Titre: Title:	Impact of Rater Style on Deep Learning Segmentation in Medical Imaging
Auteur: Author:	Olivier Vincent
Date:	2021
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Vincent, O. (2021). Impact of Rater Style on Deep Learning Segmentation in Medical Imaging [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/6283/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/6283/
Directeurs de recherche: Advisors:	Julien Cohen-Adad
Programme: Program:	Génie biomédical

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Impact of rater style on deep learning segmentation in medical imaging

OLIVIER VINCENT
Institut de génie biomédical

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie biomédical

Avril 2021

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Impact of rater style on deep learning segmentation in medical imaging

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

Samuel KADOURY, président
Julien COHEN-ADAD, membre et directeur de recherche
Christopher J. PAL, membre

DEDICATION

To Tobby, my furry friend.

ACKNOWLEDGEMENTS

I would like to thank my advisor Julien Cohen-Adad for the opportunity to be a part of his stimulating and innovative Neuropoly lab. I'm also thankful to my colleagues in the lab, in particular Charley Gros, Andréanne Lemay and Lucas Rouhier for their help, insightful comments and team spirit.

I would also like to thank my friends Jérémie, Lucas and Patrick for our interesting discussions, about life, the Universe, and everything.

Finally, I'm grateful for the support thought the ups and downs from my parents, sister and girlfriend, without which this wouldn't have been possible.

RÉSUMÉ

La sclérose en plaques est la maladie auto-immune la plus courante du système nerveux central. Elle se caractérise par la présence de lésions dans le cerveau et la moelle épinière, visibles en imagerie par résonance magnétique (IRM). Cependant, pour extraire des informations utiles des images, il est nécessaire de segmenter les lésions sur ces images, ce qui est un processus long et coûteux lorsqu'il est réalisé manuellement par un expert tel qu'un radiologue. L'objectif de ce projet est d'utiliser des méthodes innovantes d'apprentissage profond pour améliorer la segmentation des images médicales.

Premièrement, pour aborder la généralisation à travers différents contrastes dans la segmentation de la sclérose en plaques de la moelle épinière, nous mettons en œuvre la modulation linéaire par caractéristique (FiLM) pour tirer parti de la grande variété de paramètres d'acquisitions IRM dans le modèle de segmentation, en apprenant les caractéristiques de chaque contraste. Fait intéressant, un U-Net bien optimisé a atteint les mêmes performances que notre FiLMed-Unet sur un ensemble de données à contrastes multiples (0,72 de score Dice), ce qui suggère qu'il existe un goulot d'étranglement dans cette tâche, qui n'est pas la généralisation à travers différents contrastes. Ce goulot d'étranglement provient probablement de la variabilité interexperts, qui est estimée à 0,61 de score Dice dans notre ensemble de données.

Deuxièmement, afin de s'attaquer à ce goulot d'étranglement, nous quantifions le style des experts qui annotent les données sous forme de biais et de consistance. Cela nous permet ensuite d'explorer l'impact des styles d'annotations sur les modèles d'apprentissage profond. Deux ensembles de données publics multiévaluateurs et multicentriques sont utilisés, un de lésions de sclérose en plaques cérébrales et un de segmentation de la matière grise de la moelle épinière. Sur les deux ensembles de données, les résultats montrent une corrélation ($R^2 = 0,60$ et $0,93$) entre le biais de l'expert et l'incertitude du modèle d'apprentissage profond. L'impact de la fusion d'annotations des experts sur cette relation est également étudié, et nous montrons que les consensus multicentriques sont plus efficaces que les consensus monocentriques pour réduire l'incertitude, car le style de l'expert n'est pas spécifique à l'individu, mais principalement au centre.

ABSTRACT

Multiple sclerosis is the most common autoimmune disease of the central nervous system. It is characterized by the presence of lesions in the brain and spinal cord, which are visible in magnetic resonance imaging (MRI). However, to extract useful information from the images, it is necessary to segment the lesions on these images, which is a long and expensive process when performed manually by an expert such as a radiologist. The goal of this project is to use innovative deep learning methods to improve segmentation of medical images.

First, to tackle generalization across imaging contrasts in spinal cord multiple sclerosis segmentation we implement Feature-wise Linear Modulation (FiLM) to leverage physics knowledge within the segmentation model and learn the characteristics of each contrast. Interestingly, a well-optimized U-Net reached the same performance as our FiLMed-Unet on a multi-contrast dataset (0.72 of Dice score), which suggests that there is a bottleneck in spinal MS lesion segmentation different from the generalization across varying contrasts. This bottleneck likely stems from inter-rater variability, which is estimated at 0.61 of Dice score in our dataset.

Second, as a follow-up we quantify rater style in the form of bias and consistency and explore the impacts on deep learning models. Two multi-rater and multi-center public datasets are used, consisting of brain multiple sclerosis lesion and spinal cord grey matter segmentation. On both datasets, results show a correlation ($R^2 = 0.60$ and 0.93) between rater bias and deep learning uncertainty. The impact of label fusion between raters' annotations on this relationship is also explored, and we show that multi-center consensuses are more effective than single-center consensuses to reduce uncertainty, since rater style is not individual-specific but mostly center-specific.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

AI	Artifical intelligence
ANN	Artificial neural network
CNN	Convolutional neural network
CT	Computed tomography
DL	Deep learning
FiLM	Feature-wise Linear Modulation
GM	Gray matter
MLP	Multilayer perceptron
MRI	Magnetic resonance imaging
MC	Monte Carlo
MS	Multiple sclerosis
PVE	Partial Volume Effect
ReLU	Rectified linear unit
SC	Spinal cord
SCT	Spinal cord toolbox
STAPLE	Simultaneous Truth and Performance Level Estimation
SVM	Support vector machine

CHAPTER 1 INTRODUCTION

Multiple Sclerosis (MS) is the most common autoimmune disease of the central nervous system (CNS), affecting 90 000 Canadians [1]. MS is characterized by lesions in the CNS, which reflect damage to patches of myelin. Since the role of myelin is to act as a protective sheath around axons, this loss of myelin disrupts the flow of electrical impulses, causing a wide array of symptoms (sensorial, cognitive, motor, visual) that depends on which part of the CNS is affected.

One of the main diagnosis criteria in MS is the quantification of lesion load in the brain and spinal cord [2]. This information is also useful for categorizing patients under different MS subtype (phenotyping), and monitoring the progression of the lesions. However, quantifying lesion load requires a precise segmentation (delineation) of lesions on MRI images, which is a time-consuming and expensive process when done manually by an expert, typically a radiologist. While some recent segmentation methods based on deep learning show promising results [3], they still face challenges. This study will explore two of these challenges.

1.1 Problem statement & research objectives

1.1.1 Generalizing across contrasts

The first issue is the apparent inability of CNN models to generalize towards real-world data, which has greater variability than small carefully crafted research datasets. This greater variability is due to factors such as different imaging device vendors and varying acquisition parameters, yielding different contrasts than those seen by the CNN model during training, which in turn results in poor segmentation performance [4]. Methods such as transfer learning or domain adaptation [5] are difficult to apply in a real-world scenario, where clinicians from other centers might not have the ability or expertise to retrain a new model specific to their data.

The goal of this first project is to explore innovative deep learning architectures that enable a model to generalize over a variety of image contrasts to ensure high performance of medical image segmentation. To achieve this goal, I did the following:

- Explore FiLM, a novel deep learning method [6]. In particular, I worked on learning contrast characteristics using image metadata to modulate segmentation.
- Compare this new approach with traditional CNN methods (e.g., U-Net), in data from

MS patients acquired by collaborative clinical centers.

1.1.2 Rater variability and uncertainty

The second issue is that data heterogeneity poses challenges not only for CNNs but also to human annotators, where it appears in the form of disagreement between experts. Disagreement can be classified as intra-rater and inter-rater variability, both of which affect CNNs' performance since they are trained using expert annotations. While the output of deep learning segmentation models is often binary, these models carry some inherent uncertainty. Understanding the impact of rater variability on uncertainty is critical in order to build robust non-biased models. Trust in deep learning models can only be attained if we understand their limitations, including how human limitations are reproduced and sometimes exacerbated in deep learning.

The goal of this second project is to understand the relationship between rater style and uncertainty in deep learning. While it is obvious that not all raters generate the same level of uncertainty, what makes some rating styles more of less suitable for deep learning purposes as not been studied yet. To achieve this goal, I did the following:

- Quantify rater style in multi-rater and multi-center datasets.
- Explore the potential relationship between rater style and deep learning uncertainty.
- Compare single rater and multi-rater consensus in the context of deep learning, to understand when consensus makes sense from a rater style perspective and how it affects uncertainty.

1.1.3 ivadomed

A common objective to both projects is to implement these algorithms into ivadomed [7], an open-source project developed by my host lab to allow widespread availability and testing.

1.2 Thesis outline

The chapters of this thesis are organized as follows. Chapter 2 begins with a literature review of deep learning in a medical imaging context. Chapter 3 presents results of investigations into contrast generalization. Chapter 4 explains how shortcomings in the contrast generalization project led to the uncertainty project. Chapter 5 presents the article resulting from the uncertainty project. Finally, chapter 7 concludes with global discussion of results and future steps.

CHAPTER 2 LITERATURE REVIEW

2.1 Machine Learning

This section introduces machine learning concepts which will be useful in later sections. Machine learning is a sub-field of artificial intelligence (AI) which aims to create algorithms with the ability to learn and improve from data or experience. ML therefore includes classical algorithms such as support vector machines (SVM) and decision trees, as well more recent methods such as deep learning (DL), which will be our focus. This section aims to provide a brief overview of basic deep learning concepts. A thorough review is out of the scope of this thesis, and many resources already cover these concepts in detail. From general to specific, here are resources that cover deep learning [8], DL applied to segmentation [9], DL applied medical imaging [10] and CNNs for brain MRI analysis [11].

2.1.1 Deep learning

Deep learning is a machine learning technique which relies on artificial neural networks (ANN). These artificial neurons are essentially learned functions applied to inputs in order to compute an output. The learned parameters are called weights $\mathbf{w}$ and correspond to the "importance" given to inputs $\mathbf{x}$ when computing an output y. A bias b is then applied, followed by an activation function such as a sigmoid (σ) which acts as a non-linearity, such that :

$$y = \sigma(\mathbf{w}^T \cdot \mathbf{x} + b) \tag{2.1}$$

Artificial neurons are stacked such that the output of a group of neurons (layer) acts as the input for another group of neurons. This stacking can lead to learning complex, non-linear functions, and is called deep learning when multiple layers are stacked and connected in such a fashion.

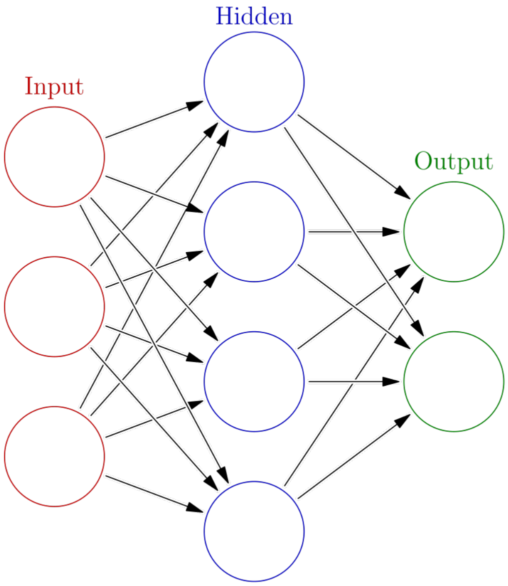


Figure 2.1 A basic MLP. Each neuron (circles) computes the weighted sum of its inputs and passes the results to next layer to which it is connected (arrows). (Image by Glosser.ca, CC BY-SA 3.0)

The idea is that as information flows through layers, each layer can create a higher level abstraction based on the information of the previous layer. This abstraction allows the extraction of complex features (in our case shapes and patterns in images) without the need to manually encode these features. Since we are in a supervised learning scheme, weights are learned in an iterative way through a process called backpropagation. It essentially consists of:

1. Propagate the input through the network to compute the output
2. Compute the difference between the obtained output and the desired output (loss)
3. Tune weights by a small amount in a way to minimise the loss

The key is the use of a differentiable loss functions, which enables backpropagation to compute how much each weight impacts to loss (partial derivative of the loss with respect to each weight) and therefore in which way they should be updated to minimise it. The desired output is known as ground truth and is usually determined by a human annotator (in our case, an expert such as a radiologist). The importance of quality ground truth cannot be understated, since it is the goal that our network tries to reach ; any error, bias, inconsistency in the ground truth will have consequences on the model.

2.1.2 CNNs

Convolutional neural networks (CNNs) are a type of ANN particularly useful to process images. There are two big differences over a standard MLP, the first one being that they use sliding window (convolution) filters to share weights. Small filters can thus be easily applied to the entire image patch-wise, keeping the number of weights to the minimum (i.e. a single neuron for a fully connected network taking a 100 × 100 image as input would need 10k weights whereas a 3 × 3 convolution filter applied patch by patch to the entire images requires only 9 weights. Each filter learns to recognize shapes or patterns (e.g. the curved boundary of an MS lesion), which increase in complexity by building upon those of previous layers. The second big characteristic of CNNs is the use of pooling layers, to decrease the size of the feature map and therefore increase the receptive field of the following layers.

2.1.3 Segmentation

A common task for CNNs is classification. It consists of determining whether a certain image corresponds to a class or not (e.g. is it the image of a MS lesion). In our case, we perform a different task, segmentation which is classification on a pixel-by-pixel basis (i.e. is this pixel part of a MS lesion or not). Therefore, the idea is for the CNN to take an image as an input and output an image of the same size, with values corresponding to the probability that each pixel belongs to belong to a certain class (which are then binarized). For the last few years U-Net [12] derivatives have been the gold standard of segmentation in medical imaging [10]. It takes its name from its U-shape, as shown in figure 2.2.

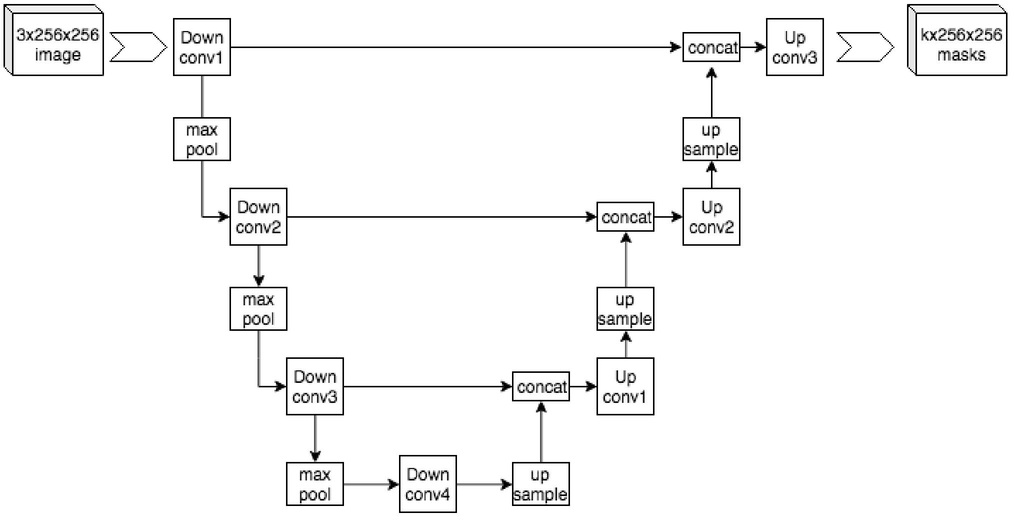


Figure 2.2 U-Net architecture (Image by Mehrdad Yazdani, CC BY-SA 4.0)

Each convolutional blocks in a U-Net is composed of a succession of layers:

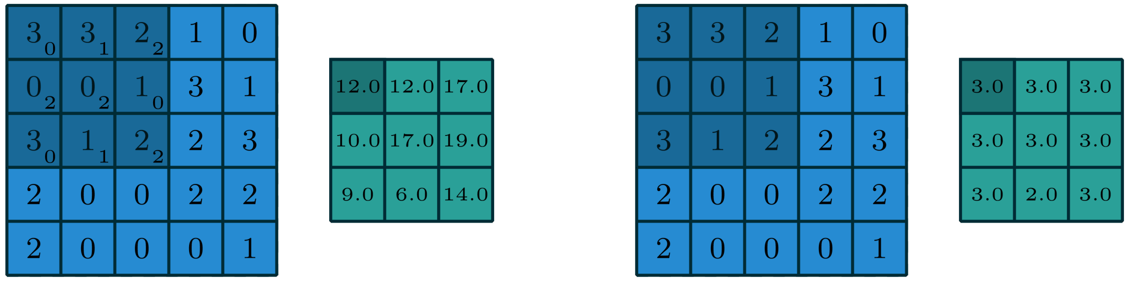
- **Convolutional layers** are the sliding filters that learn shapes and patterns that are applied to the images. They are applied by sliding dot products, with the outputs corresponding to feature maps that indicate where the learned shapes are positioned in the images.

