and passes through successive convolutional blocks that extract image features at different scales (first half of the network). The extracted features are then upsampled to restore the original resolution and output a high-resolution segmentation (right side). The skip connections (gray arrows) allow the recovery of high-resolution details during the feature decoding. This figure was extracted from [18].

with binary GT. However, they are not optimal for soft labels. For instance, a soft GT with a value of 0.5 and a prediction of 0.5 will not lead to the maximal loss scores, i.e., 0 for cross-entropy and -1 for Dice loss, even though both values are equal. However, a regression loss such as mean squared error (MSE) would generate the minimal loss value of 0.

$$CE_{loss} = -\sum_{c=0}^{C}\sum_{i=0}^{N} y_i^c log(\hat{y}_i^c) \tag{2.1}$$

$$Dice_{loss} = -\frac{2\sum_{i=0}^{N} y_i \hat{y}_i}{\sum_{i=0}^{N} y_i + \sum_{i=0}^{N} \hat{y}_i} \tag{2.2}$$

where N is the total number of voxels in the image, y_i is the ground truth (GT), $\hat{y}_i$ is the prediction, and C is the total number of classes.

DL models trained with Dice loss yield predictions with sharp edges [30,40,41] hindering non-binary predictions. These almost binary predictions were shown to lead to volumetric bias due to overestimation of the structures volume [42]. Moreover, Dice loss predictions do not reflect the model's uncertainty and are overconfident, even for on misclassified regions [21]. While cross-entropy is associated with better-calibrated predictions [21], and softer edges [30], it is sensitive to class imbalance, which can be mitigated by weighting under-represented classes. However, Dice loss outperforms cross-entropy loss on many medical segmentation tasks [21, 43, 44] even when leveraging balancing strategies.

2.1.3 Final activation

After the model prediction for a segmentation task, a final activation function is applied to the output to constrain the values between 0 and 1. For multi-class segmentation, a softmax (Equation 2.3) is usually applied, while the sigmoid (Equation 2.4), is usually the choice for a binary segmentation. Due to exponential functions, these activations are non-linear and plateau to 0 or 1 for extreme values (see Figure 2.2), which can lead to information loss from the raw output. Sigmoid and softmax were designed for classification as most raw outputs will result in a value near 0 or 1. Other final activation function has been explored such a ReLU [45] for classification [46] leading to value from 0 to infinity. However, few work suggest or study other alternative to the popular sigmoid and softmax functions.

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{2.3}$$

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{2.4}$$

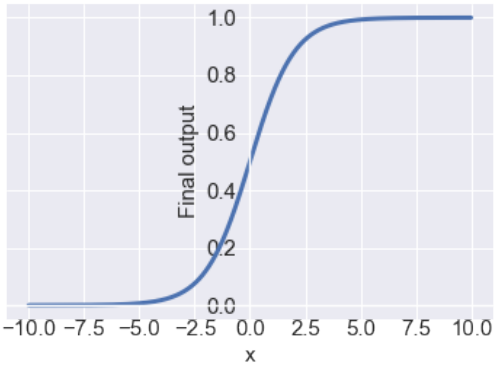
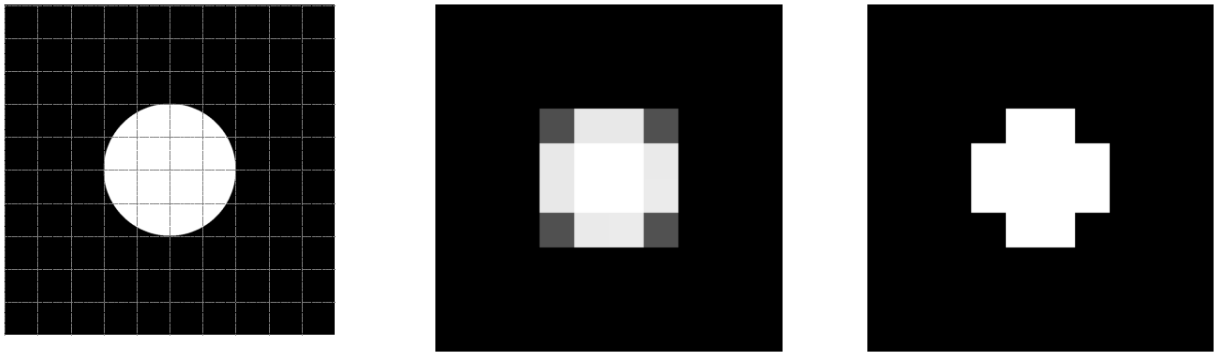


Figure 2.2 Sigmoid function.

2.2 Partial volume effect

Medical imaging aims at capturing anatomical structures, but the imaging of these structures is limited by the native spatial resolution of the acquisition. Small structures, such as MS lesions, are susceptible to partial volume effect (PVE), but higher resolution mitigates this effect. When multiple tissue types are present in a voxel, the resulting value will, proportionally to their presence, represent the properties of the tissues composing the voxel of each tissue [47]. Figure 2.3 illustrated the PVE. Ignoring the partial volume effect by using binary segmentation values causes error in quantitative volumetric measurements [47]. The binarization causes information loss at the object boundary and does not preserve the intensity value sum of the image as seen on Figure 2.3. Few deep learning models take into account PVE [48–50], however, to reach perfect classification quantifying the presence of overlapping tissues in a voxel is necessary [49].



(a) Object to image. Area: 125629 pixels.

(b) Imaged object with coarse resolution. Area: 125629 pixels.

(c) Binarized imaged object with coarse resolution. Area: 120000 pixels.

Figure 2.3 Partial volume effect.

(a) represents the object to image and the dashed lines indicates the resolution of the imaging. (b) illustrates the partial volume effect when the spatial resolution is coarse relative to the size of the object. (c) highlights the precision loss at the object boundary when using a binarized representation. The area is measured by summing the pixels' intensity. The area of the binarized object (c) is different from the area from the original object (a).

2.3 Uncertainty and Calibration

With growing concerns about the lack of transparency of DL models, rising interest has been given to uncertainty in artificial intelligence [9, 12, 51–53]. Reliable uncertainty highlights

predictions prone to model failure. Various factors cause uncertainty, such as lack of training data (epistemic uncertainty), poor image quality (aleatoric uncertainty), or inter-rater variability. Uncertainty can be obtained via Bayesian methods, which generate a model likelihood or through approximates of Bayesian [51]. Examples of Bayesian approximation are uncertainty derived from Monte Carlo iterations [54] or ensembles [55]. Multiple predictions are generated and compared to determine the level of certainty of the model. However, these methods are computationally expensive. While the raw predictions of a model is arguably an indication of model uncertainty [52], numerous authors [9, 56–58] consider it as a baseline uncertainty metric. In this work, we focused on uncertainty directly generated by the model as it is the easiest to obtain. The predictions encode the voxel-wise uncertainty while the predictive entropy can represent the overall uncertainty of an image [9, 56].

Calibration can be used as a surrogate to uncertainty reliability [56, 59]. A calibrated model will yield predictions corresponding to the actual probability of this outcome. For instance, a prediction of 0.8 should correspond to the class 80% of the time. Guo et al. [12] demonstrated that modern neural networks are overconfident and showed that temperature scaling as a post-processing step could improve the calibration. However, modifying the processing using and yielding soft labels has the potential to make the model inherently more calibrated without the need of extra processing [60–62] (see Section 2.5).

2.4 Inter-rater variability and label fusion

Medical imaging tasks such as segmentation or classification are prone to inter-rater variability. The causes of this disagreement arise from experts' experience, and training, poor image quality, or guideline clarity [7, 63]. A common method to address expert disagreement is to use labels from more than one expert to reduce the bias associated with a single rater. Multiple strategies exist to combine multiple expert labels. Figure 2.4 summarizes the most popular label fusion methods.

2.4.1 Hard fusion

The most common ones are the hard fusion methods and include STAPLE (simultaneous truth and performance level estimation) [64], majority voting, intersection, and union [9]. STAPLE is an expectation-maximization algorithm that generates a consensus segmentation based on all annotations. Majority voting looks at each voxel individually and associates a label based on what most raters chose. Intersection generates labels that include all the voxels segmented by the raters, while union includes only the voxels that all the raters annotated.

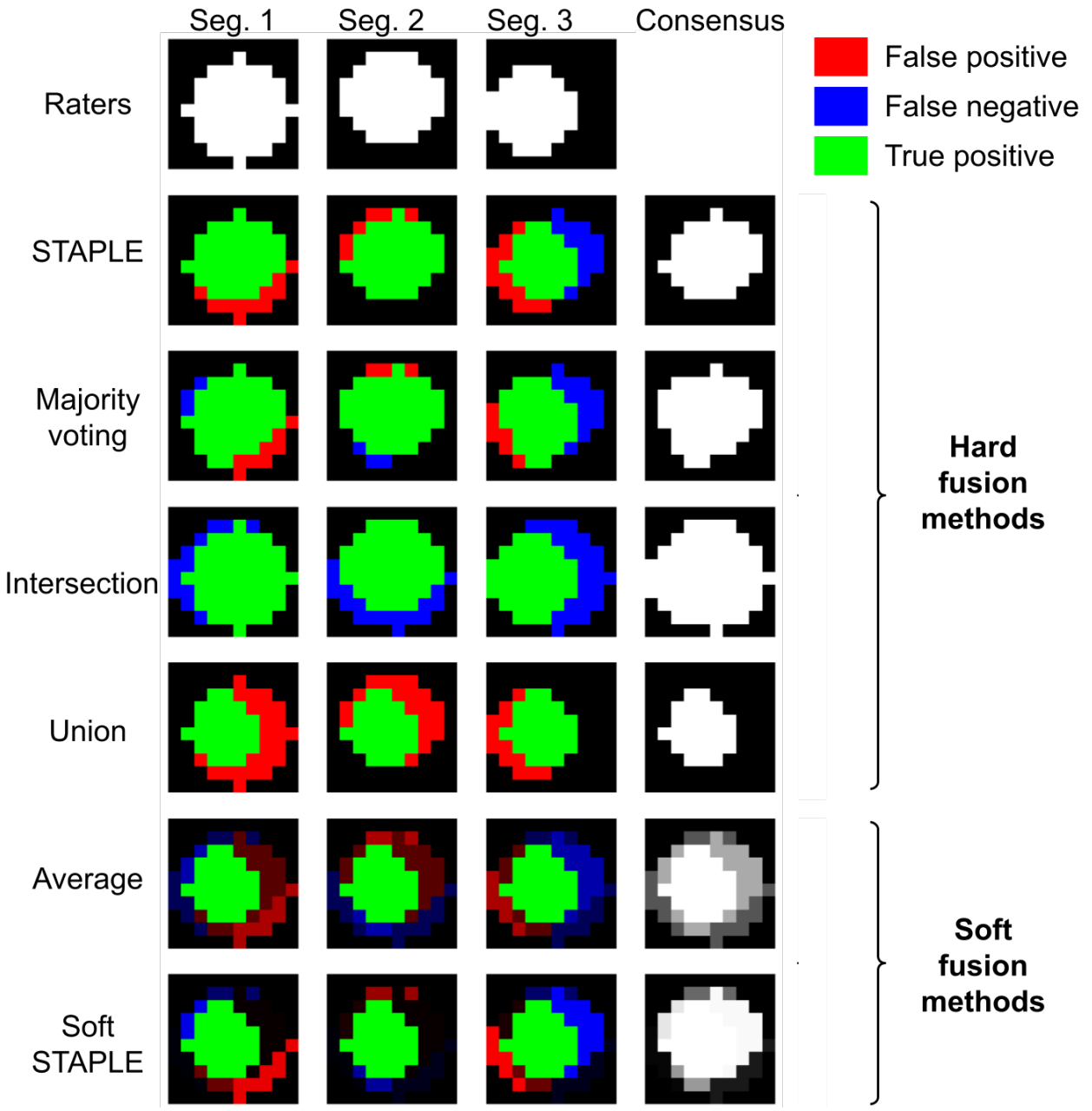


Figure 2.4 Summary of hard and soft label fusion methods.

The first three columns represent segmentation from distinct raters while the last column contains the consensus label. Rows 2 to 7 each illustrate one label fusion method. Red, blue, and green repesent the false positive, false negative, and true positive voxels compared to the consensus labels, respectively. Abbreviations: Seg.: Segmentation.

2.4.2 Soft fusion

Soft fusion methods are less popular due to the binary nature of segmentation. Kats et al. proposed a soft version of STAPLE, soft-STAPLE, as additional anatomical information can be stored with intermediate values [65]. Kats et al. improved the segmentation performance with soft-STAPLE as GT for MS brain segmentation. Averaging the expert labels is another more straightforward soft fusion label method.

2.4.3 No fusion

Finally, multiple works focused on methods to leverage multiple expert labels without fusing them [8, 9, 63]. Jungo et al. [9] and Jensen et al. [8] both studied random sampling and reported advantages to this method compared with hard fusion methods. One annotation is randomly selected at each epoch during training to expose the model to all the labels. Jungo et al. observed that this method represented the more truthfully the inter-rater variability for the synthetic data and brain tumor patients with high Dice scores [9]. Jensen et al. noted better calibration when random sampling labels for a classification task [8]. Other methods than random sampling exist, such as deep ensembles where each model is trained with annotations from one rater [63].

2.5 Label softening

Previous work demonstrated the positive impacts of performing label smoothing on DL models [60–62]. Muller et al. [60] and Pham et al. [61] focus on label softening of classification labels which reduced model overconfidence and improved generalization. Li et al. [62] propose a smoothing method for segmentation labels to account for the uncertainty at the boundaries. Their approach, which combines the original hard label and the softened label, improved the Dice score on brain and optical coherence tomography segmentation tasks.

CHAPTER 3 METHODOLOGY

This section will overview the role of each article presented in this work in achieving the research objectives described in Section 1.2.

3.1 Objective 1: Proposition of a new non-binary approach

The first objective of this Master's thesis was to propose a new method that would take into account phenomena that are disregarded by the conventional binary training framework. This approach should have a better uncertainty representation, reflect inter-rater variability, and account for PVE and ill-defined boundaries due to poor image quality. Chapter 4 aims at answering this objective. In the article "SoftSeg: Advantages of soft versus binary training for image segmentation", we targeted and explored different features from the conventional training pipeline that caused binary output and were responsible for the potential information loss [13]. The features targeted were the label processing, the final activation, and the loss function. The conventional approach is characterized by the use of binary labels, sigmoid as final activation (or softmax for multiclass), and Dice loss function. In this study, we modified these three parameters, which we called the SoftSeg approach, and compared the results to a conventionally trained model. SoftSeg uses soft labels derived from the data processing during resampling or augmentation, normalized ReLU as final activation, and Adaptive Wing loss as regression loss function. Moreover, we did an ablation study starting from SoftSeg and setting each of these three parameters, one at a time, to the conventional option to highlight the effect of removing each SoftSeg characteristic.

3.1.1 Label processing

Since classification losses require a categorical GT, labels are usually binarized after the data processing or data augmentation steps (see Section 2.1.1). However, resampling the data or applying affine transformations such as rotation and scaling for data augmentation requires an interpolation step. Using nearest neighbor interpolation or a higher order interpolation followed by a binarization step suppresses valuable partial volume information [66, 67]. To mitigate the information loss and encourage the output of soft labels, we propose to use interpolation of order one or more during the label processing and augmentation, which generates non-binary GT. Moreover, this opens the door to inherently soft GTs, e.g., derived from the average of multiple expert labels.

3.1.2 Final activation

Sigmoid and softmax are the most common activation function for the last layer on DL models (see Section 2.1.3) and convert all extreme values to near 0 or 1 values. Very different raw output values such as 10 and 100 will lead to a difference of only 5e-5 when a sigmoid function is applied. To take into account the different values learned by the model, the final activation was changed to a normalized version of ReLU to ensure the linearity of the positive raw output values. The output is normalized to ensure values inclusively between 0 and 1.

3.1.3 Loss function

Classification losses such as cross-entropy or Dice cannot be paired with soft labels as they hinder the prediction of ambivalent values (see Section 2.1.2). Training models with a regression loss would enable the use of soft labels derived from data resampling, data augmentation, or labels from multiple raters and would bolster the generation of values from 0 to 1. Adaptive Wing loss was introduced by [68] to train regression heatmap for facial landmark localization (see Equation 3.1). This regression loss function was chosen for SoftSeg due to its state-of-the-art performance for this task and its robustness to class imbalance [68]. More traditional regression loss such as mean squared error (MSE), the weight of foreground and background voxels is the same. However, in many medical segmentation tasks, most voxels are background. The Adaptive Wing loss gives more weight to foreground voxels to mitigate issues related to class imbalance. Another strength of this loss is that the loss is continuous and smooth around $|y-\hat{y}| = \theta$. The parameter θ serves as threshold to switch between linear and non-linear section of the loss. The exponential term $\alpha - y$ shapes the loss to y and insure smoothness. The parameters ω and ϵ control the importance given to small errors. For more details on optimal hyperparameters and their behavior, refer to the original work [68].

$$\text{AdaptiveWingLoss} = \begin{cases} \omega ln(1 + |\frac{y-\hat{y}}{\epsilon}|^{\alpha-y}), & \text{if } |(y-\hat{y})| < \theta \\ A|y-\hat{y}| - C, & \text{else} \end{cases} \tag{3.1}$$

where $A = \frac{\omega(\alpha-y)(\frac{\theta}{\epsilon})^{\alpha-y-1}}{\epsilon(1+(\frac{\theta}{\epsilon})^{\alpha-y})}$ and $C = \theta A - \omega ln(1 + (\frac{\theta}{\epsilon})^{\alpha-y})$.

3.2 Objective 2: Validate the output quality of SoftSeg

The second objective of this project was to assess the quality of the predictions generated by SoftSeg models. This new approach was evaluated on three publicly-available datasets: spinal

cord gray (and white) matter (SCGM) challenge [69], MS brain lesion challenge [70], and the brain tumor segmentation challenge 2019 (BraTS 2019). While segmentation performance is crucial, other considerations such as reliability and transparency through faithful uncertainty representation are essential to integrating DL models in clinical settings. In the SCGM and MS brain lesion datasets, four and seven expert annotations were available, respectively, for each image, making this kind of analysis possible. In both articles presented in this work, segmentation performance was evaluated through common classification metrics. In the second article, "Label fusion and training methods for reliable representation of inter-rater uncertainty", the preservation of the inter-rater variability and calibration were quantified for SoftSeg models and different label fusion methods [14].