- **Activation layers** introduce a non-linearity in the neural network. CNNs typically use a rectified linear unit (ReLU) which is $f(x) = \max(0, x)$. ReLU avoids the vanishing gradient issues caused by functions such as sigmoid in deep networks, while being fast to compute.

- **Batch normalization** is used to normalize layer output to means of 0 and variance of 1 [13]. While the cause of its effectiveness is debated, batch normalization enables a faster and more stable training.

- **Dropout** is the process of ignoring some nodes at random during training. Dropout increases model robustness and reduces overfitting, by preventing the neurons from relying too much on any individual neighbour or path [14].

The main strength of U-Net is the way it combines low-level details and high-level abstractions with the so-called "skip connections" connecting downsampling and upsampling blocks. The first half of the network performs downsampling, reducing image size and the second half performs upsampling, to recreate an output with the original image size, which is required by the segmentation task. In the first half of U-Net, downsampling is performed by pooling blocks. They reduce the dimensions of the feature maps, aiming to capture global information by discarding local details. There are different ways to achieve this dimension reduction, but they all rely on the idea of taking a block of the feature map and keeping only a single value to represent the block. Max pooling is the most commonly used, but there are other possibilities such as average pooling. In the second half of U-Net convolutional blocks are followed by upsampling blocks, which unlike fixed function like max pooling, are learned during training. This is due the fact that the way to combine optimally the high-level information (from the U part) and low-level information (from the skip connections) is task-dependent. Convolution and pooling operations are illustrated in figure 2.3.



(a) Convolution filter

(b) Max pooling filter

Figure 2.3
a) the convolution filter outputs (dark green) the dot product between the input (dark blue) and the weights (lower right corner of each box)
b) the max pooling filter outputs (dark green) the maximum value from the input region (dark blue)

Most state-of-the-art medical image segmentation models use a variant of the U-Net, such as : 3D U-Net [15] which can capture the information from multiple MRI slices at once and Attention U-Net [16] which has the ability to focus on a region of interest. These networks are fairly general architectures which can be adapted to almost any segmentation task and are not specific to neuroimaging.

Dice score

The Dice score [17] is the most commonly used metric to measure segmentation performance. It's computed as the intersection over union between two images A, B:

$$\text{Dice } = \frac{2|A \cap B|}{|A| + |B|}$$

Where $|X|$ is the number of voxels in X. Dice is also used as a loss function [18].

2.1.4 Uncertainty

One of the practical concerns of applying deep learning in medical settings is the "black box" nature of deep learning creates some concern due to the difficulty to interpret the model's decision [19]. Uncertainty is critical when making a diagnosis [20], and while an expert can explain why a region of an image should or shouldn't be classified as a lesion, it's a lot harder to interpret why a model with thousands of weights and artificial neurons made a decision. This is where deep learning uncertainty can be useful, to identify samples were the model output is potentially wrong or out of the training distribution.

Types

The two main types of uncertainty used in deep learning are aleatoric and epistemic [21, 22]. Epistemic uncertainty corresponds to uncertainty in the model parameters. This type of uncertainty is a result of the lack of knowledge about the data. This means that epistemic uncertainty can be lessened by using more data to fine-tune parameters. Aleatoric uncertainty corresponds to uncertainty inherent to the data (e.g. random noise). This type of uncertainty does not go away no matter how much data is used, since the data is inherently unpredictable [22].

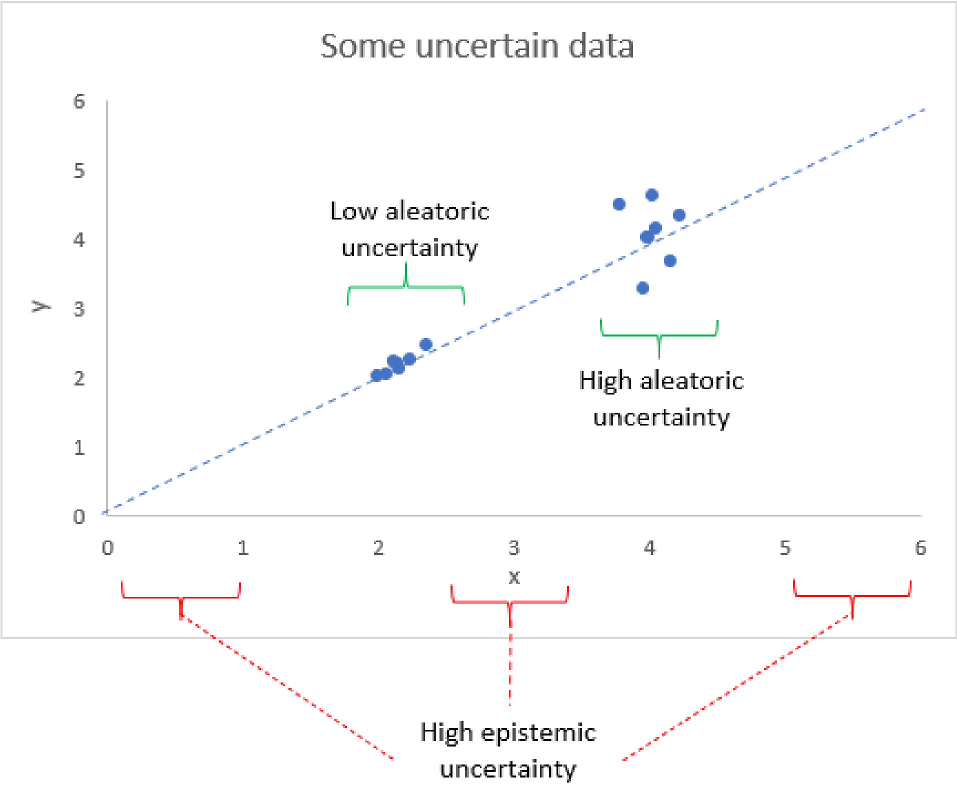


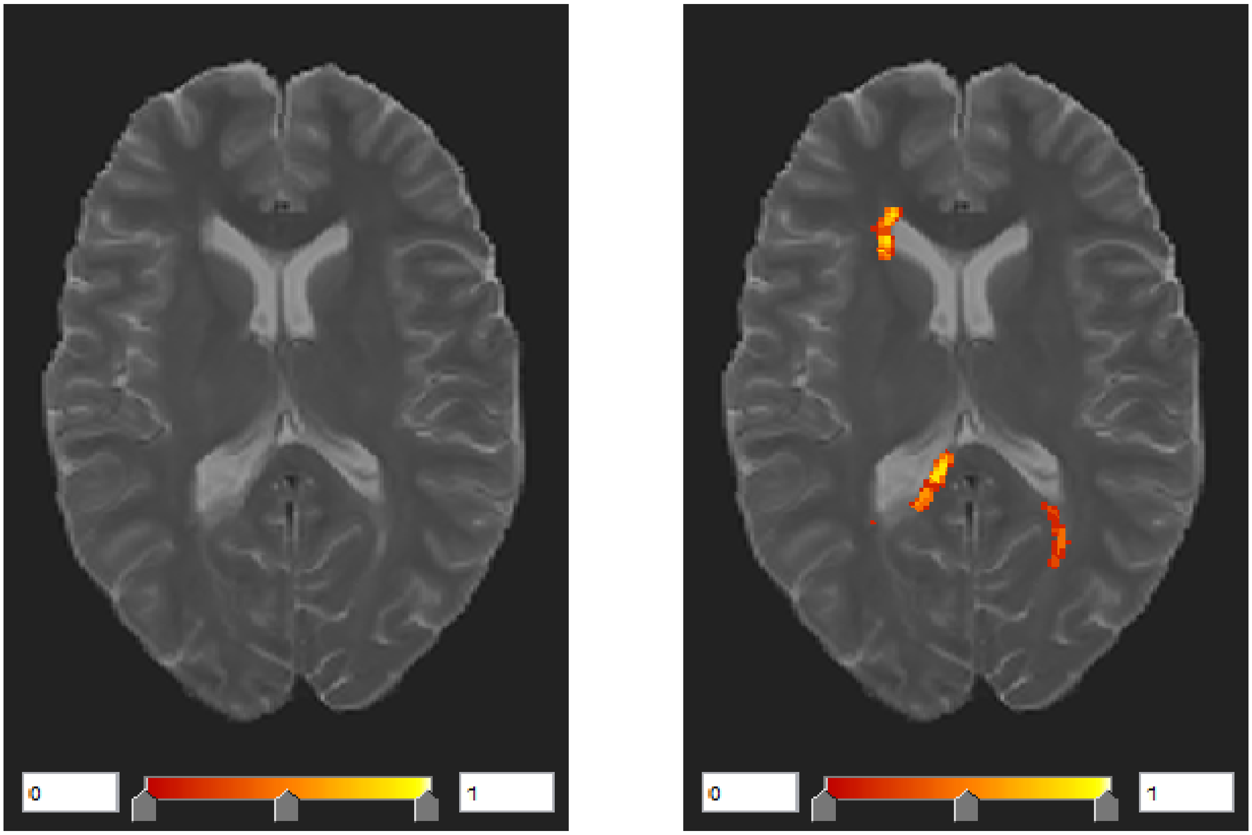
Figure 2.4 Difference between aleatoric and epistemic uncertainty illustrated. Variable x is the input data and y is what we want to predict. The dashed line corresponds to an (unknown) function and dots corresponds to observations (data) fed to the model.

These two types of uncertainty are illustrated in figure 2.4. In the regions with few data points (left, middle and right gaps), epistemic uncertainty is high since many functions (model parameters) could satisfy the given ground truth. Aleatoric uncertainty, on the other hand, is where data is present but scattered (right cluster vs left cluster).

Computation

Uncertainty is estimated by performing multiple inferences for each input, which generates so-called Monte Carlo (MC) samples. These Monte Carlo samples are then combined in a way to highlight differences (e.g. entropy, variance) [23, 24] to generate a voxel-wise uncertainty map. On the epistemic side, there are different ways to generate MC samples, including using dropout [25, 26], batch normalization [27] and ensemble models [28]. I chose to focus on studying test-time dropout since it has already been applied successfully to MS lesions segmentation [24]. Aleatoric uncertainty, on the other hand, can be either estimated when combined with epistemic (since the former tends to dominate the latter on large datasets) [22] or separately using test-time data augmentation [29]. I chose to study it separately and focus on test-time data augmentation, since this approach has shown useful in brain MRI tumour segmentation [30].

Aleatoric uncertainty samples are therefore generated using test-time data augmentations [29, 30], while epistemic uncertainty uses test-time dropout [25, 26]. Both aleatoric and epistemic uncertainties allow to create multiple unique outputs from the same input image, using parts of the pipeline containing randomness, and that are usually only used for model training. The way of combining these MC samples can give rise to different uncertainty metrics [23], the two main one being voxel-wise and structure-wise (in the case of MS lesions, a structure would be a lesion). Since the notion of structure is task-dependant, the information given by structure-wise uncertainty is task specific, whereas voxel-wise uncertainty is more general.

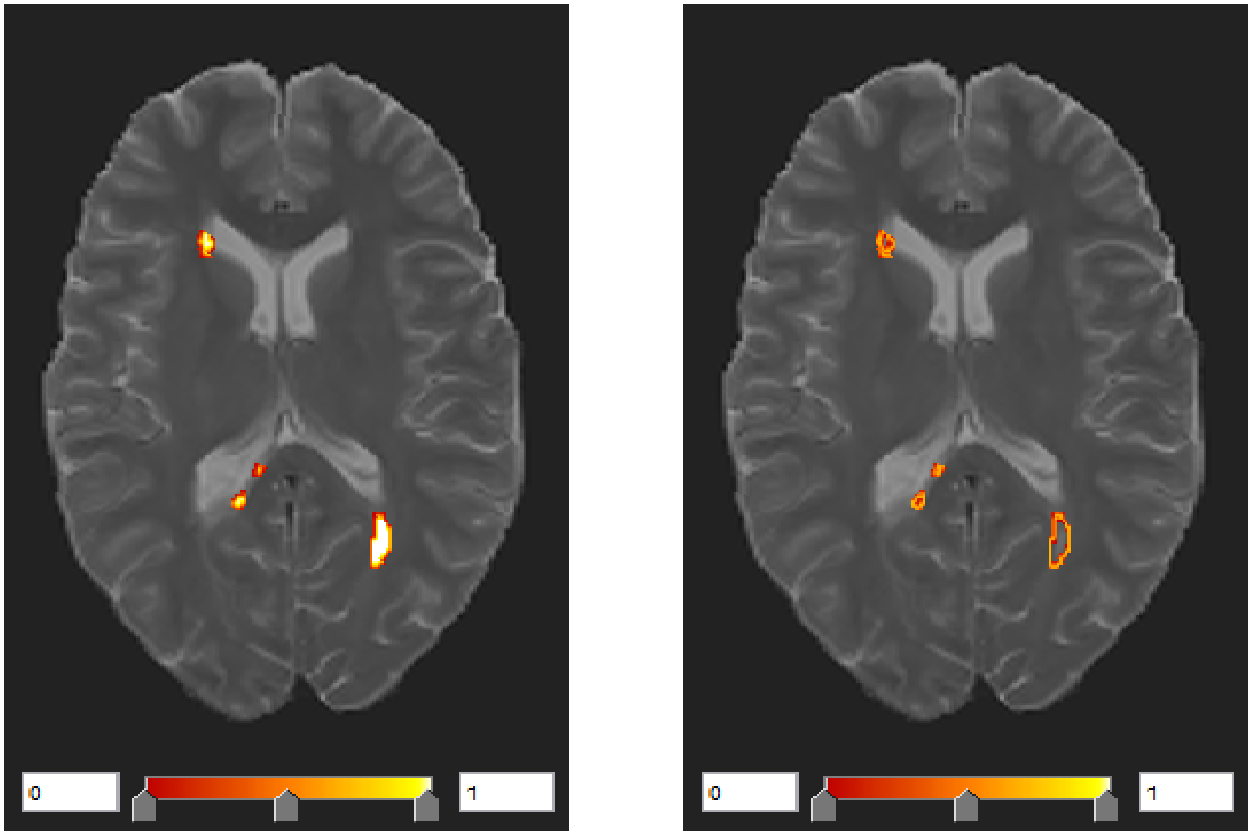


(a) Input image (MS brain)

(b) Ground truth (average segmentation of all raters)

Figure 2.5 Input image and average GT

The following figures are examples of uncertainty. Figure 2.5 shows an example of input and GT from the MS brain dataset used in section 5. 7 raters are averaged, and we can see that disagreement occurs mostly at the boundaries of lesions. This disagreement is reflected in figure 2.6, where we see that aleatoric uncertainty is present at the boundary of large lesions, but their center is mostly certain (roughly 1 in the soft segmentation and 0 uncertainty). However, for smaller lesions, the entire lesions including the center can be uncertain (such as the 2nd lesion from the bottom left of the image).