3.2.1 Segmentation performance

Most classification metrics require binary predictions. Hence, the segmentation performance was separated into the evaluation of the soft predictions and the binarized predictions. The soft predictions were compared to the GT using the Brier score, i.e., MSE of the prediction maps. For the binarized predictions, the output quality was assessed with the Dice score, precision, recall, absolute volume difference (AVD), and relative volume difference (RVD). Moreover, for the MS lesions segmentation, the false detection rate (LFDR) and lesion true positive rate (LTPR) were computed.

3.2.2 Inter-rater variability preservation

The inter-rater variability preservation was studied on the two datasets containing labels generated by multiple raters. The GT derived from the average of all annotations was considered the gold standard for inter-rater disagreement representation. The image-wise uncertainty was measured by computing the entropy of the predicted or GT average segmentation. We expect that the entropy, i.e., uncertainty proxy, of the prediction and inter-rater disagreement would match. We measured the uncertainty correspondence with the mean absolute error. Since this metric does not consider spatial coherence of the GT average and predictions, we also calculate the Brier score, which essentially computes the distance between both probability maps.

3.2.3 Calibration

In the medical field, errors can have an impact on diagnosis or medical decisions. Identifying model misclassification can help mitigate silent failures. Calibration is a reliability metric

since it describes how truthful the prediction is in terms of classification error. With calibrated models, regions with low-probability voxels indicate structures or lesions that are likely misclassified or prone to inter-rater disagreement. Calibration was assessed using reliability diagrams and expected calibration error (see Section 5.3.3).

3.3 Objective 3: Implement and give open-source access to SoftSeg

The final objective was to document and make the research accessible. Alongside the research, ivadomed [35], a medical image analysis framework powered by DL, was developed to facilitate the reproducibility of the results. This library includes preprocessing, data loading, training, post-processing, data augmentation, and data analysis tools to facilitate the use of DL in medical image segmentation. I collaborated on this open-source project with other members of the NeuroPoly lab. All the tools and code to reproduce and analyze the results presented in this work are available at `https://ivadomed.org`.

CHAPTER 4 ARTICLE 1: SOFTSEG: ADVANTAGES OF SOFT VERSUS BINARY TRAINING FOR IMAGE SEGMENTATION

Published in Medical Image Analysis Vol. 71, p. 102038, 2021

C. Gros, A. Lemay, and J. Cohen-Adad, "Softseg: Advantages of soft versus binary training for image segmentation," *Medical Image Analysis*, vol. 71, p. 102038, 2021.

Title SoftSeg: Advantages of soft versus binary training for image segmentation

Authors Charley Gros* [1,2], Andreanne Lemay*[1,2], Julien Cohen-Adad[1,2,3]
* These authors equally contributed to this work.

Affiliations
[1] NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada
[2] Mila, Quebec AI Institute, Montreal, Qc, Canada
[3] Functional Neuroimaging Unit, CRIUGM, Université de Montréal, Montreal, QC, Canada

Corresponding author
Julien Cohen-Adad
Dept. Genie Electrique, L5610
Ecole Polytechnique
2900 Edouard-Montpetit Bld
Montreal, QC, H3T 1J4, Canada
Phone: 514 340 5121 (office: 2264);
e-mail: jcohen@polymtl.ca

Abbreviations
BraTS: brain tumor segmentation
GT: ground truth
MS: multiple sclerosis
MSE: mean squared error

PVE: partial volume effect
ReLU: rectified linear function
RPI: right-to-left, posterior-to-anterior, inferior-to-superior
SCGM: spinal cord gray matter

4.1 Abstract

Most image segmentation algorithms are trained on binary masks formulated as a classification task per pixel. However, in applications such as medical imaging, this "black-and-white" approach is too constraining because the contrast between two tissues is often ill-defined, i.e., the voxels located on objects' edges contain a mixture of tissues (a partial volume effect). Consequently, assigning a single "hard" label can result in a detrimental approximation. Instead, a soft prediction containing non-binary values would overcome that limitation. In this study, we introduce SoftSeg, a deep learning training approach that takes advantage of soft ground truth labels, and is not bound to binary predictions. SoftSeg aims at solving a regression instead of a classification problem. This is achieved by using (i) no binarization after preprocessing and data augmentation, (ii) a normalized ReLU final activation layer (instead of sigmoid), and (iii) a regression loss function (instead of the traditional Dice loss). We assess the impact of these three features on three open-source MRI segmentation datasets from the spinal cord gray matter, the multiple sclerosis brain lesion, and the multimodal brain tumor segmentation challenges. Across multiple cross-validation iterations, SoftSeg outperformed the conventional approach, leading to an increase in Dice score of 2.0% on the gray matter dataset ($p = 0.001$), 3.3% for the brain lesions, and 6.5% for the brain tumors. SoftSeg produces consistent soft predictions at a tissues' interfaces and shows an increased sensitivity for small objects (e.g., multiple sclerosis lesions). The richness of soft labels could represent the inter-expert variability, the partial volume effect, and complement the model uncertainty estimation, which is typically unclear with binary predictions. The developed training pipeline can easily be incorporated into most of the existing deep learning architectures. It is already implemented in the freely-available deep learning toolbox ivadomed (`https://ivadomed.org`).

Keywords Segmentation, Deep learning, Soft training, Partial volume effect, Label smoothing, Soft mask

4.2 Introduction

Medical image analysis is at a turning point as a growing number of clinical studies are fully embracing automated processing, thanks to the recent ground-breaking performances of deep learning [71–73]. A popular medical application of deep learning is image segmentation, whereby voxels are assigned a label (e.g., 1 if pertaining to the tissue of interest, 0 otherwise). This binary approach to tissue classification is limited in that it does not allow the model to exploit the rich information present in the expert annotation or in the input image. This richness could take the form of inter-expert representation (in case a ground truth is created by several experts) [74], level of uncertainty (e.g., a ground truth could take the value 0.5 instead of 1, if the expert is unsure a voxel belongs to a lesion) [75], pathology severity (e.g., the signal intensity in multiple sclerosis lesions is associated with tissue damage [76]) , or partial volume effect (PVE) [77]. PVE is characterized by the mixing of signals coming from different tissue types, and usually happens at their interfaces. For example, if tissue A has the intensity 50 on a MRI scan and tissue B the intensity 100, voxels at their interface exhibit values between 50 and 100, depending on the volume fraction occupied by each tissue. PVE is a well-known problem in computer vision, and it can notably be handled by Gaussian mixture modeling to estimate the true fraction of underlying tissue signals [78, 79] or integrated into classical probabilistic Markov Random Fields [80, 81] or fuzzy sets based [82] segmentation methods. However, PVE is rarely accounted for in conventional deep learning segmentation methods [48–50]. Instead, most deep learning segmentation pipelines are trained on binary data, with value 0 (outside the tissue) or 1 (inside the tissue), and therefore produce uncalibrated output probabilities. Ideally, segmentation methods would encode predictions as "50 shades of gray", representing partial volume information of the segmented tissue. Hence, there is a strong rationale for inputting/outputting "soft" labels in a deep learning segmentation pipeline to better calibrate the model confidence.

4.2.1 Related works

Soft labels have led to a better generalization, faster learning speed, and mitigation of network over-confidence [60]. Label smoothing was investigated in image classification [61, 83], style transfer [84], speech recognition [85], and language translation [86]. To segment multiple sclerosis lesions on MRI data, a recent study proposed to train a model using soft masks to account for the high uncertainty in lesion borders' delineation [65]. The soft masks were generated from the binary masks using morphological dilations. For the loss function, the authors used the soft version of the Dice loss [34]. This study reported an improved performance (+1.8% of Dice on the ISBI 2015 dataset) when using soft vs. binary masks. Another

study suggested proposed another ground truth softening method using over-segmentation and smoothing based on the distance to an annotated boundary, and also reported better performance over hard labels (+0.7% of Dice on the MRBrainS18 dataset) [62]. However, according to the authors, the performance improvements were conditioned by optimizing some hyper-parameters (e.g., number of super-pixels, beta), suggesting a potential limitation to generalize to new datasets and tasks. In the studies of Li et al. and Kats. et al., alteration of ground truth was based on arbitrary modifications of the input mask (mathematical morphology) and might not truly represent the underlying PVE. Moreover, even if the network is fed with soft ground truths, this rich information somewhat vanishes down the line in the training pipeline by the use of sharp activation functions (e.g., sigmoid) and classification-based loss functions (e.g., Dice loss) [40, 41].

4.2.2 Study outline

In this work, we explore training models using soft segmentations, both as input and output. While manual soft ground-truth generation is costly and highly time-consuming, we obtain soft inputs "for free" from binary ground truth data by skipping the binarization step that typically follows preprocessing and data augmentation. We focus on three key features: (i) training on soft (vs. hard) ground truth masks, (ii) the activation function used at the last layer (normalized ReLU vs. sigmoid), (iii) the use of a regression loss (vs. Dice loss) to favor soft predictions. We perform ablation studies for these three training features, whose combination is called SoftSeg, against the conventional training scheme on three open-source segmentation datasets: the spinal cord gray matter (SCGM) challenge [69], the multiple sclerosis (MS) brain lesion challenge [70], and the multimodal brain tumor segmentation (BraTS) challenge 2019 (BraTS 2019). In the following sections, the differences between SoftSeg and the conventional training pipeline will be detailed, along with the evaluation framework we used to compare them. Second, the results of the comparison on the three datasets will be presented from different perspectives: (i) the training process, (ii) the qualitative aspect of the segmentation, and (iii) the quantitative performances. Finally, the key contributions of SoftSeg and perspectives will be discussed.

4.3 Material and methods

4.3.1 Proposed method

The comparison between a conventional training pipeline and our proposed approach, SoftSeg, is illustrated in Figure 4.1. The key differences involve the binarization of the input

ground truth, the activation function, and the loss function. These differences are detailed in this section.

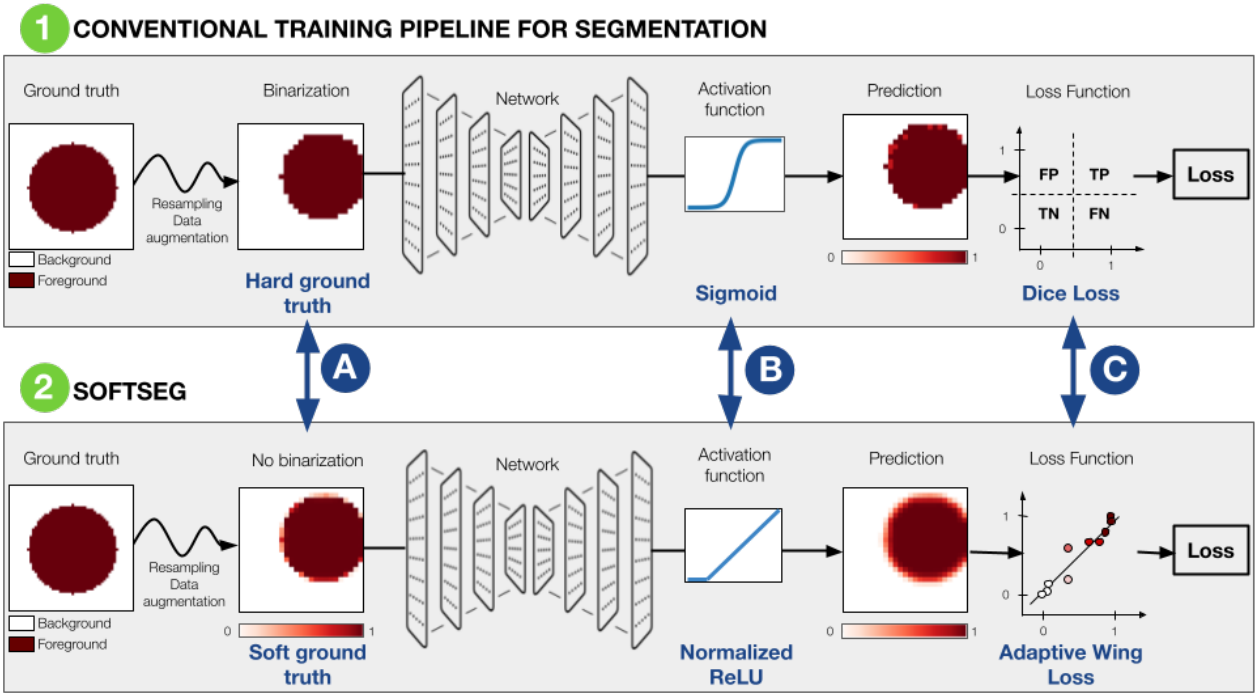


Figure 4.1 Training pipelines for segmentation.

1: Conventional training pipeline; 2: Our proposed approach (SoftSeg). The main differences are: (A) No binarization of the ground truth after the preprocessing and data augmentation operations; (B) A linear activation function is used instead of a sigmoid activation; (C) The loss function aims at solving a regression problem instead of a classification task.

Hard vs. Soft ground truth masks

Ground truth masks received by the network are conventionally binary, i.e., zeros and ones only, so-called "hard" ground truth. Although rarely specified, it is common to binarize the ground truth after applying preprocessing and data augmentation operations before feeding the network. Binarization is an approximation and a loss of information, especially for voxels at the border between two tissue types. To prevent such approximations, we propose to use soft (i.e., continuous values between 0 and 1) instead of hard masks, as illustrated in Figure 4.1A which are the result of the preprocessing and data augmentation without binarizing prior to the network. Soft masks used in this work notably aim at preserving partial volume information throughout the learning process, without applying complex label smoothing methods [62, 65] or resorting to costly soft ground-truths (e.g., from multiple experts).

Sigmoid vs. Linear activation function

The sigmoid activation function is popular in binary image segmentation models. Often used as the final activation, this non-linear activation is appropriate for classification since most values lie near 0 and 1, yielding a quasi-binary output. However, in the context of soft prediction, the sigmoid function undesirably narrows the range of soft values that potentially carry valuable PVE information. Although it can be partially addressed by increasing the temperature to make the active region larger, the use of other final activation functions (e.g., ReLU) has been recently explored, see for instance the comparison between CNN-Softmax and CNN-ReLU for classification tasks [46]. To avoid the polarizing effect in voxels observed when using the sigmoid, we propose to change the final activation of the segmentation model from the sigmoid function to a normalized rectified linear function (ReLU, see Figure 4.1B). A ReLU activation is applied to the model's output to set all negative values to 0 [45]. The result is then normalized by the maximum value to have a final output between 0 and 1, leading to a linear activation for the positive values and therefore highlighting the full range of prediction values from the model:

$$\text{NormReLU} \equiv \begin{cases} \frac{ReLU(X)}{max\{ReLU(X)\}}, & \text{if } \max\{\text{ReLU(X)}\} \neq 0 \\ 0, & \text{else} \end{cases} \tag{4.1}$$

where X represents the matrix output of the model before the final activation.

Classification vs. regression loss function

Segmentation is often considered a classification task where each voxel is assigned to one class. In that context, classification loss functions are commonly prioritized for segmentation tasks, such as the binary cross-entropy or the Dice loss functions. Although widely used with medical data [17, 37–39], the Dice loss yields sharp segmentation edges [40, 41], hindering predictions of non-binary values and can lead to a volumetric bias [42]. In contrast, the training approach we suggest is closer to a regression task in that the output prediction represents the input with high fidelity (e.g., an input voxel composed of 70% of the class of interest would produce an output prediction of 0.7). Consequently, we suggest using a regression loss function to train our network instead of a classification loss function (see Figure 4.1C). In this paper, we use the Adaptive Wing loss [68], which has shown fast convergence and efficient mitigation against class imbalance [87].

4.3.2 Datasets

To compare our approach with the conventional pipeline, we selected three publicly-available datasets: the SCGM challenge [69], the MS brain lesion challenge [70], and the multimodal BraTS challenge 2019 (BraTS 2019).

Spinal cord gray matter challenge

The SCGM dataset contains 80 MRI T2*-weighted 3D images of cervical spinal cord, evenly-acquired in four centers with different MR protocols and 3T scanners (Philips Achieva, Siemens Trio, Siemens Skyra). Demographics of the scanned subjects and acquisition parameters can be found in [69]. The gray matter was manually segmented on each 3D image by four independent experts (inter-expert Dice score ranging from 89% to 93% when compared to majority voting). The binary ground truth used in our experiments was generated with voxel-wise majority voting across all four experts. The dataset totalizes 940 cross-sectional 2D slices, whose resolution varies across centers: from 0.25x0.25 mm^2 to 0.5x0.5 mm^2.

MS brain lesion challenge

The MS brain lesion dataset was presented during the MICCAI 2016 challenge. It includes MRI scans of 15 subjects with five contrasts: T1-weighted, T1-weighted Gadolinium-enhanced, T2-weighted, PD T2-weighted, and FLAIR. The data was evenly acquired from three different centers and scanners: Philips Ingenia (3T), Siemens Aera (1.5T), and Siemens Verio (3T). MS lesions were manually segmented by seven experts. A consensus segmentation obtained with the Logarithmic Opinion Pool Based STAPLE algorithm [88] is used as ground truth in our experiments. The Dice score fluctuates between 69% and 77% when comparing each expert segmentation with the consensus ground truth. Moreover, the resolution varies from one center to another: 1x0.5x0.5 mm^3, 1.25x1x1 mm^3, and 0.7x0.75x0.75 mm^3 (right-to-left, posterior-to-anterior, inferior-to-superior). The provided dataset was already preprocessed as follows: denoising with the non-local means algorithm [89], rigid registration [90] on the FLAIR contrast, brain extraction, and bias correction with N4 algorithm [91].