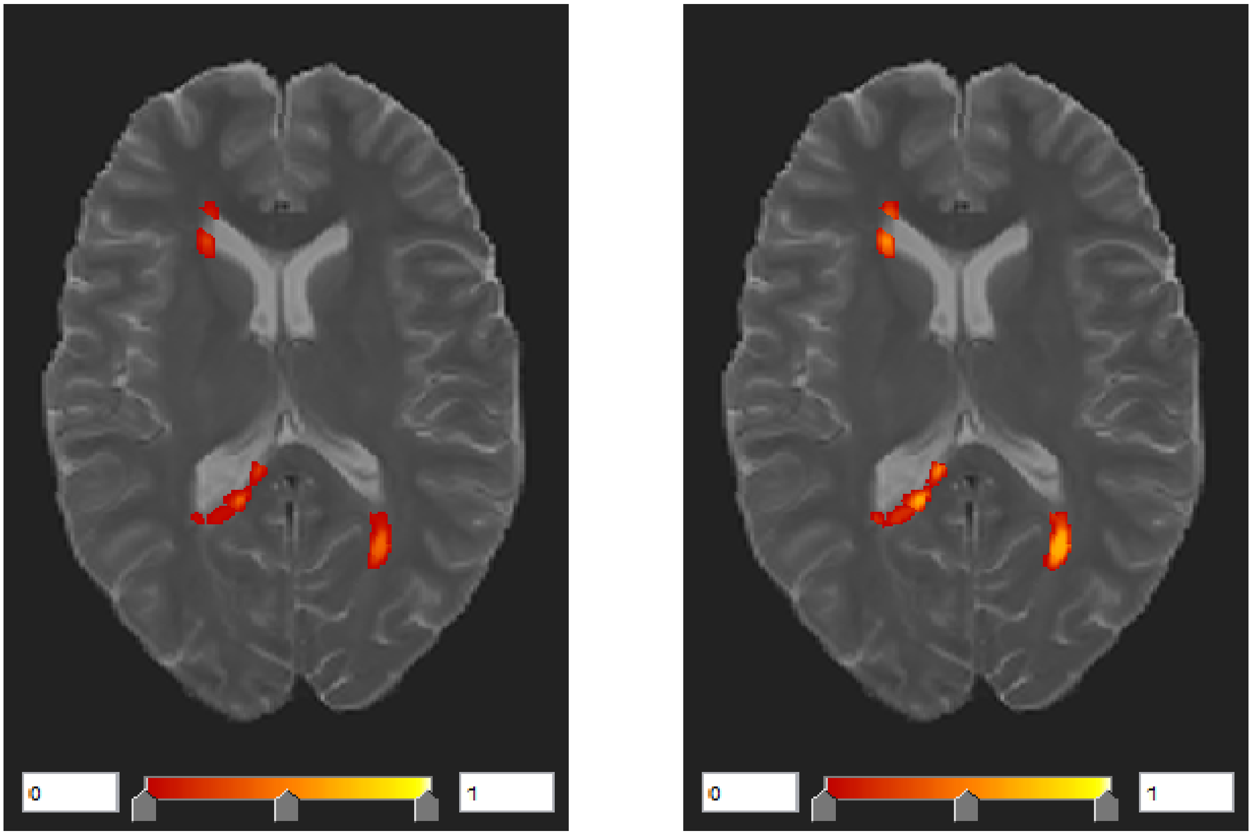


(a) Soft segmentation (average of MC samples)

(b) Voxel-wise uncertainty (entropy of MC samples)

Figure 2.6 Aleatoric uncertainty

Figure 2.7 shows that epistemic uncertainty is less of a clean cut; the centers of large lesions are a lot less certain than for aleatoric uncertainty. Even stranger, with epistemic uncertainty the center of lesions is more uncertain than the boundary. This is possibly due to the high class imbalance, and the fact that the model sees very few lesions compared to non-lesions, which means sparsity in the training dataset and therefore epistemic uncertainty as discussed in figure 2.4. In general we observe that epistemic type of uncertainty is smoother and more continuous than aleatoric uncertainty, and is present everywhere in and around lesions. The characteristics of aleatoric uncertainty therefore correspond more to what we expect from rater disagreement [24], which is why it is the type of uncertainty used later on, in chapter 5.



(a) Soft segmentation (average of MC samples)

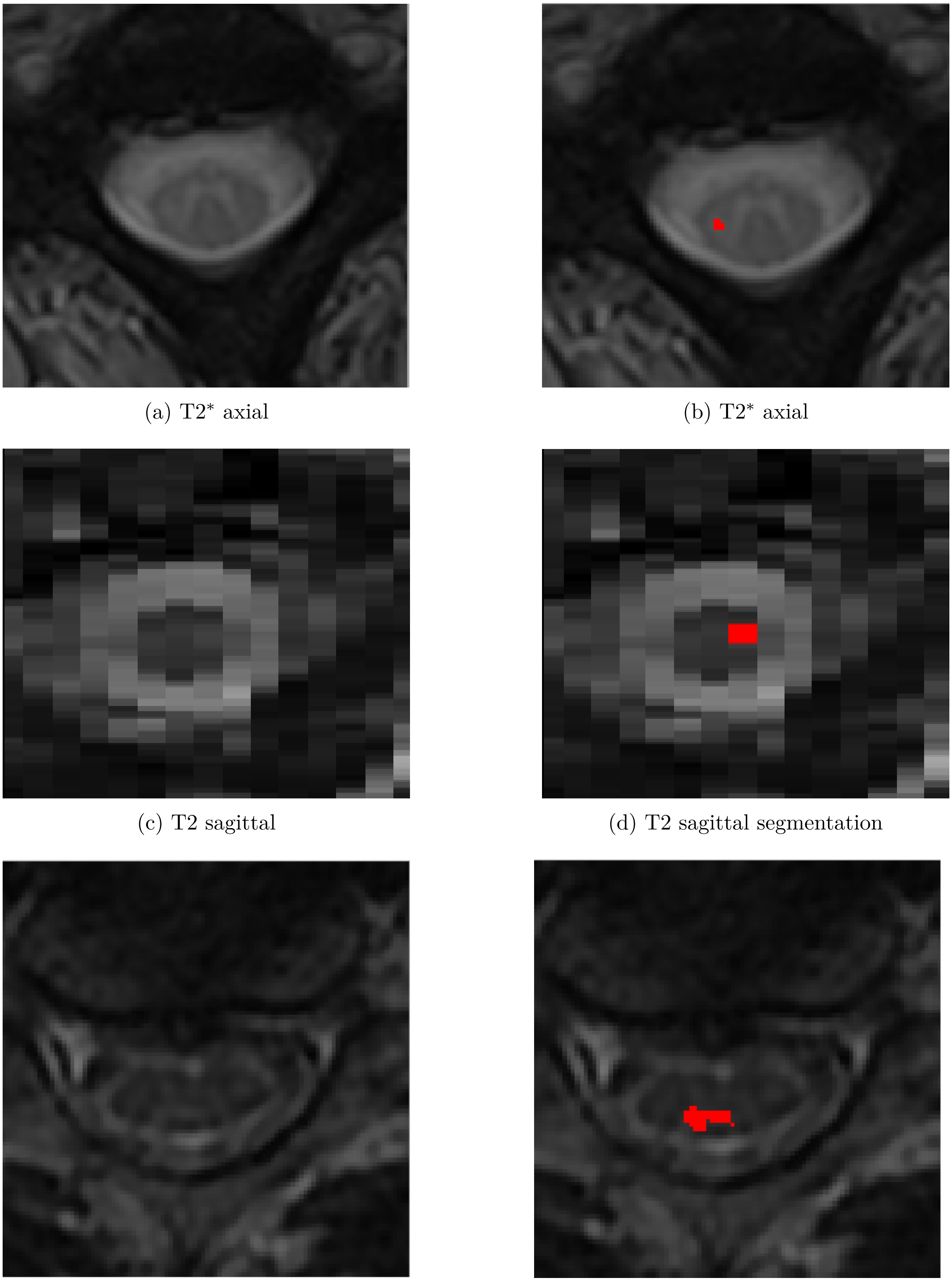
(b) Voxel-wise uncertainty (entropy of MC samples)

Figure 2.7 Epistemic uncertainty

2.2 Challenges in MS lesions segmentation

MS lesions segmentation is one of the most challenging segmentation tasks for both humans and neural networks. This is due in part to the fact that class imbalance is really high since. Lesions are often small, and the spinal cord occupies an even smaller fraction of the image when compared to the brain. Therefore in spinal cord MRI, less than 0.1% of pixels are part of a lesion [3]. It is not uncommon that the majority of slices in an MRI volume contain no lesion at all. Small lesions and lesion boundaries are difficult to identify, which is why inter-rater agreement is low and has been measured at Dice scores of 0.63 for the brain [31] and 0.61 for the spinal cord [3]. Low inter-rater agreement results in high DL uncertainty and it has been shown that filtering segmentation based on uncertainty can improve segmentation performance [24]. Since both uncertainty and inter-rater variability are high in MS lesions segmentation, it is an ideal task for exploring the relationship between these two phenomena.

Figure 2.8 shows some examples of contrasts and resolutions. MS lesions can be difficult to identify when 1) lesions are small, sometimes a single pixel as is the case in subfigure 2.8c and 2) the lesion boundary is ill-defined such as in 2.8a.



(a) T2* axial

(b) T2* axial

(c) T2 sagittal

(d) T2 sagittal segmentation

(e) T2 axial

(f) T2 axial segmentation

Figure 2.8 Example of different MRI contrasts with MS lesions and the corresponding segmentations

2.2.1 Partial Volume Effect

Part of the reason MS lesions boundaries are ill-defined is the partial volume effect (PVE). Briefly, if an MRI voxel is located at the interface between two different tissues (such lesion and non-lesion) it will have a value (intensity) that lies somewhere in between the values of the voxels of these individual tissues. It's a common phenomenon in medical imaging and techniques such as Gaussian mixture models can be used to model it [32, 33]. Partial volume effect is, however, not typically taken into account in deep learning segmentation models [34–36].

2.2.2 Contrast generalization

MRI contrast depends on hardware (manufacturer, model) and acquisition parameters (e.g. repetition time, echo time, flip angle) meaning that it is almost impossible to cover all possible contrasts in a training set. Domain adaptation methods such as transfer learning [4, 5] can be used to fine-tune pre-trained models. A common way to do this is freezing all layers except the N lasts and, retraining these layers with only a few samples from the new domain, the idea being that general patterns learned by the first layers are still applicable and therefore less data should be needed since there are fewer weights to train. An interesting approach that could potentially be useful to avoid the necessity of re-training (since clinicians might not have the ability or expertise to retrain or fine-tune a model specific to their data) is called Feature-wise linear modulation (FiLM) [6]. FiLM consists in adding layers which modulate the output of the CNN layers (in our case U-Net) based on non-image data. Therefore, while U-Net learns how to segment based only on the input image, FiLM layers learn how metadata (such as information about the contrast) should modify that segmentation in order to better predict the ground truth. FiLM's modulation being linear, for each feature in the CNN feature map it outputs parameters named (γ, β) such that a feature x becomes :

$$x' = \gamma \cdot x + \beta \tag{2.2}$$

2.2.3 Rater variability

While errors in ground truth (GT) appear in all domains, they are particularly important in medical imaging. All radiologists have a different rating "style" and even though standard rating scales exists, there will always be intra and inter-rater disagreement due to the subjective nature of the task [37–40]. This is due to variations in rater expertise, image quality and the fact that best practices can be center-dependent. This is an issue outside of deep learning, since all automated methods need to be ultimately compared to some kind of human benchmark to assess their performance. Errors in ground truth are, however, an even bigger issue in deep learning than with other automated methods, since models learn directly from rater data, and there is no handcrafting of features or human judgment [10].

2.2.4 Consensus

One way to avoid individual rater biases is to combine ratings from different raters in order for errors to "cancel out". There are multiple ways to achieve this, with the best known being majority voting and STAPLE [40]. The combined rating is known as consensus, and can be used to train models that are hopefully more accurate. A study on cardiovascular MRI has shown that raters from different centers can deviate from consensus in multiple ways: some raters have a small bias, meaning they are on average close to consensus, but at the price of a large variance (tend to alternate between over and underestimating the volume) while others tend to be consistently far from consensus (small variance, but large bias) [41].

2.2.5 Impact on uncertainty

Multiple recent studies have shown the impact of inter-rater variability on deep learning in medical imaging. In particular, it has been shown that there is a correlation between areas of an image that have high uncertainty and high rater disagreement in lung nodule CT scan segmentation [42]. Rater variability is therefore a factor that affects uncertainty, but the full extent is not known.

Moreover, training on consensus is not a silver bullet since, as shown by a recent study on brain tumor MRI segmentation: this practice can lead to models becoming overconfident and uncertainty being underestimated [43]. This overconfidence is due to the fact to models only see the consensus and are not aware of the underlying data where there is disagreement. A suggested alternative is to instead use label sampling, which is to select an annotation randomly from one of the raters at each epoch. This leads to more questions and potential pitfalls, since not all raters are equal some are necessarily more experienced and "better" than others. While these rater disparities are not considered in this sampling scheme, they are considered in some consensus approaches such as STAPLE, an algorithm that gives different weight raters to minimise the impact of outliers. In the same spirit, another study on synthetic data [44] shows that training using all available annotations yields uncertainty maps that are more representative of inter-rater disagreement when compared to consensus methods, while having similar segmentation performance. It was also shown that models can learn and amplify rater style, meaning a model's output tends to exacerbate characteristics found in the rater that annotated the training data [45].

CHAPTER 3 ARTICLE 1 : AUTOMATIC SEGMENTATION OF SPINAL MULTIPLE SCLEROSIS LESIONS : HOW TO GENERALIZE ACROSS MRI CONTRASTS?

This study was presented at OHBM 2020 and at the 7th Spinal Cord MRI workshop. It is available on arxiv.

3.1 Title

Automatic segmentation of spinal multiple sclerosis lesions: How to generalize across MRI contrasts?

3.2 Authors

Olivier Vincent[1]	ovincent.poly@gmail.com
Charley Gros[1]	charley.gros@polymtl.ca
Joseph Paul Cohen[2]	joseph@josephpcohen.com
Julien Cohen-Adad[1,3]	jcohen@polymtl.ca

[1] NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Canada
[2] Mila, University of Montreal, Canada
[3] Functional Neuroimaging Unit, CRIUGM, University of Montreal, Canada

3.3 Abstract

Despite recent improvements in medical image segmentation, the ability to generalize across imaging contrasts remains an open issue. To tackle this challenge, we implement Feature-wise Linear Modulation (FiLM) to leverage physics knowledge within the segmentation model and learn the characteristics of each contrast. Interestingly, a well-optimised U-Net reached the same performance as our FiLMed-Unet on a multi-contrast dataset (0.72 of Dice score), which suggests that there is a bottleneck in spinal MS lesion segmentation different from the generalization across varying contrasts. This bottleneck likely stems from inter-rater variability, which is estimated at 0.61 of Dice score in our dataset.

3.4 Keywords

Deep Learning, Segmentation, MRI, Spinal cord, Multiple Sclerosis

3.5 Introduction

Multiple Sclerosis (MS) is the most prevalent autoimmune disease of the central nervous system [46]. Lesion quantification on both brain and spinal cord MRI data is part of the diagnosis criteria for MS, [47] and has been extensively used in clinical studies [48]. Although recent methods based on deep convolutional neural networks (CNNs) showed promising results [3], they are hampered by major issues [10]. One of the issues is the inability of CNN models to generalize to heterogeneous imaging parameters (e.g. MR field strength or manufacturer, image contrast, resolution and field of view) that were not represented in the training data.

3.6 Methods

To address the generalization problem, we adapted the Feature-wise Linear Modulation[1] (FiLM) [6, 49] approach to the segmentation task. FiLM enables us to modulate CNNs features based on non-image metadata as illustrated in figure 3.1. In order to facilitate the model generalization, we input the MRI contrast (e.g. T2-weighted) in the FiLM generator, instead of directly inputting MR acquisition parameters which, based on preliminary investigations, would produce too many degrees of freedom and non-linearity issues across parameters. Each FiLM generator (multi-layer perceptron) optimises γ, β based on the contrast information ($\boldsymbol{z}$). Each U-Net feature map ($\boldsymbol{x}$) is then linearly-modulated by these FiLM parameters, such that:

$$\text{FiLM}(\boldsymbol{z}) = \gamma(\boldsymbol{z}) \circ \boldsymbol{x} + \beta(\boldsymbol{z}) \tag{3.1}$$

[1] Additional details on the FiLM architecture are available in section 4

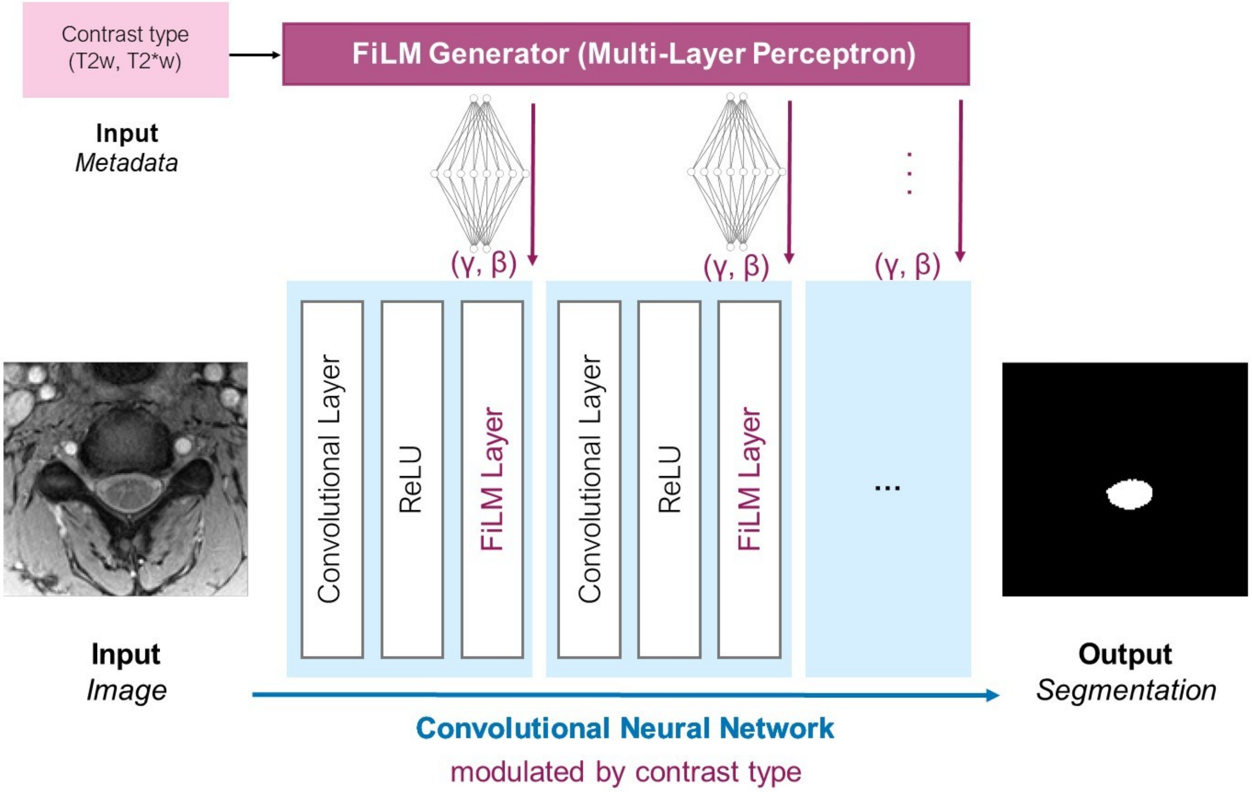


Figure 3.1 FiLM architecture using MRI contrast type in input. A FiLM layer is added after each convolutional block of the U-Net to modulate the CNN feature maps. The FiLMedUnet is trained end-to-end.

We compared this new approach with a traditional U-Net [12], on 'real-world' data from 642 MS patients, acquired by thirteen centers, yielding 2, 549 MR volumes (T2-weighted or T2*-weighted, 38, 855 axial slices in total), spanning a large range of acquisition parameters (e.g. resolution, orientation, field of view). To alleviate the issue of class imbalance, slices were cropped around the region of interest using the spinal cord segmentation (48x48 pixels). Data augmentation was performed on both the MRI data (random affine and elastic transformations) and by altering the ground truth segmentation via a series of morphological and affine realistic operations. This was done to simulate rater uncertainty at the boundary of lesions and have the network learn these uncertainties.

Training was done on axial slices using Dice loss [18] with the Adam optimizer [50] and a 60/20/20% training / validation / testing random split of the dataset. Hyperparameters such as learning rate scheduler, batch size and U-Net depth were optimised using a grid search. Models were implemented in PyTorch 1.2 [51] and trained on an NVIDIA P100 GPU, which took 7h. The implementation is open source and available on Github: ivadomed.

3.7 Results

As shown in table 3.1, models have almost indistinguishable performance when optimised. U-Net yielded negligible performance difference when trained and tested on single-contrast dataset (i.e. T2w or T2*w) compared with multi-contrast dataset (i.e. T2w and T2*w), suggesting that contrast generalization is not a bottleneck, at least in this dataset and for this MS lesion segmentation task. This observation is consistent with the fact that FiLM does not reach higher performance than an optimised U-Net on this dataset.

Training configuration	**Dice score (higher is better)**
U-Net T2w only	0.72
U-Net T2*w only	0.73
U-Net T2w + T2*w	0.72
FiLMed-Unet T2w + T2*w	0.72

Table 3.1 Results comparison between the U-Net and our FiLMed-U-Net in terms of Dice on the testing dataset, including T2w (top row) or T2*w (second row) or T2w and T2*w (last rows) data.

3.8 Conclusion

In this paper we implemented FiLM to modulate U-Net segmentation based on MRI contrast type. Results show that a simple U-Net can achieve the same performance as FiLM, both on single and multi-contrast datasets. This result however highlights a bottleneck in spinal MS lesion automatic segmentation, and likely in medical image segmentation in general: a high inter-rater variability, as also been reported in brain studies [31]. Inter-rater variability had a Dice of 0.61 in our dataset [3], which is lower than our results of 0.72. The difference could possibly be explained by some overfitting on certain rater styles. Future work will encode the rater identity into the CNN learning in order to account for rater style and expertise. This would enable the model to learn the difference between multiple rating styles, and to choose a desired style at inference time.

3.9 Acknowledgements

The authors would like to thank Lucas Rouhier, Anthime Bucquet, Valentine Louis-Lucas and Christian Perone from the IVADO medical imaging team for helpful discussions. Funded by IVADO [EX-2018-4], 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 Fonds de Recherche du Québec - Nature et Technologies [2015-PR-182754], the FRQNT Strategic Clusters Program (2020-RS4-265502 - Centre UNIQUE - Union Neurosciences & Artificial Intelligence – Quebec, 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], Spinal Research and Wings for Life (INSPIRED project).

CHAPTER 4 ADDITIONAL INFORMATION ON THE FILM PROJECT

Due to length constraints arising from the OHBM abstract format, we present here supplementary information regarding the methods and results of the FiLM project.

4.1 FiLM architecture

4.1.1 Input encoding

As described in equation 3.1, image metadata is being fed into the FiLM generator. All metadata, both categorical (e.g. contrast type such as $T2$ and $T2^*$) and continuous (acquisition parameters such as TE, TR, and flip angle) are one-hot encoded in order to have a single format of input for the FiLM generator. To achieve this, continuous data was categorized, using kernel density estimation (KDE) to perform clustering. While some information is lost in the process, this discretization enables faster prototyping since the same pipeline can be used for both types of data. This one-hot vector is the fed into what is called the FiLM generator, which is a simple neural network (MLP) with the task of predicting a pair of values (γ, β) for each feature in the feature maps of the U-Net. These (γ, β) pairs then modulate the values in the features maps in a linear fashion, such that a feature x becomes :

$$x' = \gamma \cdot x + \beta \tag{4.1}$$

4.1.2 Architectural details

U-Net

U-Net was composed of 8 layers (meaning a U of height 3) with FiLM applied after each layer. The rationale behind this is that while FiLM may potentially be more useful after some layers than others, this optimal location should be learned instead of hard-coded. Adding more layers (thus more downsampling) to the U-Net was not helpful due to the small image size (48×48). Kernels were 3×3 with a stride and padding of 1. A cosine annealing learning rate (from 10^{-2} to 10^{-3}) was applied with an Adam optimizer. Dropout was applied with a rate of 0.3 and batch normalization with a momentum of 0.1. Models were trained from scratch (no pre-training) and both networks were trained at the same time (FiLM and U-Net) using the same Dice loss function.

FiLM

The FiLM generator is composed of a MLP. Multiple variations of FiLM generators (dark pink block at the top in figure 3.1) were tested, with the major ones being :

- Width of hidden layers (from 32 to 256 for the first layer)
- Depth of the network (from 2 to 5 layers)
- Non-linearity functions
  - ReLU
  - Leaky-ReLU
  - Sigmoid
  - Tanh
- Loss function
  - Dice loss
  - Cross-entropy loss
  - Focal loss
  - Generalized Dice loss
  - Focal Dice loss

First, when gradually increasing first layer width (w_1) and network depth (d), performance plateaued at a width of 64 and a depth of 3 layers. These were therefore the chosen values for the network size, since additional complexity proved unnecessary. Width for subsequent layer is $w_n = w_{n-1}/4$. The output layer has two outputs for each convolutional filter in the U-Net (e.g. $2 \times n_{\text{channel}}$). The architecture is illustrated in figure 4.1.

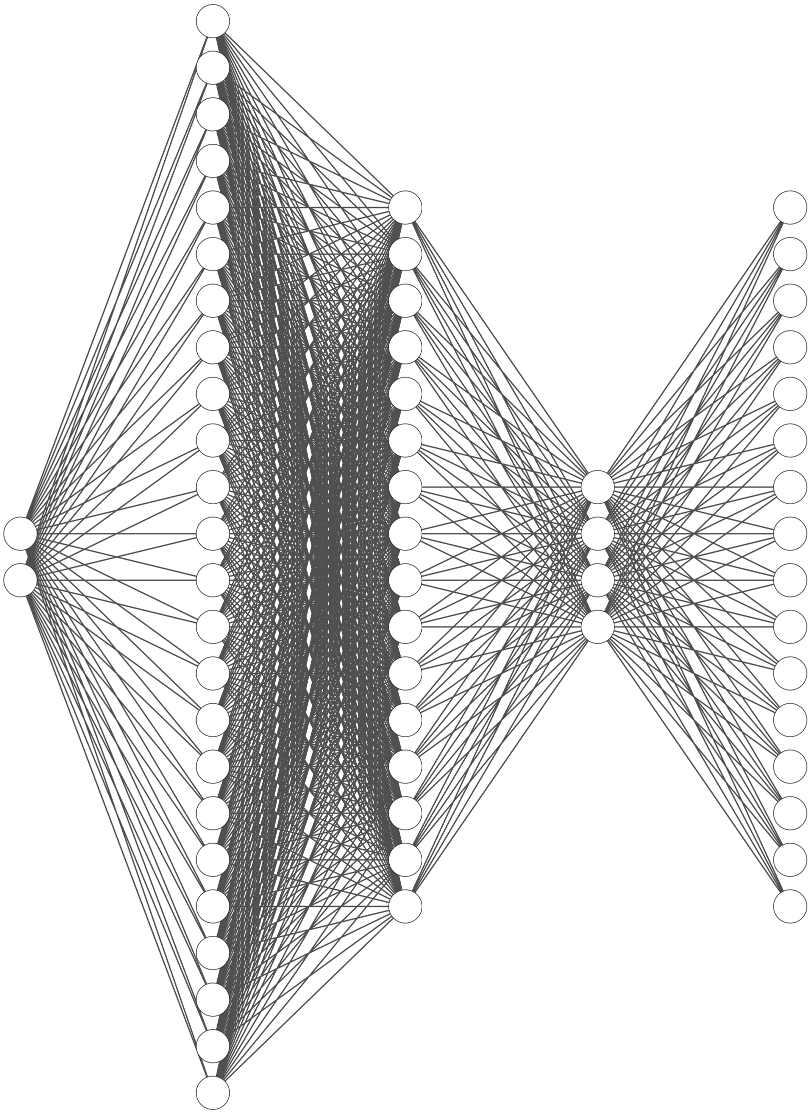


Figure 4.1 Simplified illustration of the FiLM generator architecture (from left to right : input layer, 3 hidden layers and output layer). 2 input neurons correspond to the simplest case, which is a single metadata type with a 2D one-hot vector (e.g. contrast type with only 2 available contrasts). The full width of the first hidden layer is not shown here to achieve a reasonable image aspect ratio (i.e. Full width of 64 is reduced). Finally, in practice, there are 2 output neurons per U-Net channel (varies for each U-Net layer, but there are lot more than illustrated here).

Next, for the activation functions, I found that ReLU, Leaky-ReLU and Tanh all offered similar performance. However, sigmoid gave 5% higher Dice score than ReLU as shown in table 4.1 and it was selected for this reason. Finally, various loss functions were tested to deal with the class imbalance issues, since losses such as focal loss and its variants have been shown to achieve higher performance in some segmentation datasets with small objects. However, in our case, Dice loss was the one yielding the highest Dice scores.

Lesion segmentation Dice score of U-Net vs FiLM					
Epochs		10	20	50	100
Dice score	U-Net	0.53	0.53	0.53	0.55
	Base (ReLU) FiLM	0.52	0.50	0.50	0.46
	Sigmoid FiLM	0.57	0.57	0.56	0.60

Table 4.1 Comparison between the U-net and our FiLMed-U-net (ReLU vs sigmoid for the FiLM generator MLP) in terms of Dice on the testing dataset

4.2 FiLM preliminary results

The original goal of this project was to improve MS lesion segmentation. Chapter 3 presented the final results of efforts to improve segmentation of MS lesions by using FiLM [6]. However, not shown are preliminary results which motivated the project. FiLM initially showed a 5% increase higher Dice in FiLMed U-Net when compared to baseline U-Net, as shown in table 4.2. This difference disappeared after performing an extensive hyperparameter optimization (batch size, learning rate, scheduler function, U-Net depth, loss function, positioning of FiLM layers), and fine-tuning of preprocessing (cropping around the spinal cord, binarizing ground truth) the performance of both models is increased such that there is no difference between the two as was described. Multiple variations were tried (in terms of network architecture, of where to apply the modulation and which metadata to feed) without success. As an example, I tried directly input the values of main acquisition parameters (TR, TE, flip angle), since they offer more granularity contrast types, but results were similar. This suggested that a bottleneck that looked data dependant instead of architecture dependant, since a theoretically more advanced architecture which had proven to be better on this task in preliminary tests was faced with the same hurdle as a basic U-Net.