BraTS challenge 2019

The BraTS challenge 2019 includes 335 subjects with high grade or low grade gliomas acquired from 19 different centers with varying acquisition protocols and 3T scanners (BraTS 2019). Four contrasts were provided: T1-weighted, T1-weighted Gadolinium-enhanced,

T2-weighted, and FLAIR. The peritumoral edema, the Gadolinium-enhancing tumor, and the necrotic and non-enhancing tumor core were manually segmented by one to four expert neuro-radiologists according to a common protocol. Rigid registration to a common anatomical template, skull-stripping, and 1 mm isotropic resampling was performed on the provided dataset. 20 subjects with high grade gliomas were randomly chosen from the dataset to perform multiple trainings within a reasonable time, while allowing proper cross-validation between them. The 20 subjects selected are listed in the 'brats_subjects.txt' file (`https://github.com/ivadomed/article-softseg`). As our study focuses on the comparison between soft and hard segmentation, we did not perform multi-class training. Hence, a single label was retained for the experiments: the tumor core composed of the necrotic and enhancing tumor.

4.3.3 Training protocol

Different training protocols were selected for each dataset based on initial hyperparameter exploration (Table 4.1).

Training / validation / testing split

For the SCGM challenge dataset, the four centers with their associated data were randomly split into groups of size two / one / one to compose the training, validation, and testing sets, respectively. We split the SCGM dataset according to the acquisition center to assess the approaches' ability to generalize to new acquisition parameters. For the MS brain lesion and BraTS segmentation tasks, we trained the networks on 60% of the patients, with 20% held out for validation and 20% for testing. Center-wise splitting was not possible for the MS brain or BraTS datasets as the origin of images was not directly available.

Preprocessing

All data were resampled to a common dataset-specific resolution (see Table 4.1), using spline interpolation (2^{nd} order) for the images and linear interpolation for the ground truths. The 2^{nd} order interpolation was chosen to preserve higher spatial frequency content in the images, while the 1st order for the labels was selected to avoid high frequency oscillations at the interface of the binary segmentation. Cross-sectional slices were subsequently center-cropped to a common size specific to each dataset (see Table 4.1).

Table 4.1 Training parameters for each dataset.

For all training parameters, please see configuration files: `https://github.com/ivadomed/article-softseg/tree/main/config`. 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	BraTS dataset 2019
Preprocessing	**Resample**	0.25 × 0.25 × 2 mm^3 (RPI)	1 mm isotropic	1 mm isotropic
	Batch format	2D axial slices		
	Crop	128 × 128 $pixels^2$	160 × 124 $pixels^2$	210 × 210 $pixels^2$
Data Augmentation	**Rotation**	± 20 degrees		
	Translation	± 3%		
	Scale	± 10%		
Batch Size		8	24	24
U-Net Depth		3	4	4
Dropout Rate		30%		
Learning Rate	**Initial**	0.001	0.00005	0.0001
	Scheduler	Cosine Annealing		
Adaptive Wing Loss		ϵ=1; α=2.1; θ=0.5; ω=8		
Early Stopping		Patience: 50 epochs ; ϵ: 0.001		
Maximum Number of Epochs		200		

Data augmentation

For data augmentation, affine transformations were randomly applied to all training samples using linear interpolation (see Table 4.1 for details). Segmentation labels from the conventional approach (i.e., hard training) were binarized after applying data augmentation, while soft training candidates were untouched to preserve the softness of their augmented masks. We assessed the impact of binarized augmented masks (i.e., hard ground truth) compared to

non-binarized augmented masks (i.e., soft ground truth), see section 4.3.4 for more details.

Intensity normalization

The intensities of each image were standardized by mean centering and standard deviation normalization. When several contrasts were available (MS brain, BraTS), this normalization was done on each contrast separately.

Iterations

All models were trained with a patience of 50 epochs and a maximum epoch count of 200. Batch sizes of 8, 24, and 24 were respectively used for the SCGM, brain MS, and BraTS datasets

Optimization

The learning rate was modified throughout the training according to the cosine annealing scheduler with an initial value of 0.001 for the SCGM dataset, 0.0005 for the MS dataset, and 0.0001 for the BraTS dataset.

Network architecture

For all experiments, we used a U-Net architecture [18] with a depth (i.e., number of down-sampling layers) of 3 for the SCGM challenge and of 4 for the brain MS lesion challenge and BraTS data (see section 4.3.4 for details). The choice of depth was based on preliminary hyperparameters optimization. Batch normalization [92], ReLU function, and dropout [93] followed each convolution layer. Convolution layers had standard 3×3 2D convolutions filters and a padding size of 1.

Activation function

Two different activation functions were tested on the model's output: sigmoid or normalized ReLU function (Figure 4.1B). Throughout the experiments, we assessed the characteristics exhibited by the model's predictions when using either sigmoid or normalized ReLU.

Loss function

We compared the use of a regression loss function to a standard classification loss function for segmentation tasks (see Figure 4.1C), using the Adaptive Wing loss [68] vs. the Dice

loss [34]. The Adaptive Wing loss, initially introduced for heatmap regression for labeling facial key points, was chosen for its ability to propagate and predict soft values, but the proposed approach could work with other regression losses. For the Adaptive Wing loss, preliminary experiments led to the hyperparameters indicated in Table 4.1.

Implementation

Implementation and model training was done with ivadomed v2.2.1 [35]. ivadomed is a Python-based open-source framework for deep learning applied to medical imaging (`https://ivadomed.org/`). To promote the reproducibility of our experiments, all configuration files can be found at `https://github.com/ivadomed/article-softseg`.

4.3.4 Evaluation

Evaluation protocol

To isolate the specific impact of each explored feature (hard/soft mask, activation function, loss), five candidates were compared (see Table 4.2). *Hard-Sig-Dice* represents the conventional deep learning candidate using binarization with a sigmoid activation function and Dice loss (Figure 4.1, panel 1). Our proposed hypothetically-best candidate is *Soft-ReLU-Wing* (Figure 4.1, panel 2, *SoftSeg*). *Hard-ReLU-Wing*, *Soft-ReLU-Dice*, and *Soft-Sig-Wing* each has only one feature changed from our proposed candidate.

Cross-validation was applied to each model candidate. For the SCGM datasets, each model was trained 40 times, 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.). For the MS and BraTS datasets, each model was trained 10 and 15 times respectively, with a different dataset split for each model. For each of the evaluation metrics (see 4.3.4), a non-parametric 2-sided Wilcoxon signed-rank test compared the *Soft-ReLU-Wing* candidate with every other candidate. A p-value inferior or equal to 0.05 was considered significant.

Evaluation metrics

Before computing the evaluation metrics, the network predictions were resampled to the native resolution (i.e., resolution of the native ground truth) and binarized. The threshold used to binarize the predictions was determined by searching for the optimal value (between 0 and 1 with an increment of 0.05) in terms of Dice score when using the trained model on the training and validation images. The metrics include: (i) Dice score, (ii) precision, (iii)

Table 4.2 Candidates description.

Each row represents a candidate (i.e. a training approach), whose features are detailed in the columns. Abbreviations: GT: ground truth.

	Binary GT after Data Augmentation	Activation function	Loss function
Hard-Sig-Dice (Conventional)	Yes	Sigmoid	Dice
Hard-ReLU-Wing	Yes	NormReLU	Adaptive Wing
Soft-Sig-Wing	No	Sigmoid	Adaptive Wing
Soft-ReLU-Dice	No	NormReLU	Dice
Soft-ReLU-Wing (SoftSeg)	No	NormReLU	Adaptive Wing

recall, (iv) absolute volume difference (absolute volume difference between the ground truth and prediction, divided by the ground truth volume), (v) relative volume difference, and (vi) mean squared error (MSE). All metrics are expressed in percentages. For the MS lesion segmentation task, we also included lesion detection metrics which are clinically relevant: the lesion true positive rate (LTPR) and false detection rate (LFDR) as defined in [75]. These detection metrics were not used for the other datasets (SCGM and BraTS), because in these cases there was always only one 3D target object per MRI volume.

4.4 Results

In the following sections, we compare how the features illustrated in Figure 4.1 influence the training process (section 4.4.1), the prediction values dynamic (section 4.4.2), and the overall model performance on the testing dataset (section 4.4.3).

4.4.1 Training process

Figure 4.2 shows the evolution of the training process across different model configurations. The conventional approach (*Hard-Sig-Dice*) yielded quasi-binary predictions from the very early stages of the training. Conversely, the other candidates produced predictions with low values on the gray matter surrounding at the early stages (see epoch #5 and 10), while at

later stages the object is delineated with a soft segmentation (i.e., high prediction values within the object core and lower values on the edges). Among the three proposed training schemes (bottom rows), the candidate *Soft-ReLU-Dice* produced high prediction values (i.e., red voxels in Figure 4.2) earlier in the training process than the other two. Although the conventional candidate yielded high prediction values earlier, it did not necessarily trigger an "early-stopping" of the training earlier than the proposed candidates. The mean early stopping epochs were 123 and 128 for the conventional and the proposed approach, respectively. This means that training time was not importantly impacted when performing soft training. Unlike the output of the conventional candidate, the edges of the segmented object with soft training remained soft even at the final stages (see "Last epoch" in Figure 4.2). This was particularly the case for the *Soft-ReLU-Wing* candidate. Results of the *Soft-Sig-Wing* candidate are not depicted here because the model training did not converge during this experiment (see Table 4.3 for overall quantitative results).

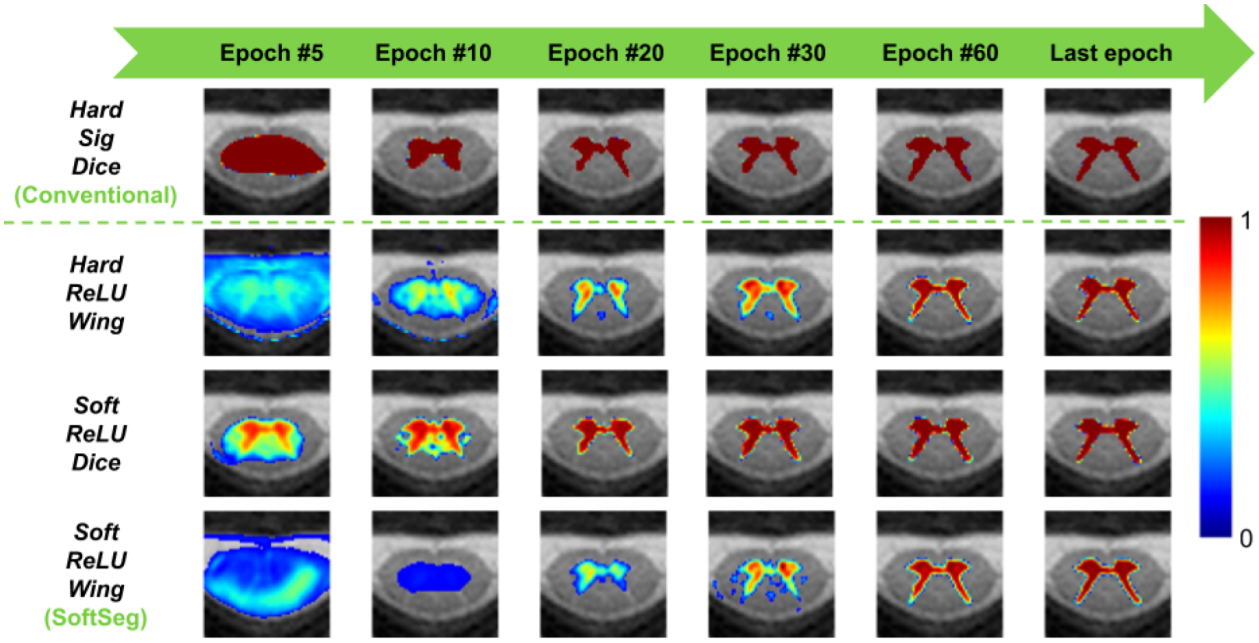


Figure 4.2 Learning progression through epochs for different training schemes on the SCGM dataset.

Each row represents a training scheme, while each column shows the model prediction on a validation slice at a particular training epoch. The last epoch (right column) varied across approaches because of the early stopping feature. Predictions are overlaid on the anatomical data and range from 0 (transparent) to 1 (Red). *Soft-Sig-Wing* predictions are not shown here since the model training did not converge.

4.4.2 Output softness

To compare the performance of the different model configurations, we binarized the model predictions before computing the evaluation metrics. The binarization threshold was optimized by finding the value (between 0.05 and 0.95, with an incremental step of 0.05) that maximizes the Dice score when inferring on the training and validation dataset. Figure 4.3 shows the results of this optimization for each candidate (rows) and each iteration (purple dots). One notable observation is the large min-max Dice range across threshold values (up to 34% for *Soft-ReLU-Wing*), confirming the importance of this threshold optimization step. Conversely, the Dice range is more modest for the conventional candidate (9% for *Hard-Sig-Dice*), which is a direct consequence of the greater number of polarized values around 0 and 1. The loss function had the greatest impact on the min-max Dice range: it dropped from 34% to 13% when switching Adaptive Wing loss to Dice loss functions. This result highlights the importance of threshold fine-tuning when using a regression loss.

Figure 4.4 represents the voxels intensity distribution across the tested candidates and datasets. For the SCGM dataset (Figure 4.4A), all candidates yielded predictions with values concentrated around 1. *Soft-ReLU-Wing* intensity distribution is more spread out compared to other candidates and therefore its predictions could be considered being the least binarized. On the brain MS dataset (Figure 4.4B) and the BraTS brain tumor dataset (Figure 4.4C), two groups of candidates stand out: the "hard" group *Hard-Sig-Dice*, *Soft-Sig-Wing* and the "soft" group *Soft-ReLU-Dice*, *Hard-ReLU-Dice*, *Soft-ReLU-Wing*. In the "hard" group both candidates exhibit polarized predictions near 0 or 1. Conversely, the "soft" group values are more spread out in the]0, 1] range (for the MS dataset) and]0, 0.5] range (for the BraTS dataset). In the MS dataset, *Soft-ReLU-Wing* and *Soft-ReLU-Dice* are almost superimposed and yielded more non-zero values than *Hard-ReLU-Wing*. Across the three datasets, the "soft" group exhibits a higher number of non-zero predictions (higher area under curve). Overall, Figure 4.4 shows that candidates using the ReLU activation function (vs. sigmoid) are associated with softer predictions.

Figure 4.5 illustrates the performance of each training scheme for the SCGM dataset in one representative subject per center. From this figure, one can appreciate the variability in terms of image resolution, white-to-gray matter contrast, and signal-to-noise ratio. Image heterogeneity had a notable impact on candidates' performance across test centers. On average, across all iterations, the candidates presented in Figure 4.5 obtained a Dice score of 86.2%, 81.2%, 88.6%, and 78.0% for centers 1, 2, 3, and 4, respectively. When compared with the conventional candidate, *Soft-ReLU-Wing* showed the highest Dice score for all test centers except for center #4 (77.4% for *Soft-ReLU-Wing* vs. 79.0% for *Hard-Sig-Dice*). Interestingly,

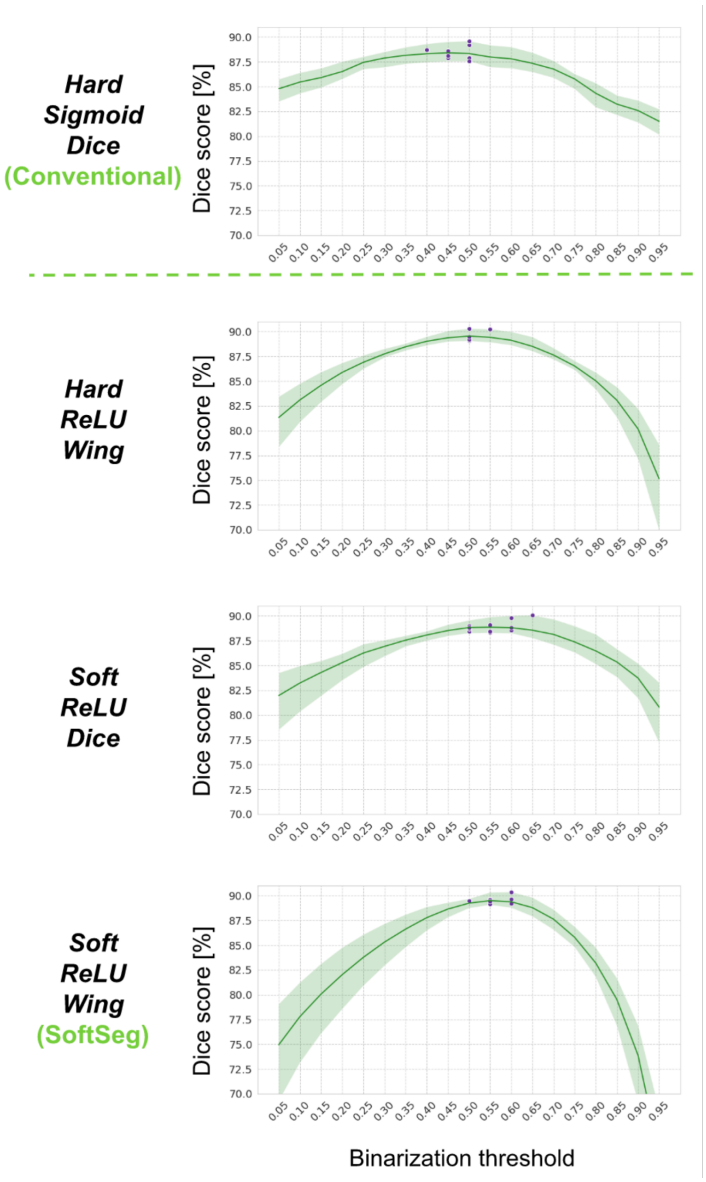


Figure 4.3 Optimization of the binarization threshold for model prediction.

For each threshold value (between 0.05 and 0.95, with an incremental step of 0.05), the Dice score was computed on the trained model predictions for the training and validation SCGM data. The thick green line represents the average value while the green shaded area represents the min/max range of values. Purple dots represent the threshold that maximizes the Dice score, for each iteration. For the sake of comparison, the y-scale was kept the same across the four candidates. The lowest value for the *Soft-ReLU-Wing* (which is not shown due to cropping) is 56, and 70 for *Hard-ReLU-Wing*. *Soft-Sig-Wing* graph is not shown here since the model training did not converge.