	Unoptimized model (Dice score)	Optimized model (Dice score)
U-Net T2w only	0.552	0.721
U-Net T2*w only	0.585	0.726
U-Net T2w + T2*w	0.545	0.724
FiLMed-Unet T2w + T2*w	0.598	0.723

Table 4.2 Comparison between the U-net and our FiLMed-U-net in terms of Dice on the testing dataset, including T2w (top row) or T2*w (second row) or T2w and T2*w (last rows) data.

4.3 From improving MS lesions segmentation to understanding the impact of rater variability

This bottleneck is what prompted investigations into the rater variability phenomenon. Models were already achieving score higher than inter-rater variability, and therefore it appeared as the logical bottleneck. The focus thus shifted from simply getting a higher Dice to understanding what is the impact of rater variability on deep learning segmentation. As discussed in Chapter 2, some studies have looked at the impact of consensus on models' confidence, but in the real world, obtaining ratings from multiple raters if often not possible. Therefore, some of the remaining questions are :

- How can we quantify what rating style is more suitable for deep learning purposes?
- Which factors in rater style impact model performance and uncertainty?
- Is it possible to identify a less suitable rating styles for deep learning purposes without knowing the consensus?
- Does any consensus systematically lower uncertainty or are some combinations of raters better than other when creating consensus?

These questions appeared following the FiLM project, and are what the article presented in Chapter 5 aims to answer. The idea is that uncertainty could be a good indicator of the bias/consistency of a rater, and we aim to bridge the gap between investigations on rater style [41] and those on uncertainty [42–45]

4.4 Shift from spinal cord MS to brain

Datasets for the inter-rater projects were different. The FiLM project used a large private spinal cord MS lesion dataset that encompassed a variety of "real-life" contrast, and there are no comparable public datasets. With the constraint of generalization removed (since the bottleneck is elsewhere), we decided to use public datasets for the inter-rater project. While these are in general smaller and more carefully crafted (less representative of real-world data) it allows for better reproducibility, since both the code and data are open. This is why in chapter 5 we use an MS brain dataset instead of a spinal cord dataset. However, we still wanted to include a spinal cord aspect, which is why a second dataset, gray matter segmentation (also publicly available), was used to replicate some findings.

CHAPTER 5 ARTICLE 2 : IMPACT OF INDIVIDUAL RATER STYLE ON DEEP LEARNING UNCERTAINTY IN MEDICAL IMAGING SEGMENTATION

This study has been submitted at the Journal of Machine Learning for Biomedical Imaging (MELBA).

5.1 Title

Impact of individual rater style on deep learning uncertainty in medical imaging segmentation

5.2 Authors

Olivier Vincent[1,2] ovincent.poly@gmail.com
Charley Gros[1,2] charley.gros@gmail.com
Julien Cohen-Adad[1,2,3] jcohen@polymtl.ca

[1] NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Canada
[2] Mila, University of Montreal, Canada
[3] Functional Neuroimaging Unit, CRIUGM, University of Montreal, Canada

5.3 Abstract

While multiple studies have explored the relation between inter-rater variability and deep learning model uncertainty in medical segmentation tasks, little is known about the impact of individual rater style. This study quantifies rater style in the form of bias and consistency and explores their impacts when used to train deep learning models. Two multi-rater public datasets were used, consisting of brain multiple sclerosis lesion and spinal cord grey matter segmentation. On both datasets, results show a correlation ($R^2 = 0.60$ and 0.93) between rater bias and deep learning uncertainty. The impact of label fusion between raters' annotations on this relationship is also explored, and we show that multi-center consensuses are more effective than single-center consensuses to reduce uncertainty, since rater style is mostly center-specific.

5.4 Keywords

Deep learning, Segmentation, Uncertainty, Rater style

5.5 Introduction

Inter-rater variability limits the achievable segmentation performance of deep learning segmentation by introducing human error to the ground truth [31]. Tasks such as multiple sclerosis (MS) lesions segmentation are highly challenging due to the smallness of lesions and the poorly defined borders, leading to a low inter-rater agreement and high deep learning model uncertainty [3, 24]. For instance, some experts tend to over-segment, others under-segment, yielding "confusion" for the segmentation model trained on data labelled by different raters [44]. Understanding the rater style could allow for better performance of models, e.g., by integrating this knowledge within the deep learning training scheme.

5.5.1 Related Works

Previous studies have shown that the rater style can be learned [45], and therefore the inter-rater disagreement patterns could potentially also be learned by the model [42]. There has also been work on jointly learning individual rater characteristic at the same time as "true" consensus segmentation, in classification [52], segmentation [53] , and object detection [54]. Also shown was that the method used to generate the ground truth from multiple rater annotations, e.g., label fusion [44] and label sampling [43] largely impacts the model uncertainty.

5.5.2 Contribution

While many studies have addressed the uncertainty introduced by multiple raters, less work addressed the uncertainty introduced by a single rater. A model trained with data from a single rater will still exhibit some level of uncertainty due to rater style, and our goal is therefore to find what factors in a rater's style generate uncertainty. Those factors could include tendency to under/over-segment, consistency across images, non-independence of raters (e.g. influence of the expert who trained the rater). Intuitively, a non-biased and highly consistent rater would be the ideal candidate for training a deep learning model. We therefore expect a correlation between a rater's bias/consistency and the uncertainty of the model trained with their annotations. This would mean that characterization of the rater's bias could eventually be incorporated as prior knowledge within the learning scheme or in the reporting of uncertainty (post-processing).

5.6 Material and Methods

5.6.1 Datasets

Two public MRI datasets with multiple raters annotations were used. The first is a brain multiple sclerosis (MS) lesion dataset introduced at a MICCAI 2016 challenge [55]. It consists of 15 subjects each annotated by seven raters from three different centers. The second dataset is a spinal cord (SC) gray matter (GM) introduced at a segmentation challenge [56], which includes 40 subjects with annotations from four raters (all raters from a different center).

5.6.2 Metrics

In this paper, we characterize rater's style using rater bias and consistency. Since the consensus of all raters is the closest we have to the real ground truth, we define a rater's bias to be the average difference (in terms of positive voxels count) between the rater's annotation and the consensus across all volumes:

$$\text{bias} = \frac{\sum_{\text{images}} n_{\text{rater}} - n_{\text{consensus}}}{\#\text{ of images}} \tag{5.1}$$

With n_{XX} the number of positive voxels in a segmentation mask (i.e., belong to the target segmentation class). Consensus is defined by majority voting, as explained in section 5.6.3. A positive or a negative bias therefore measures if a given rater has a tendency to over- or under-segment, respectively. Images refer to 3D volumes, but using 2D slices as a basis instead would give the same results up to a constant factor, since bias is an average.

Similarly, we define rater consistency as the standard deviation of the difference (in terms of positive voxel count) between the rater's annotation and the consensus across all volumes:

$$\text{consistency} = \sqrt{\frac{\sum_{\text{images}} (n_{\text{rater}} - n_{\text{consensus}} - \text{bias})^2}{\#\text{ of images}}} \tag{5.2}$$

Consistency therefore measures whether a rater is either always over-segmenting or under-segmenting (consistent, close to zero) or if they are doing a bit of both (inconsistent: higher values).

We choose to use an absolute bias metric as opposed to a relative one since we do not think all slices deserved the same weight. For example, it would be unfair to penalize a rater by the same amount for a 10% error on a slice showing only a single 10-voxel lesion, versus for a 10% error on a slice with multiple lesions totalling hundreds of voxels. The error in the former case is likely negligible, whereas the error in the latter case is large and systematic (multiple lesions), but both would have the same impact on the computed bias if we had used a relative metric. We however considered using relative instead of absolute metrics, by normalizing the difference used in bias and consistency by the number of positive voxels in the consensus in each image. Results of these investigations are in appendix 5.10, and show that the bias/uncertainty relationship in figure 5.2 and 5.3 still holds when using relative bias

5.6.3 Processing

Images were resampled ($1 \times 1 \times 1mm^3$ for MS brain and $0.25 \times 0.25 \times 2mm^3$ for SC GM) and cropped (respectively 160×224 and 128×128) before being fed to the models. Data augmentation (rotation, translation, scaling) was applied slice-wise. Datasets were split 60/10/30 randomly for training/validation/testing respectively. 2D U-Nets [12] were trained slice-wise with the annotation of each individual rater. While it is no more state of the art, a 2D U-Net is sufficient since it can achieve near inter-rater variability levels of performance [3, 57]. Additional performance would not be beneficial since the main goal is to study uncertainty and not segmentation performance. A more advanced architecture would probably only result in overfitting on some rater styles. Training was done on NVIDIA P100 GPUs using the open source framework ivadomed [1] $v2.1.0$ [7] which is based on PyTorch [51]. Configuration files containing all hyperparameters for both datasets are also available here [2]. Models were trained using a Dice loss [18]. Inference was then done on the test set to measure model's performance (Dice score) and aleatoric uncertainty [30]. Uncertainty is estimated using test-time data augmentation (rotation, translation, scaling), and is computed as the entropy of 10 Monte Carlo samples for each image. The exact settings for the transforms are described in the config files linked above. The choice of aleatoric uncertainty was made because it is considered as being representative of "inherent" uncertainty in the data, whereas epistemic uncertainty is considered dependent on the model parameters (i.e. it could go away with more data) [21, 22]. All the previous steps (pre-processing, data augmentation, training, evaluation and uncertainty computation) were done with ivadomed. Preliminary experiments on the MS brain dataset showed that generating ground truth with STAPLE [40] yielded similar results

[1] `http://ivadomed.org/`

[2] `https://github.com/olix86/paper_rater_uncertainty`

in terms of rater style (bias & consistency) compared to majority voting. The only difference was a constant offset to all raters bias, meaning that majority voting has a tendency to over-segment when compared to STAPLE. Since this affects all raters and doesn't have an impact when comparing styles between raters, ground truths were generated using majority voting due to it being easier to interpret (i.e. consensus voxel = 1 if at least 50% of raters voted 1). By default, the term "consensus" will refer to this combination of all raters for a given dataset, however, single-center consensuses were also computed using the same method and will be compared to the global consensus.

5.7 Results

5.7.1 Rater style

We first examine rater style in the form of bias and consistency relative to consensus for MS brain. Styles are shown in Figure 5.1 and solely depend on the ground truths from each rater; they do not involve any deep learning model.

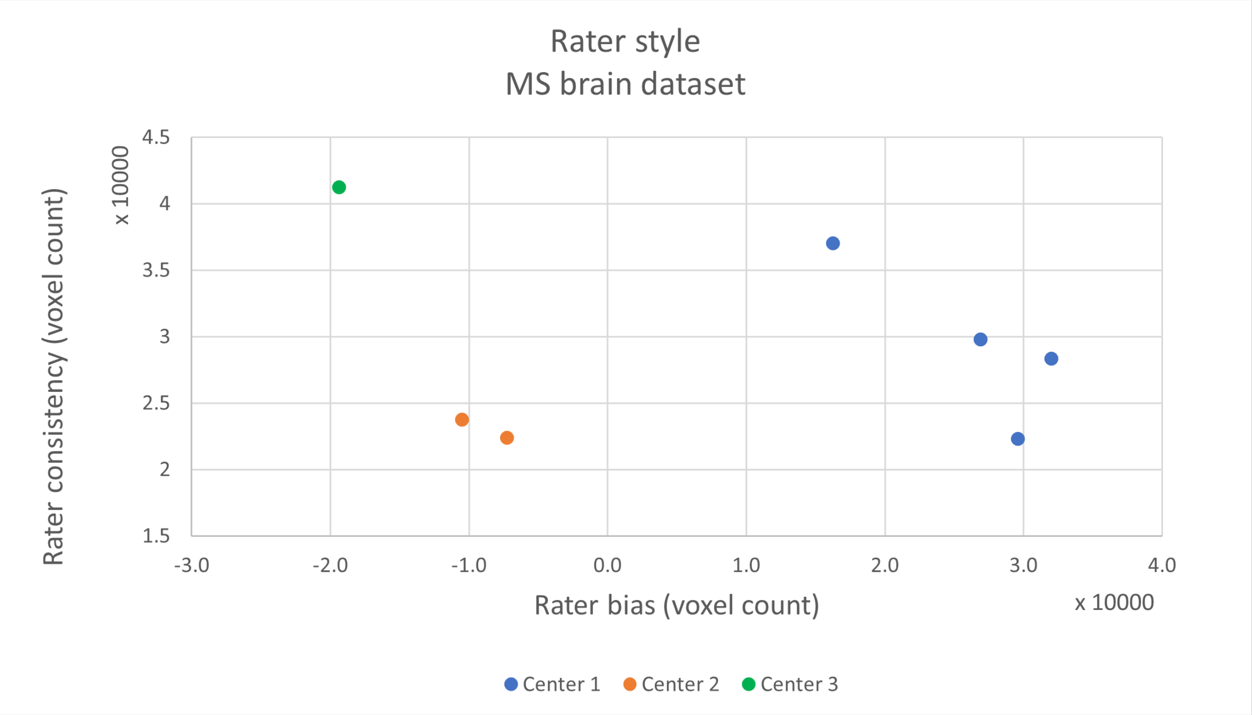


Figure 5.1 Rater style is characterised by consistency and bias. Dots colour corresponds to each rater's center.

We notice 3 clusters which are clearly delimited by the center to which raters belong. This implies that rater style depends a lot more on the rater's center than its individual characteristics. Indeed, cluster radii of $[0, 1.8, 12] \times 10^3$ are a lot smaller than the distances between pairs of clusters centroid $[21, 36, 47] \times 10^3$. To assess the quality of the clustering we use the Davies-Bouldin index [58], a metric which quantifies the quality of clustering through ratios of intra-cluster scatter to inter-cluster distance (lower is better). Here, $DB = 0.21$, meaning that intra-cluster scatter is quite lower than inter-cluster distances. Our hypothesis is that uncertainty for individual raters should follow a similar center-centric pattern assuming it depends on the rater style. This does not apply for the GM dataset since it contains only a single rater per center.

5.7.2 Uncertainty

Next, we look at how uncertainty in models trained separately for each rater relates to rater style for both datasets, in Figure 5.2 and 5.3.

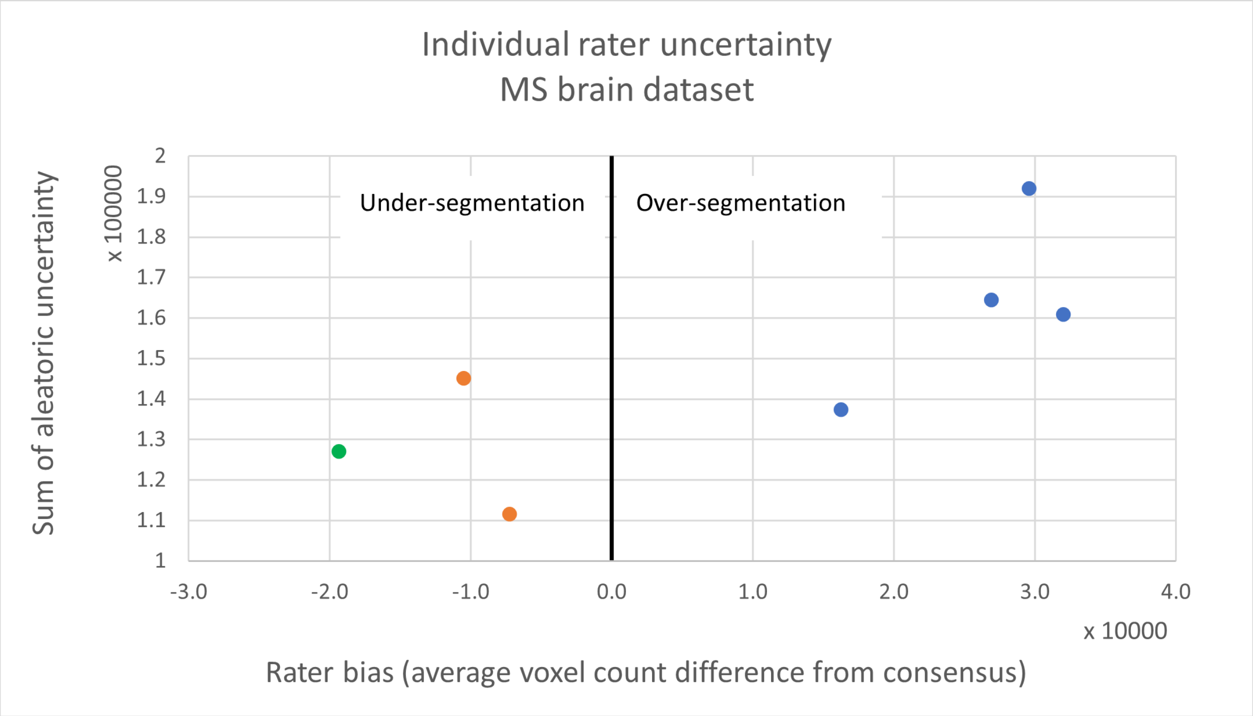


Figure 5.2 Relationship between the uncertainty of models trained for each rater and the bias of the corresponding rater for the MS brain dataset. Each colour corresponds to a center. $R^2 = 0.60$

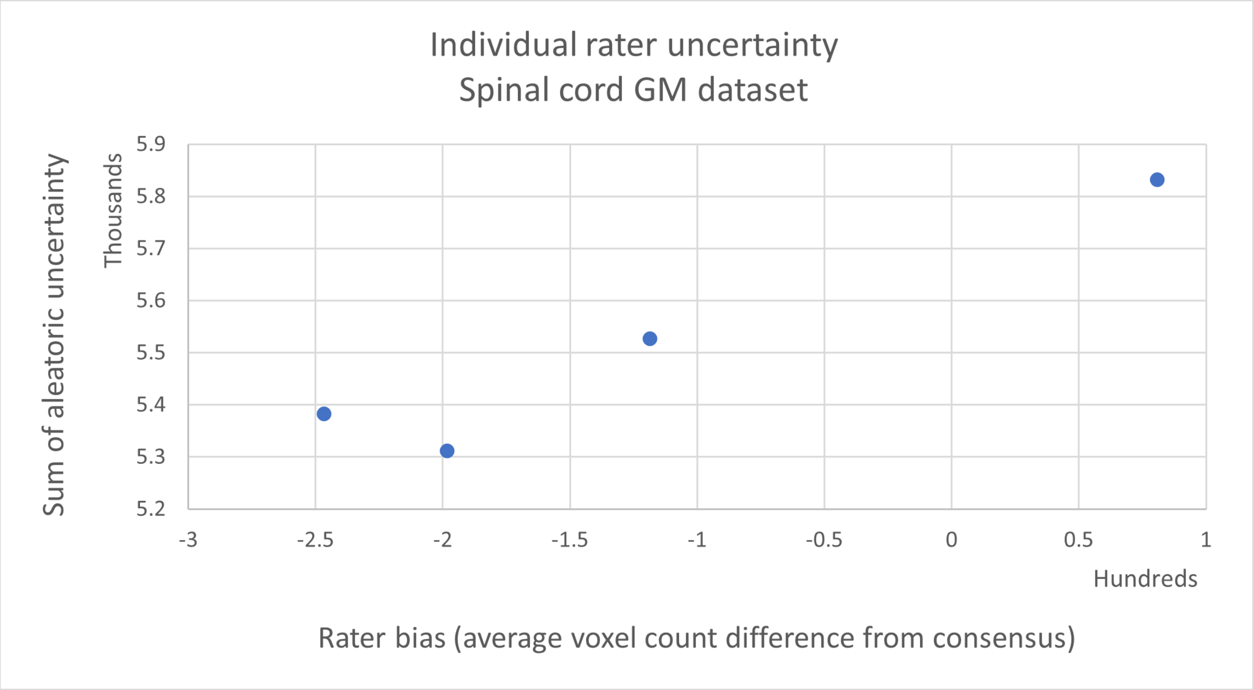


Figure 5.3 Relationship between the uncertainty of models trained for each rater and the bias of the corresponding rater for the SC GM dataset. $R^2 = 0.93$

In both datasets raters with a higher bias also have higher uncertainty. Over-segmentation (bias > 0) seems to be associated with higher uncertainty than under-segmentation (bias < 0). Raters are also clustered by center for the MS brain dataset, but in this case the distance between clusters is smaller than in the rater-style graph, since other factors also influence uncertainty, such as noise in data and the limited size of the training set.

It is interesting to note that while a higher rater bias produces higher uncertainty, it does not affect model performance as assessed by the Dice score ($R^2 = 0.07$), as shown in Figure 5.4.

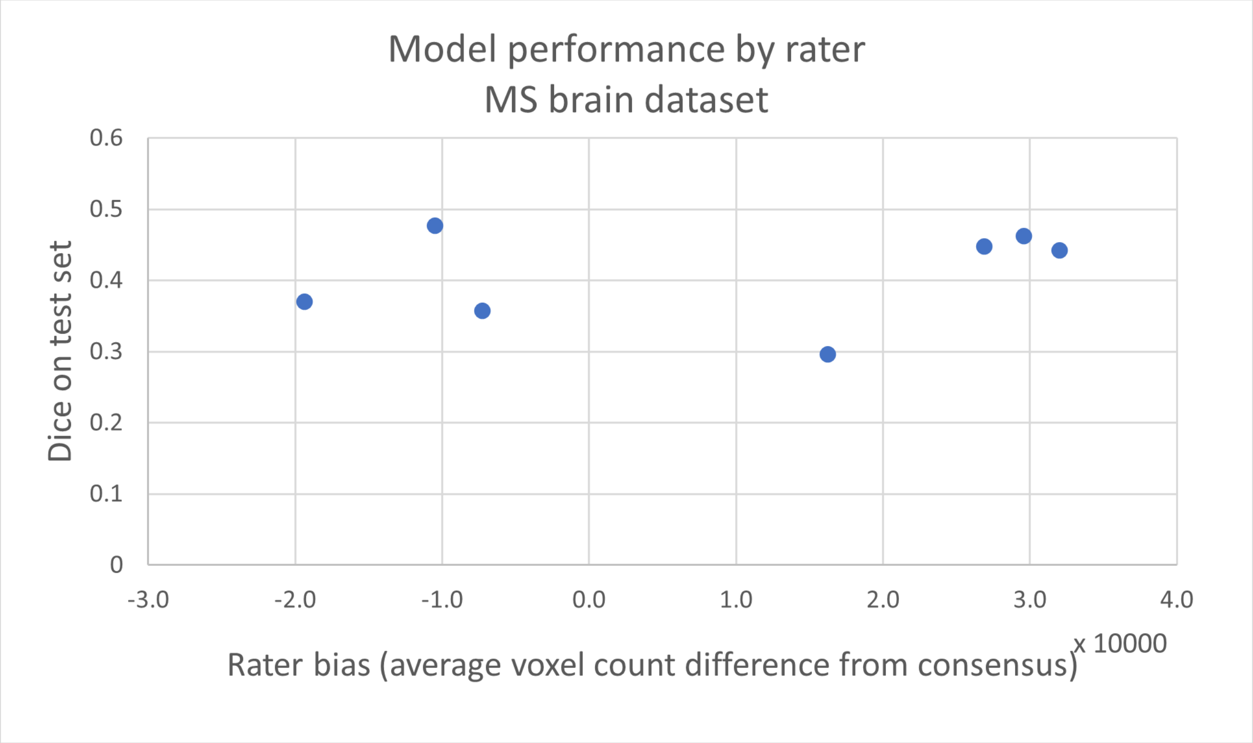


Figure 5.4 Relationship between the Dice score on the test set of models trained for each rater and the bias of the corresponding rater ($R^2 = 0.07$). Dice score is computed for each model on the same ground truth that was used for training (e.g. model trained on data from rater X is evaluated with respect to ground truth unseen during training from the same rater).

5.7.3 Consensus

All raters exhibit some level of bias and as we saw earlier, bias is correlated with uncertainty (Figures 5.2-5.3). We now investigate whether combining raters through consensus would lower uncertainty when compared to single-rater training. Results of this investigation are shown in Figure 5.5, highlighting a consensus uncertainty 30% lower than the average across individual raters.

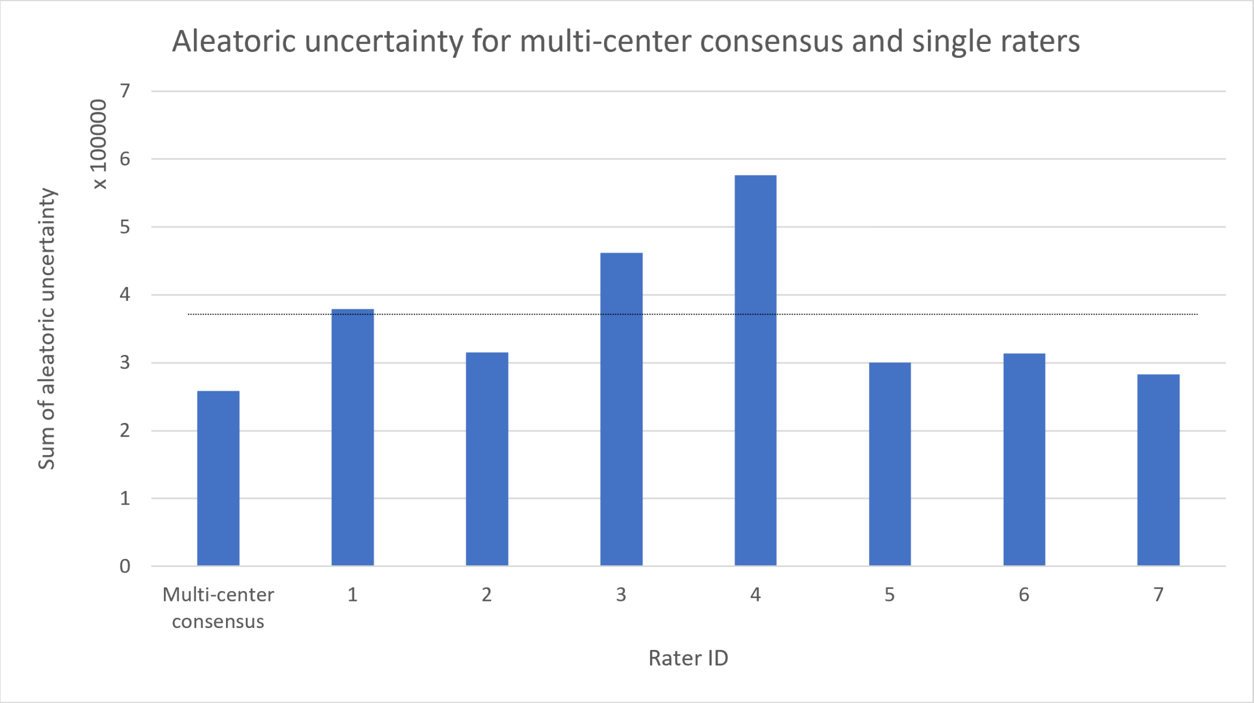


Figure 5.5 Comparison of uncertainty for individual raters and multi-center consensus for the MS brain dataset. Dotted line is the average of single raters uncertainty.

Center-wise consensuses do not, however, exhibit the same characteristic, as they have higher uncertainty than the global (multi-center) consensus and are comparable (lower for center 2, slightly higher for center 1, and irrelevant for the single rater of center 3) to the average uncertainty of their raters used individually (Figure 5.6)

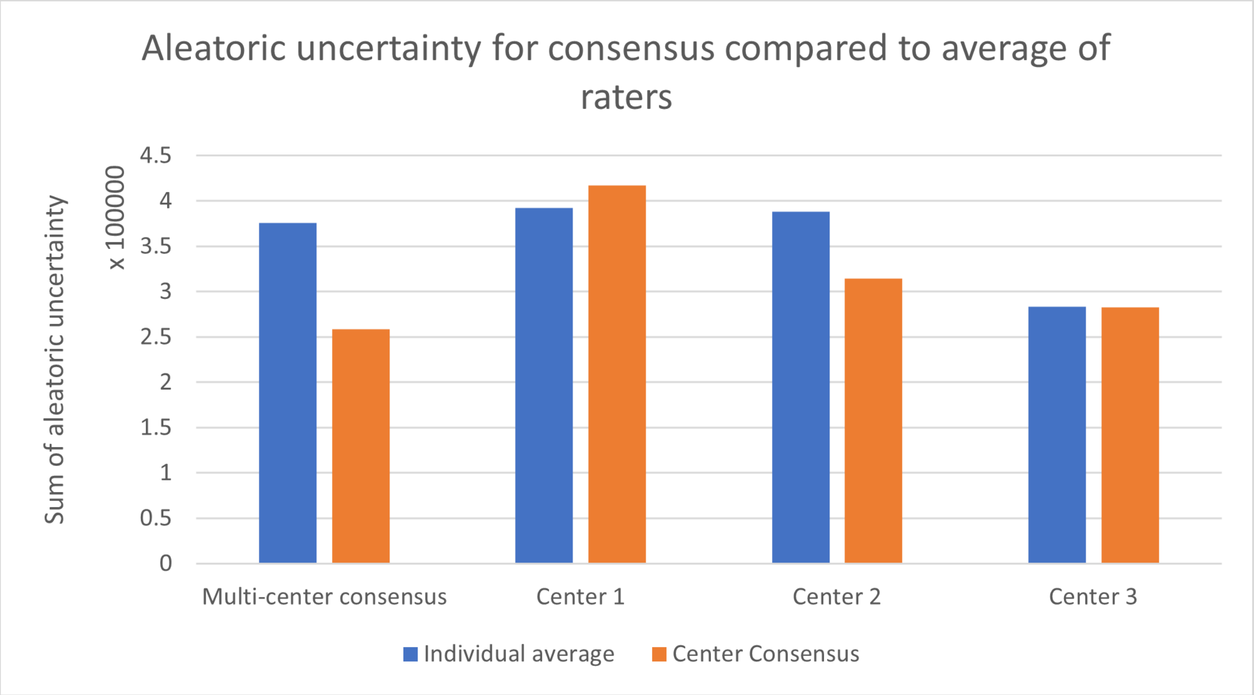


Figure 5.6 Per-center comparison of average rater uncertainty and consensuses uncertainty for the MS brain dataset.

Finally, the previous results are also reflected in the performance (Dice score) of models with the global consensus scoring a full 0.1 above the average of individual raters, and scoring $[0.05 - 0.09]$ above single center consensuses as shown in table 5.1. Thus, it seems that combining raters from different centers has a more positive impact on uncertainty and Dice than combining raters from the same center.

Dice score for consensus	
Raters average	0.42
Center 1 consensus	0.47
Center 2 consensus	0.46
Center 3 consensus	0.43
Multi-center consensus	0.52

Table 5.1 Dice score for different combinations of raters for the MS brain dataset. Dice score is computed for each model on the same ground truth that was used for training (e.g. model trained on data from center X's consensus is evaluated with respect to ground truth unseen during training from the same center).