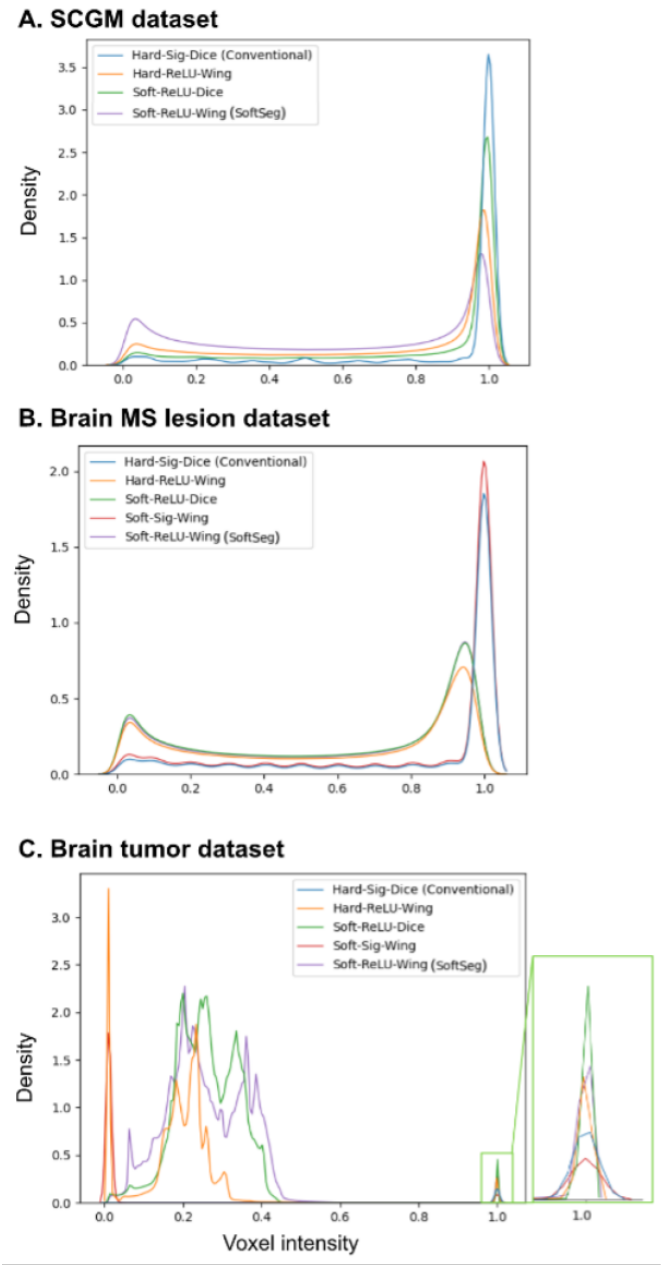


Figure 4.4 Distribution of non-zero prediction voxels for each candidate on SCGM (A), MS brain lesions (B), and BraTS (C) datasets.

Distributions are computed using the kernel density estimation method and normalized so the area under the curve sums up to 1 for all curves. Training of the *Soft-Sig-Wing* model did not converge and is therefore not shown. The *Soft-ReLU-Dice* (green) and *Soft-ReLU-Wing* (purple) curves are almost perfectly superimposed on B. Because of the density estimation, the curves slightly extend outside of the prediction values (below 0 and above 1). Abbreviations: MS: multiple sclerosis ; SCGM: spinal cord gray matter.

in centers 1, 2, and 3 where images have the lowest resolution in the cross-sectional plane (0.3, 0.5, and 0.5 mm isotropic for centers 1, 2, and 3 respectively vs. 0.25 mm isotropic for center 4), the softer candidates segmented more truthfully the gray matter with an average improvement of 3.2% Dice score. This observation is in line with the hypothesis that soft training is well suited for mitigating PVE, i.e, the benefits are more considerable in images with lower spatial resolution.

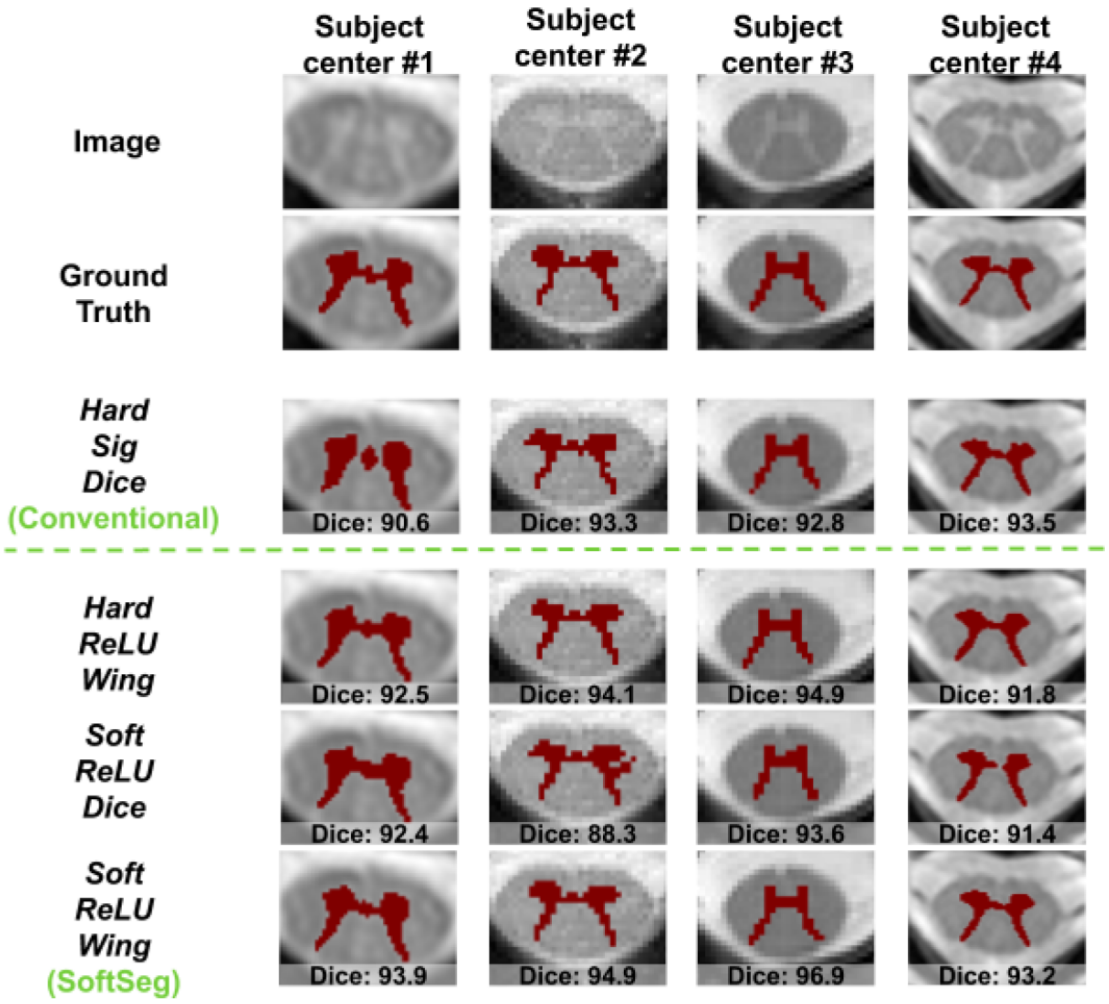


Figure 4.5 Example of segmentation result for the SCGM dataset, across the four centers (columns) and the four candidates.

The first row shows the input 2D slice, the second row shows the manual ground truth. Rows 3-6 correspond to specific training schemes (see Table 4.2 for details). Predictions were binarized as described in section 4.3.4. *Soft-Sig-Wing* predictions are not shown here since the model training did not converge.

Figure 4.6 depicts MS lesion predictions across the five candidates. MS lesion predictions present two patterns of softness among approaches. *Hard-Sig-Dice* and *Soft-Sig-Wing* predict

mostly values around 1 ("hard" group, as defined in Figure 4.4 description), whereas *Soft-ReLU-Dice*, *Hard-ReLU-Dice*, and *Soft-ReLU-Wing* display a broader range of prediction values ("soft" group, as defined in Figure 4.4 description). The final activation distinguishes the two groups; the candidates displaying softer outputs had a normalized ReLU activation function, while the other candidates predicting more binarized values used a sigmoid as final activation. The candidates from the "hard" group, *Hard-Sig-Dice* and *Soft-Sig-Wing*, show overall less true positives (and consequently less false positives). Conversely, the softer candidates, *Soft-ReLU-Dice*, *Hard-ReLU-Dice*, and *Soft-ReLU-Wing*, are associated with a higher true lesions positive rate. On the close-ups from Figure 4.6 (left column), the candidates from the "hard" group show a single segmented lesion (two are missing), while all candidates from the "soft" group exhibit three distinct true positives.

Figure 4.7 illustrates segmentation results for the BraTS dataset. As observed in Figure 4.4 and 4.6, the same two groups with differing softness patterns can be isolated: *Hard-Sig-Dice* and *Soft-Sig-Wing* ("hard" group), and *Soft-ReLU-Dice*, *Hard-ReLU-Dice*, and *Soft-ReLU-Wing* ("soft" group). The "hard" group presents over-segmentation of the tumor core. Even on the false positive voxels, the raw prediction of the model yields a value of 1. Conversely, the soft group exhibits a ranging value of confidence around the borders of the tumor cores. The blue background on the *Soft-Sig-Wing* candidate is caused by most values being near 0 (not exactly 0). This candidate showed instability during training and did not reach convergence on every cross-validation. Like for the SCGM and MS lesion brain datasets, candidates from the "soft group" produce soft edges, and consistent shapes.

4.4.3 Segmentation performance

SCGM

Table 4.3 summarizes the segmentation performance metrics for the five candidates on the SCGM dataset. *Soft-ReLU-Wing* yielded the highest Dice, precision, recall, absolute volume difference, and MSE scores compared to the conventional and other proposed approaches. When considering only the Dice score, there is a statistical difference between *Soft-ReLU-Wing* vs. *Hard-Sig-Dice* (p-value=0.0011), *Hard-ReLU-Wing* (p-value=0.0385), and *Soft-Sig-Wing* (p-value=1.10e-7). *Soft-Sig-Wing* did not converge with the GM dataset on all iterations which explains the low performances compared to the other candidates.

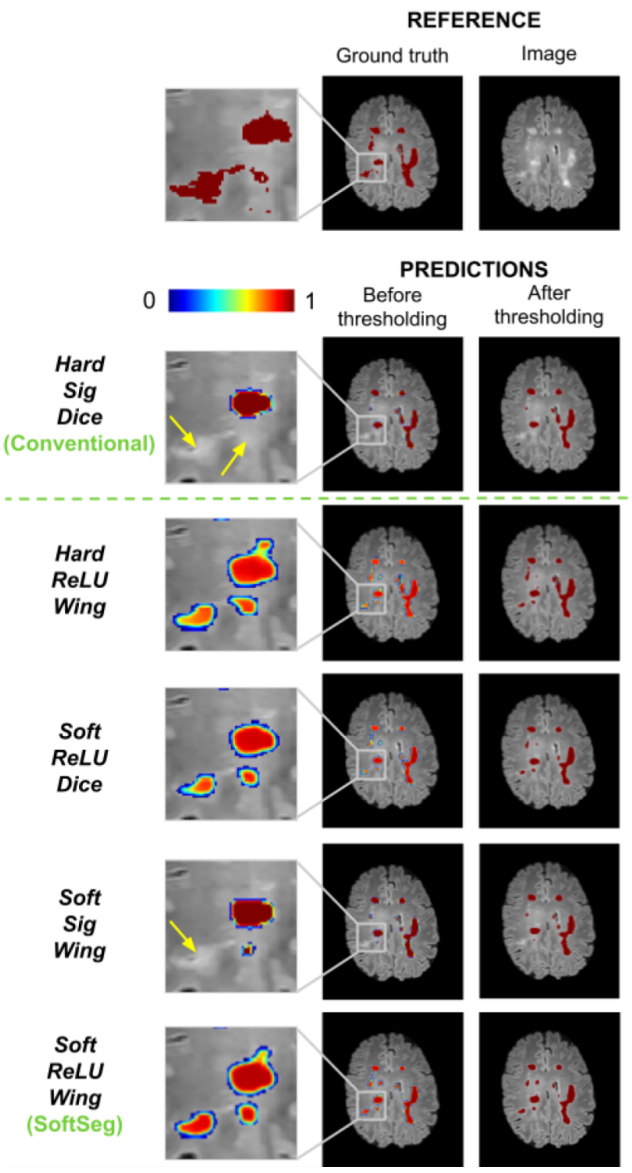


Figure 4.6 Brain MS lesions segmentation for the five candidates.

The first row represents the input image and the consensus segmentation from the seven experts. For the remaining rows, the second column presents the raw predictions and the third column contains the binarized predictions. Predictions are overlaid on the anatomical data and range from 0 (transparent) to 1 (Red)

Brain MS lesions

Table 4.4 presents the candidates performance metrics on the MS lesions dataset. As observed on the SCGM dataset, *Soft-ReLU-Wing*, had the highest Dice score, recall, and LTPR. *Soft-Sig-Wing* predicted less false positives compared to the other candidates illustrated by the

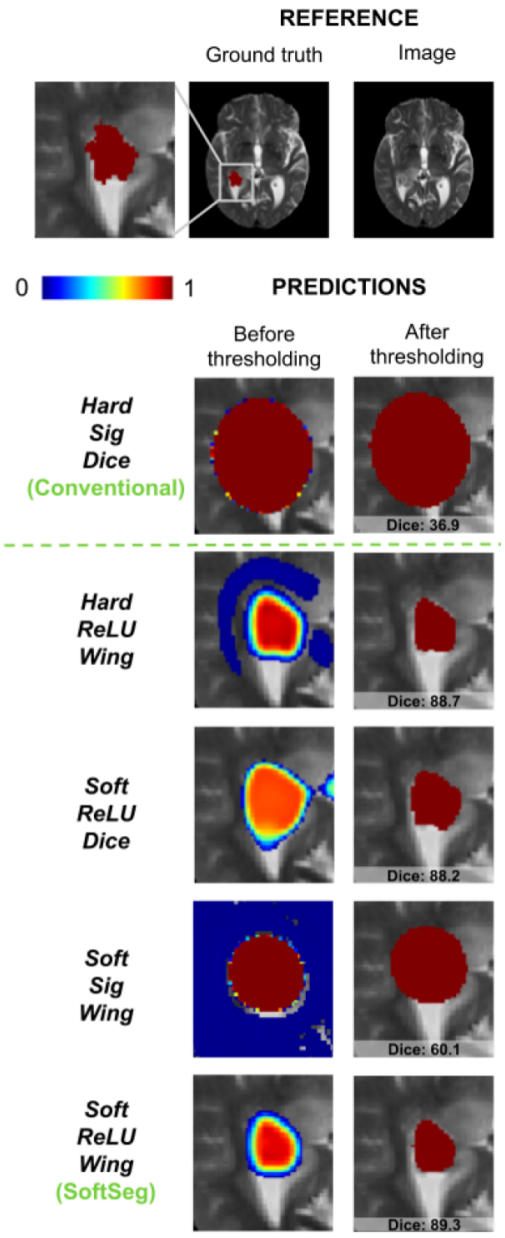


Figure 4.7 Segmentation of brain tumor core for the five candidates.

The first row represents the input image and the ground truth with a close-up of the tumor segmentation. For the remaining rows, the left image represents the raw core tumor segmentation prediction from the model and the right the binarized prediction. Predictions are overlaid on the anatomical data and range from 0 (transparent) to 1 (Red).

highest precision score and the lowest LFPR. No statistical differences were observed between groups, probably due to large standard deviation between iterations on the MS dataset (testing set: n=3).

Table 4.3 Gray matter segmentation performance metrics for the five candidates.

The error represents the standard deviation from 40 trainings (MEAN ± STD). The optimal score value is indicated under each metric name. Rows identify the five candidates (see Table 4.2 for candidates description). Columns represent the metrics. ** $p - value < 0.05$ for 2-sided Wilcoxon signed-rank test compared to the *Soft-ReLU-Wing* candidate. Abbreviations: MSE: mean squared error ; Opt: optimal.

	Dice [%] *Opt. value: 100*	**Precision [%]** *Opt. value: 100*	**Recall [%]** *Opt. value: 100*	**Absolute Volume Difference [%]** *Opt. value: 0*	**Relative Volume Difference [%]** *Opt. value: 0*	**MSE [%]** *Opt. value: 0*
Hard-Sig-Dice **(Conventional)**	82.3 ± 5.0 **	84.4 ± 9.2 **	83.3 ± 8.1	17.6 ± 11.7	**-1.1 ± 20.3**	0.290 ± 0.058 **
Hard-ReLU-Wing	83.7 ± 4.9 **	85.5 ± 11.3	84.5 ± 6.1	18.6 ± 13.8 **	-2.0 ± 22.6	0.275 ± 0.082
Soft-ReLU-Dice	83.7 ± 5.2	85.7 ± 10.0	84.6 ± 7.6	17.6 ± 12.8	-1.3 ± 21.2	0.269 ± 0.066
Soft-Sig-Wing	52.9 ± 36.7 **	73.1 ± 18.0 **	48.5 ± 39.1 **	53.0 ± 36.7 **	34.3 ± 54.3 **	0.611 ± 0.306 **
Soft-ReLU-Wing **(SoftSeg)**	**84.3 ± 4.7**	**85.8 ± 10.8**	**84.9 ± 5.0**	**16.9 ± 12.1**	-1.6 ± 19.9	**0.268 ± 0.083**

Brain tumors

Table 4.5 reports the segmentation performance metrics of the candidates on the BraTS dataset. *Soft-ReLU-Wing* is associated with the highest Dice score, precision, relative volume difference, and MSE (Table 4.5). This candidate, when compared with the conventional candidate, reached statistical differences for precision (p-value=0.039), and MSE (p-value=0.024). The "soft" group, composed of *Soft-ReLU-Wing*, *Hard-ReLU-Wing*, and *Soft-*

Table 4.4 Brain MS lesion segmentation performance metrics for the five candidates.

The error represents the standard deviation from 10 trainings (MEAN ± STD). The optimal score value is indicated under each metric name. Rows identify the five candidates (see Table 4.2 for candidates description). Columns represent the metrics. Abbreviations: LFDR: lesion false detection rate ; LTPR: lesion true positive rate ; Opt: optimal.