5.8 Discussion

This study shows that rater style can be characterised by measuring rater consistency and bias. Moreover, results from the brain MS dataset suggest that rater style is mostly center-specific instead of rater-specific. Results also show on both MS brain and SC GM that when using annotations from a single rater to train a deep learning model, a high rater bias leads to high model uncertainty. This is interesting since these models are trained on annotations from a single rater and therefore have never "seen" the consensus although bias relative to consensus still impacts uncertainty. While rating style impacts the amount of uncertainty, bias doesn't directly affect the average performance (Dice score) of the model meaning that the rater style can be learned by the model regardless of uncertainty. A mechanism that could potentially explain why oversegmentation leads to higher uncertainty is partial volume effect. Indeed, a rater that undersegments (e.g. labelling only voxels that contain 100% lesion tissue, and not those at the boundary that contain some other tissue) would give an easier task to the model; voxels labelled as lesions are homogeneous, and simple to identify. At the opposite, a rater that oversegments also includes voxels containing a varying percentage of lesion tissue, which is potentially harder since there is less homogeneity, therefore yielding more uncertainty.

Another interesting result is that uncertainty was lower for the global consensus model (i.e., when fusing all raters' annotations into a single binary annotation) than for models trained using annotations from a single rater. We hypothesize this phenomenon originates from the biases of individual raters which get smoothed away when combining raters from different centers which have different styles. This is also probably why combining raters' annotations from a single center (center-wise consensuses) does not reduce a model's uncertainty : individual bias can't cancel out since we combine raters with similar styles and shortcomings. A single rater, such as the one from Center #3, can therefore have lower uncertainty than the consensus from the four raters of Center #1 due to their higher bias. Multi-center consensus could therefore be a mechanism to lower the impact of rater style.

This lower uncertainty for the global consensus, however, opens up questions regarding the impact of inter-rater variability on uncertainty. Inter-rater variability by definition is not present for single-rater models, but is present in center-wise consensuses, and is at its highest for the global consensus since it combines raters with diverging styles. Our results therefore suggest that the reduction in rater bias when going from single rater to global consensus has a bigger impact on uncertainty than the addition of inter-rater variability. It is therefore possible that inter-rater variability is indeed present but relatively constant throughout the dataset, thus not generating much uncertainty. A limitation of this study is that the number of raters (7 for the MS dataset and 4 for the SC GM dataset) is relatively small, therefore our results would benefit from further validations in datasets with larger pools of raters from different centers.

5.8.1 Impact and perspectives

Raters used in the MS brain study were junior raters trained by senior raters from their center [55], therefore the mutual influence among raters during the learning and segmentation process probably drives the similarities in rating style. This center-wise rater style pattern raises a few questions concerning label fusion, which is largely used in deep learning medical imaging studies. Indeed, in the case of the MS brain dataset, since the split between centers is 4-2-1, if one uses a majority voting consensus it essentially becomes the vote of the four raters from a single center, negating the benefits of having two additional centers with raters in the study. It is doubtful that STAPLE and its variants could really solve the issue since it is based on majority voting, only with weights updated iteratively. If the four raters from one center dominate during the first iteration, the remaining raters will see their weighting be progressively reduced until convergence.

Future studies should therefore consider whether raters from the same center can really be considered independent, or if voting should be weighted by centers instead of raters. Weights of raters could also be considered as hyperparameters that can be optimised in order to minimize uncertainty. Alternatively, raters weight could be incorporated into the input ground truth segmentation using a “soft training” pipeline [59]. While our rater style was defined as simple metrics independent of deep learning models, it would be interesting to see if learned rater style approaches [52–54] show a similar relationship to uncertainty.

Other potentially interesting metrics include measuring boundary difference instead of volume difference. An example would be the average symmetric surface distance (ASSD) which computes the average Euclidean distance between the object boundaries across raters. This metric would be particularly relevant for the MS lesions task where there is a large heterogeneity of object shape, and therefore it could be interesting to complement the volume difference analysis with some shape analysis. Indeed, an increase of lesion radius (e.g. evenly adding 1 voxel along the lesion boundary) would have a different impact on the relative increase of the lesion volume if the lesion is small or large (e.g., 10-voxels vs. 100-voxels lesion). Therefore, from a "radius segmentation style" perspective, it could be said that our absolute metric overweights large lesions at the expense of small ones, whereas it would be the opposite for our relative metric. However, measures based on boundaries also have drawbacks. It is possible that two raters segment the same lesion volume with a slightly different boundary (translation, change of shape, etc.). Therefore such a metric would measure a bias even though there is none (there is indeed a disagreement, but not in the form of over/under-segmentation that we are looking for). To summarize, the main drawback of our volumetric bias is the possibility that it turns out to be non-linearly dependant on some other underlying bias (e.g. if it in facts depends on the radius). While we present two bias metrics here, a detailed comparison with other relevant metrics would be interesting to explore in future research.

Finally, uncertainty could have potential applications for quality control such as identifying biased raters when there are not many ratings available. As an example, a rater generating significantly higher than expected uncertainty for a given task could be excluded as an outlier. Conversely, rating style could be used as a pre-processing step to “correct” biases on an individual rater basis or incorporated as a prior in future deep learning architecture. Model segmentation could be modulated using metrics about the rater style (e.g used as inputs for FiLM layers [6, 60]).

5.9 Acknowledgements

The authors would like to thank Andréanne Lemay and Lucas Rouhier from the IVADO medical imaging team for helpful discussions. Funded by the Institut de valorisation des données (IVADO), 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]. FRQNT Strategic Clusters Program (2020-RS4-265502 - Centre UNIQUE -Union Neurosciences & Artificial Intelligence –Quebec, Canada First Research Excellence Fund through the TransMedTech Institute. C.G has a fellowship from IVADOMED [EX-2018-4], O.V. has a fellowship from NSERC, FRQNT and UNIQUE.

5.10 Appendix A. Relative bias

In this appendix we present the equivalent of figure 5.2 and 5.3 using relative instead of absolute bias. Relative bias is defined in equation 5.3 in a similar way as absolute bias in equation 5.1, with the only change being the fact that we normalize the difference between rater and consensus by the number of positive voxels in the consensus in each image, therefore ensuring no volume has a disproportionate weight.

$$\text{relative bias} = \frac{\sum_{\text{images}} \dfrac{n_{\text{rater}} - n_{\text{consensus}}}{n_{\text{consensus}}}}{\#\text{ of images}} \tag{5.3}$$

Figure 5.7 and 5.8 show that on both datasets, the relationship between uncertainty and bias is still present using relative bias. Correlation is slightly stronger (0.64 vs 0.60) for MS and identical for (0.93) for GM than when using relative bias compared to absolute. On the qualitative side, for MS lesions we observe in figure 5.7 that one rater is an outlier (blue dot close to orange ones). It already was relatively far from its peer in figure 5.2, but this is exacerbated here. On GM segmentation, figure 5.8 shows that switching from relative to absolute bias makes pretty much no difference. The rater distribution is almost identical to figure 5.3. Overall GM bias is a lot lower than MS in both the absolute and relative cases, since the task is easier there is less disagreement between raters. The lack of difference between the relative and absolute bias for GM is potentially explained by the fact that the GM volume varies a lot less across slices and subjects than MS lesions. It is inline with our expectations that relative bias is useful to accentuate the errors on very small lesions, which is why this re-weighting affects mostly the MS dataset.

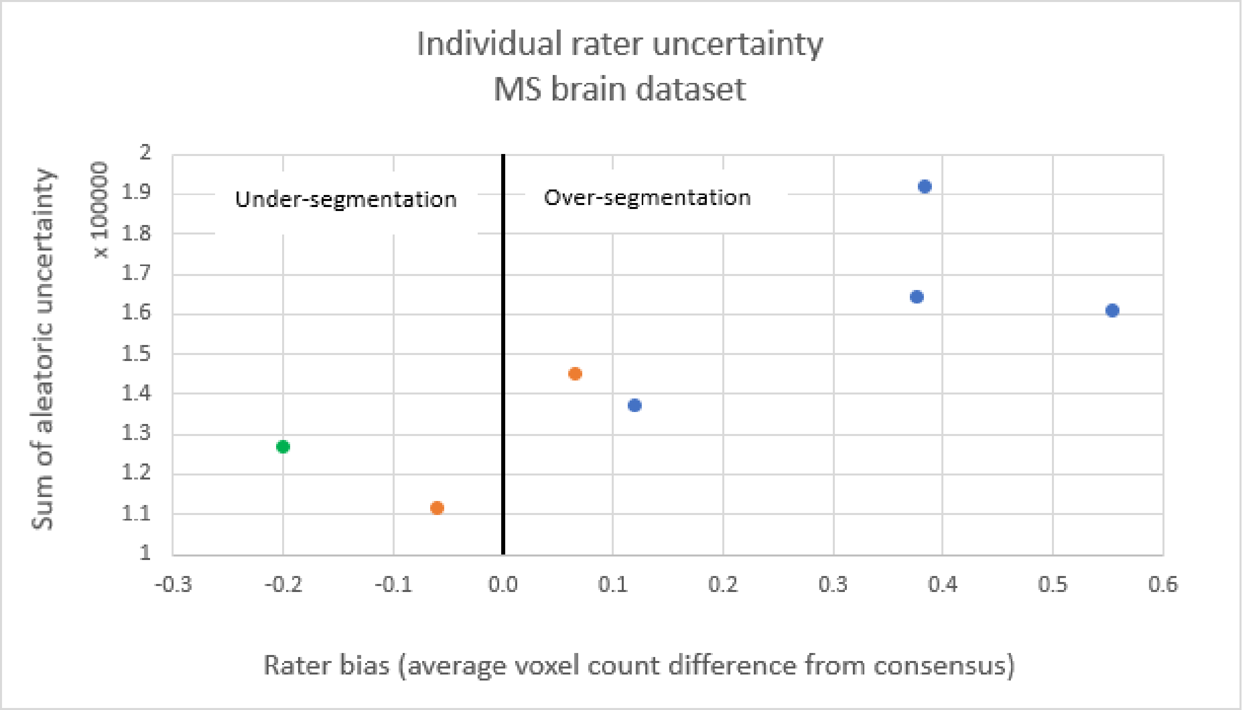


Figure 5.7 Relationship between the uncertainty of models trained for each rater and the relative bias of the corresponding rater for the MS brain dataset. Each colour corresponds to a center. $R^2 = 0.64$

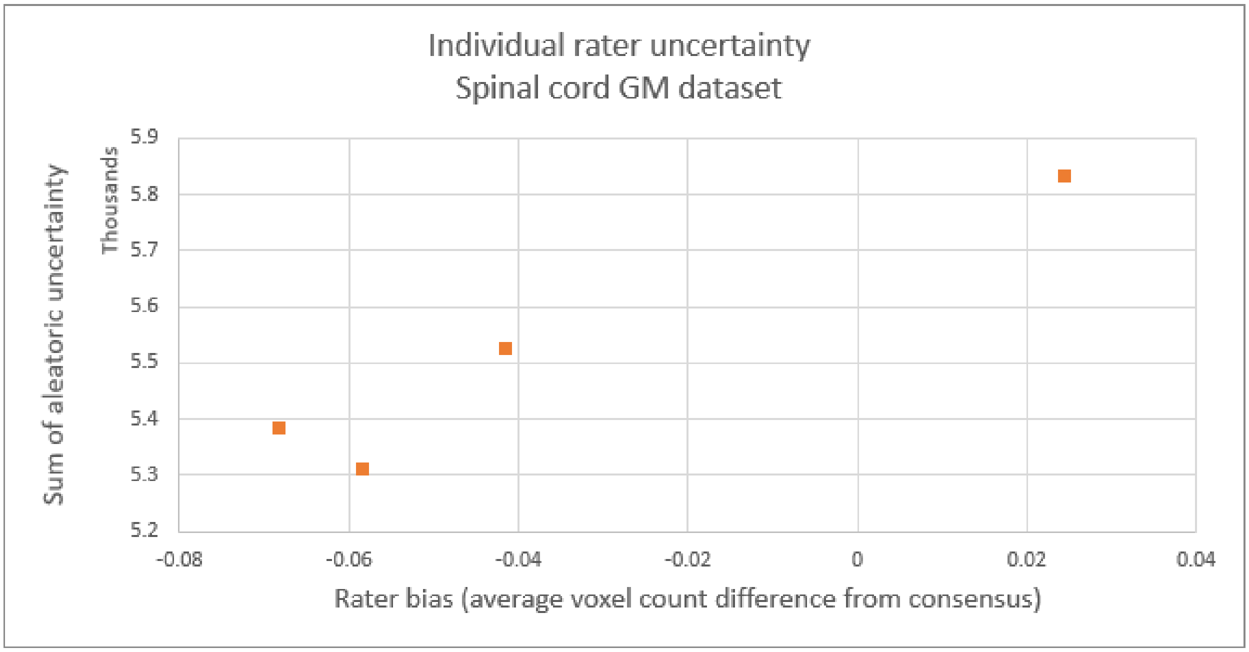


Figure 5.8 Relationship between the uncertainty of models trained for each rater and the relative bias of the corresponding rater for the SC GM dataset. $R^2 = 0.93$

CHAPTER 6 GENERAL DISCUSSION

In summary, our results suggest that contrast generalization is not necessarily what limits performance in MS lesions segmentation. Instead, this limitation likely stems from rater style variability. We have shown that rater bias is correlated with aleatoric uncertainty. Rater style (bias and consistency) mostly varies across centers and less across individuals, as initially thought. Therefore, it follows that consensus can reduce uncertainty, but only when combining raters with different styles (centers). This center-centric rating style is further confirmed by the fact that U-Net has similar performance (Dice) on single-rater and single-center whereas performance is better on multi-center consensus. We also found no correlation between rater bias and single rater Dice score, suggesting that U-Net has no problem learning single-rater style, which is consistent with a previous study [45]. Consensus is therefore a tool among others, but in no way silver bullet. Having many raters annotate a dataset is not automatically "better" when training a DL model, especially if they are from the same center or have had similar training. Consensus doesn't guarantee lower uncertainty and better Dice, since a single low-bias rater can fare better than multiple combined high-bias raters. The center-centric style also highlights potential improvements in rating protocols. A recent review of clinical guidelines highlights these flaws which lead to subjective assessments:

> “Current MRI criteria for multiple sclerosis are based on imaging features that are characteristic of the disease, but are not sufficiently specific. Over time, revisions of the multiple sclerosis diagnostic criteria have improved the sensitivity, particularly adding the capability to confirm the diagnosis at first clinical presentation. However little attention has been given to describing the imaging features included in these criteria in detail, and guiding neurologists and neuroradiologists in correctly interpreting them.” [61]

A high sensitivity but low specificity means in our terms a high positive bias, which is what we have shown to be correlated with uncertainty. Therefore, improvements in image interpretation and diagnostic criteria would likely benefit deep learning segmentation, yielding less oversegmentation and lower uncertainty.

It is also interesting that this bias vs. uncertainty correlation was replicated in gray matter segmentation (large objects relative to image size, systematic presence of the object across all slices) which is very different from MS lesions segmentation (small objects relative to the image size, variable object sizes, non-systematic presence of the object across slices, very large class imbalance). GM segmentation is a task that is a lot easier, with a lot less uncertainty and rater disagreement. While this bias vs. uncertainty relationship needs to be replicated in other multi-rater datasets, the fact that it is present in two very different datasets suggests it could also apply to other segmentation tasks.

CHAPTER 7 CONCLUSION

In this work we explored two topics in MS lesions segmentation. The first, FiLM, aimed to improve generalization across contrasts but failed to improve segmentation performance. The second, uncertainty and its relation to rater style, yielded some interesting insight on what rater characteristics influence uncertainty. The main finding was that rater bias is correlated with aleatoric uncertainty, and that the main variations in rater style (bias and consistency) are across centers rather than individuals. Both of these projects left some open questions for future research, which are discussed in this section.

7.0.1 FiLM

The FiLM project was continued by my colleague Andréanne Lemay who went on to show that that FiLM can improve segmentation of spinal cord tumors using the tumor type as metadata instead of contrast type [60]. It therefore looks plausible that MS lesion was simply a bad task to utilize this architecture, since simpler models can already attain inter-rater levels of performance. She achieves a 5.1% higher Dice using FiLM, which is consistent with my 5% higher Dice in "non-optimized" MS lesions results presented in chapter 4. While here FiLM was used with contrast type as metadata, there are other applications such as using it to deal with multi-class datasets with missing labels (by specifying which labels are present/missing as a metadata). Models can still learn from data with missing labels since model's weights are shared between tasks. FiLM is therefore an architecture which could have many uses, even though it was ineffective here [60].

7.0.2 Uncertainty

While many studies focus on developing new and more complex neural network architectures, we have to wonder if the higher theoretical performance makes sense. Indeed, Dice scores and performance metrics are not everything and they are often only the tip of the iceberg. In the case of MS lesions, attaining scores higher than the inter-rater variability makes no practical sense, since it probably only means overfitting on one or more rater styles. Tools such as soft segmentation [59] and uncertainty provide valuable information absent from binary approaches.

In the end, there is not a unique "true" binary segmentation for a given image. Considering the partial volume effect at the boundary of lesions, there is some inherent ambiguity in the data that and 2 raters could potentially disagree without committing an "error". Therefore, another way to look at the problem would be to see centers with different styles as different "domains" or style, with neither of them being wrong. We could imagine two-steps approaches were a model predicts a segmentation and another one translates the segmentation into the desired center style. A U-Net trained on X center could deploy this model in center Y by adding a post-processing step that converts between the two centers' style. This is similar as transfer between artistic styles which has been studied [49], and rating style could be viewed in the same way since it contains a subjective component.

7.0.3 ivadomed

All the processing for both projects was done using ivadomed [7]. While basic tools such as pre-processing transforms and some models such as U-Net were already implemented before the beginning of my master, many others were implemented during the projects. A particularly useful tool I developed in ivadomed is *automate_training*, an end-to-end multi-GPU training and evaluation pipeline which enables quicker experiments with reproducible of results. Other models, losses and improvements were implemented along the way, as the need arose. These tools will be useful for other MRI segmentation and classification projects since ivadomed is now the backbone of Spinal Cord Toolbox (SCT) [62]. In the future other imaging methods will also benefit from them since the ivadomed team is now working towards histology and EEG compatibility.

REFERENCES

[1] "MS Society of Canada." [Online]. Available: https://mssociety.ca/

[2] I. K. Sand, "Classification, diagnosis, and differential diagnosis of multiple sclerosis," *Current Opinion in Neurology*, vol. 28, no. 3, pp. 193–205, Jun. 2015. [Online]. Available: insights.ovid.com

[3] C. Gros *et al.*, "Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks," *NeuroImage*, vol. 184, pp. 901–915, Jan. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811918319578

[4] C. S. Perone and J. Cohen-Adad, "Promises and limitations of deep learning for medical image segmentation," *Journal of Medical Artificial Intelligence*, vol. 2, pp. 1–1, Jan. 2019. [Online]. Available: http://jmai.amegroups.com/article/view/4659/html

[5] M. Ghafoorian *et al.*, "Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, ser. Lecture Notes in Computer Science, M. Descoteaux *et al.*, Eds. Cham: Springer International Publishing, 2017, pp. 516–524.

[6] E. Perez *et al.*, "FiLM: Visual Reasoning with a General Conditioning Layer," *arXiv:1709.07871 [cs, stat]*, Dec. 2017, arXiv: 1709.07871. [Online]. Available: http://arxiv.org/abs/1709.07871

[7] C. Gros *et al.*, "ivadomed: A Medical Imaging Deep Learning Toolbox," *Journal of Open Source Software*, vol. 6, no. 58, p. 2868, Feb. 2021. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.02868

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[9] S. A. Taghanaki *et al.*, "Deep Semantic Segmentation of Natural and Medical Images: A Review," *arXiv:1910.07655 [cs, eess]*, Nov. 2019, arXiv: 1910.07655. [Online]. Available: http://arxiv.org/abs/1910.07655

[10] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1361841517301135

[11] J. Bernal *et al.*, “Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review,” *Artificial Intelligence in Medicine*, vol. 95, pp. 64–81, Apr. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365716305206

[12] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *arXiv:1505.04597 [cs]*, May 2015, arXiv: 1505.04597. [Online]. Available: http://arxiv.org/abs/1505.04597

[13] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *arXiv:1502.03167 [cs]*, Mar. 2015, arXiv: 1502.03167. [Online]. Available: http://arxiv.org/abs/1502.03167

[14] G. E. Hinton *et al.*, “Improving neural networks by preventing co-adaptation of feature detectors,” *arXiv:1207.0580 [cs]*, Jul. 2012, arXiv: 1207.0580. [Online]. Available: http://arxiv.org/abs/1207.0580

[15] O. Çiçek *et al.*, “3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation,” *arXiv:1606.06650 [cs]*, Jun. 2016, arXiv: 1606.06650. [Online]. Available: http://arxiv.org/abs/1606.06650

[16] O. Oktay *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” *arXiv:1804.03999 [cs]*, May 2018, arXiv: 1804.03999. [Online]. Available: http://arxiv.org/abs/1804.03999

[17] L. R. Dice, “Measures of the Amount of Ecologic Association Between Species,” *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945. [Online]. Available: http://doi.wiley.com/10.2307/1932409

[18] F. Milletari, N. Navab, and S. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,” in *2016 Fourth International Conference on 3D Vision (3DV)*, Oct. 2016, pp. 565–571.

[19] S. K. Zhou *et al.*, “A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises,” *Proceedings of the IEEE*, pp. 1–19, 2021, arXiv: 2008.09104. [Online]. Available: http://arxiv.org/abs/2008.09104

[20] V. Bhise *et al.*, “Defining and Measuring Diagnostic Uncertainty in Medicine: A Systematic Review,” *Journal of General Internal Medicine*, vol. 33, no. 1, pp. 103–115, Jan. 2018. [Online]. Available: https://doi.org/10.1007/s11606-017-4164-1

[21] A. D. Kiureghian and O. Ditlevsen, “Aleatory or epistemic? Does it matter?” *Structural Safety*, vol. 31, no. 2, pp. 105–112, Mar. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167473008000556

[22] A. Kendall and Y. Gal, “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” *arXiv:1703.04977 [cs]*, Oct. 2017, arXiv: 1703.04977. [Online]. Available: http://arxiv.org/abs/1703.04977

[23] A. G. Roy *et al.*, “Bayesian QuickNAT: Model Uncertainty in Deep Whole-Brain Segmentation for Structure-wise Quality Control,” *arXiv:1811.09800 [cs]*, Nov. 2018, arXiv: 1811.09800. [Online]. Available: http://arxiv.org/abs/1811.09800

[24] T. Nair *et al.*, “Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation,” *Medical Image Analysis*, vol. 59, p. 101557, Jan. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841519300994

[25] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” *ICML’16: Proceedings of the 33rd International Conference on International Conference on Machine Learning*, vol. 48, p. 10, 2016.

[26] ——, “Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference,” *arXiv:1506.02158 [cs, stat]*, Jan. 2016, arXiv: 1506.02158. [Online]. Available: http://arxiv.org/abs/1506.02158

[27] M. Teye, H. Azizpour, and K. Smith, “Bayesian Uncertainty Estimation for Batch Normalized Deep Networks,” in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, Jul. 2018, pp. 4907–4916. [Online]. Available: http://proceedings.mlr.press/v80/teye18a.html

[28] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles,” *arXiv:1612.01474 [cs, stat]*, Nov. 2017, arXiv: 1612.01474. [Online]. Available: http://arxiv.org/abs/1612.01474

[29] M. S. Ayhan and P. Berens, “Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks,” *1st Conference on Medical Imaging with Deep Learning (MIDL 2018),*, p. 9, 2018.

[30] G. Wang *et al.*, “Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks,” *Neurocomputing*, vol.

338, pp. 34–45, Apr. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231219301961

[31] A. Carass *et al.*, "Longitudinal Multiple Sclerosis Lesion Segmentation: Resource & Challenge," *NeuroImage*, vol. 148, pp. 77–102, Mar. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5344762/

[32] S. Lévy *et al.*, "White matter atlas of the human spinal cord with estimation of partial volume effect," *NeuroImage*, vol. 119, pp. 262–271, Oct. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811915005431

[33] 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, Sep. 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811904002745

[34] Z. Akkus *et al.*, "Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions," *Journal of Digital Imaging*, vol. 30, no. 4, pp. 449–459, Aug. 2017. [Online]. Available: https://doi.org/10.1007/s10278-017-9983-4

[35] C. F. Baumgartner *et al.*, "PHiSeg: Capturing Uncertainty in Medical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, ser. Lecture Notes in Computer Science, D. Shen *et al.*, Eds. Cham: Springer International Publishing, 2019, pp. 119–127.

[36] B. Billot *et al.*, "Partial Volume Segmentation of Brain MRI Scans of Any Resolution and Contrast," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, ser. Lecture Notes in Computer Science, A. L. Martel *et al.*, Eds. Cham: Springer International Publishing, 2020, pp. 177–187.

[37] T. Watadani *et al.*, "Interobserver Variability in the CT Assessment of Honeycombing in the Lungs," *Radiology*, vol. 266, no. 3, pp. 936–944, Mar. 2013, publisher: Radiological Society of North America. [Online]. Available: https://pubs.rsna.org/doi/10.1148/radiol.12112516

[38] A. B. Rosenkrantz *et al.*, "Comparison of Interreader Reproducibility of the Prostate Imaging Reporting and Data System and Likert Scales for Evaluation of Multiparametric Prostate MRI," *American Journal of Roentgenology*, vol. 201, no. 4, pp. W612–W618, Oct. 2013. [Online]. Available: http://www.ajronline.org/doi/10.2214/AJR.12.10173

[39] E. Lazarus *et al.*, “BI-RADS Lexicon for US and Mammography: Interobserver Variability and Positive Predictive Value,” *Radiology*, vol. 239, no. 2, pp. 385–391, May 2006, publisher: Radiological Society of North America. [Online]. Available: https://pubs.rsna.org/doi/abs/10.1148/radiol.2392042127

[40] 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, Jul. 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1283110/

[41] A. Suinesiaputra *et al.*, “Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours,” *Journal of Cardiovascular Magnetic Resonance*, vol. 17, no. 1, p. 63, Jul. 2015. [Online]. Available: https://doi.org/10.1186/s12968-015-0170-9

[42] E. Chotzoglou and B. Kainz, “Exploring the Relationship Between Segmentation Uncertainty, Segmentation Performance and Inter-observer Variability with Probabilistic Networks,” in *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*, L. Zhou *et al.*, Eds. Cham: Springer International Publishing, 2019, vol. 11851, pp. 51–60. [Online]. Available: http://link.springer.com/10.1007/978-3-030-33642-4_6

[43] M. H. Jensen *et al.*, “Improving Uncertainty Estimation in Convolutional Neural Networks Using Inter-rater Agreement,” in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019*, D. Shen *et al.*, Eds. Cham: Springer International Publishing, 2019, vol. 11767, pp. 540–548. [Online]. Available: http://link.springer.com/10.1007/978-3-030-32251-9_59

[44] A. Jungo *et al.*, “On the Effect of Inter-observer Variability for a Reliable Estimation of Uncertainty of Medical Image Segmentation,” *arXiv:1806.02562 [cs]*, Jun. 2018, arXiv: 1806.02562. [Online]. Available: http://arxiv.org/abs/1806.02562

[45] O. Shwartzman *et al.*, “The Impact of an Inter-rater Bias on Neural Network Training,” *arXiv:1906.11872 [cs, eess]*, Jun. 2019, arXiv: 1906.11872. [Online]. Available: http://arxiv.org/abs/1906.11872

[46] K. Berer and G. Krishnamoorthy, “Microbial view of central nervous system autoimmunity,” *FEBS Letters*, vol. 588, no. 22, pp. 4207–4213, Nov. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0014579314002932

[47] A. J. Thompson *et al.*, “Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria,” *The Lancet. Neurology*, vol. 17, no. 2, pp. 162–173, Feb. 2018.

[48] H. Kearney, D. H. Miller, and O. Ciccarelli, “Spinal cord MRI in multiple sclerosis–diagnostic, prognostic and clinical value,” *Nature Reviews. Neurology*, vol. 11, no. 6, pp. 327–338, Jun. 2015.

[49] V. Dumoulin, J. Shlens, and M. Kudlur, “A Learned Representation For Artistic Style,” *arXiv:1610.07629 [cs]*, Feb. 2017, arXiv: 1610.07629. [Online]. Available: http://arxiv.org/abs/1610.07629

[50] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980 [cs]*, Jan. 2017, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[51] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” *arXiv:1912.01703 [cs, stat]*, Dec. 2019, arXiv: 1912.01703. [Online]. Available: http://arxiv.org/abs/1912.01703

[52] L. Zhang *et al.*, “Disentangling Human Error from the Ground Truth in Segmentation of Medical Images,” *arXiv:2007.15963 [cs]*, Oct. 2020, arXiv: 2007.15963. [Online]. Available: http://arxiv.org/abs/2007.15963

[53] R. Tanno *et al.*, “Learning From Noisy Labels by Regularized Estimation of Annotator Confusion,” 2019, pp. 11 244–11 253. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Tanno_Learning_From_Noisy_Labels_by_Regularized_Estimation_of_Annotator_Confusion_CVPR_2019_paper.html

[54] C. H. Sudre *et al.*, “Let’s agree to disagree: learning highly debatable multirater labelling,” *arXiv:1909.01891 [cs]*, Sep. 2019, arXiv: 1909.01891. [Online]. Available: http://arxiv.org/abs/1909.01891

[55] O. Commowick *et al.*, “Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure,” *Scientific Reports*, vol. 8, no. 1, p. 13650, Sep. 2018, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-018-31911-7

[56] F. Prados *et al.*, “Spinal cord grey matter segmentation challenge,” *NeuroImage*, vol. 152, pp. 312–329, May 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811917302185

[57] O. Vincent *et al.*, “Automatic segmentation of spinal multiple sclerosis lesions: How to generalize across MRI contrasts?” *arXiv:2003.04377 [cs, eess]*, Mar. 2020, arXiv: 2003.04377. [Online]. Available: http://arxiv.org/abs/2003.04377

[58] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[59] C. Gros, A. Lemay, and J. Cohen-Adad, “SoftSeg: Advantages of soft versus binary training for image segmentation,” *arXiv:2011.09041 [cs, eess]*, Nov. 2020, arXiv: 2011.09041. [Online]. Available: http://arxiv.org/abs/2011.09041

[60] A. Lemay *et al.*, “Benefits of Linear Conditioning for Segmentation using Metadata,” *arXiv:2102.09582 [cs, eess]*, Feb. 2021, arXiv: 2102.09582. [Online]. Available: http://arxiv.org/abs/2102.09582

[61] M. Filippi *et al.*, “Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines,” *Brain*, vol. 142, no. 7, pp. 1858–1875, Jul. 2019. [Online]. Available: https://academic.oup.com/brain/article/142/7/1858/5519813

[62] B. De Leener *et al.*, “SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data,” *NeuroImage*, vol. 145, pp. 24–43, Jan. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811916305560