	Dice [%] *Opt. value: 100*	**Precision [%]** *Opt. value: 100*	**Recall [%]** *Opt. value: 100*	**LFDR [%]** *Opt. value: 0*	**LTPR [%]** *Opt. value: 100*
Hard-Sig-Dice **(Conventional)**	42.7 ± 14.5	58.3 ± 13.1	41.4 ± 17.1	61.9 ± 13.3	34.1 ± 16.7
Hard-ReLU-Wing	45.1 ± 13.0	55.5 ± 17.8	44.0 ± 15.5	65.6 ± 10.1	37.0 ± 14.4
Soft-ReLU-Dice	45.3 ± 14.1	56.6 ± 20.4	46.1 ±17.1	64.1 ± 12.0	36.5 ± 15.2
Soft-Sig-Wing	45.1 ± 12.2	**59.8 ± 13.2**	43.4 ± 13.2	**57.7 ± 17.3**	34.8 ± 15.6
Soft-ReLU-Wing **(SoftSeg)**	**46.0 ± 12.2**	55.2 ± 17.6	**46.7 ± 13.8**	63.0 ± 15.4	**38.6 ± 14.8**

ReLU-Dice, presented similar Dice, precision, recall, MSE scores. *Soft-ReLU-Wing* yielded the highest recall score and *Hard-ReLU-Wing* the best absolute volume difference. As previously observed with the SCGM dataset, the candidate *Soft-Sig-Wing* did not converge on every iteration leading to lower segmentation scores. The conventional approach largely over-segmented tumor cores yielding an average relative volume difference of -29.9% and an average absolute volume difference of 67.1%, as illustrated in Figure 4.7.

Table 4.5 Brain tumor segmentation performance metrics for the five candidates.

The error represents the standard deviation from 15 trainings (MEAN ± STD) on 20 randomly-selected subjects from the 2019 BraTS dataset. The optimal score value is indicated under each metric name. Rows identify the five candidates (see Table 4.2 for candidates description). Columns represent the metrics. ** $p - value < 0.05$ for 2-sided Wilcoxon signed-rank test compared to the *Soft-ReLU-Wing* candidate. Abbreviations: Opt: optimal.

	Dice [%] *Opt. value: 100*	**Precision [%]** *Opt. value: 100*	**Recall [%]** *Opt. value: 100*	**Absolute Volume Difference [%]** *Opt. value: 0*	**Relative Volume Difference [%]** *Opt. value: 0*	**MSE [%]** *Opt. value: 0*
Hard-Sig-Dice (Conventional)	63.6 ± 28.7	66.1 ± 29.0 **	70.9 ± 30.8	67.1 ± 132.3	-29.9 ± 145.6	40.1 ± 44.7 **
Hard-ReLU-Wing	57.2 ± 28.5 **	61.8 ± 32.0 **	70.0 ± 26.9	527.5 ± 164.3 **	-490.2 ± 1654.8	92.2 ± 167.5 **
Soft-ReLU-Dice	69.8 ± 26.4	**72.6 ± 28.8**	**73.2 ± 26.0**	49.7 ± 83.1	-20.7 ± 94.8	29.8 ± 35.0
Soft-Sig-Wing	55.7 ± 27.3 **	66.7 ± 30.3	60.1 ± 30.8 **	98.2 ± 193.5 **	-45.9 ± 212.5	43.3 ± 43.8 **
Soft-ReLU-Wing (SoftSeg)	**70.1 ± 23.2**	71.9 ± 25.1	72.8 ± 25.0	**38.6 ± 64.5**	**-8.1 ± 74.9**	**29.7 ± 38.0**

4.5 Discussion

We introduced an alternative approach, SoftSeg, to train deep learning models for image segmentation. We demonstrate the application of SoftSeg in three different and publicly-available medical imaging datasets. The proposed training scheme is based on prediction labels with continuous ("soft") rather than binary values. The benefits of soft segmentation

include: a better precision when computing segmentation-based morphometric measurements (e.g., tumor size), the possibility to encode partial volume information, and other useful information that are discussed in the perspectives section (4.5.5). These soft segmentations are obtained for free as a side effect of not binarizing after data augmentation. To allow soft label propagation through the network training process, we modified the conventional training pipeline by using (i) soft ground truth masks, (ii) a normalized ReLU final activation layer, and (iii) a regression loss function (Adaptive Wing loss). Overall, the combination of these three features outperformed the conventional candidate on the three tested datasets (see Tables 4.3 to 4.5). Besides, this candidate yields soft predictions, especially at object boundaries or on small objects such as MS lesions. These soft predictions provide relevant insights on the model's confidence and allow meaningful automated post-processing. In particular, the proposed approach has an increased sensitivity (e.g., identify a higher number of lesions), which is desired by radiologists. The developed training pipeline is freely available as part of ivadomed [35].

4.5.1 Impact of the soft features for training

The three soft features differing from the conventional approach are a soft input, the final activation, and the loss function. These features had an overall positive impact on segmentation performance. Taken separately or altogether, they yielded the highest Dice score and best output softness for each of the three tested datasets. Removing one soft feature from the fully soft candidate (*Soft-ReLU-Wing*) slightly lowered the Dice score for the candidates that reached convergence. On the brain MS and BraTS datasets, the final activation had the greatest impact on the predictions' softness. Two different behaviors were clearly distinguishable when changing the final activation. The group associated with the normalized ReLU activation function ("soft" group) yielded softer predictions that can be assessed quantitatively (Figure 4.4) and qualitatively (Figure 4.6), when compared to the group with a sigmoid as final activation ("hard" group). In Table 4.4, the "soft" group can be associated with higher true positive detection rates (better recall and LTPR) and the "hard" group with less false positives (better precision and LFDR). This comparison cannot be made for the SCGM dataset, since the candidate with the conventional final activation did not converge. Similarly, the loss function and the use of hard vs. soft ground truths had overall a positive impact on the segmentation performance. Both features had an average Dice score drop of 0.8% across datasets when using their hard versions compared to the fully soft candidate, *Soft-ReLU-Wing*.

Future investigations could look at the potential benefits of other loss and last activation

functions in combination with the SoftSeg framework. For instance, the recently proposed "Log-Cosh-Dice loss" could be of interest since its log-cosh transformation has been successfully employed in regression tasks for smoothing purposes [31]. Regarding the last activation function, one can consider "softplus", which is a smooth version of the ReLU activation function [94]. A multiclass version of NormReLU could also be investigated since the current version does not guarantee the classes to be mutually exclusive (e.g., normalize NormReLU output by the sum of predictions along the class axis).

4.5.2 Non convergence of Soft-Sig-Wing

Soft-Sig-Wing performed poorly on SCGM and BraTS datasets (Tables 4.3 and 4.5). Some training runs from this candidate did not reach convergence while others did. Since *Soft-ReLU-Wing* always converged in our experiments, the instability of *Soft-Sig-Wing* may be attributed to the use of the sigmoid with a soft training approach. Since the sigmoid function tends to classify voxels (i.e., almost binary outputs), it may not be suitable to use in combination with a regression loss function which is not designed for polarized inputs. Consequently, the association of these two features could hinder training convergence.

4.5.3 Thresholding the output prediction

Given that the ground truths used in this study are binary, we were bound to use evaluation metrics that accommodate binary inputs (e.g., Dice score, prediction, recall). Moreover, thresholding the prediction was necessary because these metrics penalize soft predictions. For instance, a soft prediction of 0.51 leads to a Dice score of 0.675. The same prediction undergoing binarization at a 0.5 threshold would produce a Dice of 1.0. Note that, when considering a regression-type metric like MSE, SoftSeg still outperformed the conventional approach without thresholding the output predictions. For example on the SCGM dataset, MSE was 0.215 ± 0.070 for SoftSeg vs. 0.251 ± 0.064 for the conventional approach.

SoftSeg generates more distributed values between 0 and 1, hence this method is more sensitive to the selected threshold. Nevertheless, SoftSeg proved to generalize well to new data. It yielded the best performance on the SCGM cross-validation where testing data originated from unseen centers with different acquisition parameters even though the threshold was optimized on the training/validation sets.

4.5.4 Repeatability and statistical differences

Although rarely employed in deep learning model evaluation, we performed a cross-validation statistical analysis for each datasets. We used 40 folds on the SCGM dataset (10 per center), 10 folds on the brain MS dataset, and 15 folds on the BraTS dataset. For each dataset, the number of iterations for the cross-validation was determined by the typical training time while allowing us to run the different experiments in a reasonable time (12 hours/training for the BraTS dataset on a single NVIDIA Tesla P100 GPU). Resorting to cross-validation to evaluate our approaches is particularly relevant for the brain MS and brain tumor datasets due to the small number of subjects. Also, the heterogeneity of lesion load and tumor core size led to high variations in performance across iterations (mean Dice standard deviation: 13.2% on MS lesions and 25.9% for brain tumors). Statistical difference was not reached for most metrics of MS and brain tumor dataset. The absence of statistical difference can be explained by the large standard deviations due to a wide performance range from one subject to another. The MS lesion and brain tumor datasets included 15 and 20 subjects leading to only 3 and 4 testing subjects respectively. Also due to the size of the dataset, only 10 (MS lesions) and 15 iterations (brain tumors) were performed on these datasets. Datasets with more patients leading to smaller standard deviations and more iterations would help in getting statistical differences.

4.5.5 Perspectives

Partial volume effect accountability

Morphometric analyses in MRI aim at measuring shape and/or volumes from anatomical (e.g., brain, spinal cord) or pathological structures (e.g., tumors, MS lesions). These measures are traditionally computed from binary segmentations produced manually or (semi-)automatically. As a result, their precision is inherently limited by the native spatial resolution (set during image acquisition) relative to the size of the object [77, 95]. A strong motivation for this work was to introduce a means to produce soft segmentations faithful to the partial volume information. We show that SoftSeg does produce soft segmentations (Figure 4.4) while maintaining good performances on traditional metrics (Tables 4.3-4.5). The next step is to confirm/infirm that SoftSeg can produce accurate partial volume estimations. In order to do so, one needs a ground truth that encodes such information, unlike the datasets used in the present study where ground truths were binary. A possible approach would be to synthesize a dataset at various resolutions from an analytical model of tissue distributions [96–98], train a model with SoftSeg and validate the estimated tissue class fraction

voxel-wise.

Encoding expert confidence during training

Manual segmentation of medical image segmentation is highly challenging and is prone to intra-expert variability. For instance, experts usually have a difficult time precisely delineating very small lesions [74]. This challenge is partly due to them being required to decide if a voxel pertains to a lesion or not. The need for this binary decision has been driven by traditional training approaches, which require a binary ground truth as input for the model. With the SoftSeg method proposed here, expert raters will have the possibility to modulate their manual rating and assign values that reflect their level of confidence, e.g., 0.5, 1, and 2 would be respectively associated with a low, medium, and high confidence about the presence of a lesion. Although more time-consuming than binary manual segmentation, encoding expert confidence in neural networks via the generation of soft ground truths would likely have a positive impact on segmentation performance.

Preserving inter-expert variability

High inter-expert variability is a widespread challenge in medical image segmentation, resulting from factors such as image quality, expert training/experience [74,99,100]. Some datasets provide the segmentation from multiple experts to account for this variability. However, these manual segmentations are usually merged into a single binary mask using label fusion methods, e.g., majority voting, STAPLE [64]. Recent studies highlighted the negative effect of label fusion methods to obtain reliable estimates of segmentation uncertainty as inter-expert variability is ignored when models are trained on the resulting binary masks [8, 9]. The SoftSeg method introduced here could elegantly account for the inter-expert variability and calibrate the model confidence by inputting soft ground truths that incorporate information about experts' disagreement. It is however unclear how this initial soft segmentation should be obtained, e.g., by averaging the expert segmentations, or using the recently-published soft STAPLE approach [65]. Moreover, validating the specific benefits of encoding richer inter-rater information into the ground truth masks would pose additional challenges since intra-rater variability also exists. Future investigations should assess the relevance of soft labels, both in terms of segmentation performance and uncertainty estimation.

Combining soft segmentation with uncertainty estimation

Estimation of deep learning model uncertainty is an active field of research [101–103] in medical image segmentation, following the seminal works from [52]. Whether they are based on output probability calibration [21, 101], ensemble methods [101] or Bayesian models [75], all these approaches provide a representation of how trustful a prediction is. A common denominator of the recent investigations on uncertainty applied to segmentation tasks, is that they have relied on the conventional "hard" training, which produces highly polarized predictions, and as such might not be the most adequate for representing the rich spectrum of uncertainty values on the prediction. The conventional segmentation pipeline tends to yield overconfident predictions, even on misclassified voxels, leading to poor interpretation of the model's output [21], which is well illustrated in Figure 4.7. A more comprehensive interpretation of deep learning model outputs would be achievable by estimating uncertainty on soft segmentation instead. Soft segmentation could also alleviate issues encountered with some uncertainty metrics which are sensitive to binary outputs [101]. Future investigations could evaluate the benefits of soft training used in combination with uncertainty estimation.

4.6 Conclusion

We introduced SoftSeg, a deep learning training method that can produce soft segmentations instead of the traditional binary segmentations. SoftSeg leads to informative and relevant soft outputs well calibrated while demonstrating an increase of performance on three open-source medical imaging segmentation tasks. Although used here with a simple 2D U-net as a proof-of-concept, SoftSeg can easily be integrated within already-existing deep learning architectures. Besides, SoftSeg could be leveraged to exploit a lossless combination of ground truth from multiple expert raters or to incorporate uncertainty estimation into an end-to-end soft framework.

4.7 Acknowledgements

The authors thank Joseph Paul Cohen, Olivier Vincent, Lucas Rouhier, Marie-Hélène Bourget, and Leander Van Eekelen for useful discussions and proof-reading this manuscript. The authors also wish to thank the associate editor and two anonymous reviewers for their constructive comments that improved this manuscript. This study was funded by the Canada Research Chair in Quantitative Magnetic Resonance Imaging [950-230815], the Canadian Institute of Health Research [CIHR FDN-143263], the Canada Foundation for Innovation [32454, 34824], the Fonds de Recherche du Québec - Santé [28826], the Natural Sciences and

Engineering Research Council of Canada [RGPIN-2019-07244], the Canada First Research Excellence Fund (IVADO and TransMedTech), the Courtois NeuroMod project and the Quebec BioImaging Network [5886, 35450]. C.G has a fellowship from IVADO [EX-2018-4], A.L. has a fellowship from NSERC and FRQNT. The authors thank the NVIDIA Corporation for the donation of a Titan X GPU and Compute Canada for granting access to its GPU infrastructure.

CHAPTER 5 ARTICLE 2: LABEL FUSION AND TRAINING METHODS FOR RELIABLE REPRESENTATION OF INTER-RATER UNCERTAINTY

Submitted to Journal of Machine Learning for Biomedical Imaging

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

Authors Andreanne Lemay[1,2], Charley Gros[1,2], Julien Cohen-Adad[1,2,3]

Affiliations

[1] NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada

[2] Mila, Quebec AI Institute, Montreal, Qc, Canada

[3] Functional Neuroimaging Unit, CRIUGM, Université de Montréal, Montreal, QC, Canada

Corresponding author

Julien Cohen-Adad
Dept. Genie Electrique, L5610
Ecole Polytechnique
2900 Edouard-Montpetit Bld
Montreal, QC, H3T 1J4, Canada
Phone: 514 340 5121 (office: 2264);
e-mail: jcohen@polymtl.ca

Abbreviations

BraTS: brain tumor segmentation
AVD: absolute volume difference
ECE: expected calibration error
GT: ground truth
MAE: mean absolute error
MRI: magnetic resonance imaging
MS: multiple sclerosis
ReLU: rectified linear unit
RVD: relative volume difference
SCGM: spinal cord gray matter
STAPLE: simultaneous truth and performance level estimation

5.1 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. Our code is available at `https://ivadomed.org`.

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

5.2 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 [7] as it hinders the definition of

ground truth (GT) annotation [63, 99, 104]. 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 [105]. This variability can arise from many factors, including image quality, expert experience, or guidelines clarity [7,63]. To mitigate this issue along with speeding annotating time and enhancing reproducibility, a large number of automatic annotation algorithms have been proposed [11, 23, 54, 106]. 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 [107]. Therefore, it is common practice to provide, for each image, annotations from multiple experts [7,63,69,70]. 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 [63].

5.2.1 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) [64], 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) [13]. 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 [8, 9, 63].

5.2.2 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) [8]. 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 [9]. 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.

5.2.3 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 [104] 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 [18], 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 5.1), are compared on two MRI open-source datasets, the spinal cord gray matter (SCGM) challenge [69] and multiple sclerosis (MS) brain lesion challenge [70], in terms of (i) preservation of the inter-rater variability, (ii) model calibration and, (iii) segmentation performance.

5.3 Method and Material

5.3.1 Method

Label fusion

Three methods to exploit multiple rater labels were studied: STAPLE [64], average across GTs, and random sampling of one annotation during training without fusion [9]. STAPLE is an expectation-maximization algorithm widely used for label fusion in medical imaging [49,70,108]. 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 [13] 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. Therefore, the random sampling method uses binary segmentations.

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 [13]. The final activation and the regression loss function are normalized ReLU and Adaptive Wing loss [68] respectively as defined in [13]. 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.

Table 5.1 Candidates' description.

The columns indicate the label fusion method while the rows present the training framework.

		Label fusion method		
		STAPLE	**Average**	**Random Sampling**
Training framework	**Conventional**	Conv-STAPLE	Conv-Average	Conv-RandomSampling
	SoftSeg	SoftSeg-STAPLE	SoftSeg-Average	SoftSeg-RandomSampling

Training protocol

All candidates were trained on 2D U-Net models. Training parameters for this work were the same as the one described in [13] for the SCGM and MS brain lesion datasets. The processing, training and evaluation pipeline is based on the open-source framework `ivadomed.org` [35].

5.3.2 Datasets

Two publicly available datasets with multiple raters were used to study label fusion: the SCGM challenge [69] and MS brain lesion challenge [70].

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 [69].

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 [70].

5.3.3 Evaluation

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.

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 5.1) [9] which can be directly compared with the entropy associated to the multiple rater segmentation (GT average). 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.

$$H = -\sum_{i=0}^{N_{vox}} p_i log(p_i) \tag{5.1}$$

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 5.2) 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.

$$Brierscore = \frac{1}{N_{vox}} \sum_{i=0}^{N_{vox}} (y_i - \hat{y}_i)^2 \tag{5.2}$$

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

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 [109] and the expected calibration error (ECE) [110] were computed with the code from google-research repository[1] as used in [12] 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 5.3) in relation to the model's confidence (Equation 5.4). 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 ($\frac{1}{K}$). 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 $[\frac{k}{K}, \frac{k+1}{K})$ [12]. This accuracy is then compared to the average

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

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.

$$Accuracy(B_k) = \frac{\sum_{i \in B_k} 1(y_i = \hat{y}_i)}{\#B_k} \tag{5.3}$$

$$Confidence(B_k) = \frac{\sum_{i \in B_k} \hat{y}_i}{\#B_k} \tag{5.4}$$

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 5.5) 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 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.

$$ECE = \sum_{k=0}^{K} \frac{\#B_k}{N_{vox}} \Big| Accuracy(B_k) - Confidence(B_k) \Big| \tag{5.5}$$

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

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. 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

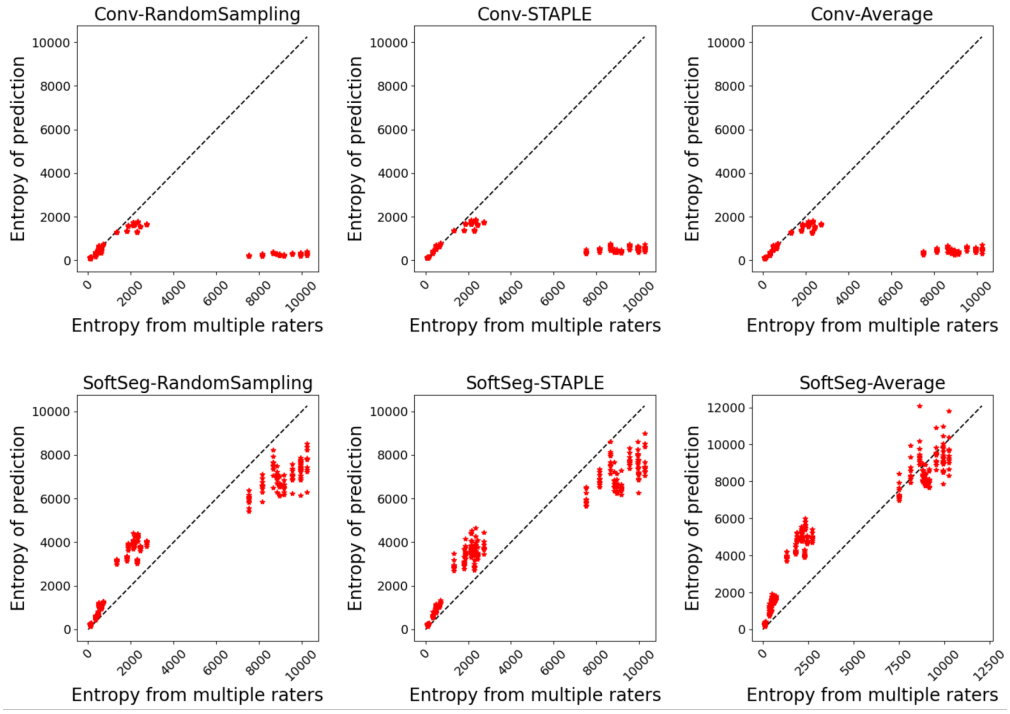
with an increment of 0.05) in terms of Dice score as done in [13]. For the SCGM dataset, there are three classes: gray matter, white matter and background. The predicted class is obtained by selecting the maximum prediction across the three classes.

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).

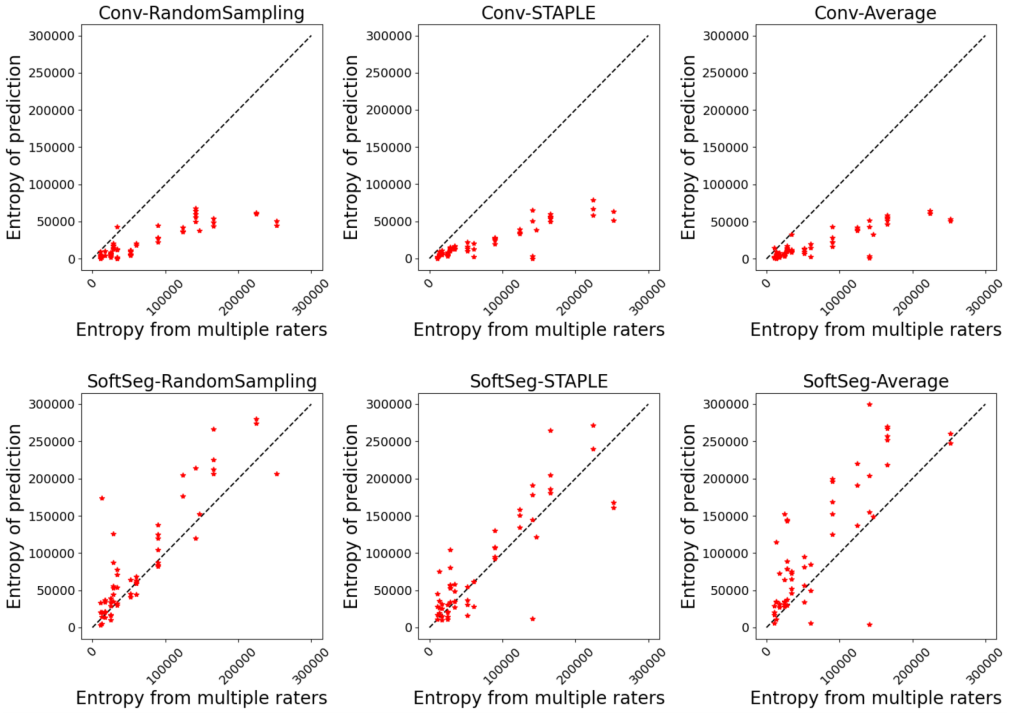
5.4 Results

5.4.1 Inter-rater uncertainty

The predicted segmentation should ideally reflect the uncertainty associated with the disagreement between experts. Figure 5.1 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 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 5.1a and are associated with the different data centers of the SCGM dataset.



(a) SCGM



(b) MS brain

Figure 5.1 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.

Table 5.2 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 220% and 34% higher compared with SoftSeg models for the SCGM and MS brain datasets respectively.

Table 5.2 Quantitative assessment of the inter-rater variability preservation on the SCGM and MS brain datasets (MEAN ± 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 (x10^3) *Opt. value: 0*		**MAE** (x10^2) *Opt. value: 0*	**Brier score** (x10^3) *Opt. value: 0*	**MAE** (x10^4) *Opt. value: 0*
		GM	**WM**	**Entire image**	**MS lesions**	**Entire image**
Conventional	**STAPLE (ref)**	1.59 ± 1.19	4.86 ± 4.31	23.7 ± 0.2	39.9 ± 30.1	5.3 ± 2.9
	Average	1.70 ± 1.05 **	4.71 ± 4.17 **	23.3 ± 0.1 **	40.5 ± 26.0	5.4 ± 3.1
	Random Sampling	1.46 ± 0.89	4.36 ± 3.93 **	23.7 ± 0.1 **	42.4 ± 19.8	4.9 ± 2.5
SoftSeg	**STAPLE**	1.33 ± 1.01 **	4.22 ± 3.82 **	**9.8 ± 0.6 ****	29.2 ± 13.8 **	**3.0 ± 2.6 ****
	Average	1.49 ± 0.86	3.96 ± 3.66 **	11.3 ± 0.9 **	28.4 ± 14.7 **	5.2 ± 3.2
	Random Sampling	**1.21 ± 0.76 ****	**3.83 ± 3.71 ****	10.8 ± 0.8 **	**25.7 ± 12.2 ****	3.4 ± 3.9 **

5.4.2 Visual assessment

Figure 5.2 and Figure 5.3 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 5.2 at the extremity of the dorsal horns and near the central canal (black arrows) and in Figure 5.3 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.

5.4.3 Calibration

Figure 5.4 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 other SoftSeg candidates due to the overly soft predictions encouraged by the non-binary GT.

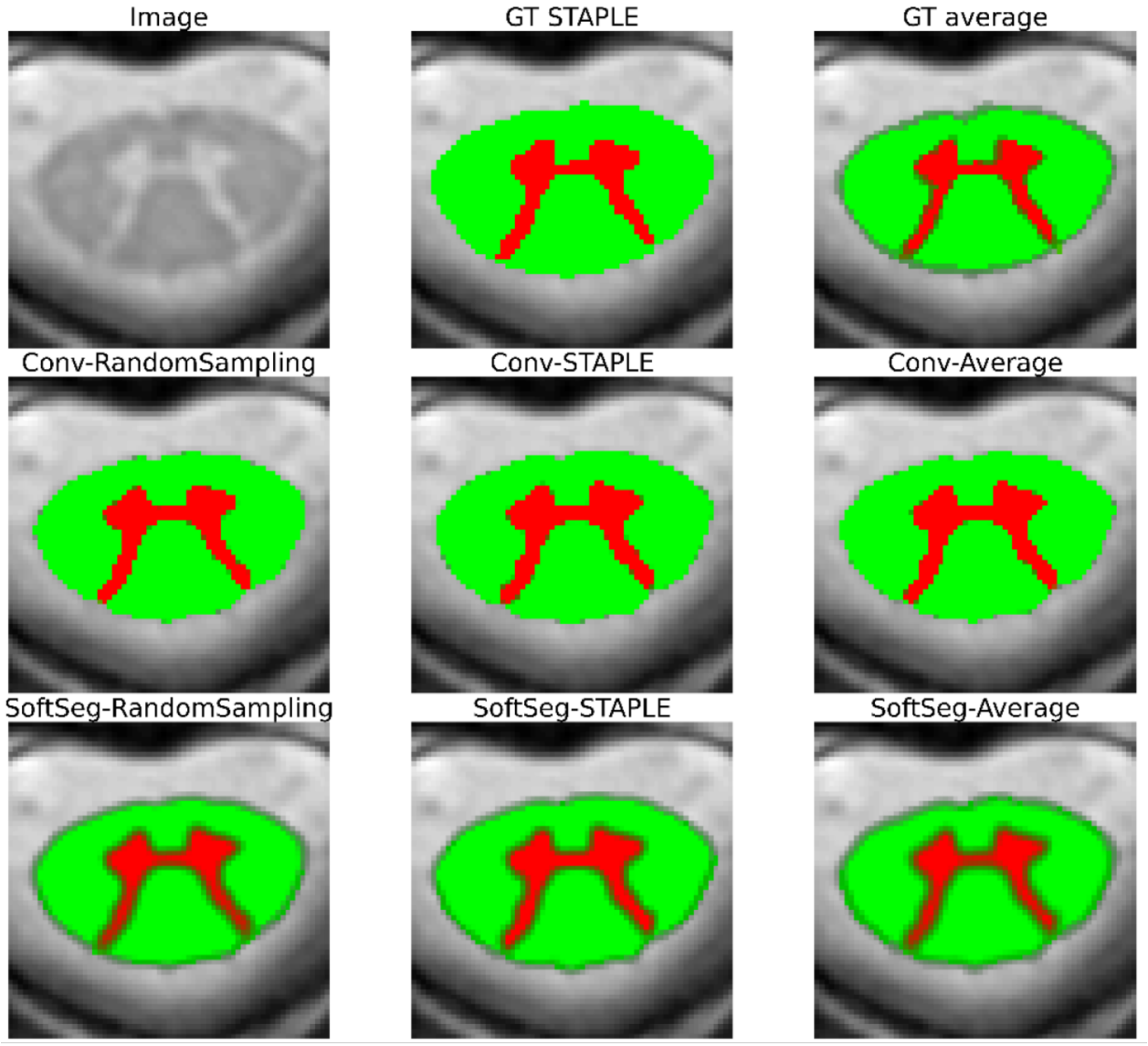


Figure 5.2 Visual assessment of STAPLE and average GTs and predictions from the six candidates on spinal gray and white matter segmentation.

Abbreviations: GT: ground truth.

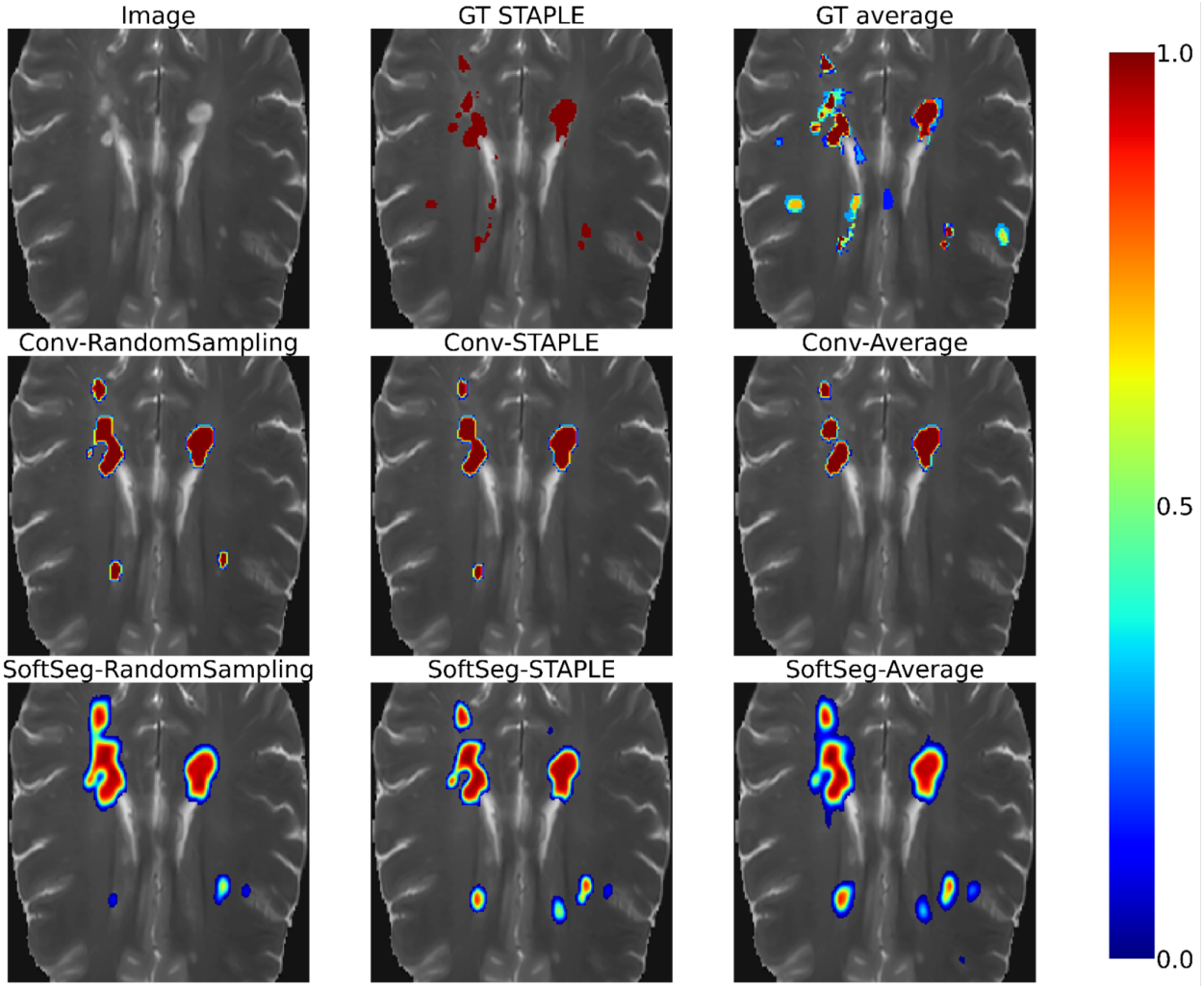
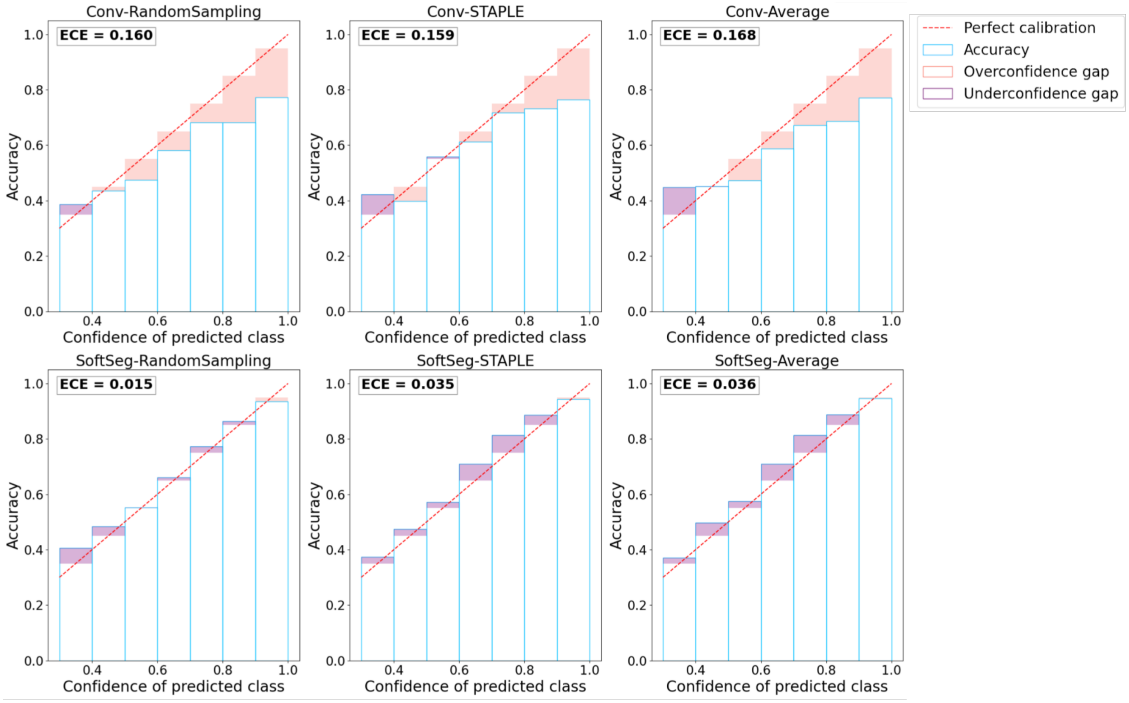
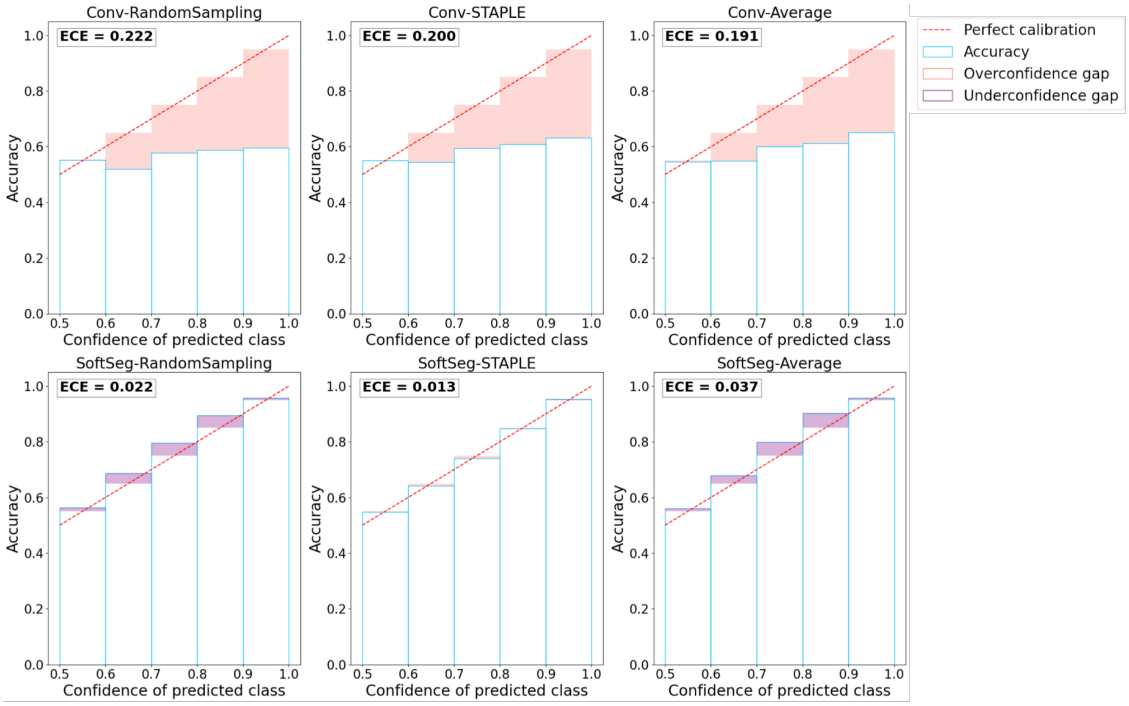


Figure 5.3 Visual assessment of STAPLE and GT average and predictions from the six candidates on MS brain segmentation.

Red: Gray matter. Green: White matter. Voxels at tissue boundaries represent values between 0 and 1. Abbreviations: GT: ground truth.



(a) SCGM



(b) MS brain

Figure 5.4 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.

5.4.4 Segmentation accuracy

Table 5.3a 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 5.5a 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 5.5a). 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 5.3b 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 5.5b summarizes the metrics presented in Table 5.3b 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.

5.5 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

		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 **

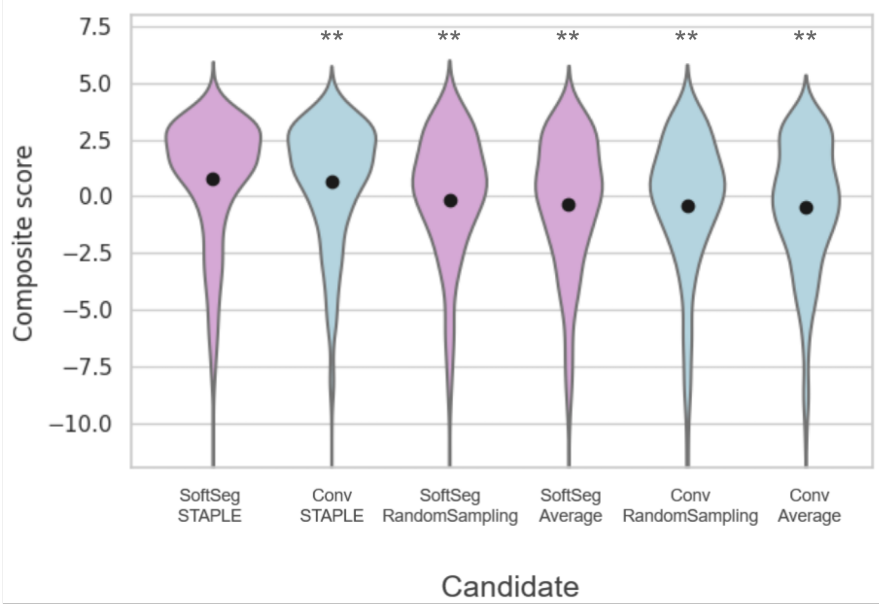
(a) SCGM

		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 ± 13.0 **	42.0 ± 57.1	-11.8 ± 65.5
	Random Sampling	50.6 ± 11.8 **	61.4 ± 13.4	52.9 ± 12.6	54.8 ± 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 ± 9.4 **	39.7 ± 50.2	-8.3 ± 59.2

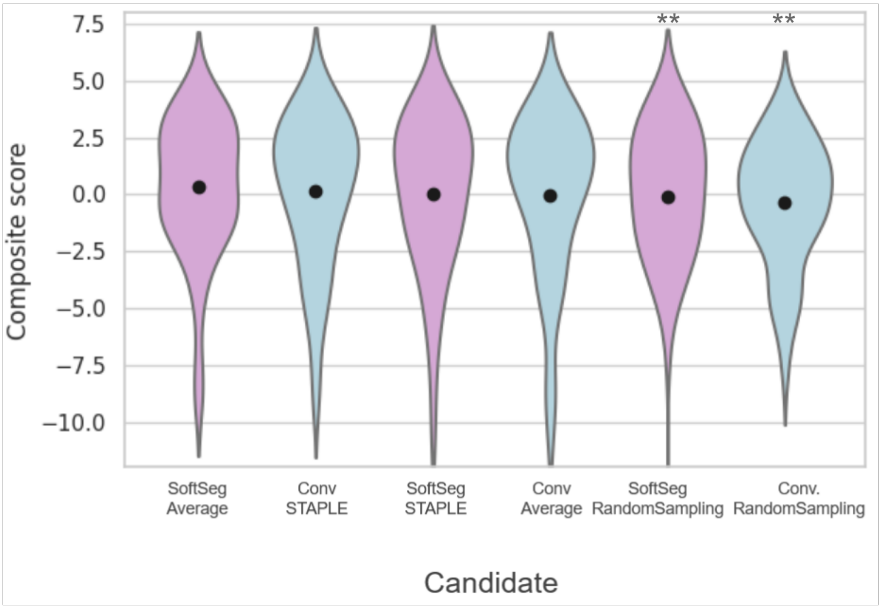
(b) MS brain

Table 5.3 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



(b) MS brain

Figure 5.5 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 spittings. 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 (**: $p < 0.05$).

discuss the importance of repeatability in medical deep learning research.

5.5.1 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.

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 5.2 and Figure 5.3), the larger underconfidence gaps on reliability diagrams and the associated higher ECE (Figure 5.4), and uncertainty correspondence plots (Figure 5.1). 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 [8, 111]. 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 5.4), 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 5.1). 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 [9]. 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 [8, 9, 63], 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. Future studies could also consider other label fusion methods, such as recently proposed deep-learning approaches to explicitly model the consensus process [104,112].

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. 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 [36], or by the use of expertise-aware inferring module [111].

5.5.2 A multifaceted evaluation with model training repetitions

While it is common to select the best model solely based on segmentation accuracy considerations [69, 70, 106], 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 5.5), but is not properly calibrated as it showed an important overconfident gap (see Figure 5.4). A multifaceted evaluation scheme has the potential to facilitate model acceptance and integration in the clinical routine, which still remains limited [9]. Some avenues are discussed below.

The ongoing research around uncertainty and calibration estimation

In the same way that there are numerous segmentation accuracy metrics [113], there are many ways to assess the model uncertainty and calibration. For instance, recent studies suggested the voxel-wise aleatoric [114] and epistemic [115] uncertainties, or the structure-wise uncertainty [116], 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 [12, 51, 52]. We acknowledge the exhaustive comparison performed by [53] 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 [12]. While multiple studies suggest post-hoc strategies to improve calibration [12, 100, 117], we suggest a training strategy that directly generates calibrated outputs without the need of extra computation or hyperparameters.

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. Another avenue would be to represent the performance across the different criteria using a radar visualization, e.g., used by [118].

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.

5.6 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, SoftSeg was systematically superior or equal in terms of segmentation performance and had the best calibration and preservation of the inter-rater variability. 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.

5.7 Acknowledgements

The authors thank Marie-Hélène Bourget for her methodological insights and Yang Ding, Nick Guenther, Joshua Newton, Ainsleigh Hill, and Alexandru Foias for helping with ivadomed maintenance. Funded by the Canada Research Chair in Quantitative Magnetic Resonance Imaging [950-230815], the Canadian Institute of Health Research [CIHR FDN-143263], the Canada Foundation for Innovation [32454, 34824], the Fonds de Recherche du Québec - Santé [322736], the Natural Sciences and Engineering Research Council of Canada [RGPIN-2019-07244], the Canada First Research Excellence Fund (IVADO and TransMedTech), the Courtois NeuroMod project, the Quebec BioImaging Network [5886, 35450], INSPIRED (Spinal Research, UK; Wings for Life, Austria; Craig H. Neilsen Foundation, USA), Mila - Tech Transfer Funding Program. A.L. has a fellowship from Centre UNIQUE, NSERC and FRQNT. C.G. has a fellowship from IVADO [EX-2018-4], The authors thank the NVIDIA Corporation for the donation of a Titan X GPU.

CHAPTER 6 GENERAL DISCUSSION

Chapter 4 and 5 presented SoftSeg and analyzed its relevance for medical imaging segmentation. The following chapter will highlight the clinical usefulness of this approach, some limitations of the work, and finally, perspectives arising from this Master's project.

6.1 Clinical usefulness

6.1.1 Inter-rater and uncertainty representation

Several aspects of SoftSeg make this approach clinically relevant compared with the conventional training framework. First of all, when trained on labels derived from multiple raters, the SoftSeg models' segmentation will preserve the total and spatial inter-rater uncertainty more truthfully. Even for labels with a single rater, SoftSeg models were better-calibrated, indicating more truthfully uncertain areas. Highlighting structures or lesions with high uncertainty, i.e., which corresponds to voxels with low values, enables clinicians to identify cases where the model is likely wrong or prone to inter-rater disagreement. For instance, predicted MS lesions with uncertainty over a certain threshold should be isolated and verified by experts.

6.1.2 Mitigation of volumetric bias for morphometric measures

Secondly, outputting soft predictions can reduce the volumetric bias for morphometric measurements [42]. Voxels at the junction of different tissue types should be labeled 0 or 1, but rather a value representing the tissue ratio. This feature can help having more precise measurements for disease monitoring such as tumor growth or MS lesions spread.

6.1.3 Possibility of using soft labels encoding expert uncertainty

Thirdly, SoftSeg enables the use of soft labels, which can be optimally used for training due to the regression loss, in which the experts' confidence can be encoded. Examples of soft GT could be the average of several expert labels as described in Chapter 5 or annotations determined by the expert between 0 and 1. For instance, for MS lesions segmentation, the radiologist could annotate a lesion with 0.5 to express the expert's uncertainty.

6.1.4 Segmentation performance

Finally, the results from Chapter 4 and 5 indicate improved segmentation performance for raw and binarized predictions when looking at numerous metrics such as Dice, absolute and relative volume differences, lesion false and positive detection rates, precision, and recall. The calibration, volumetric and uncertainty representation advantages of SoftSeg come at no cost of segmentation performance, often even increasing it.

6.2 Limitations

6.2.1 Small and limited datasets

The conclusions presented in the two articles were drawn from only three MRI datasets with relatively few patients, i.e., from 15 to 80. Expanding the study to more datasets with more patients from other medical modalities such as CT or non-medical tasks would confirm the robustness of SoftSeg. This method could potentially improve segmentation performance in non-medical segmentation tasks even though it was developed to address specific issues faced in the medical field.

6.2.2 Analysis focused on Dice loss

In this work, the conventional framework was associated with the Dice loss. While this loss function is widely used for segmentation models, cross-entropy, for instance, is also very popular [29–33]. Dice loss is known for its very sharp edges, and its high segmentation performance [30, 40, 41]. Comparing SoftSeg to the conventional approach with the cross-entropy loss would be interesting. Other authors reported softer edges with cross-entropy [30]. The main reason motivating the Dice loss choice was superior segmentation performance for the task studied. Uncertainty and calibration metrics might be better with cross-entropy than with the Dice loss. However, even with cross-entropy loss, the use of soft labels is not optimal due to the classification loss (see Section 2.1.2).

6.3 Perspectives

Due to time and resource limitations, some avenues were not explored but would be worth studying.

6.3.1 Volumetric analysis

As mentioned throughout this work, one motivation for introducing SoftSeg was accountability for PVE. However, volume preservation was merely studied. AVD and RVD, which are metrics to compute volume differences, were calculated on binarized predictions in both Chapter 4 and 5. However, no volumetric analysis was performed on soft labels. One hypothesis is that soft labels could represent the partial volume information, but this research avenue was not fully explored. To achieve this type of analysis, a dataset of structures with known volumes, e.g., phantoms, [47] would be ideal. The true volume could be compared to the predicted volume from SoftSeg or conventional models. Another approach would be to estimate the PVE using statistical models [47, 119] and then compare volumetric differences. This analysis is less precise but does not require a dataset with known volumes.

6.3.2 Modification of the SoftSeg framework

In Chapter 5, training with soft labels from the average of expert labels, i.e., GT average, was explored. However, results suggest visually too much softness compared to the GT average and underconfidence from the model. The GT was processed with a first-order interpolation which is known to create a blurring effect [66]. The SoftSeg pipeline could be improved by choosing another interpolation algorithm, such as cubic spline or nearest neighbor, that would reduce the smoothing effect for labels that are already soft. Moreover, only one regression has been tested for the SoftSeg framework, the Adaptive Wing loss. The MSE, Wing, or weighted loss map could be explored as alternatives to the Adaptive Wing loss chosen for this work.

CHAPTER 7 CONCLUSION

7.1 Summary of works

In this work, we proposed a new segmentation training framework named SoftSeg to improve segmentation performance and uncertainty representation. Three main components of the conventional training framework were modified (i) the use of soft labels for training, (ii) the final activation to a normalized linear function, and (iii) the training of the model with a regression loss. Combining these three modifications outperformed the conventional approach in terms of segmentation performance. A second article presented in Chapter 5 compared label fusion methods combined with SoftSeg or the conventional training framework. SoftSeg models were systematically better-calibrated and preserved more truthfully the inter-rater variability while having improved, or minimally equivalent, segmentation performances. No label fusion method consistently obtained the best performance across datasets or metrics. While SoftSeg has the potential to reduce volumetric bias by representing partial volume effect, this avenue was not fully explored and should be the subject of future work. The SoftSeg approach was implemented in the DL image analysis framework ivadomed (`https://ivadomed.org`) [35]. As part of the MICCAI MS new lesions challenge [120], Macar et al. proposed segmentation models using SoftSeg features for detection of new MS lesions on longitudinal images of the same patient [121].

The work done during my Master's was not restricted to the articles presented in this thesis, but due to space limitation, some research projects were not included. Here is a summary of the publications or submitted work completed during my graduate studies:

Submitted with NeuroPoly

- A. Lemay, C. Gros, and J. Cohen-Adad, "Label fusion and training methods for reliable representation of inter-rater uncertainty," *arXiv preprint arXiv:2202.07550*, 2022.
- A. Lemay, C. Gros, O. Vincent, Y. Liu, J. Cohen, and J. Cohen-Adad, "Benefits of Linear Conditioning for Segmentation using Metadata," in *Medical Imaging with Deep Learning (MIDL)*. PMLR, 2021, pp. 416-430.
- A. Lemay, C. Gros, Z. Zhuo, J. Zhang, Y. Duan, J. Cohen-Adad, and Y. Liu, "Automatic multiclass intramedullary spinal cord tumor segmentation on mri with deep learning," *NeuroImage: Clinical*, vol. 31, p. 102766, 2021.

- . Gros, A. Lemay, O. Vincent, L. Rouhier, M.-H. Bourget, A. Bucquet, J. P. Cohen, and J. Cohen-Adad, “ivadomed: A medical imaging deep learning toolbox,” *Journal of Open Source Software*, vol. 6, no. 58, p. 2868, 2021. [Online]. Available: https://doi.org/10.21105/joss.02868
- C. Gros, A. Lemay, and J. Cohen-Adad, “Softseg: Advantages of soft versus binary training for image segmentation,” *Medical Image Analysis*, vol. 71, p. 102038, 2021.
- U. Macar, E. N. Karthik, C. Gros, A. Lemay, and J. Cohen-Adad, “Team neuropoly: Description of the pipelines for the miccai 2021 ms new lesions segmentation challenge,” arXiv preprint arXiv:2109.05409, 2021.

Submitted with Harvard University

- A. Lemay, K. Hoebel, C. Bridge, B. Befano, S. De Sanjose, D. Egemen, A. Rodriguez, M. Schiffman, J. Campbell, and J. Kalpathy-Cramer. "Improving the repeatability of deep learning models with Monte Carlo dropout," *arXiv preprint arXiv:2202. 07562*, 2022.
- C. Lu, A. Lemay, K. Chang, K. Hoebel, and J. Kalpathy-Cramer. "Fair Conformal Predictors for Applications in Medical Imaging," in *AAAI Workshops*, 2022.
- K. Hoebel, C. Bridge, A. Lemay, K. Chang, J. Patel, B. Rosen, and J. Kalpathy-Cramer. "Do I know this? segmentation uncertainty under domain shift," in *SPIE Medical Imaging*, 2022.
- A. Lemay, K. Hoebel, C. Bridge, D. Egemen, A. Rodriguez, M. Schiffman, J. Campbell, and J. Kalpathy-Cramer. "Monte Carlo dropout increases model repeatability," in *Machine Learning for Health (ML4H) at NeurIPS*, 2021.
- C. Lu, A. Lemay, K. Hoebel, and J. Kalpathy-Cramer. "Evaluating subgroup disparity using epistemic uncertainty in mammography," in *International Conference on Machine Learning (ICML): Workshop on Interpretable Machine Learning in Healthcare*, 2021.

7.2 Recommendations

In light of this work, we recommend testing SoftSeg for segmentation neural networks even for applications requiring binarized output, as both articles suggest better performance for binary masks. However, SoftSeg is more interesting on soft predictions as there is uncertainty,

and potentially, partial volume information. A final recommendation would be to encourage experiment repetition, i.e., training a model multiple times with the same parameters, to evaluate performance variation and ensure statistical differences since high variation in metrics can arise, especially with small datasets.

REFERENCES

[1] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[2] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[3] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–9, 2021.

[4] S. Yeasmin, "Benefits of artificial intelligence in medicine," in *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*. IEEE, 2019, pp. 1–6.

[5] S. Yadav, A. Kaushik, and S. Sharma, "Simplify the difficult: Artificial intelligence and cloud computing in healthcare," *IoT and Cloud Computing for Societal Good*, pp. 101–124, 2022.

[6] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[7] M. Schaekermann, G. Beaton, M. Habib, A. Lim, K. Larson, and E. Law, "Understanding expert disagreement in medical data analysis through structured adjudication," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–23, 2019.

[8] M. H. Jensen, D. R. Jørgensen, R. Jalaboi, M. E. Hansen, and M. A. Olsen, "Improving uncertainty estimation in convolutional neural networks using inter-rater agreement," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 540–548.

[9] A. Jungo, R. Meier, E. Ermis, M. Blatti-Moreno, E. Herrmann, R. Wiest, and M. 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*. Springer, 2018, pp. 682–690.

[10] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," *Annual review of biomedical engineering*, vol. 2, no. 1, pp. 315–337, 2000.

[11] A. Lemay, C. Gros, Z. Zhuo, J. Zhang, Y. Duan, J. Cohen-Adad, and Y. Liu, "Automatic multiclass intramedullary spinal cord tumor segmentation on mri with deep learning," *NeuroImage: Clinical*, vol. 31, p. 102766, 2021.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.

[13] C. Gros, A. Lemay, and J. Cohen-Adad, "Softseg: Advantages of soft versus binary training for image segmentation," *Medical image analysis*, vol. 71, p. 102038, 2021.

[14] A. Lemay, C. Gros, and J. Cohen-Adad, "Label fusion and training methods for reliable representation of inter-rater uncertainty," *arXiv preprint arXiv:2202.07550*, 2022.

[15] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[16] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, A. Oliver, and X. Lladó, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, 2017.

[17] C. S. Perone, E. Calabrese, and J. Cohen-Adad, "Spinal cord gray matter segmentation using deep dilated convolutions," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[19] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of digital imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[20] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Scientific Reports*, vol. 10, no. 1, pp. 1–16, 2020.

[21] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[22] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnu-net for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2020, pp. 118–132.

[23] C. Gros, B. De Leener, A. Badji, J. Maranzano, D. Eden, S. M. Dupont, J. Talbott, R. Zhuoquiong, Y. Liu, T. Granberg *et al.*, "Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks," *Neuroimage*, vol. 184, pp. 901–915, 2019.

[24] M. Livne, J. Rieger, O. U. Aydin, A. A. Taha, E. M. Akay, T. Kossen, J. Sobesky, J. D. Kelleher, K. Hildebrand, D. Frey *et al.*, "A u-net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease," *Frontiers in neuroscience*, p. 97, 2019.

[25] Z. Eaton-Rosen, F. Bragman, S. Ourselin, and M. J. Cardoso, "Improving data augmentation for medical image segmentation," 2018.

[26] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE transactions on medical imaging*, vol. 38, no. 4, pp. 1016–1025, 2018.

[27] Z. Zhao, L. Yang, H. Zheng, I. H. Guldner, S. Zhang, and D. Z. Chen, "Deep learning based instance segmentation in 3d biomedical images using weak annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 352–360.

[28] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[29] M. Akil, R. Saouli, R. Kachouri *et al.*, "Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy," *Medical image analysis*, vol. 63, p. 101692, 2020.

[30] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep learning and data labeling for medical applications*. Springer, 2016, pp. 179–187.

[31] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2020, pp. 1–7.

[32] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, p. 102026, 2022.

[33] X. Chen, B. M. Williams, S. R. Vallabhaneni, G. Czanner, R. Williams, and Y. Zheng, "Learning active contour models for medical image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 632–11 640.

[34] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[35] C. Gros, A. Lemay, O. Vincent, L. Rouhier, M.-H. Bourget, A. Bucquet, J. P. Cohen, and J. Cohen-Adad, "ivadomed: A medical imaging deep learning toolbox," *Journal of Open Source Software*, vol. 6, no. 58, p. 2868, 2021. [Online]. Available: https://doi.org/10.21105/joss.02868

[36] A. Lemay, C. Gros, O. Vincent, Y. Liu, J. P. Cohen, and J. Cohen-Adad, "Benefits of linear conditioning for segmentation using metadata," in *Medical Imaging with Deep Learning*. PMLR, 2021, pp. 416–430.

[37] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu *et al.*, "Niftynet: a deep-learning platform for medical imaging," *Computer methods and programs in biomedicine*, vol. 158, pp. 113–122, 2018.

[38] C. Shen, H. R. Roth, H. Oda, M. Oda, Y. Hayashi, K. Misawa, and K. Mori, "On the influence of dice loss function in multi-class organ segmentation of abdominal ct using 3d fully convolutional networks," *arXiv preprint arXiv:1801.05912*, 2018.

[39] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in

Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, 2017, pp. 240–248.

[40] R. Deng, C. Shen, S. Liu, H. Wang, and X. Liu, "Learning to predict crisp boundaries," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 562–578.

[41] S. Jia, A. Despinasse, Z. Wang, H. Delingette, X. Pennec, P. Jaïs, H. Cochet, and M. Sermesant, "Automatically segmenting the left atrium from cardiac images using successive 3d u-nets and a contour loss," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 221–229.

[42] J. Bertels, D. Robben, D. Vandermeulen, and P. Suetens, "Optimization with soft dice can lead to a volumetric bias," in *International MICCAI Brainlesion Workshop*. Springer, 2019, pp. 89–97.

[43] J. Zhang, X. Shen, T. Zhuo, and H. Zhou, "Brain tumor segmentation based on refined fully convolutional neural networks with a hierarchical dice loss," *arXiv preprint arXiv:1712.09093*, 2017.

[44] O. Kodym, M. Španěl, and A. Herout, "Segmentation of head and neck organs at risk using cnn with batch dice loss," in *German conference on pattern recognition*. Springer, 2018, pp. 105–114.

[45] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[46] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[47] M. A. G. Ballester, A. P. Zisserman, and M. Brady, "Estimation of the partial volume effect in mri," *Medical image analysis*, vol. 6, no. 4, pp. 389–405, 2002.

[48] B. Billot, E. Robinson, A. V. Dalca, and J. E. Iglesias, "Partial volume segmentation of brain mri scans of any resolution and contrast," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2020, pp. 177–187.

[49] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep learning for brain mri segmentation: state of the art and future directions," *Journal of digital imaging*, vol. 30, no. 4, pp. 449–459, 2017.

[50] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu, "Phiseg: Capturing uncertainty in medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 119–127.

[51] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[52] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[53] A. Jungo, F. Balsiger, and M. Reyes, "Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation," *Frontiers in neuroscience*, vol. 14, p. 282, 2020.

[54] R. E. Gabr, I. Coronado, M. Robinson, S. J. Sujit, S. Datta, X. Sun, W. J. Allen, F. D. Lublin, J. S. Wolinsky, and P. A. Narayana, "Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study," *Multiple Sclerosis Journal*, vol. 26, no. 10, pp. 1217–1226, 2020.

[55] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[56] A. Jungo and M. Reyes, "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 48–56.

[57] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in neural information processing systems*, vol. 32, 2019.

[58] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[59] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[60] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in neural information processing systems*, vol. 32, 2019.

[61] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest x-rays via cnns that exploit disease dependencies and uncertainty labels," *medRxiv*, p. 19013342, 2019.

[62] H. Li, D. Wei, S. Cao, K. Ma, L. Wang, and Y. Zheng, "Superpixel-guided label softening for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 227–237.

[63] Z. Mirikharaji, K. Abhishek, S. Izadi, and G. 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*, 2021, pp. 1837–1846.

[64] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE transactions on medical imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[65] E. Kats, J. Goldberger, and H. Greenspan, "A soft staple algorithm combined with anatomical knowledge," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 510–517.

[66] J. A. Parker, R. V. Kenyon, and D. E. Troxel, "Comparison of interpolating methods for image resampling," *IEEE Transactions on medical imaging*, vol. 2, no. 1, pp. 31–39, 1983.

[67] Q. Li, I. Sato, and Y. Murakami, "Interpolation effects on accuracy of mutual information based image registration," in *2006 IEEE International Symposium on Geoscience and Remote Sensing*. IEEE, 2006, pp. 180–183.

[68] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6971–6981.

[69] F. Prados, J. Ashburner, C. Blaiotta, T. Brosch, J. Carballido-Gamio, M. J. Cardoso, B. N. Conrad, E. Datta, G. Dávid, B. De Leener *et al.*, "Spinal cord grey matter segmentation challenge," *Neuroimage*, vol. 152, pp. 312–329, 2017.

[70] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. C. Pop, P. Girard, R. Ameli, J.-C. Ferré *et al.*, "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Scientific reports*, vol. 8, no. 1, pp. 1–17, 2018.

[71] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[72] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[73] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[74] A. Carass, S. Roy, A. Gherman, J. C. Reinhold, A. Jesson, T. Arbel, O. Maier, H. Handels, M. Ghafoorian, B. Platel *et al.*, "Evaluating white matter lesion segmentations with refined sørensen-dice analysis," *Scientific reports*, vol. 10, no. 1, pp. 1–19, 2020.

[75] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Medical image analysis*, vol. 59, p. 101557, 2020.

[76] R. Tam, A. Traboulsee, A. Riddehough, F. Sheikhzadeh, and D. Li, "The impact of intensity variations in t1-hypointense lesions on clinical correlations in multiple sclerosis," *Multiple Sclerosis Journal*, vol. 17, no. 8, pp. 949–957, 2011.

[77] H. Chaves, F. Dorr, M. E. Costa, M. M. Serra, D. F. Slezak, M. F. Farez, G. Sevlever, P. Yanez, and C. Cejas, "Brain volumes quantification from mri in healthy controls: Assessing correlation, agreement and robustness of a convolutional neural network-based software against freesurfer, cat12 and fsl," *Journal of Neuroradiology*, vol. 48, no. 3, pp. 147–156, 2021.

[78] S. Lévy, M. Benhamou, C. Naaman, P. Rainville, V. Callot, and J. Cohen-Adad, "White matter atlas of the human spinal cord with estimation of partial volume effect," *Neuroimage*, vol. 119, pp. 262–271, 2015.

[79] J. Tohka, A. Zijdenbos, and A. Evans, "Fast and robust parameter estimation for statistical partial volume models in brain mri," *Neuroimage*, vol. 23, no. 1, pp. 84–97, 2004.

[80] J. V. Manjón, J. Tohka, and M. Robles, "Improved estimates of partial volume coefficients from noisy brain mri using spatial context," *Neuroimage*, vol. 53, no. 2, pp. 480–490, 2010.

[81] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "A unifying framework for partial volume segmentation of brain mr images," *IEEE transactions on medical imaging*, vol. 22, no. 1, pp. 105–119, 2003.

[82] X. Li, L. Li, H. Lu, and Z. Liang, "Partial volume segmentation of brain magnetic resonance images based on maximum a posteriori probability," *Medical Physics*, vol. 32, no. 7Part1, pp. 2337–2345, 2005.

[83] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[84] H.-H. Zhao, P. L. Rosin, Y.-K. Lai, and Y.-N. Wang, "Automatic semantic style transfer using deep convolutional neural networks and soft masks," *The Visual Computer*, vol. 36, no. 7, pp. 1307–1324, 2020.

[85] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *arXiv preprint arXiv:1612.02695*, 2016.

[86] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[87] C. Kaul, N. Pears, H. Dai, R. Murray-Smith, and S. Manandhar, "Penalizing small errors using an adaptive logarithmic loss," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 368–375.

[88] A. Akhondi-Asl, L. Hoyte, M. E. Lockhart, and S. K. Warfield, "A logarithmic opinion pool based staple algorithm for the fusion of segmentations with associated reliability weights," *IEEE transactions on medical imaging*, vol. 33, no. 10, pp. 1997–2009, 2014.

[89] P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot, "An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images," *IEEE transactions on medical imaging*, vol. 27, no. 4, pp. 425–441, 2008.

[90] O. Commowick, N. Wiest-Daesslé, and S. Prima, "Block-matching strategies for rigid registration of multimodal medical images," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2012, pp. 700–703.

[91] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: improved n3 bias correction," *IEEE transactions on medical imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[92] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[93] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[94] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[95] M. Moccia, F. Prados, M. Filippi, M. A. Rocca, P. Valsasina, W. J. Brownlee, C. Zecca, A. Gallo, A. Rovira, A. Gass *et al.*, "Longitudinal spinal cord atrophy in multiple sclerosis using the generalized boundary shift integral," *Annals of Neurology*, vol. 86, no. 5, pp. 704–713, 2019.

[96] S. Abbasi-Sureshjani, S. Amirrajab, C. Lorenz, J. Weese, J. Pluim, and M. Breeuwer, "4d semantic cardiac magnetic resonance image synthesis on xcat anatomical model," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 6–18.

[97] A. Jog, A. Hoopes, D. N. Greve, K. Van Leemput, and B. Fischl, "Psacnn: Pulse sequence adaptive fast whole brain segmentation," *NeuroImage*, vol. 199, pp. 553–569, 2019.

[98] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *International workshop on simulation and synthesis in medical imaging*. Springer, 2018, pp. 1–11.

[99] O. Shwartzman, H. Gazit, I. Shelef, and T. Riklin-Raviv, "The worrisome impact of an inter-rater bias on neural network training," *arXiv preprint arXiv:1906.11872*, 2019.

[100] L. Zhang, R. Tanno, K. Bronik, C. Jin, P. Nachev, F. Barkhof, O. Ciccarelli, and D. C. Alexander, "Learning to segment when experts disagree," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 179–190.

[101] R. Camarasa, D. Bos, J. Hendrikse, P. Nederkoorn, E. Kooi, A. v. d. Lugt, and M. d. Bruijne, "Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*. Springer, 2020, pp. 32–41.

[102] A. Loquercio, M. Segu, and D. Scaramuzza, "A general framework for uncertainty estimation in deep learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3153–3160, 2020.

[103] R. Mehta, A. Filos, Y. Gal, and T. Arbel, "Uncertainty evaluation metric for brain tumour segmentation," *arXiv preprint arXiv:2005.14262*, 2020.

[104] S. Yu, H.-Y. Zhou, K. Ma, C. Bian, C. Chu, H. Liu, and Y. Zheng, "Difficulty-aware glaucoma classification with multi-rater consensus modeling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 741–750.

[105] O. Commowick, M. Kain, R. Casey, R. Ameli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, S. Camarasu-Pop, T. Glatard *et al.*, "Multiple sclerosis lesions segmentation from multiple experts: The miccai 2016 challenge dataset," *NeuroImage*, vol. 244, p. 118589, 2021.

[106] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 287–297.

[107] O. Vincent, C. Gros, and J. Cohen-Adad, "Impact of individual rater style on deep learning uncertainty in medical imaging segmentation," *arXiv preprint arXiv:2105.02197*, 2021.

[108] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen *et al.*, "Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri," *Medical image analysis*, vol. 35, pp. 250–269, 2017.

[109] M. H. DeGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 32, no. 1-2, pp. 12–22, 1983.

[110] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[111] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng, "Learning calibrated medical image segmentation via multi-rater agreement modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 341–12 351.

[112] B. Nichyporuk, J. Szeto, D. L. Arnold, and T. Arbel, "Optimizing operating points for high performance lesion detection and segmentation using lesion size reweighting," *arXiv preprint arXiv:2107.12978*, 2021.

[113] V. Yeghiazaryan and I. Voiculescu, "An overview of current evaluation methods used in medical image segmentation," *Department of Computer Science, University of Oxford*, 2015.

[114] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, 2019.

[115] Y. Xia, D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, and H. Roth, "Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation," *Medical Image Analysis*, vol. 65, p. 101766, 2020.

[116] A. G. Roy, S. Conjeti, N. Navab, C. Wachinger, A. D. N. Initiative *et al.*, "Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control," *NeuroImage*, vol. 195, pp. 11–22, 2019.

[117] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," in *International conference on machine learning*. PMLR, 2018, pp. 2796–2804.

[118] G. Placidi, L. Cinque, F. Mignosi, and M. Polsinelli, "Multiple sclerosis lesions identification/segmentation in magnetic resonance imaging using ensemble cnn and uncertainty classification," *arXiv preprint arXiv:2108.11791*, 2021.

[119] J. Dehmeshki, X. Ye, H. Amin, M. Abaei, X. Lin, and S. D. Qanadli, "Volumetric quantification of atherosclerotic plaque in ct considering partial volume effect," *IEEE Transactions on Medical Imaging*, vol. 26, no. 3, pp. 273–282, 2007.

[120] O. Commowick, F. Cervenansky, F. Cotton, and M. Dojat, "Msseg-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI 2021-24th International Conference on Medical Image Computing and Computer Assisted Intervention*, 2021, pp. 1–118.

[121] U. Macar, E. N. Karthik, C. Gros, A. Lemay, and J. Cohen-Adad, "Team neuropoly: Description of the pipelines for the miccai 2021 ms new lesions segmentation challenge," *arXiv preprint arXiv:2109.05409*, 2021.